

# Inference in Incomplete Models

Alfred Galichon and Marc Henry

Harvard University and Columbia University

First draft: September 15, 2005

This draft<sup>1</sup>: May 26, 2006

## Abstract

We provide a test for the specification of a structural model without identifying assumptions. We show the equivalence of several natural formulations of correct specification, which we take as our null hypothesis. From a natural empirical version of the latter, we derive a Kolmogorov-Smirnov statistic for Choquet capacity functionals, which we use to construct our test. We derive the limiting distribution of our test statistic under the null, and show that our test is consistent against certain classes of alternatives. 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 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, Core, selections, plausibility constraint, Monge-Kantorovich mass transportation problem, Kolmogorov-Smirnov test for capacity functionals.

---

<sup>1</sup>This research was carried out while the first author was visiting the Bendheim Center for Finance, Princeton University and financial support from NSF grant SES 0350770 to Princeton University, from the Program for Economic Research at Columbia University and from the Conseil Général des Mines is gratefully acknowledged. The authors also thank Gary Chamberlain, Xiaohong Chen, Victor Chernozhukov, Pierre-André Chiappori, Ronald Gallant, Peter Hansen, Han Hong, Guido Imbens, Michael Jansson, Massimo Marinacci, Rosa Matzkin, Francesca Molinari, Ulrich Müller, Alexei Onatski, Ariel Pakes, Victor de la Peña, Jim Powell, Peter Robinson, Bernard Salanié, Thomas Sargent, José Scheinkman, Jay Sethuraman, Azeem Shaikh, Chris Sims, Kyungchul Song and Edward Vytlacil and seminar participants at Berkeley, Columbia, École polytechnique, Harvard, MIT, NYU, Princeton, SAMSI and Stanford for helpful comments (with the usual disclaimer). Correspondence addresses: Department of Economics, Harvard University, Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, USA. galichon@fas.harvard.edu and 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 assumptions 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 many-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 many-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: 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 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 sub-sampling 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. 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 (2003) 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.

Finally, other papers considering inference in partially identified models include Shaikh and Vytlacil (2005), Magnac and Maurin (2005) and Blundell, Browning, and Crawford (2005).

The inference procedure proposed here is in the same spirit as the Andrews, Berry, and Jia (2003) 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.

After a prelude to expound the ideas developed here in the familiar case of Kolmogorov-Smirnov specification testing, the general set-up is described (with some examples) in section 1. It comprises the specification of a structure (in the Koopmans terminology) with observable and unobservable variables (unobservable to the analyst but not necessarily

to the economic agents) related by a many-to-many mapping as opposed to the one-to-one mapping required for identification. The structure is defined by the many-to-many mapping (which can comprise rationality constraints as before, as well as any constraints that are plausible within the theory) and a hypothesized distribution for the unobserved variables. To fix ideas, we call  $\Gamma$  the many-to-many mapping defining the structure,  $\nu$  a hypothesized distribution of unobservables and  $P$  the true distribution of observables.

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

Second, in section 2, we show that the characterizations of correct specification of the structure 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  $\nu$  with zero-one cost associated with violation of the constraints implied by  $\Gamma$ . 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  $P_n$ ). Empirical formulations pertaining to the different characterizations of correct specification of the structure are compared, and several are found to be equivalent, whereas others differ according to the choice of probability metric. 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  $\Gamma$  is one-to-one.

The properties of this statistic are examined in section 3. The classical Kolmogorov-Smirnov statistic tests the equality of two probability measures by checking their difference on a *good* class of sets (large enough to be convergence-determining, but small enough to allow asymptotic treatment). 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 (2003) approach. Hence the dual empirical problem provides us with a computable test

statistic, and a distribution to compare it to, and a parallel with the classical case.

We derive the asymptotic distribution of our test statistic and describe how classes of alternatives against which our test has power are related to what we call core-determining classes of sets.

Finally, the fourth 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  $\Gamma$  and  $\nu$  are specified in parametric form. 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. 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 the end of providing confidence regions that cover the elements of the identified set with pre-determined probability, or confidence regions that cover the identified set itself.

At the end of this section, we discuss semi-nonparametric extensions of our approach to include models which do not specify a parametric family of hypothesized data generating processes for the unobservable variables. This includes as a special case models defined by moment inequalities, the full treatment of which is the subject of the companion paper Galichon and Henry (2006).

The last section of the main text concludes; whereas 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.

For ease of notation, 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” structure as a pair  $(\nu, \gamma)$ , where

$\nu$  is a data generating process for the unobservables, and  $\gamma$  is a bijection from the set of observables to the set of unobservables, as in figure 1.

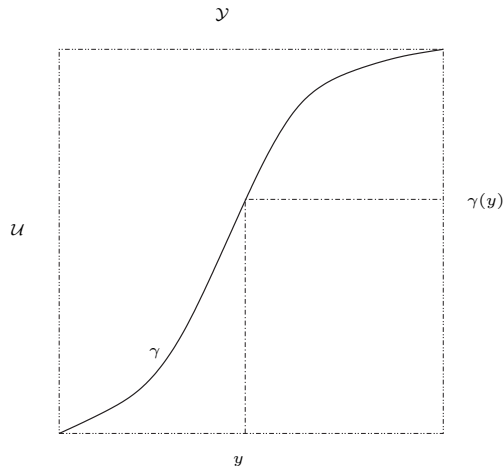


Figure 1: Bijective structure

If we call  $P$  the true data-generating process for the observables, we say that the complete structure is well specified if  $P(A) = \nu(\gamma(A))$  for all Borel sets  $A$ , 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

$$\sup_{A \in \mathcal{C}} (P(A) - \nu(\gamma(A))) = 0 \quad (1)$$

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

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

$$\inf_{\pi \in \mathcal{M}(P, \nu)} \int_{\mathbb{R}^2} 1_{\{u \neq \gamma(y)\}} \pi(dy, du) = 0, \quad (2)$$

where  $1_{\{x \in A\}}$  denotes the indicator function of the set  $A$ , and the infimum is taken over all joint probability measures with marginals  $P$  and  $\nu$ . 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 constraint  $u = \gamma(y)$ .

This formulation can be interpreted as the existence of a probability that is concentrated on the structure, or alternatively, to the existence of a coupling between the random variable  $Y$  with law  $P$  and the random variable  $U$  with law  $\nu$ , i.e. the existence of  $\pi$  with marginals

$P$  and  $\nu$  such that

$$\pi(U \neq \gamma(Y)) = 0. \quad (3)$$

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

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

$$\inf_{\pi \in \mathcal{M}(P, \nu)} \int_{\mathbb{R}^2} 1_{\{u \neq \gamma(y)\}} \pi(dy, du), \quad (4)$$

where the infimum is taken over probabilities  $\pi$  with marginals  $P_n$  and  $\nu$ . By the above mentioned duality, the latter is equal to

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

with  $\mathcal{B}_y$  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  $\nu(\gamma(\cdot))$  over all Borel sets are determined by the latter's values on the restricted class. The class  $\mathcal{C}$  satisfies both requirements, and the resulting test statistic is

$$\sup_{A \in \mathcal{C}} (P_n(A) - \nu(\gamma(A))) = \sup_{y \in \mathbb{R}} |P_n(-\infty, y] - \nu(\gamma(-\infty, y])|, \quad (5)$$

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  $\gamma$  is replaced by a correspondence  $\Gamma$ , as in figure 2. Then we shall consider parameterized versions of the structure where both  $\Gamma$  and  $\nu$  depend on a parameter  $\theta$ , and form confidence regions with all values of  $\theta$  such that the specification of model  $(\Gamma_\theta, \nu_\theta)$  is not rejected.

## 1 Incomplete model specifications

We consider a very general econometric 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.

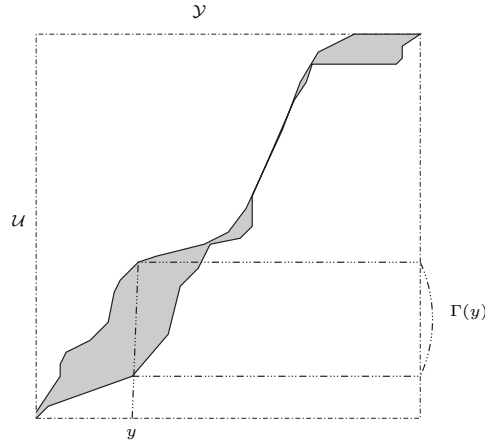


Figure 2: Incomplete structure

- 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.

The Borel sigma-algebras of  $\mathcal{Y}$  and  $\mathcal{U}$  will be respectively denoted  $\mathcal{B}_y$  and  $\mathcal{B}_u$ . Call  $P$  the Borel probability measure that represents the true data generating process for the observable variables, and  $\nu$  the hypothesized data generating processes for the latent variables. The structure 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  $\Gamma$ . Finally, the set of Borel probability measures on  $(\mathcal{Y} \times \mathcal{U}, \sigma(\mathcal{B}_y \times \mathcal{B}_u))$  with marginals  $P$  and  $\nu$  is denoted by  $\mathcal{M}(P, \nu)$ . Whenever there is no ambiguity, we shall adopt the de Finetti notation  $\mu f$  to denote the integral of  $f$  with respect to  $\mu$ .

## 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 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:

$$Y = DY_1 + (1 - D)Y_0,$$

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  $\Gamma$  from observable to unobservables in the following way:

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

**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 observable variables are the pairs  $(Y_*, Y^*, X)$  of upper and lower values for the dependent variable, and the explanatory variables. The correspondence describing the structure 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  $j$ ,  $j = 1, \dots, J$  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 determinants of the payoff. Finally  $\theta$  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  $\Gamma_\theta$  from unobservable player characteristics to observable variables  $(S, X)$ .

**Example 5: Entry models.** Consider the special case of example 4 proposed by Jovanovic (1989). The payoff functions are

$$\begin{aligned} \Pi_1(x_1, x_2, u) &= (\lambda x_2 - u)1_{\{x_1=1\}}, \\ \Pi_2(x_1, x_2, u) &= (\lambda x_1 - u)1_{\{x_2=1\}}, \end{aligned}$$

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  $\lambda$  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 structure 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  $[0, 1]$ .

