Inference in Incomplete Models

Alfred Galichon and Marc Henry

Harvard University and Columbia University

First draft: September 15, 2005
This draft[1]: March 15, 2006

Abstract

The aim of this study is to provide a test for the specification of a structural model without identifying assumptions. We show the equivalence of three natural definitions of correct specification, which we take as our null hypothesis. Using a representation of the null hypothesis as a Monge-Kantorovich optimal mass transportation, we show that the natural test statistic is a form of Kolmogorov-Smirnov statistic for Choquet capacities. When the model is given in parametric form, the test can be inverted to yield confidence regions for the identified parameter set. The approach can be applied to areas as diverse as the estimation of models with sample selection, censored observables and to games with multiple equilibria.

JEL Classification: C10, C12, C13, C14, C52, C61

Keywords: partial identification, specification test, random correspondences, selections, plausibility constraint, Monge-Kantorovich mass transportation problem, Kolmogorov-Smirnov test for capacities.

[1] Preliminary and incomplete. Helpful discussions with Xiaohong Chen, Victor Chernozhukov, Pierre-André Chiappori, Ronald Gallant, Han Hong, Massimo Marinacci, Rosa Matzkin, Francesca Molinari, Alexei Onatski, Ariel Pakes, Jim Powell, Peter Robinson, Bernard Salanié, Thomas Sargent, Jay Sethuraman, Azeem Shaikh, Chris Sims, Kyungchul Song and Edward Vytlacil are gratefully acknowledged. The authors are also grateful for comments from seminar audiences at Columbia, École polytechnique, MIT and the Research Triangle (with the usual disclaimer). Correspondence address: Department of Economics, Columbia University, 420 W 118th Street, New York, NY 10027, USA. mh530@columbia.edu.

Introduction

In many contexts, the ability of econometric models to identify, hence estimate from observed frequencies, the distribution of residual uncertainty often rests on strong prior assumption that are difficult to substantiate and even to analyze within the economic decision problem.

A recent approach, pioneered by Manski has been to forego such prior assumptions, thus giving up the ability to identify a single probability distribution for residual uncertainty, and allow instead for a set of distributions compatible with the empirical setup. A variety of models have been analyzed in this way, whether partial identification stems from incompletely specified models (typically models with multiple equilibria) or from structural data insufficiencies (typically cases of data censoring). See Manski (2005) for an up-to-date survey on the topic.

All these models with incomplete identification share the basic fundamental structure that the residual uncertainty and the relevant observable quantities are linked by a one-to-many mapping instead of a one-to-one mapping as in the case of identification.

In this paper, we propose a general framework for conducting inference without additional assumptions such as equilibrium selection mechanisms necessary to identify the model (i.e. to ensure that the one-to-many mapping is actually one-to-one). The usual terminology for such models is “incomplete” or “partially identified.”

In a parametric setting, the objective of inference in partially identified models is the estimation of the set of parameters (hereafter called *identified set*) which are compatible with the distribution of the observed data and an assessment of the quality of that estimation. For the latter objective, two routes have been taken.

Chernozhukov, Hong, and Tamer (2002) initiated research to obtain regions that cover the identified set with a prescribed probability. They propose an M-estimation approach with a sub-sampling procedure to approximate quantiles of the supremum of the criterion function over the identified set. Shaikh (2005) proposes an alternative M-estimation with subsampling procedure that nests the Chernozhukov, Hong, and Tamer (2002) proposal. M-estimation with subsampling is the only general proposal to date that does not rely on a conservative testing procedure, but the choice of criterion function in the M-estimation procedure is arbitrary, and may have a large effect on the confidence regions.

In related research, a more direct application of random set methods has been taken to achieve the goal of constructing confidence regions for the identified set: Shaikh and Vytlacil (2005) consider a special model where the identified set is a deterministic mapping of a collection of expectations, and base inference on the sample analogs of these expectations. Beresteanu and Molinari (2006) propose the use of central limit theorems for random sets to conduct inference in models with set valued data. However, the adaptation of delta theorems for random sets (as in King (1989)) is required for this approach to attain its full potential.

The second route was initiated by Imbens and Manski (2004) who considered the different problem of covering each element of the identified set, and demanded uniform coverage. Shaikh (2005) shows that the M-estimation with subsampling procedure can also be applied to uniform coverage of elements of the identified set. Pakes, Porter, Ho, and Ishii (2004) consider models that are defined by moment inequalities and propose a conservative procedure to form a confidence region for all parameters in the identified set based on inequalities testing ideas put forward by Gouriéroux, Holly, and Monfort (1982). The procedure is conservative since the limiting distribution of the test statistic depends on the number of constraints that are actually binding, and unlike in the special one dimensional treatment response case analyzed by Imbens and Manski (2004), no superefficient pre-test is available.

Still in the latter spirit, Andrews, Berry, and Jia (2004) consider entry games (and more generally games with discrete strategies) and propose a conservative procedure to form a confidence region for all parameters in the identified set based on the idea that the probability of a certain outcome is no larger than the probability that necessary conditions (such as Nash rationality constraints) are met.

The inference procedure proposed here is in the same spirit as this latter contribution, but it gives a full formalization of the idea in a very general framework, does not restrict the class of distributions of observables (hence allows estimation of games with continuous strategies as well as entry games), does not rely on resampling procedures (though they may be used as alternative quantile approximation devices), and provides an exact test as opposed to the conservative procedures considered above.

The general set-up comprises a model specification with observable and unobservable variables (unobservable to the analyst but not necessarily to the economic agents) related by a one-to-many mapping as opposed to the one-to-one mapping required for identification. The model is defined by the one-to-many mapping (which can comprise rationality constraints as before, as well as any constraints that are plausible within the theory) and a family of hypothesized distributions for the unobserved variables. To fix ideas, we call Γ the one-to-many mapping defining the model, ν a hypothesized distribution of unobservables and P the true distribution of observables.

First, a characterization is given of what we mean by correct specification, viz. compatibility of the model with the distribution of the observable variables, and it is shown that several natural ways of defining compatibility are in fact equivalent. They include a compatibility notion based on selections γ of Γ (i.e. functions such that $\gamma \in \Gamma$), a notion based on the existence of a joint probability that admits ν and P as marginals and is supported on the region where the constraints implied by model Γ are satisfied, and finally the notion of maximum plausibility introduced by Dempster (1967). It is the topic of section 1.

Second, we show that the characterizations of correct specification of the model are equivalent to the existence of a zero cost solution to a Monge-Kantorovich mass transportation problem, where mass is transported between distribution P and distribution ν with zero-one cost associated with violation of the constraints implied by the model Γ. This is the topic of section 2. Note that a special case of Monge-Kantorovich transportation problem is the well-know matching problem.

Third, still in section 2, this observation allows us to conduct inference using the empirical version of the mass transportation problem (with the unknown P replaced by the empirical distribution P_n). It turns out that the dual of the empirical problem yields a statistic that reduces to the familiar Kolmogorov-Smirnov specification test statistic in the identified case where Γ is one-to-one. The Kolmogorov-Smirnov statistic tests the equality of two probability measures by checking their difference on a *good* class of sets (large enough to be value-determining, but small enough to make things tractable). Here our test statistic checks that $P(A)$ is no larger than $\nu(\Gamma(A))$ for all A in a similar class of sets. Since $\nu(\Gamma(A))$ is the probability of the sufficient conditions implied by A, we see the strong similarity with the Andrews, Berry, and Jia (2004) approach. Hence the dual empirical problem provides us with an easy to compute test statistic, and a distribution to compare it to, and a parallel with the classical case.

Finally, the last section shows simple implementation procedures, and the inversion of the test to construct a confidence region for the elements of the identified set of parameters when both Γ and ν are specified in parametric form. We argue, hence that our approach answers the relevant questions pertaining to both the research programs described above: if one is interested in testing structural hypotheses such as extra constraints implied by theory, within the framework of a partially identified model, the constraints should be rejected if the region they imply on the parameter set does not intersect with the identified set, hence coverage of the identified set as a whole is what matters. Here the question can be answered directly by incorporating the extra constraints in the model and testing the restricted specification. If, on the other hand, one is interested in reporting parameter value estimates with confidence bounds for policy analysis, the specification test can be inverted to that end.

Proofs and additional results are collected in the appendix.