We now turn to the definition of the null hypothesis of correct specification and its empirical counterparts (in section 2), the analysis of the properties of the test statistic (in section 3) and the implementation and applications of the test (in section 4).

## 1.2 Null hypothesis of correct specification

We wish to develop a procedure to detect whether the structure  $(\Gamma, \nu)$  and the distribution of observables are compatible. First we explain what we mean by *compatible*. We start by

taking  $P$ ,  $\Gamma$  and  $\nu$  as given and by considering three natural formalizations of compatibility, a first representation based on measurable selections of  $\Gamma$ , the second based on the existence of a suitable probability measure with marginals  $P$  and  $\nu$  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  $\Gamma$  between latent and observable variables is parametric and  $\Gamma = \gamma$  is measurable and single valued. Defining the image measure of  $P$  by  $\gamma$  by

$$P\gamma^{-1}(A) = P\{y \in \mathcal{Y} \mid \gamma(y) \in A\}, \quad (6)$$

for all  $A \in \mathcal{B}_{\mathcal{U}}$ , we say that the structure is well specified if and only if  $\nu = P\gamma^{-1}$ . In the general case considered here,  $\Gamma$  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  $\Gamma$ , we can construct an analogue of the image measure, which will now be a set  $\text{Core}(\Gamma, P)$  of Borel probability measures on  $\mathcal{U}$  (defined by (10)), and the hypothesis of *compatibility* of the restrictions on latent variable distributions and on the structures linking latent and observable variables will naturally take the form

$$H_0 : \nu \in \text{Core}(\Gamma, P). \quad (7)$$

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

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  $\Gamma$ , as in figure 3.

A measurable selection  $\gamma$  of  $\Gamma$  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  $\Gamma$  that satisfies Assumption 1 is denoted  $\text{Sel}(\Gamma)$  (which is known to be non-empty by the Rokhlin-Kuratowsky-Ryll-Nardzewski Theorem). To each selection  $\gamma$  of  $\Gamma$ , 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  $\gamma$  in  $\text{Sel}(\Gamma)$  is such that  $\nu = P\gamma^{-1}$ . However, such a requirement implies

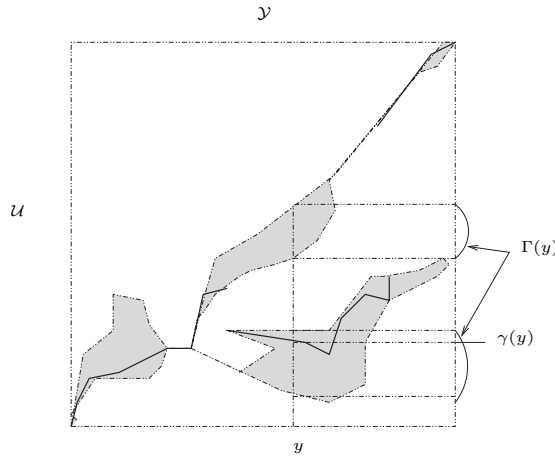


Figure 3: Selection of a correspondence

that  $\gamma$  corresponds to the equilibrium that is always selected. Under such a requirement, if for a given observable value the structure 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 structure 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  $\gamma$  spans  $\text{Sel}(\Gamma)$  (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 specification in leaving it completely unspecified.

Hence, we argue that the correct reformulation of the compatibility condition is that  $\nu$  can be written as a mixture of probability measures of the form  $P\gamma^{-1}$ , where  $\gamma$  ranges over  $\text{Sel}(\Gamma)$ . However, as the following example show, even for 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  $\Gamma$  is indexed by the class of Borel subsets of  $[0, 1]$ . Indeed, a representative measurable selection of  $\Gamma$  is

$\gamma_B$ , such that  $\gamma_B(x) = x1_{\{x \in B\}}$  for any Borel subset  $B$  of  $[0, 1]$ , where  $1_{\{x \in B\}}$  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 structure  $(\Gamma, \nu)$  is based on the existence of probability measures on the product  $\mathcal{Y} \times \mathcal{U}$  that admit  $P$  and  $\nu$  as marginals.

In the benchmark case of  $\Gamma = \gamma$  one-to-one, the structure 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  $\gamma$ , i.e. the set

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

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  $\gamma$  (i.e. that gives probability zero outside the constrained region defined by the structure) and admits  $P$  and  $\nu$  as marginals.

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

$$\text{Graph } \Gamma = \{(y, u) \in \mathcal{Y} \times \mathcal{U} : u \in \Gamma(y)\}, \tag{8}$$

in other words, a probability measure that admits  $P$  and  $\nu$  as marginals and gives probability zero to the event  $U \notin \Gamma(Y)$ , where  $U$  and  $Y$  are random elements with probability law  $\nu$  and  $P$  respectively (namely (12) below).

### 1.2.3 Dempster plausibility

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

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

the two being linked by the relation

$$\bar{P}(A) = 1 - \underline{P}(A^c), \quad (9)$$

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

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 (9), are larger than the smallest reliability associated with a set. We thus form the *core* of the belief function<sup>1</sup>:

$$\begin{aligned} \text{Core}(\Gamma, P) &= \{\mu \in \Delta(\mathcal{U}) \mid \forall B \in \mathcal{B}_{\mathcal{U}}, \mu(B) \geq \underline{P}(B)\} \\ &= \{\mu \in \Delta(\mathcal{U}) \mid \forall B \in \mathcal{B}_{\mathcal{U}}, \mu(B) \leq \bar{P}(B)\} \end{aligned} \quad (10)$$

where the first equality can be taken as a definition, and the second follows immediately from (9). 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 structure  $(\Gamma, \nu)$  is that  $\nu$  belongs to  $\text{Core}(\Gamma, P)$ , in other words, that  $\nu$  satisfies  $\nu(B) \leq P(\{y \in \mathcal{Y} : \Gamma(y) \cap B \neq \emptyset\})$  for all  $B \in \mathcal{B}_{\mathcal{U}}$ . Figure 4 illustrates this requirement in the case of finite sets.

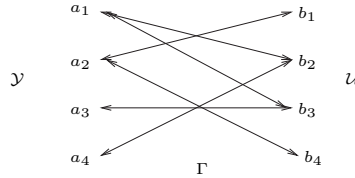


Figure 4: Graph of the correspondence  $\Gamma$  in a finite case. The event  $\{a_3\}$  always gives rise to the event  $\{b_3, b_4\}$ , whereas event  $\{a_4\}$  never does, so it is natural to constrain the probability of the event  $\{b_3, b_4\}$  by the upper bound  $P(\{a_1, a_2, a_3\})$  and the lower bound  $P(\{a_3\})$ .

---

<sup>1</sup>The name Core is standard in the literature to denote the set of probability measures satisfying (13). 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  $\Gamma$  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)  $\nu$  is a mixture of images of  $P$  by measurable selections of  $\Gamma$ , (i.e.  $\nu$  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  $\pi_\nu(y, \cdot)$  on  $\mathcal{U}$  with support  $\Gamma(y)$ , such that

$$\nu(B) = \int_{\mathcal{Y}} \pi_\nu(y, B) P(dy), \text{ all } B \in \mathcal{B}_{\mathcal{U}}. \quad (11)$$

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

$$\pi(U \notin \Gamma(Y)) = 0. \quad (12)$$

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

$$\nu(B) \leq P(\{y \in \mathcal{Y} : \Gamma(y) \cap B \neq \emptyset\}) \quad (13)$$

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

$$P(A) \leq \nu(\Gamma(A)). \quad (14)$$

**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  $\{P\gamma^{-1}; \gamma \in \text{Sel}(\Gamma)\}$ .

**Remark 2:** Notice that (11) looks like a disintegration of  $\nu$ , and indeed, when  $\Gamma$  is the inverse image of a single-valued measurable function (i.e. when the structure is given by

a single-valued measurable function from latent to observable variables), the probability kernel  $\pi_\nu$  is exactly the  $(P, \Gamma^{-1})$ -disintegration of  $\nu$ , in other words,  $\pi_\nu(y, \cdot)$  is the conditional probability measure on  $\mathcal{U}$  under the condition  $\Gamma^{-1}(u) = \{y\}$ . Hence (11) has the interpretation that a random element with distribution  $\nu$  can be generated as a draw from  $\pi_\nu(y, \cdot)$  where  $y$  is a realization of a random element with distribution  $P$ .

**Remark 3:** 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.

**Remark 4:** 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.

Note that since as a multivalued function,  $\Gamma$  is always invertible, and Assumption 1 holds for  $\Gamma$  if and only if it holds for  $\Gamma^{-1}$ , the roles of  $P$  and  $\nu$  can be interchanged in the formulations. In some cases, the symmetric formulation, with the roles of  $P$  and  $\nu$  interchanged, is useful, so we state it for completeness below:

**Theorem 1':** Under assumption 1, the following statements are equivalent, and are also equivalent to each of the statements in Theorem 1:

- (i')  $P$  is a mixture of images of  $\nu$  by measurable selections of  $\Gamma^{-1}$ , (i.e.  $P$  is in the weak closed convex hull of  $\{\nu\gamma^{-1}; \gamma \in \text{Sel}(\Gamma^{-1})\}$ ).
- (ii') There exists for  $\nu$ -almost all  $u \in \mathcal{U}$  a probability measure  $\pi_P(u, \cdot)$  on  $\mathcal{Y}$  with support  $\Gamma^{-1}(u)$ , such that

$$P(A) = \int_{\mathcal{U}} \pi_P(u, A) \nu(du), \text{ all } A \in \mathcal{B}_{\mathcal{Y}}. \quad (15)$$

(iii') is identical to Theorem 1(iii).

(iv') The probability assigned by  $P$  to an event in  $A \in \mathcal{B}_{\mathcal{Y}}$  is no greater than the largest

plausibility associated with  $A$  given  $\nu$  and  $\Gamma^{-1}$ , i.e.

$$P(A) \leq \nu(\{u \in \mathcal{U} : \Gamma^{-1}(u) \cap A \neq \emptyset\}) \quad (16)$$

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

$$\nu(B) \leq P(\Gamma^{-1}(B)). \quad (17)$$

**Remark 1:** The reason for giving this second theorem is that some of the new formulations will more amenable to forming empirical counterparts.

## 2 Empirical formulations

Each of the theoretical formulations of correct specification of the structure given in Theorems 1 and 1' has empirical counterparts, obtained essentially by replacing  $P$  by an estimate such as  $P_n$  in the formulations. The equivalence of the theoretical formulations does not necessarily entail equivalence of the empirical counterparts, especially in the cases where they rely on a choice of distance on the (metrizable) space of probability measures on  $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$  or  $(\mathcal{U}, \mathcal{B}_{\mathcal{U}})$ . Hence we need to consider the relations existing between the different empirical counterparts. We shall form our test statistic based on the empirical formulation relative to (v), so the reader may jump to section 2.4 without loss of continuity.

### 2.1 Empirical representations relative to (i)

For this empirical formulation, we consider (i') from Theorem 1'. We denote  $\text{Core}(\Gamma^{-1}, \nu)$  the set of arbitrary mixtures of  $\nu\gamma^{-1}$  when  $\gamma$  spans  $\text{Sel}(\Gamma^{-1})$ , and denoting by  $d$  a choice of metric on the space of probability measures on  $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ , the null can be reformulated as

$$d(P, \text{Core}(\Gamma^{-1}, \nu)) := \inf_{\mu \in \text{Core}(\Gamma^{-1}, \nu)} d(P, \mu) = 0.$$

Hence the empirical version is obtained by replacing  $P$  by an estimate such as  $P_n$  to yield

$$d(P_n, \text{Core}(\Gamma^{-1}, \nu)).$$

It will naturally depend on the specific choice of metric.

To see the relation between this and other empirical formulations, consider the Kolmogorov-Smirnov metric defined by

$$d_{KS}(\mu_1, \mu_2) = \sup_{A \in \mathcal{B}_Y} (\mu_1(A) - \mu_2(A))$$

for any two probability measures  $\mu_1$  and  $\mu_2$  on  $(\mathcal{Y}, \mathcal{B}_Y)$ . With this choice of metric, we can derive conditions under which the equalities

$$\begin{aligned} d_{KS}(P_n, \text{Core}(\Gamma^{-1}, \nu)) &= \inf_{\gamma \in \text{Sel}(\Gamma^{-1})} \sup_{A \in \mathcal{B}_Y} (P_n(A) - \nu\gamma^{-1}(A)) \\ &= \sup_{A \in \mathcal{B}_Y} \inf_{\gamma \in \text{Sel}(\Gamma)} (P_n(A) - \nu\gamma(A)) \\ &= \sup_{A \in \mathcal{B}_Y} (P_n(A) - \nu(\Gamma(A))) \end{aligned}$$

hold, and therefore this empirical formulation is equivalent to empirical formulations based on (iii), (iv), and (v) below.

## 2.2 Empirical representations relative to (ii)

We consider (ii) from Theorem 1 and  $d$  a metric on the space of probability measures on  $(\mathcal{U}, \mathcal{B}_U)$ . Under the null hypothesis, let  $\pi_\nu$  be the family of kernels defined in (ii) of Theorem 1. Denoting  $\mu f$  the integral of a function  $f$  by a measure  $\mu$ , we can write (ii) as  $d(\nu, P\pi_\nu) = 0$ , which admits  $d(\nu, P_n\pi_\nu)$  as empirical counterpart, and the latter is equal to  $d(P\pi_\nu, P_n\pi_\nu)$ . A notable aspect of this empirical formulation is that for many choices of metric  $d$  or indeed pseudo-metric (such as relative entropy), it will take the form of a functional of the empirical process  $\mathbb{G}_n := \sqrt{n}(P_n - P)$  applied to the functions  $y \mapsto \pi_\nu(y)$ . Different Goodness-of-fit tests can therefore be generalized within a single framework. The difficulty here of course is that the kernel  $\pi_\nu$  depends on the unknown  $P$  in a complicated way through the integral equation (11).

## 2.3 Empirical representation relative to (iii)

In view of representation (iii) of Theorem 1, i.e. equation (12), the null can be reformulated as the following Monge-Kantorovich mass transportation problem

$$\min_{\pi \in \mathcal{M}(P, \nu)} \int_{\mathcal{Y} \times \mathcal{U}} 1_{\{u \notin \Gamma(y)\}} \pi(dy, du) = 0, \quad (18)$$

where the transportation cost function  $1_{\{u \notin \Gamma(y)\}}$  is an indicator penalty for violation of the structure.

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

$$T^*(P_n, \Gamma, \nu) = \min_{\pi \in \mathcal{M}(P_n, \nu)} \int_{\mathcal{Y} \times \mathcal{U}} 1_{\{u \notin \Gamma(y)\}} \pi(dy, du). \quad (19)$$

We shall see below that it is equal to the empirical formulations relative to (iv) and (v).

## 2.4 Empirical representation relative to (iv) and (v)

Since formulations (iv) and (v) from Theorem 1 can be rewritten

$$\sup_{A \in \mathcal{B}_Y} (P(A) - \nu(\Gamma(A))) = 0,$$

the following empirical formulation seems the most natural:

$$\sup_{A \in \mathcal{B}_Y} (P_n(A) - \nu(\Gamma(A))).$$

The following Theorem states the equivalence between the latter and the empirical formulation derived from (iii):

**Theorem 2:** The following equalities hold:

$$T^*(P_n, \Gamma, \nu) = \max_{f \oplus g \leq \varphi} (P_n f + \nu g) \quad (20)$$

$$= \sup_{A \in \mathcal{B}_Y} (P_n(A) - \nu(\Gamma(A))), \quad (21)$$

where  $\varphi(y, u) = 1_{\{u \notin \Gamma(y)\}}$ , and  $f \oplus g \leq \varphi$  signifies that the maximum in (20) is taken over all measurable functions  $f$  on  $\mathcal{Y}$  and  $g$  on  $\mathcal{U}$  such that for all  $(y, u)$ ,  $f(y) + g(u) \leq \varphi(y, u)$ .

We shall therefore take  $T^*(P_n, \Gamma, \nu)$  as our starting point to construct a test statistic in the following section.

## 3 Specification test

We propose to adopt a test statistic based on the dual Monge-Kantorovich formulation (21), in other words a statistic that penalizes large values of (21). However,  $T^*(P_n, \Gamma, \nu)$  seemingly involves checking condition (14) on all sets in  $\mathcal{B}_Y$ . We need to elicit a reduced

class of sets on which to check condition (14). Call such a reduced class  $\mathcal{S}$ , and the resulting statistic is

$$T_{\mathcal{S}}(P_n, \Gamma, \nu) = \sup_{A \in \mathcal{S}} (P_n(A) - \nu(\Gamma(A))). \quad (22)$$

$\mathcal{S}$  is the result of a formal trade-off: it needs to be small enough to allow us to derive a limiting distribution for a suitable re-scaling of  $T(P_n, \Gamma, \nu)$ , and large enough to determine the direction of the inequality  $P - \nu\Gamma$ , which corresponds to a requirement that our test retain power against fixed alternatives.

To illustrate these requirements, we start by considering two simple types of structures to be tested. First we shall consider bijective structures (which correspond to our “prelude”), then the case where  $\mathcal{Y}$  is finite.

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

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

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. Hence the choice of  $\mathcal{C}$  for our reduced class  $\mathcal{S}$  is suitable on both counts: we know, as was discussed in the prelude, that  $\mathcal{C}$  is a value-determining class for probability measures, hence checking the inequality  $P - \nu\gamma$  on the reduced class is equivalent to checking it on all measurable sets. In addition, from Appendix A1, we know that this class is Vapnik-Červonenkis, and hence that  $\sqrt{n}T_{\mathcal{C}}(P_n, \gamma, \nu) = \sup_{A \in \mathcal{C}} \mathbb{G}_n(A)$  converges weakly to the supremum of a  $P$ -Brownian bridge, and the test of specification can be constructed based on approximations of the quantiles through simulations of the Brownian bridge or the bootstrap.

- **Discrete observables:** In the case where the observables belong to a finite set, the power set  $2^{\mathcal{Y}}$  is finite, hence Vapnik-Červonenkis. This will be sufficient to derive the limiting distribution of  $\sqrt{n}T_{2^{\mathcal{Y}}}(P_n, \Gamma, \nu) = \sqrt{n} \sup_{A \in 2^{\mathcal{Y}}} (P_n(A) - \nu(\Gamma(A)))$ . Since class of whole subsets is used, we do not need to worry about the competing requirements that the class determine the direction of the inequality  $P - \nu\Gamma$ .

We shall consider the two requirements on the class of sets  $\mathcal{S}$  sequentially. First, in the next subsection, we derive the asymptotic distribution of  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$  for a given choice of  $\mathcal{S}$ . Then, in the following subsection, we examine the power of the test based on  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$ , which amounts to linking the choice of the class of sets  $\mathcal{S}$  with classes of alternatives.

### 3.1 Asymptotic analysis

We start with a short heuristic description of the behaviour of  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$  which will motivate some definitions and constructions. We then give specific sets of conditions for the asymptotic results to hold.

#### 3.1.1 Heuristic description of asymptotic behaviour

Under the null hypothesis  $H_0$ , we have  $P(A) - \nu(\Gamma(A)) \leq 0$  for all  $A \in \mathcal{B}_y$ . Recalling that  $\mathbb{G}_n$  is the empirical process  $\sqrt{n}(P_n - P)$ , we have

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

Unlike the case of the classical Kolmogorov-Smirnov test, the second term in the previous display does not vanish under the null, since the “regions of indeterminacy” allow  $\delta(A) := P(A) - \nu(\Gamma(A))$  to be strictly negative for some sets  $A \in \mathcal{S}$ . What we know at this stage is that under the null, we have

$$\sqrt{n} T_{\mathcal{S}}(P_n, \Gamma, \nu) = \sup_{A \in \mathcal{S}} (\mathbb{G}_n(A) + \sqrt{n}(P(A) - \nu(\Gamma(A)))) \leq \sup_{A \in \mathcal{S}} \mathbb{G}_n(A),$$

but relying on this bound may lead to very conservative inference.