Prelude: complete model benchmark

Before we define incomplete model specifications, we give a short heuristic univariate description of the benchmark that we use and discuss the Kolmogorov-Smirnov specification test statistic that we are effectively generalizing in this paper.

We consider observables $y \in \mathbb{R}$ and unobservables $u \in \mathbb{R}$ (also called "unobserved shocks", "latent variables", etc...). Abstracting from dependence on an unknown deterministic parameter, we define a "complete" model as a pair (ν, γ), where ν is a data generating process for the unobservables, and γ is a bijection from the set of observables to the set of unobservables.

If we call P the true data-generating process for the observables, we say that the complete model is well specified if $P(A) = \nu(\gamma(A))$ for all Borel set A, which, by Dynkin's lemma, is equivalent to $P(A) = \nu(\gamma(A))$ for all cells A of the form $(-\infty, y]$, $y \in \mathbb{R}$, which is immediately seen to be equivalent to

$$\sup_{A \in \mathcal{S}} \left(P(A) - \nu(\gamma(A))\right) = 0 \tag{1}$$

where $\mathcal{S} = \{(-\infty, y_1], (y_2, \infty) : \ (y_1, y_2) \in \mathbb{R}^2\}$.

(1) is a programming problem, and it will turn out to be very fruitful to consider its dual formulation

$$\inf_{\pi} \int_{\mathbb{R}^2} \{u \neq \gamma(y)\} \ \pi(dy, du) = 0, \tag{2}$$

where $\{x \in A\}$ denotes the indicator function of the set A, and the supremum is taken over all joint probability measures with marginals P and ν. The latter is a mass transportation (or "generalized matching") problem, where mass is transported from the set of observables to the set of unobservables with zero-one cost of transportation associated with violations of the model constraint $u = \gamma(y)$.

This formulation can be interpreted as the existence of a probability that is "concentrated on the model", or alternatively, to the existence of a strong coupling between the random variable Y with law P and the random variable U with law ν, i.e. the existence of π with marginals P and ν such that

$$\pi(U \neq \gamma(Y)) = 0. \tag{3}$$

We shall show that this dual representation of the hypothesis of correct specification has a natural generalization to the case of incomplete models.

Turning to empirical versions of the problem, we can consider the statistic obtained by replacing P by the empirical distribution P_n of a sample of independent and identically distributed variables with law P, we obtain

$$\inf_{\pi} \int_{\mathbb{R}^2} \{u \neq \gamma(y)\} \ \pi(dy, du), \tag{4}$$

where the infimum is taken over probabilities π with marginals P_n and ν. By the above mentioned duality, the latter is equal to

$$\sup_{A \in \mathcal{B}} \left(P_n(A) - \nu(\gamma(A))\right),$$

with $\mathcal{B}$ the class of Borel sets.

The last step is to determine a class of sets that is small enough to allow determination of the limiting behaviour of the statistic, i.e. we need to class of sets to be P-Donsker, and large enough that the values of $\nu(\Gamma(.))$ over all Borel sets are determined by the latter's values on the restricted class. The class $\mathcal{S}$ satisfies both requirements, and the resulting test statistic is

$$\sup_{A \in \mathcal{S}} (P_n(A) - \nu(\gamma(A))), \tag{5}$$

which is exactly the Kolmogorov-Smirnov specification test statistic.

We shall essentially follow these same steps to show equivalence between formulations of the hypothesis of correct specification and to derive a test of specification when the bijection γ is replaced by a correspondence Γ. Then we shall consider parameterized versions of the model where both Γ and ν depend on a parameter θ, and form confidence regions with all values of θ such that the specification of model $(\Gamma_\theta, \nu_\theta)$ is not rejected.

1 Incomplete model specifications

We consider a very general econometric model specification, thereby posing the problem exactly as in Jovanovic (1989) which was an inspiration for this work. Variables under consideration are divided into two groups.

- Latent variables, $u \in \mathcal{U}$. The vector u is not observed by the analyst, but some of its components may be observed by the economic actors. $\mathcal{U}$ is a complete, metrizable and separable topological space (i.e. a Polish space).
- Observable variables, $y \in \mathcal{Y} = \mathbb{R}^{d_y}$. The vector y is observed by the analyst[1].

The Borel sigma-algebras of $\mathcal{Y}$ and $\mathcal{U}$ will be respectively denoted $\mathcal{B}_\mathcal{Y}$ and $\mathcal{B}_\mathcal{U}$. Call P the Borel probability measure that represents the true data generating process for the observable variables, and $\mathcal{V}$ a family of Borel probability measures that are hypothesized to be possible data generating processes for the latent variables. Finally, the economic model is given by a relation between observable and latent variables, i.e. a subset of $\mathcal{Y} \times \mathcal{U}$, which we shall write as a multi-valued mapping from $\mathcal{Y}$ to $\mathcal{U}$ denoted by Γ. Finally, the set of Borel probability

[1] Theorem 1 holds more generally when $\mathcal{Y}$ is a convex metrizable subset of a locally convex topological vector space.

measures on $(\mathcal{Y}\times\mathcal{U}, \sigma(\mathcal{B}_{\mathcal{Y}}\times\mathcal{B}_{\mathcal{U}}))$ with marginals P and ν is denoted by $\mathcal{M}(P,\nu)$. Whenever there is no ambiguity, we shall adopt the de Finetti notation μf to denote the integral of f with respect to μ.

1.1 Examples

Example 1: Sample selection and other models with missing counterfactuals. The typical Heckman sample selection models require very strong and often implausible assumptions to guarantee identification. Weaker assumptions, such as certain forms of monotonicity are plausible and restrict significantly the identified set without reducing it to a singleton. As an illustration of our model formulation in this case, consider for instance the classical set-up in Heckman and Vytlacil (2001). We observe (Y, D, W), where Y is the outcome variable, D is an indicator variable for the receipt of treatment, and Z is a vector of instruments (we implicitly condition the model on exogenous observable covariates). The outcome variable is generated as follows:

$$Y = DY_1 + (1-D)Y_2,$$

where Y_0 is the binary potential outcome if the individual does not receive treatment, and Y_1 is the binary potential outcome if the individual does receive treatment. The model is completed with the specification of D as follows:

$$D = 1_{\{g(Z)\geq U\}},$$

where g is a measurable function and U is uniformly distributed on $[0,1]$ (without loss of generality). The model can be written in the form of a multi-valued mapping Γ from observable to unobservables in the following way:

$$\begin{array}{ccl}
(y,d,z) & \longmapsto & \{(u,y_1,y_0)\in\Gamma(y,d,z)\} \\
(1,1,z) & \longmapsto & [0,g(z)]\times\{1\}\times\{0,1\} \\
(1,0,z) & \longmapsto & (g(z),1]\times\{0,1\}\times\{1\} \\
(0,1,z) & \longmapsto & [0,g(z)]\times\{0\}\times\{0,1\} \\
(1,1,z) & \longmapsto & (g(z),1]\times\{0,1\}\times\{0\}
\end{array}$$

Example 2: Returns to schooling. Consider a general specification for the returns to education, where income Y is a function of years of education E, other observable characteristics X and unobserved ability U as $Y = G(E, X, U)$. G can be inverted as a multi-valued mapping to yield a correspondence $U = \Gamma(Y, E, X)$.

Example 3: Censored data structures. Models with top-censoring or positive censoring such as Tobit models fall in this class. A classic problem where identification fails is regression with interval censored outcomes: the observables variables are the pairs (Y_*, Y^*, X) of upper and lower values for the dependent variable, and the explanatory variables. The model correspondence is

$$\Gamma_\theta(y_*, y^*, x) = [y_* - x'\theta, y^* + x'\theta].$$

Example 4: Games with multiple equilibria. Very large classes of economic models become estimable with this approach, when one allows the object of interest to be the identified set of parameters as opposed to single parameter values. A simple class of examples is that of models defined by a set of Nash rationality constraints. Suppose the payoff function for player j, $j = 1, \ldots, J$ is given by

$$\Pi_j(S_j, S_{-j}, X_j, U_j; \theta),$$

where S_j is player j's strategy and S_{-j} is their opponents' strategies. X_j is a vector of observable characteristics of player j and U_j a vector of unobservable determinents of the payoff. Finally θ is a vector of parameters. Pure strategy Nash equilibrium conditions

$$\Pi_j(S_j, S_{-j}, X_j, U_j; \theta) \geq \Pi_j(S, S_{-j}, X_j, U_j; \theta), \text{ for all } S$$

define a correspondence Γ_θ from unobservable player characteristics to observable variables (S, X), and if the unobservable player characteristics, interpreted as types of the players are supposed uniformly distributed on the relevant domain, then $\mathcal{V}$ is a singleton.