Note that  $\delta$  is independent of  $n$ , so that the scaling factor  $\sqrt{n}$  will pull the second term in the previous display to  $-\infty$  for all the sets where the inequality is strict. This prompts the following definition, illustrated in figure 5:

**Definition 3.1:** We denote the subclass of sets from  $\mathcal{S}$  where  $P = \nu\Gamma$  by  $\mathcal{S}_b$ , i.e.

$$\mathcal{S}_b := \{A \in \mathcal{S} : P(A) = \nu(\Gamma(A))\}.$$

If the class  $\mathcal{S}$  is a Vapnik-Červonenkis class of sets, the empirical process converges weakly to the  $P$ -Brownian bridge  $\mathbb{G}$ , i.e. a tight centered Gaussian stochastic process with

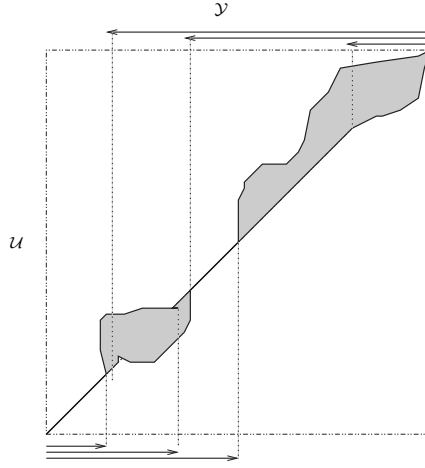


Figure 5: Examples of sets in  $\mathcal{C}_b$  (symbolized by the arrows) in a correctly specified case ( $P$  and  $\nu$  are uniform, hence correct specification corresponds to the graph of  $\Gamma$  containing the diagonal).

variance-covariance defined by

$$\mathbb{E}\mathbb{G}(A_1)\mathbb{G}(A_2) = P(A_1 \cap A_2) - P(A_1)P(A_2),$$

and the convergence is uniform over the class  $\mathcal{S}$  (i.e. the convergence is in  $l^\infty(\mathcal{F})$ , where  $\mathcal{F}$  is the class of indicator functions of sets in  $\mathcal{S}$ ), so that by the continuous mapping theorem, the supremum of the empirical process converges weakly to the supremum of the Brownian bridge (for a detail of the proof, see Appendix A1).

Under (mild) conditions that ensure that the function  $\delta$  “takes off” frankly from zero on  $\mathcal{S}_b$  to negative values on  $\mathcal{S} \setminus \mathcal{S}_b$ , the term  $\sqrt{n}\delta$  dominates the oscillations of the empirical process, and the sets in  $\mathcal{S} \setminus \mathcal{S}_b$  drop out from the supremum in the asymptotic expression, so that

$$\sqrt{n}T_{\mathcal{S}}(P_n, \Gamma, \nu) \rightsquigarrow \sup_{A \in \mathcal{S}_b} \mathbb{G}(A), \quad (23)$$

where  $\rightsquigarrow$  denotes weak convergence. Naturally, since  $\mathcal{S}_b$  depends on the unknown  $P$ , we need to find a data dependent class of sets to approximate  $\mathcal{S}_b$ . By the Law of Iterated Logarithm (see for instance page 476 of Dudley (2003)), we know that the empirical process  $\mathbb{G}_n$  is uniformly  $O_p(\sqrt{\ln \ln n})$ , so that if we construct the data dependent class as in definition 2 below with a bandwidth sequence  $h = h_n > 0$  satisfying

$$h_n + h_n^{-1} \sqrt{\frac{\ln \ln n}{n}} \rightarrow 0, \quad (24)$$

we shall pick out the sets in  $\mathcal{S}_b$  asymptotically.

**Definition 3.2:** We denote the data dependent subclass of sets from  $\mathcal{S}$  where  $P_n \geq \nu\Gamma - h$  by  $\hat{\mathcal{S}}_{b,h}$ , i.e.

$$\hat{\mathcal{S}}_{b,h} := \{A \in \mathcal{S} : P_n(A) \geq \nu(\Gamma(A)) - h\}.$$

This data dependent class of sets allows us to approximate the distribution of  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$  based on the following limiting result

$$\sup_{A \in \hat{\mathcal{S}}_{b,h_n}} \mathbb{G}(A) \rightsquigarrow \sup_{A \in \mathcal{S}_b} \mathbb{G}(A) \quad (25)$$

under requirement (24) on the bandwidth sequence  $h_n$ , and the additional requirement that

$$h_n(\ln \ln n) \rightarrow 0, \quad (26)$$

which allows to control local oscillations of the empirical process as well. Note that (24) and (26) are very mild, as they are both satisfied whenever

$$h_n n^{-\zeta} + h_n^{-1} n^\eta \rightarrow 0, \text{ for some } -1/2 < \eta \leq \zeta < 0. \quad (27)$$

Hence we shall be able to choose between the following methods for approximating quantiles of the distribution of  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$  and constructing rejection regions for our test statistic:

- We can simulate the Brownian bridge and compute the quantiles of the distribution of its supremum over the data dependent class  $\hat{\mathcal{S}}_{b,h_n}$  for some choice of  $h_n$ .
- We can use a subsampling approximation of the quantiles of the distribution of  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$ . Indeed,  $\sup_{A \in \mathcal{S}_b} \mathbb{G}(A)$  has continuous distribution function on  $[0, +\infty)$ , hence the subsampling approximation of quantiles is valid.

Before moving on to specific asymptotic results, we close this heuristic description with a discussion of the cases where the class of saturated sets  $\mathcal{S}_b$  is the trivial class  $\{\emptyset, \mathcal{Y}\}$ . In such cases, the test statistic converges to zero if one chooses the scaling factor  $\sqrt{n}$ . A refinement of the test will therefore involve a faster rate of convergence, determined through the construction of a local empirical process tailored to the shape of  $\nu\Gamma$  close to  $\emptyset$  and to  $\mathcal{Y}$ .

### 3.1.2 Specific asymptotic results

We now turn to specific conditions on the structure  $(\Gamma, \nu)$  and the law  $P$  of the observables such that results (23) which allows the subsampling approach, and (25) which then also allows the simulation approach, hold.

- (a) Case where  $\mathcal{Y}$  is finite and  $\mathcal{S}$  is the class of all subsets  $\mathcal{S} = 2^{\mathcal{Y}}$ .

In that case, we show in Theorem 3a below that both approaches to inference are valid.

**Theorem 3a:** If  $\mathcal{Y}$  is finite and  $\mathcal{S} = 2^{\mathcal{Y}}$ , (23) and (25) hold.

- (b) Case where  $\mathcal{Y} = \mathbb{R}^{d_y}$ ,  $P$  is absolutely continuous with respect to Lebesgue measure and  $\mathcal{S} = \{(y_1, z_1) \times \dots \times (y_{d_y}, z_{d_y}) : y_1, \dots, y_{d_y}, z_1, \dots, z_{d_y} \in \overline{\mathbb{R}}\}$  or any subclass, such as the class  $\mathcal{C}$  defined above<sup>2</sup>.

As indicated above, the asymptotic results are derived under assumptions such that the function  $\delta$  “takes off” frankly from zero. To make this precise, we introduce the following “frank separation” assumption. Recall that if  $d$  is the Euclidean metric on  $\mathcal{Y}$ , the Hausdorff metric  $d_H$  between two sets  $A_1$  and  $A_2$  is defined by

$$d_H(A_1, A_2) = \max \left( \sup_{y \in A_1} \inf_{z \in A_2} d(y, z), \sup_{z \in A_2} \inf_{y \in A_1} d(y, z) \right).$$

We need to ensure that on sets that are sufficiently distant from sets in  $\mathcal{S}_b$  (where the inequality is binding), then  $\delta$  is sufficiently negative so that it dominates local oscillations of the empirical process. To formalize this, we define the subclass of  $\mathcal{S}$  of sets such that the inequality is nearly binding.

**Definition 3.3:** We denote the subclass of sets from  $\mathcal{S}$  where  $P \geq \nu\Gamma - h$  by  $\mathcal{S}_{b,h}$ , i.e.

$$\mathcal{S}_{b,h} := \{A \in \mathcal{S} : P(A) \geq \nu(\Gamma(A)) - h\}.$$

We can now state

**Assumption FS (Frank Separation):** There exists  $K > 0$  and  $0 < \eta < 1$  such that for all  $A \in \mathcal{S}_{b,h}$ , for  $h > 0$  sufficiently small, there exists an  $A_b \in \mathcal{S}_b$  such that  $A_b \subseteq A$  and  $d_H(A, A_b) \leq Kh^\eta$ .

---

<sup>2</sup>Note that since  $P$  is absolutely continuous, considering only open intervals is without loss of generality.

**Remark 1:** Assumption is very mild, in the sense that it fails only in pathological cases, such as the case where  $\mathcal{Y} = \mathbb{R}$ ,  $\mathcal{S} = \mathcal{C}$ , and  $y \mapsto P((-\infty, y]) - \nu(\Gamma((-\infty, y]))$  is  $C^\infty$  with all derivatives equal to zero at some  $y = y_0$  such that  $(-\infty, y_0] \in \mathcal{C}$ .

Then, we have:

**Theorem 3b:** Suppose assumptions FS and (27) hold and that  $P$  is absolutely continuous with respect to Lebesgue measure. Then (23) and (25) hold.

The proof is based on the following lemma,

**Lemma 3a:** Under the conditions of Theorem 3b, we have

$$\sup_{A \in \mathcal{S}_{b, h_n}} \mathbb{G}_n(A) \rightsquigarrow \sup_{A \in \mathcal{S}_b} \mathbb{G}(A),$$

which involves bounds on oscillations of the empirical process.

## 3.2 Power of the test

As mentioned before, to ensure consistency of our specification test statistic, we need to derive conditions on the structure  $(\Gamma, \nu)$  and the law  $P$  of observables such that all violations of the inequality  $P \leq \nu\Gamma$  will be detected asymptotically with a test based on the statistic  $T_S(P_n, \Gamma, \nu)$ .

Before giving specific results, we shall try to convey the extent of the difficulties involved, in comparison with the case of the classical Kolmogorov-Smirnov test which was developed in our prelude.

When testing the equality of two probability measures, as in the Kolmogorov-Smirnov test, we need a class of sets that will determine the value of the law  $P$ , since it will ensure that if the equality holds on this class of sets, it holds everywhere. To be more precise, we need a convergence determining class (see section 2.6 page 18 of van der Vaart (1998)) since our test is asymptotic.

When testing the inequality  $P \leq \nu\Gamma$ , the situation is complicated in two ways. First,  $\nu\Gamma$  is a set function, but it is generally not additive unless  $\Gamma$  is bijective, and a convergence determining class is much harder come by. Second, determining the value of  $\nu\Gamma$  may not be sufficient, since it may not guarantee that the direction of the inequality  $P \leq \nu\Gamma$  will

be maintained from the reduced convergence determining class to all measurable sets. We discuss these two points in the following subsections.

### 3.2.1 Convergence determining classes for $\nu\Gamma$ :

The set function  $A \mapsto \nu(\Gamma(A))$  is a Choquet capacity functional (for definitions and properties, see Appendix A2), and the following lemma (lemma 1.14 of Salinetti and Wets (1986)) provides a convergence determining class in great generality. Recall that a closed ball  $B(y, \eta)$  with center  $y$  and radius  $\eta$  is the sets of points in  $\mathcal{Y}$  whose distance to  $y$  is lower or equal to  $\eta$ . Define  $\mathcal{S}_{\text{SW}}$  as the class of compact subsets of  $\mathcal{Y}$  with the following two properties:

(C1) Elements of  $\mathcal{S}_{\text{SW}}$  are finite unions of closed balls with positive radii,

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

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

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

Then we have:

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

The class  $\mathcal{S}_{\text{SW}}$  is not a Vapnik-Červonenkis class of sets since for any finite collection of points, there is a collection of finite union of balls that shatters it (see appendix A1). However, there is a natural restriction of this class which is. In the case where  $\mathcal{Y} = \mathbb{R}^{d_y}$ ,  $\mathcal{S}_{\text{SW}}$  can be redefined with rectangles instead of balls. Take an integer  $K$ . Define the class of finite unions of *at most*  $K$  rectangles:

$$\mathcal{S}_K = \left\{ \bigcup_{k \leq K} (y_k, z_k) : (y_k, z_k) \in \mathbb{R}^{2d_y} \right\}.$$

Then we have

**Lemma 3b:**  $\mathcal{S}_K$  is a Vapnik-Červonenkis class of sets.

Hence this class is amenable to asymptotic treatment.

### 3.2.2 Core determining classes for $\nu\Gamma$

The requirement, that we call ‘‘Core determining’’, on the class  $\mathcal{S}$  that  $P(A) \leq \nu(\Gamma(A))$  for all  $A \in \mathcal{S}$  imply  $P(A) \leq \nu(\Gamma(A))$  for all measurable  $A$  is apparently more stringent than the requirement that the values of the set function  $\nu(\Gamma(\cdot))$  on all measurable sets be determined by its values on  $\mathcal{S}$ .

**Definition 3.4:** A class  $\mathcal{S}$  of subsets of  $\mathcal{Y}$  is core determining for  $(\Gamma, \nu)$  if

$$\sup_{\mathcal{S}} (P - \nu\Gamma) = 0 \implies \sup_{\mathcal{B}_{\mathcal{Y}}} (P - \nu\Gamma) = 0$$

We have noted already the obvious fact:

**Fact 1:**  $\mathcal{S} = 2^{\mathcal{Y}}$  is core determining for observables on a finite set  $\mathcal{Y}$ .

A close inspection of the proof of Theorem 2 shows the following fact:

**Fact 2:** The class  $\mathcal{F}_{\mathcal{Y}}$  of closed subsets of  $\mathcal{Y}$  is core determining.

We now show that we can actually say much more by linking the core determining property with the convergence determining property, and showing that the class  $\tilde{\mathcal{S}}_{\text{SW}}$  of finite unions of open balls with positive raddii (or alternatively the class finite unions of open rectangles) is core determining.

First, we need to consider the following assumptions on the structure:

**Assumption (CD1):**  $\mathcal{Y}$  is a compact subset of  $\mathbb{R}^{d_{\mathcal{Y}}}$ , and  $\mathcal{U}$  is a compact subset of  $\mathbb{R}^{d_{\mathcal{U}}}$ .

**Assumption (CD2):**  $P$  and  $\nu$  are absolutely continuous with respect to Lebesgue measure.

**Assumption (CD3):** There exists  $\gamma_0 \in \text{Sel}(\Gamma)$  such that  $P(A) \rightarrow 0$  implies  $\nu(\gamma_0(A)) \rightarrow 0$ .

Note that assumption (CD3) is satisfied if either of the following hold:

- There exists  $\gamma_0 \in \text{Sel}(\Gamma)$  injective, such that  $\nu\gamma_0$  (now a probability measure) is absolutely continuous with respect to  $P$ .

- There exists  $\gamma_0 \in \text{Sel}(\Gamma)$  and  $\alpha > 0$  such that  $\nu(\gamma_0(A)) \leq \alpha P(A)$  for all  $A$  measurable.

**Assumption (CD4):**  $\Gamma$  is convex-valued, i.e.  $\Gamma(y)$  is a convex set for all  $y \in \mathcal{Y}$ .

This assumption rules out some interesting cases, for instance when the graph of  $\Gamma$  (defined in (8)) is the union of the graphs of two functions. However, our conditions are not minimal, and such cases could be treated under a different set of conditions.

We define the upper and lower envelopes of the Graph of  $\Gamma$  by

**Definition 3.5:** The upper (resp. lower) envelope of Graph  $\Gamma$  is the function  $y \mapsto u(y) = \sup \{\Gamma(y)\}$  (resp.  $y \mapsto l(y) = \inf \{\Gamma(y)\}$ ).

**Assumption (CD5):** The upper and lower envelopes  $u$  and  $l$  of the graph of  $\Gamma$  are Lipschitz, i.e. there exists  $\kappa \geq 0$  such that for all  $y_1, y_2 \in \mathcal{Y}$ ,

$$\max(|u(y_1) - u(y_2)|, |l(y_1) - l(y_2)|) \leq \kappa|y_1 - y_2|.$$

To state our last assumption, we need an extra definition:

**Definition 3.6:** A forking point of  $\Gamma$  is a  $y_0$  such that for any  $\epsilon > 0$ , there exists  $y_1$  and  $y_2$  in the open ball  $B(y_0, \epsilon)$  such that  $\Gamma(y_1)$  is a singleton, and  $\Gamma(y_2)$  is not.

**Assumption (CD6):**  $\Gamma$  has at most a finite number of forking points.

Note that this is a technical assumption that is violated only in pathological cases, and that is akin to the Frank Separation Assumption (FS).

We can now state the result:

**Theorem 3c:** Under assumption (CD1)-(CD6), the class  $\tilde{\mathcal{S}}_{\text{SW}}$  of finite unions of open balls with positive radii (or alternatively the class finite unions of open rectangles) is core determining.

This result is fundamental in that it reduces the problem of checking consistency of the test based on the statistic  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$  to the problem of checking whether  $P(A) \leq \nu(\Gamma(A))$  for  $A$  a finite union of balls (or rectangles) in  $\mathbb{R}^{d_y}$  whenever  $P \leq \nu\Gamma$  on  $\mathcal{S}$ .

We shall now apply this reasoning to give some conditions on the structure  $(\Gamma, \nu)$  under which the test based on statistic  $T_{\mathcal{S}}(P_n, \Gamma, \nu)$  is consistent with  $\mathcal{S} = \mathcal{C} = \{(-\infty, y], (y, \infty) : y \in \mathbb{R}\}$ , such as in figure 6, and conditions under which the class  $\mathcal{C}$  may not be core determining, but the class  $\mathcal{S} = \mathcal{R} = \{(y, z) : y, z \in \overline{\mathbb{R}}\}$  is. We thereby defining classes of alternatives that our tests based on  $T_{\mathcal{C}}(P_n, \Gamma, \nu)$  and  $T_{\mathcal{R}}(P_n, \Gamma, \nu)$  have power against in case  $\mathcal{Y} = \mathbb{R}$  and  $P$  is absolutely continuous with respect to Lebesgue measure.

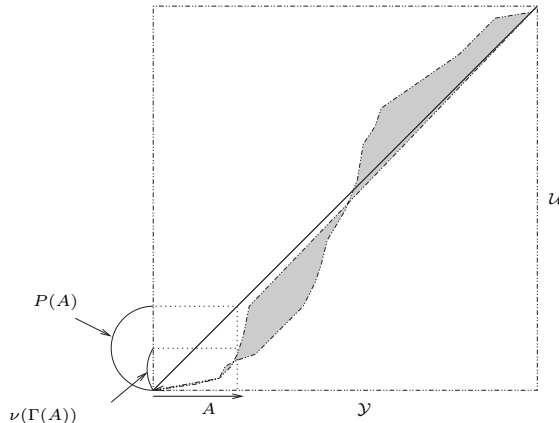


Figure 6: Violation of null that can be detected by the class of cells  $\mathcal{C}$ . Notice in particular that the inequality  $P \leq \nu\Gamma$  is violated on the set  $A$  ( $P$  and  $\nu$  are uniform).

**Theorem 3d:** If assumption (CD1) and (CD2) are satisfied, and the graph of  $\Gamma$  has increasing upper and lower envelopes, then  $\mathcal{C}$  is core determining, and hence the specification test based on the statistic  $T_{\mathcal{C}}(P_n, \Gamma, \nu)$  is consistent.