Example 5: Entry models. Consider the special case of example 3 proposed by Jovanovic (1989). The payoff functions are

$$\Pi_1(x_1, x_2, u) = (\lambda x_2 - u)I_{\{x_1=1\}},$$
$$\Pi_2(x_1, x_2, u) = (\lambda x_1 - u)I_{\{x_2=1\}},$$

where $x_i \in \{0,1\}$ is firm i's action, and u is an exogenous cost. The firms know their cost; the analyst, however, knows only that $u \in [0,1]$, and that the structural parameter λ is in $(0,1]$. There are two pure strategy Nash equilibria. The first is $x_1 = x_2 = 0$ for all $u \in [0,1]$. The second is $x_1 = x_2 = 1$ for all $u \in [0,\lambda]$ and zero otherwise. Since the two firms' actions are perfectly correlated, we shall denote them by a single binary variable $y = x_1 = x_2$. Hence the model is described by the multi-valued mapping: $\Gamma(1) = [0,\lambda]$ and $\Gamma(0) = [0,1]$. In this case, since y is Bernoulli, we can write $P = (1-p,p)$ with p the probability of a 1. For the distribution of u, we consider a parametric exponential family on $[0,1]$.

1.2 Null hypothesis of correct specification

We wish to develop a procedure to detect whether the model and the distribution of observables are compatible. First we explain what we mean by *compatible.* We start by taking P, Γ and ν as given and by considering three natural formalizations of compatibility, a first representation based on measurable selections of Γ, the second based on the existence of a suitable probability measure with marginals P and ν and a third based on Dempster's notion of maximal plausibility.

1.2.1 Equilibrium selections

It is very easily understood in the simple case where the link Γ between latent and observable variables is parametric and Γ is measurable and single valued. Defining the image measure of P by Γ by

$$P\Gamma^{-1}(A) = P\{y \in \mathcal{Y} |\, \Gamma(y) \in A\}, \tag{6}$$

for all $A \in \mathcal{B}_{\mathcal{U}}$, we say that the model is well specified if and only if $\nu = P\Gamma^{-1}$. In the general case considered here, Γ may not be single valued, and its images may not even be disjoint (which would be the case if it was the inverse image of a single valued mapping from $\mathcal{U}$ to $\mathcal{Y}$, i.e. a traditional function from latent to observable variables). However, under a measurability assumption on Γ, we can construct an analogue of the image measure, which will now be a set $\mathrm{Core}(\Gamma, P)$ of Borel probability measures on $\mathcal{U}$ (to be defined below), and the hypothesis of *compatibility* of the restrictions on latent variable distributions and on the models linking latent and observable variables will naturally take the form

$$\mathrm{H}_0 : \nu \in \mathrm{Core}(\Gamma, P). \tag{7}$$

Assumption 1: Γ has non-empty and closed values, and for each open set $\mathcal{O} \subseteq \mathcal{U}$, $\Gamma^{-1}(\mathcal{O}) = \{y \in \mathcal{Y} \mid \Gamma(y) \cap \mathcal{O} \neq \varnothing\} \in \mathcal{B}_{\mathcal{Y}}$.

To relate the present case to the intuition of the single-valued case, it is useful to think in terms of single-valued *selections* of the multi-valued mapping Γ. A measurable selection γ of Γ is a measurable function such that $\gamma(y) \in \Gamma(y)$ for all $y \in \mathcal{Y}$. The set of measurable selections of a multi-valued mapping Γ that satisfies Assumption 1 is denoted Sel(Γ)[2]. To each selection γ of Γ, we can associate the image measure of P, denoted $P\gamma^{-1}$, defined as in (6).

It would be tempting to reformulate the compatibility condition as the requirement that at least one selection γ in Sel(Γ) is such that $\nu = P\gamma^{-1}$. However, such a requirement implies that γ corresponds to the equilibrium that is always selected. Under such a requirement, if for a given observable value the model does not specify which value of the latent variables gave rise to it, the latter is nonetheless fixed. Hence two identical observed realizations in the sample of observations necessarily arose from the same realization of the latent variables. We argue, however, that if the model does not specify an equilibrium selection mechanism, there is no reason to assume that each observation is drawn from the same equilibrium.

Allowing endogenous equilibrium selection of unknown form is equivalent to allowing the existence of an arbitrary distribution on the set of $P\gamma^{-1}$ when γ spans Sel(Γ) (as opposed to a mass on one particular $P\gamma^{-1}$). A Bayesian formulation of the problem would entail a specification of this distribution. Here, we stick to the given model specification in leaving it completely unspecified[3].

Hence, we argue that the correct reformulation of the compatibility condition is that ν can be written as a mixture of probability measures of the form $P\gamma^{-1}$, where γ ranges over Sel(Γ). However, as the following example show, even for

[2] It is known to be non-empty since Rokhlin (1949) Part I, §2, N^o 9, Lemma 2. The commentary at the end of chapter 14 of Rockafellar and Wets (1998) sheds light on the controversy surrounding this attribution.

[3] See the first paragraph of section 5.A. of Jovanovic (1989) for a discussion of this issue.

the simplest multi-valued mapping, the set of measurable selections is very rich, let alone the set of their mixtures.

Example: Consider the multi-valued mapping

$$\Gamma : \; [0,1] \rightrightarrows [0,1]$$

defined by $\Gamma(x) = \{0, x\}$ for all x. The collection of measurable selections of Γ is indexed by the class of Borel subsets of $[0,1]$. Indeed, a representative measurable selection of Γ is γ_B, such that $\gamma_B(x) = x\{x \in B\}$ for any Borel subset B of $[0,1]$, where $\{x \in B\}$ denotes the indicator function which equals one when $x \in B$ and zero otherwise.

Hence, it will be imperative to give manageable equivalent representations of such a mixture, as is done in Theorem 1 below.

1.2.2 Existence of a suitable joint probability

The second natural representation of compatibility of the distribution P of observables and the model (Γ, ν) is based on the existence of probability measures on the product $\mathcal{Y} \times \mathcal{U}$ that admit P and ν as marginals.

In the benchmark case of $\Gamma = \gamma$ one-to-one, the model imposes a stringent constraint on pairs (y, u), namely that $u = \gamma(y)$. So the admissible region of the product space is the graph of γ, i.e. the set

$$\text{Graph}\,\gamma = \{(y, u) \in \mathcal{Y} \times \mathcal{U} : \; u \in \gamma(y)\}.$$

The compatibility condition described above, namely $P\gamma^{-1} = \nu$ is equivalent to the existence of a probability measure on the product space that is supported by Graph γ (i.e. that gives probability zero outside the constrained region defined by the model) and admits P and ν as marginals.

This generalizes immediately to the case of Γ multi-valued, as the existence of a probability measure that admits P and ν as marginals, and that is supported on the constrained region

$$\text{Graph}\,\Gamma = \{(y, u) \in \mathcal{Y} \times \mathcal{U} : \; u \in \Gamma(y)\},$$

in other words, a probability measure that admits P and ν as marginals and gives probability zero to the event $U \notin \Gamma(Y)$, where U and Y are random elements with probability law ν and P respectively.

1.2.3 Dempster plausibility

Dempster (1967) suggests to consider the smallest reliability that can be associated with the event $B \in \mathcal{B}_{\mathcal{U}}$ as the *belief function*

$$\underline{P}(A) = P\{y \in \mathcal{Y} \mid \Gamma(y) \subseteq B\}$$

and the largest plausibility that can be associated with the event B as the *plausibility function*

$$\overline{P}(A) = P\{y \in \mathcal{Y} \mid \Gamma(y) \cap B \neq \varnothing\}$$

the two being linked by the relation

$$\overline{P}(A) = 1 - \underline{P}(A^c), \tag{8}$$

which prompted some authors to call them *conjugates* or *dual* of each other[4].