In figure 7, we show a case where the null hypothesis does not hold, but a test based on  $T_{\mathcal{C}}(P_n, \Gamma, \nu)$  fails to detect it because of the lack of monotonicity of the upper envelope. In that case, we need the larger class of sets  $\mathcal{R}$  to detect the departure from the null.

## 4 Applications of the inference framework

The test of specification that we have developed can be applied to the construction of confidence regions in case the structure depends on unknown parameters. Let  $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$  be a vector of structural parameters, and let the model be given by  $(\Gamma_\theta, \nu_\theta)$ .

**Definition 4.1:** The identified set  $\Theta_I$  is defined as the set of all  $\theta \in \Theta$  such that the

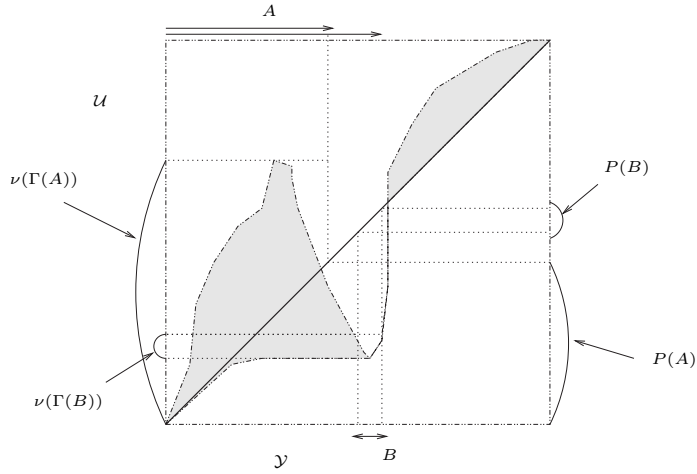


Figure 7: Violation of null that cannot be detected by the class of cells  $\mathcal{C}$ , but can be detected by the class of all intervals. Notice in particular that the inequality  $P \leq \nu\Gamma$  is violated on  $A$  but not on  $B$  ( $P$  and  $\nu$  are uniform).

null hypothesis  $H_0(\theta)$  of compatibility of  $(\Gamma_\theta, \nu_\theta)$  with  $P$  (as defined in Theorems 1 and 1') holds true.

This section is an outline of the application of our testing procedure to the construction of confidence regions for elements of the identified set and for the identified set itself.

#### 4.1 Coverage of parameters in the identified set

To form a confidence region that covers (with at least some pre-determined probability) each parameter value that makes the structure compatible with the distribution of observables, we propose to invert our test statistic to form a confidence region for elements of  $\Theta_I$ . In other words, for a given  $\alpha \in (0, 1)$ , we seek a region  $CR_n$  such that, for all  $\theta \in \Theta_I$ ,  $\liminf_n \mathbb{P}(\theta \in CR_n) \geq \alpha$ . The confidence region obtained from inverting the test has the form  $CR_n = \{\theta \in \Theta : \sqrt{n}T_{\mathcal{S}}(P_n, \Gamma_\theta, \nu_\theta) \leq \hat{Q}_\alpha(\theta)\}$  where  $\mathcal{S}$  is a class of sets which is Core determining for all  $\theta \in \Theta$  and  $\hat{Q}_\alpha(\theta)$  is an approximation of the  $\alpha$  quantile of the distribution of  $T_{\mathcal{S}}(P_n, \Gamma_\theta, \nu_\theta)$ . A valid approximation can be obtained using either one of the two methods proposed at the end of section 3.1.1.

## 4.2 Coverage of the identified set

To form a region that covers the whole identified set with pre-determined probability, we need a region  $\text{CR}_n^*$  such that  $\liminf_n \mathbb{P}(\Theta_I \subseteq \text{CR}_n^*) \geq \alpha$ . The latter can be obtained using the method proposed by Chernozhukov, Hong, and Tamer (2002) applied to the criterion function  $(\sup_{A \in \mathcal{S}} (P(A) - \nu_\theta(\Gamma_\theta(A))))^2$  with sample criterion  $T_S^2(P_n, \Gamma_\theta, \nu_\theta)$  (under the condition that C1, C2, C4 and C5 of Chernozhukov, Hong, and Tamer (2002) hold). A main contribution of this paper, therefore, is to provide the first natural and general choice of criterion function, and thereby pave the way for a comparison of criteria and a discussion of optimality.

## 4.3 Illustration

We now spell out our procedures on a very simple example: example 5 of section 1. The structure 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  $[0, 1]$ . Hence  $\nu_\phi$  has distribution function  $u^\phi$ , with  $\phi > 0$ . Our parameter vector is therefore  $\theta = (\lambda, \phi)'$ .

The null hypothesis in this case is immediately seen to be equivalent to  $p \leq \lambda^\phi$  for a given value of the parameter vector. Indeed, the easiest formulation to use is probably formulation (v) which requires that  $p = P(\{1\}) \leq \nu(\Gamma(1)) = \nu[0, \lambda] = \lambda^\phi$ . Hence  $T_{2\{0,1\}}(P_n, \Gamma_\theta, \nu_\theta) = p_n - \lambda^\phi$ . Now, if  $p = \lambda^\phi$ , then  $\mathcal{S}_b = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$  and then  $\sqrt{n}(p_n - \lambda^\phi)$  converges weakly to a normal random variable with mean zero and variance  $p(1 - p)$ , whereas if  $p < \lambda^\phi$ , then  $\mathcal{S}_b = \{\emptyset, \{0, 1\}\}$  and  $\sqrt{n}(p_n - \lambda^\phi)$  converges to zero. In either case, for a given choice of sequence  $h_n$ ,  $\hat{\mathcal{S}}_{b, h_n}$  is equal to  $\{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$  if  $p_n \geq \lambda^\phi - h_n$  and  $\{\emptyset, \{0, 1\}\}$  otherwise.

The  $\alpha$  quantile of  $\sqrt{n}T_{2\{0,1\}}(P_n, \Gamma_\theta, \nu_\theta) = \sqrt{n}(p_n - \lambda^\phi)$  can be approximated with 0 if  $p_n < \lambda^\phi - h_n$ , and with the  $\alpha$  quantile of the normal with mean zero and variance  $p_n(1 - p_n)$  if  $p_n \geq \lambda^\phi - h_n$ . Alternatively,  $Q_\alpha(\theta)$  can be approximated using subsampling (though it would be a serious case of overkill). The procedure would then be the following: Consider all (or a large number  $B_n$  of) the samples of size  $b_n$  from the sample of size  $n$  with  $1/b_n + b_n/n \rightarrow 0$  and approximate  $Q_\alpha(\theta)$  with

$$\hat{Q}_\alpha(\theta) = \inf\{x : \frac{1}{B_n} \sum_{i=1}^{B_n} \{\sqrt{b}T_S(P_b^i, \Gamma_\theta, \nu_\theta) \leq x\} \geq \alpha\}$$

where  $P_b^i$  is the empirical distribution of the  $i$ -th subsample. A confidence region is then  $\text{CR}_n = \{\theta \in [0, 1] \times (0, +\infty) : \sqrt{n}T_S(P_n, \Gamma_\theta, \nu_\theta) \leq \hat{Q}_\alpha(\theta)\}$ .

## 4.4 Semi-nonparametric extensions

Since structures are often given without a specification of the distribution of the unobservable variables, it is customary to assume only moment conditions, such as a given mean (taken to be equal to zero without loss of generality) and finite variance. This includes as special cases structures defined by moment inequality conditions.

In such cases, a similar approach can be taken where the null is defined as the existence of a joint law supported on the set  $\{u \in \Gamma_\theta(y)\}$  with marginal  $P$  on  $\mathcal{Y}$  and marginal on  $\mathcal{U}$  satisfying some moment conditions. Calling  $\mathcal{V}$  the set of laws that satisfy the said conditions, the dual formulation delivers a feasible version of the statistic

$$\inf_{\nu \in \mathcal{V}} \sup_{A \in \mathcal{S}} [P(A) - \nu(\Gamma_\theta(A))].$$

This involves a number of difficulties, which are the subject of a companion paper Galichon and Henry (2006). We only give here, as an illustration, the application of the method on a classic special case of example 3

Suppose one observes income brackets with centers in  $\mathcal{Y} = \{y_1, \dots, y_k\}$  with  $y_1 < \dots < y_k$  and width  $\delta$ . True income is unobservable, and one is interested in the mean of true income. The model correspondence is given by  $\Gamma(y) = (y - \delta/2, y + \delta/2)$ . Let  $p(y_i)$  (resp.  $p_n(y_i)$ ) denote the true (resp. empirical) probability of  $\{Y = y_i\}$ .

Consider formulation (v'):  $\nu \leq P\Gamma^{-1}$  of the null hypothesis. Denoting  $\Gamma^u(B) = \{y : \Gamma(y) \subseteq B\}$  for any  $B \in \mathcal{B}_\mathcal{U}$ , and writing  $\phi^* = P\Gamma^{-1}$  and  $\phi_* = P\Gamma^u$ , we have (using Definition A2.6 Lemma A2.2 in appendix A2) that under the null, the expectation of any measurable function  $f$  of the unobservable variables satisfies

$$\int_{\text{Ch}} f d\phi_* \leq \mathbb{E}f \leq \int_{\text{Ch}} f d\phi^*.$$

Denoting  $\phi_n^* = P_n\Gamma^{-1}$  and  $\phi_{n*} = P_n\Gamma^u$  the empirical versions of  $\phi^*$  and  $\phi_*$ , the set  $[\int_{\text{Ch}} f d\phi_{n*}, \int_{\text{Ch}} f d\phi_n^*]$  estimates the identified set  $[\int_{\text{Ch}} f d\phi_*, \int_{\text{Ch}} f d\phi^*]$ . In the case considered here, where  $f$  is the identity, this identified set equals

$$\left[ \sum_{i=1}^k (y_i - \delta/2) p(y_i), \sum_{i=1}^k (y_i + \delta/2) p(y_i) \right],$$

which is equal to

$$\left[ \sum_{i=1}^k (y_i - \delta/2) (p_n(y_i) - g_{n,i}/\sqrt{n}), \sum_{i=1}^k (y_i + \delta/2) (p_n(y_i) - g_{n,i}/\sqrt{n}) \right]$$

from which asymptotically valid confidence regions can be constructed, since  $g_n = (g_{n,1}, \dots, g_{n,k})'$ , with  $g_{n,i} = \sqrt{n}(p_n(y_i) - p(y_i))$  is asymptotically a Gaussian vector.