A natural way to construct a set of probability measures is to consider all probability measures that do not exceed the largest plausibility that can be associated with a set, and that, as a result of (8), are larger than the smallest reliability associated with a set. We thus form the *core* of the belief function[5]:

$$\begin{aligned} \text{Core}(\Gamma, P) &= \{\mu \in \Delta(\mathcal{U}) \mid \forall B \in \mathcal{B}_{\mathcal{U}},\, \mu(B) \geq \underline{P}(B)\} \\ &= \{\mu \in \Delta(\mathcal{U}) \mid \forall B \in \mathcal{B}_{\mathcal{U}},\, \mu(B) \leq \overline{P}(B)\} \end{aligned}$$

where the first equality can be taken as a definition, and the second follows immediately from (8). It is well known that $\text{Core}(\Gamma, P)$ is non-empty, and another natural representation of the compatibility of the distribution P of observables with the model (Γ, ν) is that ν belongs to $\text{Core}(\Gamma, P)$, in other words, that ν satisfies $\nu(B) \leq P(\{y \in \mathcal{Y} :\ \Gamma(y) \cap B \neq \varnothing\})$ for all $B \in \mathcal{B}_{\mathcal{U}}$.

[4] Matheron (1975) gave the first full formalization of the objects introduced by Dempster (1967).

[5] The name Core is standard in the literature to denote the set of probability measures satisfying (11). It seems to originate from D. Gillies' 1953 Princeton PhD thesis on "some theorems on n-person games." For finite sets, the core is non-empty by the Bondareva-Shapley theorem. In the present more general context, the non-emptiness of the core will follow from the equivalence of (i) and (iv) of Theorem 1 below, and the existence of measurable selections of Γ under assumption 1.

1.2.4 Equivalence of compatibility representations

The following theorem shows that the three representations discussed above are, in fact, equivalent. In addition, two more equivalent formulations are presented that will be used in the empirical formulations in the next section.

Theorem 1: Under assumption 1, the following statements are equivalent:

(i) ν is a mixture of images of P by measurable selections of Γ, (i.e. ν is in the weak closed convex hull of $\{P\gamma^{-1};\ \gamma \in \text{Sel}(\Gamma)\}$).

(ii) There exists for P-almost all $y \in \mathcal{Y}$ a probability measure $\pi_\nu(y,.)$ on $\mathcal{U}$ with support $\Gamma(y)$, such that

$$\nu(B) = \int_{\mathcal{Y}} \pi_\nu(y,B)\ P(dy), \text{ all } B \in \mathcal{B}_{\mathcal{U}}. \tag{9}$$

(iii) If U and Y are random elements with respective distributions P and ν, there exists a probability measure $\pi \in \mathcal{M}(P,\nu)$ that is supported on the admissible region, i.e. such that

$$\pi(U \notin \Gamma(Y)) = 0. \tag{10}$$

(iv) The probability assigned by ν to an event in $B \in \mathcal{B}_{\mathcal{U}}$ is no greater than the largest plausibility associated with B given P and Γ, i.e.

$$\nu(B) \leq P(\{y \in \mathcal{Y} :\ \Gamma(y) \cap B \neq \varnothing\}) \tag{11}$$

(v) For all $A \in \mathcal{B}_{\mathcal{Y}}$, we have

$$P(A) \leq \nu(\Gamma(A)). \tag{12}$$

Remark 1: The weak topology on $\Delta(\mathcal{U})$, the set of probability measures on $\mathcal{U}$, is the topology of convergence in distribution. $\Delta(\mathcal{U})$ is also Polish, and the weak closed convex hull of $\{P\gamma^{-1};\ \gamma \in \text{Sel}(\Gamma)\}$ is indeed the collection of arbitrary mixtures of elements of $\{P\gamma^{-1};\ \gamma \in \text{Sel}(\Gamma)\}$. This is a continuous version of the Birkhoff-von Neumann theorem on doubly stochastic matrices.

Remark 2: A version of representation (ii) is used in Wasserman (1990) to construct prior envelopes. Notice that (9) looks like a disintegration of ν[6],

[6] See section 3 page 116 of Pollard (2002)

and indeed, when Γ is the inverse image of a single-valued measurable function (i.e. when the model is given by a single-valued measurable function from latent to observable variables), the probability kernel π_ν is exactly the (P, Γ^{-1})-disintegration of ν, in other words, $\pi_\nu(y, .)$ is the conditional probability measure on $\mathcal{U}$ under the condition $\Gamma^{-1}(u) = \{y\}$. Hence (9) has the interpretation that a random element with distribution ν can be generated as a draw from $\pi_\nu(y, .)$ where y is a realization of a random element with distribution P.

Remark 3: We define Core(Γ, P) as the weak convex-hull of $\{P\gamma^{-1};\ \gamma \in \text{Sel}(\Gamma)\}$, or equivalently as the set of all mixtures of images of P by measurable selections of Γ. So our null hypothesis (7) is well defined.

Remark 4: Representations (iii) and (iv) are alternative natural formulations of the compatibility of the model with the observations. Representation (iii) is the existence of a probability that "lives" in the admissible region of $\mathcal{Y} \times \mathcal{U}$, and representation (iv) is a formulation of the Dempster plausibility condition (see Dempster (1967)). Hence the equivalence with representation (i) is a very desirable result.

Remark 5: As will be explained later, our test statistic will be based on violations of representation (v), which is the dual formulation of (iii) seen as a Monge-Kantorovich optimal mass transportation solution[7].

Remark 6: Equivalence of (i) and (iii) is a generalization of proposition 1 of Jovanovic (1989) to the case where P is not necessarily atomless and $\mathcal{U}$ not necessarily compact. Notice that relative to Jovanovic (1989), the roles of $\mathcal{Y}$ and $\mathcal{U}$ are reversed for the purposes of specification testing. As discussed in the second remark following proposition 1 mentioned above, atomlessness of the distribution of latent variables is innocuous as long as $\mathcal{U}$ is rich enough. However, atomlessness of the distribution of observables isn't innocuous, since it rules out many of the relevant applications.

[7] $\nu\Gamma$ is a capacity functional, and hence is alternating of order ∞ (see Choquet (1953)).

2 Empirical formulations

2.1 Empirical Monge-Kantorovich problem

In view of representation (iii) of Theorem 1, i.e. equation (10), the null can be reformulated as the following Monge-Kantorovich mass transportation problem

$$\min_{\pi \in \mathcal{M}(P,\nu)} \int_{\mathcal{Y}\times\mathcal{U}} \{u \notin \Gamma(y)\}\, \pi(dy, du) = 0, \tag{13}$$

where the transportation cost function is an indicator penalty for violation of the model (we adopt the Pollard convention and use the same notation for a set and its indicator function -see Pollard (2002)).

We now consider the empirical version of this Monge-Kantorovich problem, replacing P by the empirical distribution P_n to yield the functional

$$T^*(P_n, \Gamma, \nu) = \min_{\pi \in \mathcal{M}(P_n,\nu)} \int_{\mathcal{Y}\times\mathcal{U}} \{u \notin \Gamma(y)\}\, \pi(dy, du). \tag{14}$$

Next, we show a dual representation of the Monge-Kantorovich problem which will be instrumental in deriving our test statistic and its asymptotic properties:

Theorem 2: The following equalities hold:

$$\begin{aligned} T^*(P_n, \Gamma, \nu) &= \max_{f \oplus g \leq \varphi} \, (P_n f + \nu g) && (15) \\ &= \sup \, (P_n(A) - \nu(\Gamma(A))) \,, && (16) \end{aligned}$$

where $A \in \mathcal{B}_{\mathcal{Y}}$, $\varphi(y,u) = \{u \notin \Gamma(y)\}$, and $f \oplus g \leq \varphi$ signifies that the supremum is taken over all measureable functions f on $\mathcal{Y}$ and g on $\mathcal{U}$ such that for all (y,u), $f(y) + g(u) \leq \varphi(y,u)$.

2.2 Specification test statistic

We propose to adopt a test statistic based on the dual Monge-Kantorovich formulation (16), in other words a statistic that penalizes large values of (16). However, $T^*(P_n, \Gamma, \nu)$ seemingly involves checking condition (12) on all sets in $\mathcal{B}_{\mathcal{Y}}$, which renders computations infeasible, be it computation of the distribution of $T^*(P_n, \Gamma, \nu)$, or the computation of $T^*(P_n, \Gamma, \nu)$ itself. The objective of this section, therefore, is to elicit a reduced class of sets on which to check condition (12). Our test statistic is

$$T(P_n, \Gamma, \nu) = \sup_{A \in \mathcal{C}} \, (P_n(A) - \nu(\Gamma(A))) \tag{17}$$

and the class $\mathcal{C}$ needs to satisfy the following requirements:

(CD) $\mathcal{C}$ must be *convergence determining*, which in this case is equivalent to the property that the Choquet capacity $A \to \nu(\Gamma(A))$ is characterized by its values on all sets $A \in \mathcal{C}$. This ensures that we avoid

$$\text{limsup}\, T(P_n, \Gamma, \nu) \leq 0 \tag{18}$$

when actually $P(A) > \nu(\Gamma(A))$ for some $A \in \mathcal{B}_{\mathcal{Y}} \backslash \mathcal{C}$.

(DP) $\mathcal{C}$ must be P-Donsker, so that

$$\text{limsup}\, \sqrt{n}\, T(P_n, \Gamma, \nu) \leq \sup_{A \in \mathcal{C}} \mathbb{G}(A), \tag{19}$$

with $\mathbb{G}$ a P-Brownian bridge, provides us with a rejection region.

To be P-Donsker, the class $\mathcal{C}$ may not be too large, whereas it needs to be large enough to be convergence determining, so that the two requirements involve a formal trade-off.

We summarize the discussion above in the following theorem on asymptotic behaviour of the test statistic. First we need a definition:

Definition: (Γ, ν)**-unambiguous sets:** We call a set $A \in \mathcal{B}_{\mathcal{Y}}$ unambiguous with respect to Γ and ν when it satisfies

$$\nu(\Gamma(A)) = \nu(\{u \in \mathcal{U} :\ \Gamma^{-1}(u) \subseteq A\}).$$

It is shown in the appendix that the class $\mathcal{B}_0$ of (Γ, ν)-unambiguous sets is a σ-algebra. Hence, the definition above singles out the region of $\mathcal{Y}$ where the set function $A \to \nu(\Gamma(A))$ is actually a probability measure. Whatever the value of P, for any A in this class of sets $\mathcal{B}_0$, the null hypothesis will reduce to $P(A) = \nu(\Gamma(A))$. For any class $\mathcal{C}$ of subsets of $\mathcal{B}_{\mathcal{Y}}$, we denote $\mathcal{C}_0$ the class of unambiguous sets in $\mathcal{C}$, hence $\mathcal{C}_0 = \mathcal{C} \cap \mathcal{B}_0$.

We are now in a position to state our first result on the asymptotic behaviour of our test statistic.

Theorem 3: If $\mathcal{C}$ satisfies (DP) and the the null hypothesis H_0 holds, then the following hold almost surely:

$$\begin{aligned} \text{limsup}\sqrt{n}\, T(P_n, \Gamma, \nu) &\leq \sup_{A \in \mathcal{C}} \mathbb{G}(A) \\ \text{liminf}\sqrt{n}\, T(P_n, \Gamma, \nu) &\geq \sup_{A \in \mathcal{C}_b} \mathbb{G}(A) \end{aligned}$$

where $\mathbb{G}$ is a P-Brownian bridge and $\mathcal{C}_0 \subseteq \mathcal{C}_b \subseteq \mathcal{C}$. If $\mathcal{C}$ satisfies (CD) and the the alternative hypothesis H_a holds, then:

$$\sqrt{n}\,T(P_n, \Gamma, \nu) \rightarrow \infty$$

Theorem 4 gives bounds for the asymptotic behaviour of our test statistic. The upper bound is the supremum of a Brownian bridge on the whole class of sets $\mathcal{C}$, and the lower bound is the supremum of the same Brownian bridge over the unambiguous sets in $\mathcal{C}$. The reason is the following: The rescaled test statistic can be rewritten

$$\begin{aligned} \sqrt{n}T(P_n, \Gamma, \nu) &= \sqrt{n}\sup_{A\in\mathcal{C}}\ (P_n(A) - \nu(\Gamma(A))) \\ &= \sup_{A\in\mathcal{C}}\left(\mathbb{G}_n(A) + \sqrt{n}(P(A) - \nu(\Gamma(A)))\right). \end{aligned}$$

Define $\mathcal{C}_b$ the class of sets A in $\mathcal{C}$ such that $P(A) = \nu(\Gamma(A))$. For all $A \in \mathcal{C}\backslash\mathcal{C}_b$, we have a strict inequality, i.e. $P(A) - \nu(\Gamma(A)) < 0$, so that $\sqrt{n}(P(A) - \nu(\Gamma(A)) \rightarrow -\infty$, and those sets are not included in the supremum in the asymptotic expression. The only sets we are sure are included in $\mathcal{C}_b$ without knowledge of P are the unambiguous sets.

Obtaining an exact test requires taking the supremum over the class of sets $\mathcal{C}_b$ as opposed to $\mathcal{C}$. We propose a feasible version of this procedure using the estimator $\hat{\mathcal{C}}_b$ for $\mathcal{C}_b$ defined as follows:

$$\hat{\mathcal{C}}_b = \{A \in \mathcal{C} :\ P_n(A) > \nu(\Gamma(A)) - h_n\}$$

and h_n is a deterministic bandwidth sequence satisfying

$$h_n\sqrt{n} + \frac{1}{h_n} \rightarrow \infty.$$

Note that for a given A, there is a $\delta > 0$ such that

$$\begin{aligned} &\mathbb{P}(A \in (\mathcal{C}_b\backslash\hat{\mathcal{C}}_b) \cup (\hat{\mathcal{C}}_b\backslash\mathcal{C}_b)) \\ \leq\ \ &\mathbb{P}(P_n(A) \leq P(A) - h_n \text{ or } P_n(A) > P(A) + \delta - h_n), \end{aligned}$$

so that the latter display converges to zero, justifying the approximation of $\mathcal{C}_b$ by $\hat{\mathcal{C}}_b$.

2.3 Finding (CD) and (DP) classes:

We now discuss the determination of classes of sets that satisfy the requirements (CD) and (DP) discussed above. The following lemma (lemma 1.14 of Salinetti and Wets (1986)) provides a convergence determining class which, though still falling short of requirement (DP), allows us to restrict attention to the class of finite unions of balls with rational midpoints and positive rational radii. Define $\mathcal{C}_{\text{SW}}$ as the class of compact subsets of $\mathcal{Y}$ with the following two properties:

(C1) Elements of $\mathcal{C}_{\text{SW}}$ are finite unions of non-singleton rectangles with rational endpoints,

(C2) Elements of $\mathcal{C}_{\text{SW}}$ are continuity sets for the Choquet capacity

$$A \to \nu(\Gamma(A)),$$

in other words, if $A \in \mathcal{C}_{\text{SW}}$, then $\nu(\Gamma(\text{cl}(A))) = \nu(\Gamma(\text{int}(A)))$.

Then we have:

Lemma SW: The class $\mathcal{C}_{\text{SW}}$ is convergence determining.

Remark 1: The class $\mathcal{C}_{\text{SW}}$ is not a Vapnik-Červonenkis class of sets (hereafter VC-class) since for any finite collection of points, there is a collection of finite union of balls that shatters it (see section 2.6.1 page 134 of van der Vaart and Wellner (1996)). Though it does not follow that $\mathcal{C}_{\text{SW}}$ doesn't satisfy (DP), it seems unlikely.

2.3.1 Identified case (Γ is one-to-one):

In the identified case, where $\Gamma = \gamma$ is single-valued and one-to-one, consider the following classes of cells in $\mathbb{R}^{d_y}$:

$$\begin{aligned} \mathcal{C} &= \{(-\infty, y], (z, \infty) : \ (y, z) \in \mathbb{R}^{2d_y}\} \\ \tilde{\mathcal{C}} &= \{(-\infty, y] : \ y \in \mathbb{R}^{d_y}\}. \end{aligned}$$

Notice that

$$\sup_{A \in \mathcal{C}} \ (P_n(A) - \nu(\gamma(A))) = \sup_{A \in \tilde{\mathcal{C}}} \ |P_n(A) - \nu(\gamma(A))|$$

and the latter is the classical Kolmogorov-Smirnov specification test statistic.

2.3.2 Discrete observable distribution:

In case of observations taking values in a discrete subset $\mathcal{Y}_0$ of $\mathcal{Y}$ (as in the large class of entry models), it is immediately seen that the class $2_0^{\mathcal{Y}}$ of all subsets of the set of observable values satisfies (CD) and (DP). In view of the limiting behaviour of the test statistic, it will be seen below that in practice, the supremum is taken on a far smaller class of sets.