## Conclusion

We have provided a coherent definition of correct specification of structures with no identifying assumptions. This definition is the result of the equivalence of several natural generalizations of the hypothesis of correct specification in the identified case. These theoretical formulations of correct specification have natural empirical counterparts, several of which are also shown to be equivalent, and a test of specification is based on the latter. When the structure is parameterized, this test can be inverted to yield confidence regions for the set of structural parameters for which the null hypothesis of correct specification is satisfied.

This work has the following natural extensions: First, the whole approach is articulated around the existence of a joint measure with given marginals, hence it is essentially parametric in nature, but can be naturally extended to a problem of existence of a joint probability measure with one marginal given (the distribution of observables) and moment conditions on the other marginal (the distribution of unobservable variables). This natural extension of our work will nest structures defined by moment inequalities, and therefore deliver a way to construct confidence regions in such cases. Second, the statistic we have used to examine correct specification can be derived from the Kolmogorov-Smirnov distance between the empirical distribution and the set of data generating processes implied by the structure. Other distances and pseudo-distances will generate different specification statistics, and relative entropy may be a particularly good candidate, in that it produces optimal inference in the special case of identified structures.

## Appendix A: Additional concepts and results

### A1: Convergence of the empirical process

We give here definitions and results that we use in our asymptotic analysis. The definition of a Vapnik-Červonenkis class of sets is given in section 2.6.1 page 134 of van der Vaart and Wellner (1996) and reproduced here for the convenience of the reader.

**Definition A1.1:** Let  $\mathcal{S}$  be a collection of subsets of a set  $\mathcal{X}$ . An arbitrary set of  $n$  points  $\{x_1, \dots, x_n\}$  possesses  $2^n$  subsets. Say that  $\mathcal{C}$  picks out a certain subset from  $\{x_1, \dots, x_n\}$  if this can be formed as the set  $C \cap \{x_1, \dots, x_n\}$  for a  $C$  in  $\mathcal{S}$ . The collection  $\mathcal{S}$  is said to shatter  $\{x_1, \dots, x_n\}$  if each of its  $2^n$  subsets can be picked out in this manner. The Vapnik-Červonenkis index of the class  $\mathcal{S}$  is the smallest  $n$  for which no set of cardinality  $n$  is shattered by  $\mathcal{S}$ . A Vapnik-Červonenkis class of sets is a class with finite Vapnik-Červonenkis index.

**Fact A1:** The class of cells  $\mathcal{C}$  is a Vapnik-Červonenkis class of sets (see Example 2.6.1 page 135 of van der Vaart and Wellner (1996)).

**Definition A1.2:** The  $P$ -Brownian bridge is the tight centered Gaussian stochastic process with variance-covariance defined by  $\mathbb{E}\mathbb{G}(A_1)\mathbb{G}(A_2) = P(A_1 \cap A_2) - P(A_1)P(A_2)$ .

**Theorem A1.1:** If  $\mathcal{S}$  is a Vapnik-Červonenkis class of sets, the empirical process converges weakly to the  $P$ -Brownian bridge  $\mathbb{G}$ , and the convergence is uniform over the class  $\mathcal{S}$  (i.e. the convergence is in  $l^\infty(\mathcal{F})$ , where  $\mathcal{F}$  is the class of indicator functions of sets in  $\mathcal{S}$ ).

**Proof of Theorem A1.1:** We assume that  $\mathcal{S}$  is a Vapnik-Červonenkis class of sets. Call  $\mathcal{F}$  the class of indicator functions of sets in  $\mathcal{S}$ , and call  $V(\mathcal{F})$  the Vapnik-Červonenkis 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,  $N(\varepsilon, \mathcal{F}, \mathbb{L}_2(Q))$  satisfy

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

Hence, we have

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

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, which by definition means that  $\mathbb{G}_n$  converges in  $l^\infty(\mathcal{F})$ .

By the continuous mapping theorem, we immediately have the following corollary:

**Corollary A1.1:** If  $\mathcal{S}$  is a Vapnik-Červonenkis class of sets, then  $\sup_{\mathcal{S}} \mathbb{G}_n$  converges weakly to  $\sup_{\mathcal{S}} \mathbb{G}$ .

## A2: Choquet capacity functionals

We collect here all the definitions, equivalent representations and properties of Choquet capacity functionals (a.k.a. distributions of random sets or infinitely alternating capacities) that are useful for this paper. All the results presented here can be traced back to Choquet (1954).

Take  $\mathcal{X}$  a Polish space (complete metrizable and separable topological space) endowed with its Borel  $\sigma$ -algebra  $\mathcal{B}$ . For a sequence of numbers,  $a_n \uparrow a$  (resp.  $a_n \downarrow a$ ) denotes convergence in increasing (resp. decreasing) values, whereas for a sequence of sets, the notation  $A_n \uparrow A$  (resp.  $A_n \downarrow A$ ) denotes  $A_n \subseteq A_{n+1}$  for all  $n$  and  $A = \bigcup_n A_n$  (resp.  $A_{n+1} \subseteq A_n$  for all  $n$  and  $A = \bigcap_n A_n$ ). Finally, denote  $\mathcal{F}$  (resp.  $\mathcal{G}$ ) the set of closed (resp. open) subsets of  $\mathcal{X}$ , and for  $A \in \mathcal{B}$ ,  $\mathcal{F}_A = \{F \in \mathcal{F} : F \cap A \neq \emptyset\}$ .

**Definition A2.1:** A capacity is a set function  $\varphi : \mathcal{B} \rightarrow \mathbb{R}$  satisfying

- (i)  $\varphi(\emptyset) = 0$  and  $\varphi(\mathcal{X}) = 1$ ,
- (ii) For any two Borel sets  $A \subseteq B$ , we have  $\varphi(A) \leq \varphi(B)$ ,
- (iii) For all sequences of Borel sets  $A_n \uparrow A$ , we have  $\varphi(A_n) \uparrow \varphi(A)$ ,
- (iv) For all sequences of closed sets  $F_n \downarrow F$ , we have  $\varphi(F_n) \downarrow \varphi(F)$ .

**Definition A2.2** A capacity  $\varphi$  is called infinitely alternating if for any  $n$  and any sequence  $A_1, \dots, A_n$  of Borel sets,

$$\varphi\left(\bigcap_{i=1}^n A_i\right) \leq \sum_{\emptyset \neq I \subseteq \{1,2,\dots,n\}} (-1)^{|I|+1} \varphi\left(\bigcup_I A_i\right)$$

We call Choquet capacity functional an infinitely alternating capacity. Probability measures are special cases of Choquet capacity functionals, for which the alternating inequality of definition A2.2 holds as an equality (known as Poincaré's equality).

We now show that infinite alternation is a characteristic property of distributions of random sets (for a proof, see for instance section 2.1 of Matheron (1975)).

**Theorem A2.1:**  $\varphi$  is a Choquet capacity functional (i.e. an infinitely alternating capacity) if and only if there exists a probability measure  $\mathcal{P}$  on  $\mathcal{F}$  such that, for all  $A \in \mathcal{B}$ ,  $\varphi(A) = \mathcal{P}(\mathcal{F}_A)$ , and such a  $\mathcal{P}$  is unique.

$\varphi$  is therefore called the distribution of the random set associated with the probability measure  $\mathcal{P}$ , which allows the following definition of convergence determining classes for a Choquet capacity functional:

**Definition A2.3:** A class  $\mathcal{C}$  of Borel subsets of  $\mathcal{X}$  is called convergence determining for a Choquet capacity functional  $\varphi$  if and only if the class  $\{\mathcal{F}_A; A \in \mathcal{C}\}$  is convergence determining for the probability measure  $\mathcal{P}$  associated to  $\varphi$  as in Theorem A2.1.

We now look at the relation with measurable correspondences, defined as correspondences that satisfy Assumption 1 in the main text. Let  $(\Omega, \mathcal{B}, \mathbb{P})$  be a probability space.

**Definition A2.4:** A non-empty and closed valued correspondence  $\Gamma : \Omega \rightrightarrows \mathcal{X}$  is called a measurable correspondence if for each open set  $\mathcal{O} \subseteq \mathcal{X}$ ,  $\Gamma^{-1}(\mathcal{O}) = \{\omega \in \Omega \mid \Gamma(\omega) \cap \mathcal{O} \neq \emptyset\}$  belongs to  $\mathcal{B}$ .

If we define  $\varphi$  by  $\varphi(A) = \mathbb{P}\{\omega \in \Omega \mid \Gamma(\omega) \cap A \neq \emptyset\}$ , for all  $A \in \mathcal{B}$ , then  $\varphi$  is a Choquet capacity functional (from section 26.8 page 209 of Choquet (1954)), and its core is defined by the following:

**Definition A2.5:** the core of  $\varphi$  defined above is the set of probability measures that are set-wise dominated by  $\varphi$ , i.e.  $\text{Core}(\varphi) := \text{Core}(\Gamma, P) = \{Q : Q(A) \leq \varphi(A) \text{ all } A \text{ measurable}\}$ .

We add useful regularity properties of Choquet capacity functionals:

**Lemma A2.1:** If  $\varphi$  is a Choquet capacity functional, by the Choquet Capacitability Theorem (section 38.2 page 232 of Choquet (1954)), in addition to properties (i)-(iv) of Definition A2.1, it satisfies

$$(v) \quad \varphi(A) = \sup\{\varphi(F) : F \subseteq A, F \in \mathcal{F}\} \text{ for all } A \in \mathcal{B},$$

$$(vi) \quad \varphi(A) = \inf\{\varphi(G) : A \subseteq G, G \in \mathcal{G}\} \text{ for all } A \in \mathcal{B}.$$

Several notions extend integration in case of non-additive measures. We only use explicitly the notion of Choquet integral, which we define below.