2.3.3 Γ is convex-valued:

Suppose $\mathcal{U}$ is a convex compact subset of $\mathbb{R}^{d_u}$ (or more generally that it has a vector structure compatible with its Polish topology) and ν is the uniform distribution. Define Graph Γ as follows:

$$\text{Graph}\,\Gamma = \{(y,u) \in \mathcal{Y} \times \mathcal{U} :\ u \in \Gamma(y)\}.$$

The class $\mathcal{C}$ defined above satisfies (DP) (see for instance example 2.5.4 page 129 of van der Vaart and Wellner (1996)). We show below that it is also (CD) when Γ has convex values and Graph Γ has monotone upper and lower envelopes, or when Graph Γ is convex.

Notice that the class $\mathcal{C}$ so defined is the same as the class on which the supremum is taken to form the Kolmogorov-Smirnov statistic in the identified case as described above. It will be shown below that this case extends to when Graph Γ is the union of an arbitrary collection of *bi-separated* convex elements $(G_i)_{i\in I}$, by which we mean that for all $i,j \in I$, $G_i^{\mathcal{Y}} \cap G_j^{\mathcal{Y}}$ and $G_i^{\mathcal{U}} \cap G_j^{\mathcal{U}}$ are singletons, where

$$\begin{aligned} G^Y &= \{y \in \mathcal{Y} :\ (y,\Gamma(y)) \in G\} \\ G^U &= \{u \in \mathcal{U} :\ (\Gamma^{-1}(u),u) \in G\} \end{aligned}$$

are the traces on $\mathcal{Y}$ and $\mathcal{U}$ respectively.

2.3.4 Γ has a connected graph:

In case Graph Γ is connected, $\mathcal{C}$ defined above does not satisfy (CD) any more, but we show below that the class of rectangles

$$\mathcal{S} = \{\, [y,z] :\ (y,z) \in \mathbb{R}^{2d_y}\}$$

does. In addition, it satisfies (DP). Indeed, if $d_y = 1$, its VC-index is three, since $\mathcal{S}$ can pick out the two elements of a set of cardinality 2, but can never pick out the subset $\{x, z\}$ of a set of three elements $\{x, y, z\}$. More generally, it can be shown that the VC-index of $\mathcal{S}$ is $2d_y + 1$ (see Example 2.6.1 page 135 of van der Vaart and Wellner (1996)).

2.3.5 Graph Γ has a finite number of connected components:

Let K be the number of connected components $G_1, \ldots, G_K$ of Graph Γ, so that graph $\Gamma = \bigcup_{k=1,\ldots,K} G_k$, and call G_k^Y the trace of G_k on $\mathcal{Y}$, i.e.

$$G_k^Y = \{y \in \mathcal{Y} : \ (y, \Gamma(y)) \in G_k\}.$$

We show below that the class

$$\mathcal{S}_K \quad = \quad \{ \bigcup_{k \leq K} [y_k, z_k] : \ (y_k, z_k) \in (G_k^Y)^2 \}$$

satisfies (CD). That it satisfies (DP) follows from lemma 2.6.17(iii) page 147 of van der Vaart and Wellner (1996) and the fact that it is contained in the K-iterated union $\mathcal{S} \sqcup \ldots \sqcup \mathcal{S}$, where the "square union" of two classes of sets $\mathcal{S}_1$ and $\mathcal{S}_2$ is defined by $\mathcal{S}_1 \sqcup \mathcal{S}_2 = \{A_1 \cup A_2 : \ A_1 \in \mathcal{S}_1, \, A_2 \in \mathcal{S}_2\}$.

2.3.6 Summary of the results:

We collect the results presented above in the following theorem:

Theorem 4: If $\mathcal{U}$ is a convex compact metrizable subset of $\mathbb{R}^{d_u}$, the following hold:

- $2^{\mathcal{Y}_0}$ satisfies (CD) and (DP) when P is supported on the finite set $\mathcal{Y}_0$.
- $\mathcal{C}$ satisfies (CD) and (DP) when Graph Γ is an arbitrary union of bi-separated convex sets and ν is uniform.
- ***CONJECTURE*** $\mathcal{S}_K$ satisfies (CD) and (DP) when Graph Γ is the union of K connected components.

2.3.7 Examples:

We consider the following parametric model specifications

$$\Gamma_\theta : \ [0,1] \rightrightarrows [0,1]$$

with θ a parameter in $(0,1]$, ν_θ is $U[0,1]$.

- $\Gamma_\theta(y) = [\theta y, y]$

  This a case with convex graph. Notice that if we check the condition $P(A) \leq \nu_\theta(\Gamma_\theta(A))$ on the cells of the form $[0,y]$, for all $y \in [0,1]$, we obtain $P([0,y]) \leq y$ for all y and if we check the condition $P(A) \leq \nu_\theta(\Gamma_\theta(A))$ on the cells of the form $[y,1]$, for all $y \in [0,1]$, we obtain $P([y,1]) \geq \theta y$ for all y.

- $\Gamma_\theta(y) = y \;\text{ if }\; y > \theta \;\text{ and }\; [0,\theta] \;\text{ otherwise.}$

  This is a case where Graph Γ has two bi-separated convex components. Notice that if we check the condition $P(A) \leq \nu_\theta(\Gamma_\theta(A))$ on the cells of the form $[0,y]$, for all $y \in [0,1]$, we obtain $P([0,y]) \leq \theta$ for all $y \leq \theta$ and $P([0,y]) \leq y$ for all $y \geq \theta$. Now if we check the condition $P(A) \leq \nu_\theta(\Gamma_\theta(A))$ on the cells of the form $[y,1]$, for all $y \in [0,1]$, we obtain $P([y,1]) = 1$ for all $y \leq \theta$ and $P([y,1]) \geq y$ for all $y \geq \theta$. Hence, the proposed test is equivalent to a Kolmogorov-Smirnov test restricted to $[\theta,1]$.

- $\Gamma_\theta(y) = \{\theta y, y\}$

  This is a case where Graph Γ is connected. The image of an interval $[a,b]$ is $[\theta a, \theta b] \cup [a,b]$. For any (a,b) such that $\theta b \leq a \leq b$, we need to check $P([a,b]) \leq (1+\theta)(b-a)$ and for each (a,b) such that $a \leq \theta b$, we need to check $P([a,b]) \leq b - \theta a$.

- $\Gamma_\theta(y) = 0 \;\text{ if }\; \theta/2 < y < \theta \;\text{ and }\; [\theta y, y] \;\text{ otherwise.}$

  This is a case where Graph Γ has three connected components. For (a,b,c,d,e,f) such that $a \leq b \leq \theta/2 \leq c \leq d \leq \theta \leq e \leq f$, we need to check that $P([a,b]) + P([c,d]) + P([e,f]) \leq b - \theta a + f - \theta e$.

3 Implementation of the test

Conclusion

Appendix: Proofs of the results in the main text

Proof of Theorem 1:

(i) $\Longleftrightarrow$ (iv) $\Longleftrightarrow$ (ii)
Call $\Delta(B)$ the set of all Borel probability measures with support B. Under Assumption 1, the map $y \mapsto \Delta(\Gamma(y))$ is a map from Y to the set of all non-empty convex sets of Borel probability measures on $\mathcal{U}$ which are closed with respect to the weak topology. Moreover, for any $f \in C_b(\mathcal{U})$, the set of all continuous bounded real functions on $\mathcal{U}$, the map

$$y \longmapsto \sup\left\{\int f d\mu : \mu \in \Delta(\Gamma(y))\right\} = \max_{u\in\Gamma(y)} f(u)$$

is $\mathcal{B}_Y$-measurable, so that, by Theorem 3 of Strassen (1965), for a given $\nu \in \Delta(\mathcal{U})$, there exists π satisfying (9) with $\pi(y,.) \in \Delta(\Gamma(y))$ for P-almost all y if and only if

$$\int_{\mathcal{U}} f(u)\nu(du) \leq \int_{\mathcal{U}} \sup_{u\in\Gamma(y)} f(u)P(dy) \tag{20}$$

for all $f \in C_b(\mathcal{U})$. Now, defining $\overline{P}$ as the set function

$$\overline{P}:\ B \to P(\{y \in Y:\ \Gamma(y)\cap B \neq \varnothing\}),$$

the right-hand side of (20) is shown in the following sequence of equalities to be equal to the integral of f with respect to $\overline{P}$ in the sense of Choquet (line (21) below can be taken as a definition).

$$\begin{aligned}
& \int_Y \sup_{u\in\Gamma(y)} \{f(u)\}\ dP(y) \\
= & \int_0^\infty P\big\{y \in Y : \sup_{u\in\Gamma(y)} \{f(u)\} \geq x\big\}\,\mathrm{d}x \\
& \quad + \int_{-\infty}^0 (P\big\{y \in Y : \sup_{u\in\Gamma(y)} \{f(u)\} \geq x\big\} - 1)\,\mathrm{d}x \\
= & \int_0^\infty P\big\{y \in Y : \Gamma(y) \subseteq \{f \geq x\}\big\}\,\mathrm{d}x
\end{aligned}$$

$$+\int_{-\infty}^{0} (P\{y \in Y : \Gamma(y) \subseteq \{f \geq x\}\} - 1)\,\mathrm{d}x$$

$$= \int_{0}^{\infty} \overline{P}(\{f \geq x\})\,\mathrm{d}x + \int_{-\infty}^{0} (\overline{P}(\{f \geq x\}) - 1)\,\mathrm{d}x = \int_{\mathrm{Ch}} f\,\mathrm{d}\overline{P}. \quad (21)$$

By Theorem 1 of Castaldo, Maccheroni, and Marinacci (2004), for any $f \in C_b(\mathcal{U})$,

$$\int_{\mathrm{Ch}} f\,\mathrm{d}\overline{P} = \max_{\gamma \in \mathrm{Sel}(\Gamma)} \int_{\mathcal{U}} f(u) P\gamma^{-1}(du),$$

so that (20) is equivalent to

$$\max_{\gamma \in \mathrm{Sel}(\Gamma)} \int_{\mathcal{U}} f(u) P\gamma^{-1}(du) \geq \int_{\mathcal{U}} f(u)\nu(du) \quad (22)$$

for any $f \in C_b(\mathcal{U})$. If ν is in the weak closure of the set of convex combinations of elements of $\{P\gamma^{-1} : \gamma \in \mathrm{Sel}(\Gamma)\}$, then by linearity of the integral and the definition of weak convergence, (22) holds. Conversely, if ν satisfies (22), then it satisfies

$$\int_{\mathrm{Ch}} f\,\mathrm{d}\overline{P} \geq \int_{\mathcal{U}} f(u)\nu(du)$$

and by monotone continuity, we have for all $A \in \mathcal{B}_{\mathcal{U}}$, and I_A the indicator function,

$$\int_{\mathcal{U}} I_A(u)\nu(du) \leq \int_{\mathrm{Ch}} I_A d\overline{P}.$$

Hence $\nu(A) \leq \overline{P}(A)$ for all $A \in \mathcal{B}_{\mathcal{U}}$, which by Corollary 1 of Castaldo, Maccheroni, and Marinacci (2004) implies that ν is the weak limit of a sequence of convex combinations of elements of $\{P\gamma^{-1} : \gamma \in \mathrm{Sel}(\Gamma)\}$, hence it is a mixture in the desired sense and the proof is complete.

(iii) $\Longleftrightarrow$ (iv) $\Longleftrightarrow$ (v)
Using theorem 2 below, it suffices to show that (11) is equivalent to $\nu(\Gamma(A)) \geq P(A)$ for all $A \in \mathcal{B}_{\mathcal{Y}}$. As previously, define $\overline{P}$ as the set function on $\mathcal{B}_{\mathcal{U}}$

$$\overline{P} :\; B \rightarrow P(\{y \in Y :\; \Gamma(y) \cap B \neq \varnothing\}).$$

Define also $\underline{P}$ as the set function

$$\overline{P} :\; B \rightarrow P(\{y \in Y :\; \Gamma(y) \subseteq B\}).$$

Since $\overline{P}(B) = 1 - \underline{P}(B^c)$, we have the well known equivalence between $\nu(B) \leq \overline{P}(B)$ for all $B \in \mathcal{B}_{\mathcal{U}}$ and $\nu(B) \geq \underline{P}(B)$ for all $B \in \mathcal{B}_{\mathcal{U}}$. In particular, for $B = \Gamma(A)$ for any $A \in \mathcal{B}_{\mathcal{Y}}$, we have $\nu(B) \subseteq \{y \in \mathcal{Y} : \Gamma(y) \subseteq \Gamma(A)\}$. As $A \subseteq \{y \in \mathcal{Y} : \Gamma(y) \subseteq \Gamma(A)\}$, we have $\nu(\Gamma(A)) \geq P(B)$. Conversely, for some $B \in \mathcal{B}_{\mathcal{U}}$, call $B_* = \{y \in \mathcal{Y} : \Gamma(y) \subseteq B\}$. Then, we have $P(B_*) \leq \nu(\Gamma(B_*))$. The result follows from the observation that $\Gamma(B_*) \subseteq B$.

Lemma 1:

If $\varphi : \mathcal{Y} \times \mathcal{U} \to \mathbb{R}$ is bounded, non-negative and lower semicontinuous, then

$$\inf_{\pi \in \mathcal{M}(P,\nu)} \pi\varphi = \sup_{f \oplus g \leq \varphi} (Pf + \nu g)$$

Proof of Lemma 1:

It can be shown to be a special case of corollary (2.18) of Kellerer (1984); however, a direct proof is more transparent, so we give it here for completeness. The left-hand side is immediately seen to be always larger than the right-hand side, so we show the reverse inequality.

[a] case where φ is continuous and $\mathcal{U}$ and $\mathcal{Y}$ are compact[8].
Call G the set of functions on $\mathcal{Y} \times \mathcal{U}$ strictly dominated by φ and call H the set of functions of the form $f + g$ with f and g continuous functions on $\mathcal{Y}$ and $\mathcal{U}$ respectively. Call $s(c) = Pf + \nu g$ for $c \in H$. It is a well defined linear functional, and is not identically zero on H. G is convex and sup-norm open. Since φ is continuous on the compact $\mathcal{Y} \times \mathcal{U}$, we have

$$s(c) \leq \sup f + \sup g < \sup \varphi$$

for all $c \in G \cap H$, which is non empty and convex. Hence, by the Hahn-Banach theorem, there exists a linear functional η that extends s on the space of continuous functions such that

$$\sup_{G} \eta = \sup_{G \cap H} s.$$

By the Riesz representation theorem, there exists a unique finite non-negative measure π on $\mathcal{Y} \times \mathcal{U}$ such that $\eta(c) = \pi c$ for all continuous c. Since $\eta = s$ on

[8] This is lemma 11.8.5 of Dudley (2003). R. Dudley credits it to a private communication from J. Neveu in 1974. The proof given here fort completeness is due to N. Belili.

H, we have

$$\begin{aligned}\int_{\mathcal{Y}\times\mathcal{U}} f(y)\, d\pi(y,u) &= \int_{\mathcal{Y}} f(y)\, dP(y)\\ \int_{\mathcal{Y}\times\mathcal{U}} g(u)\, d\pi(y,u) &= \int_{\mathcal{Y}} g(u)\, d\nu(y),\end{aligned}$$

so that $\pi \in \mathcal{M}(P,\nu)$ and

$$\sup_{f\oplus g\leq\varphi} (Pf+\nu g) = \sup_{G\cap H} s = \sup_{H} \eta = \pi\varphi.$$

[b] $\mathcal{Y}$ and $\mathcal{U}$ are not necessarily compact, and φ is continuous.

For all $n > 0$, there exists compact sets K_n and L_n such that

$$\max\left(P(\mathcal{Y}\backslash K_n), \nu(\mathcal{U}\backslash L_n)\right) \leq \frac{1}{n}.$$

Let (a,b) be an element of $\mathcal{Y}\times\mathcal{U}$ and define two probability measures μ_n and ν_n with compact support by

$$\begin{aligned}\mu_n(A) &= P(A\cap K_n) + P(A\backslash K_n)\delta_a(A)\\ \nu_n(B) &= \nu(B\cap L_n) + \nu(B\backslash L_n)\delta_b(B),\end{aligned}$$

where δ denotes the Dirac measure. By [a] above, there exists π_n with marginals μ_n and ν_n such that

$$\pi_n\varphi \leq \sup_{f\oplus g\leq\varphi} (Pf+\nu g) + \frac{\varphi(a,b)}{n}.$$

Since (π_n) has weakly converging marginals, it is weakly relatively compact. Hence it contains a weakly converging subsequence with limit $\pi \in \mathcal{M}(P,\nu)$. By Skorohod's almost sure representation (see for instance theorem 11.7.2 page 415 of Dudley (2003)), there exists a sequence of random variables X_n on a probability space $(\Omega,\mathcal{A},\mathbb{P})$ with law π_n and a random variable X_0 on the same probability space with law π such that X_0 is the almost sure limit of (X_n). By Fatou's lemma, we then have

$$\begin{aligned}\text{liminf}\,\pi_n\varphi &= \text{liminf}\mathbb{E}\varphi(X_n)\\ &\geq \mathbb{E}\text{liminf}\varphi(X_n)\\ &= \mathbb{E}\varphi(X_0)\\ &= \pi\varphi.\end{aligned}$$

Hence we have the desired result.

[c] General case.

φ is the pointwise supremum of a sequence of continuous bounded functions, so the result follows from upward σ-continuity of both $\inf_{\pi\in\mathcal{M}(P,\nu)}\pi\varphi$ and $\sup_{f\oplus g\leq\varphi}(Pf+\nu g)$ on the space of lower semicontinuous functions, shown in propositions (1.21) and (1.28) of Kellerer (1984).[9]

Proof of Theorem 2:

Under assumption 1, Γ is closed valued, hence $\varphi(y,u)=\{u\notin\Gamma(y)\}$ is lower semicontinuous and (15) is a direct application of lemma 1 above. Proposition (3.3) page 424 of Kellerer (1984) shows that in the case where φ is the indicator function of a set D, the right-and side of (15) specializes to

$$\sup_{A\times B\subseteq D}(P(A)-1+\nu(B)).$$

For $D=\{(y,u):u\notin\Gamma(y)\}$, $A\times B\subseteq D$ means that if $y\in A$ and $u\in B$, then $u\notin\Gamma(y)$. In other words $u\in B$ implies $u\notin\Gamma(B_1)$, which can be written $B\subseteq\Gamma(A)^c$. Hence, the dual problem can be written

$$\sup_{\Gamma(A)\subseteq B^c}(P(A)-1+\nu(B))=\sup_{\Gamma(A)\subseteq B}(P(A)-\nu(B)).$$

and (16) follows immediately.

Proof of Theorem 3:

The convergence of the empirical process $\mathbb{G}_n$ to the P-Brownian bridge uniformely in $l^\infty(\mathcal{F})$, where $\mathcal{F}$ is the class of indicator functions of sets in $\mathcal{C}$ follows immediately from property (CD), so the convergence of the supremum of $\mathbb{G}$ follows from the continuous mapping theorem.

Proof of Theorem 4:

- Proof of property (DP):
  We have already shown in the main text that any finite class of sets, $\mathcal{C}$ and $\mathcal{S}_K$ are Vapnik-$\breve{C}$ervonenkis classes of sets (for a definition, see 2.6.1 page

[9] The duality result can be extended to Borel functions using Choquet's capacitability theorem (Choquet (1959)).

134 of van der Vaart and Wellner (1996)[10]. Call $\mathcal{F}$ the class of indicator functions of sets in $\mathcal{C}$ or $\mathcal{S}_K$, and call $V(\mathcal{F})$ the Vapnik-$\breve{C}$ervonenkis index of the corresponding class of sets. By Theorem 2.6.4 page 136, there exists a constant C such that for all probability measure Q and all $0 < \varepsilon < 1$, the covering number (see definition 2.2.3 page 98 of van der Vaart and Wellner (1996)) of $\mathcal{F}$ in $\mathbb{L}_2(Q)$ metric, $\mathrm{N}(\varepsilon, \mathcal{F}, \mathbb{L}_2(Q))$ satisfy

$$\mathrm{N}(\varepsilon, \mathcal{F}, \mathbb{L}_2(Q)) \leq C(V(\mathcal{F}))(4e)^{V(\mathcal{F})}(1/\varepsilon)^{2(V(\mathcal{F})-1)}.$$

Hence, we have

$$\int_0^\infty \sup_Q \sqrt{\ln \mathrm{N}(\varepsilon, \mathcal{F}, \mathbb{L}_2(Q))}\, d\varepsilon < \infty.$$

Since $\mathcal{F}$ is a class of indicator functions, the above suffices to satisfy conditions of Theorem 2.5.2 page 127 of van der Vaart and Wellner (1996), and $\mathcal{F}$ is P-Donsker, i.e. $\mathcal{C}$ and $\mathcal{S}_K$ satisfy (DP).

- Proof of property (CD): to be added.

References

Andrews, D., S. Berry, and P. Jia (2004): "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location," unpublished manuscript.

Beresteanu, A., and F. Molinari (2006): "Asymptotic properties for a class of partially identified models," unpublished manuscript.

Castaldo, A., F. Maccheroni, and M. Marinacci (2004): "Random sets and their distributions," *Sankhya (Series A)*, 66, 409–427.

Chernozhukov, V., H. Hong, and E. Tamer (2002): "Inference on Parameter Sets in Econometric Models," unpublished manuscript.

Choquet, G. (1953): "Théorie des Capacités," *Annales de l'Institut Fourier*, 5, 131–295.

[10] See also Vapnik (1998).

Choquet, G. (1959): "Forme abstraite du théorème de capacitabilité," *Annales de l'Institut Fourier*, 9, 83–89.

Dempster, A. P. (1967): "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, 38, 325–339.

Dudley, R. (2003): *Real Analysis and Probability.* Cambridge University Press.

Gouriéroux, C., A. Holly, and A. Monfort (1982): "Likelihood ratio test, Wald test and Kuhn-Tucker test in linear models with inequality constraints on the regression parameter," *Econometrica*, 50, 63–81.

Heckman, J., and E. Vytlacil (2001): "Instrumental variables, selection models and tight bounds on the average treatment effect," in *Econometric Evaluations of Labour Market Policies, Lechner, M., and F. Pfeiffer, eds.*, pp. 1–16. Heidelberg: Springer-Verlag.

Imbens, G., and C. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1859.

Jovanovic, B. (1989): "Observable implications of models with multiple equilibria," *Econometrica*, 57, 1431–1437.

Kellerer, H. (1984): "Duality theorems for marginal problems," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67, 399–432.

King, A. (1989): "Generalized delta theorems for multi-valued mappings and measurable selections," *Mathematics of Operations Research*, 14, 720–736.

Manski, C. (2005): "Partial identification in econometrics," forthcoming in the *New Palgrave Dictionary of Economics, 2nd Edition.*

Matheron, G. (1975): *Random Sets and Integral Geometry.* New York: Wiley.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2004): "Moment Inequalities and Their Application," unpublished manuscript.

Pollard, D. (2002): *A User's guide to measure theoretic probability.* Cambridge University Press.

Rockafellar, R. T., and R. J.-B. Wets (1998): *Variational Analysis.* Berlin: Springer.

ROKHLIN, V. (1949): "Selected topics from the metric theory of dynamical systems," *Uspekhi Matematicheskikh Nauk*, 4, 57–128, translated in *American Mathematical Society Transactions* 49(1966), 171-240.

SALINETTI, G., and R. WETS (1986): "On the convergence in distribution of measurable multifunctions (random sets), normal integrands, stochastic processes and stochastic infima," *Mathematics of Operations Research*, 11, 385–422.

SHAIKH, A. (2005): "Inference for a Class of Partially Identified Econometric Models," unpublished manuscript.

SHAIKH, A., and E. VYTLACIL (2005): "Threshhold crossing models and bounds on treatment effects: a nonparametric analysis," unpublished manuscript.

STRASSEN, V. (1965): "The existence of probability measures with given marginals," *Journal of Mathematical Statistics*, 36, 423–439.

VAN DER VAART, A., and J. WELLNER (1996): *Weak Convergence and Empirical Processes.* New York: Springer.

VAPNIK, V. (1998): *Statistical Learning Theory.* New York: Wiley.

WASSERMAN, L. (1990): "Prior envelopes based on belief functions," *Annals of Statistics*, 18, 454–464.