**Definition A2.6:** The Choquet integral of a bounded measurable function  $f$  with respect to a

capacity  $\varphi$  is defined by

$$\int_{\text{Ch}} f \, d\varphi = \int_0^\infty \varphi(\{f \geq x\}) \, dx + \int_{-\infty}^0 (\varphi(\{f \geq x\}) - 1) \, dx, . \quad (28)$$

The Choquet integral reduces to the Lebesgue integral when  $\varphi$  is a probability measure. In addition, it has a very simple expression in case  $\varphi$  is a Choquet capacity functional (see Theorem 1 of Castaldo, Maccheroni, and Marinacci (2004)).

**Lemma A2.2:** If  $\varphi$  is a Choquet capacity functional, then for all  $f$  bounded measurable, the Choquet integral of  $f$  with respect to  $\varphi$  is given by  $\int_{\text{Ch}} f \, d\varphi = \sup_{Q \in \text{Core}(\varphi)} \int f \, dQ$ .

## Appendix B: Proofs of the results in the main text

### Reader's guide to the proofs:

In the proof of Theorem 1, a result very close to (ii)  $\iff$  (iv) is stated in Wasserman (1990), but the proof is essentially omitted. The proof of (i)  $\iff$  (iii) relies on Corollary 1 of Castaldo, Maccheroni, and Marinacci (2004), which allows to generalize Proposition 1 of Jovanovic (1989). The proof of (iv)  $\iff$  (v) is straightforward, whereas the proof of (iii)  $\iff$  (v) is similar to Theorem 2. The latter is a simple application of lemma 1, which itself is a simplification of the main generalized Monge-Kantorovitch duality theorem of Kellerer (1984). Lemma 1[a] is lemma 11.8.5 of Dudley (2003). The proof given here for completeness is due to N. Belili. The rest of Theorem 2 is a specialization of the duality result to zero-one cost, which can also be proved using Proposition (3.3) page 424 of Kellerer (1984), but we give a direct proof to show that we can specialize to closed sets, a fact that we use in the discussion of the power of the test.

Theorem 3a is straightforward. Theorem 3b is structured around the inequality

$$\sup_{\mathcal{S}_b} \mathbb{G}_n \leq \sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G}_n \leq \sup_{\mathcal{S}_{b,l_n}} \mathbb{G}_n$$

which holds on an event of large enough probability, with suitable bandwidth sequences  $h_n \ll l_n$ . Then, lemma 3a shows that  $\sup_{\mathcal{S}_{b,l_n}} \mathbb{G}_n$  converges weakly to the same limit as  $\sup_{\mathcal{S}_b} \mathbb{G}_n$ , namely  $\sup_{\mathcal{S}_b} \mathbb{G}$ . Finally, the same reasoning is invoked to show that  $\sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G}$  also converges to the same limit (but for this we need to assume that the bandwidth satisfies condition (27) rather than (24) and (26)). Lemma 3a relies on the construction of a local empirical process relative to the thin sets  $A \setminus A_b$ , where  $A$  is in  $\mathcal{S}_{b,l_n}$  and  $A_b$  is in  $\mathcal{S}_b$  and is close to  $A$  in terms of Hausdorff metric (hence the term ‘‘thin’’).

Lemma 3b, like Appendix A1, brings together some facts that are scattered in van der Vaart and

Wellner (1996). Theorem 3c uses the regularity properties of Choquet capacity functionals to show that finite unions of balls are core determining. Given a closed set  $F$ , using outer regularity of  $P$  and a compactness argument, a decreasing sequence of finite unions of open balls is constructed that satisfies two requirements: it converges to  $F$  both in  $P$ -measure and in Hausdorff distance. The regularity properties of the correspondence  $\Gamma$  are then used to control the Hausdorff distance between the images by  $\Gamma$  of  $F$  and the approximating sequence. The absolute continuity of  $\nu$  is then invoked to conclude, so that the sign of the inequality is maintained by continuity. Theorem 3d ties in the problem of finding core determining classes with the Monge-Kantorovitch dual under zero-one cost: pairs  $(1_F, -1_{\Gamma(F)})$  with  $F$  in the larger class are shown to be convex combinations of pairs  $(1_A, -1_{\Gamma(A)})$  with  $A$  in the potential core determining class.

### Proof of Theorem 1:

[a] We first show equivalences (i)  $\iff$  (iv)  $\iff$  (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  $\mathcal{Y}$  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  $\mathcal{B}_{\mathcal{Y}}$ -measurable, so that, by Theorem 3 of Strassen (1965), for a given  $\nu \in \Delta(\mathcal{U})$ , there exists  $\pi$  satisfying (11) with  $\pi(y, \cdot) \in \Delta(\Gamma(y))$  for  $P$ -almost all  $y$  if and only if

$$\int_{\mathcal{U}} f(u) \nu(du) \leq \int_{\mathcal{Y}} \sup_{u \in \Gamma(y)} f(u) P(dy) \tag{29}$$

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

$$\bar{P} : B \rightarrow P(\{y \in \mathcal{Y} : \Gamma(y) \cap B \neq \emptyset\}),$$

the right-hand side of (29) is shown in the following sequence of equalities to be equal to the integral of  $f$  with respect to  $\bar{P}$  in the sense of Choquet (defined by (28)).

$$\begin{aligned} & \int_{\mathcal{Y}} \sup_{u \in \Gamma(y)} \{f(u)\} P(dy) \\ &= \int_0^\infty P\{y \in \mathcal{Y} : \sup_{u \in \Gamma(y)} \{f(u)\} \geq x\} dx + \int_{-\infty}^0 (P\{y \in \mathcal{Y} : \sup_{u \in \Gamma(y)} \{f(u)\} \geq x\} - 1) dx \\ &= \int_0^\infty P\{y \in \mathcal{Y} : \Gamma(y) \subseteq \{f \geq x\}\} dx + \int_{-\infty}^0 (P\{y \in \mathcal{Y} : \Gamma(y) \subseteq \{f \geq x\}\} - 1) dx \\ &= \int_0^\infty \bar{P}(\{f \geq x\}) dx + \int_{-\infty}^0 (\bar{P}(\{f \geq x\}) - 1) dx = \int_{\text{Ch}} f d\bar{P}. \end{aligned}$$

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

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

so that (29) is equivalent to

$$\max_{\gamma \in \text{Sel}(\Gamma)} \int_{\mathcal{U}} f(u) P\gamma^{-1}(du) \geq \int_{\mathcal{U}} f(u) \nu(du) \quad (30)$$

for any  $f \in C_b(\mathcal{U})$ . If  $\nu$  is in the weak closure of the set of convex combinations of elements of  $\{P\gamma^{-1} : \gamma \in \text{Sel}(\Gamma)\}$ , then by linearity of the integral and the definition of weak convergence, (30) holds. Conversely, if  $\nu$  satisfies (30), then it satisfies

$$\int_{\text{Ch}} f d\bar{P} \geq \int_{\mathcal{U}} f(u) \nu(du)$$

and by monotone continuity, we have for all  $A \in \mathcal{B}_{\mathcal{U}}$ , and  $1_A$  the indicator function,

$$\int_{\mathcal{U}} 1_A(u) \nu(du) \leq \int_{\text{Ch}} 1_A d\bar{P}.$$

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

[b] We now show equivalences (iii)  $\iff$  (iv)  $\iff$  (v):

Using theorem 2 below, it suffices to show that (13) is equivalent to  $\nu(\Gamma(A)) \geq P(A)$  for all  $A \in \mathcal{B}_{\mathcal{Y}}$ . As previously, define  $\bar{P}$  as the set function on  $\mathcal{B}_{\mathcal{U}}$

$$\bar{P} : B \rightarrow P(\{y \in \mathcal{Y} : \Gamma(y) \cap B \neq \emptyset\}).$$

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

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

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

### Proof of Theorem 1':

The proof completely parallels the proof of Theorem 1. The equivalence between 1(iii) and 1'(iii') drives the equivalence of each of the formulations in Theorem 1' with each of the formulations in Theorem 1.

**Lemma 1:**

If  $\varphi : \mathcal{Y} \times \mathcal{U} \rightarrow \mathbb{R}$  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  $\varphi$  is continuous and  $\mathcal{U}$  and  $\mathcal{Y}$  are compact.

Call  $G$  the set of functions on  $\mathcal{Y} \times \mathcal{U}$  strictly dominated by  $\varphi$  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  $\varphi$  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  $\eta$  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  $\pi$  on  $\mathcal{Y} \times \mathcal{U}$  such that  $\eta(c) = \pi c$  for all continuous  $c$ . Since  $\eta = s$  on  $H$ , we have

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

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  $\varphi$  is continuous.

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

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

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

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

where  $\delta$  denotes the Dirac measure. By [a] above, there exists  $\pi_n$  with marginals  $\mu_n$  and  $\nu_n$  such that

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

Since  $(\pi_n)$  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  $\pi_n$  and a random variable  $X_0$  on the same probability space with law  $\pi$  such that  $X_0$  is the almost sure limit of  $(X_n)$ . By Fatou's lemma, we then have

$$\liminf \pi_n \varphi = \liminf \mathbb{E} \varphi(X_n) \geq \mathbb{E} \liminf \varphi(X_n) = \mathbb{E} \varphi(X_0) = \pi \varphi.$$

Hence we have the desired result.

[c] General case.

$\varphi$  is the pointwise supremum of a sequence of continuous bounded functions, so the result follows from upward  $\sigma$ -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).

## Proof of Theorem 2:

Under assumption 1,  $\Gamma$  is closed valued, hence  $\varphi(y, u) = 1_{\{u \notin \Gamma(y)\}}$  is lower semicontinuous and (20) is a direct application of lemma 1 above.

We now show (21). Since the sup-norm of the cost function is 1 (the cost function is an indicator), the supremum in (20) is attained pairs of functions  $(f, g)$  in  $\mathcal{F}$ , defined by

$$\begin{aligned}\mathcal{F} &= \{(f, g) \in \mathbb{L}^1(P) \times \mathbb{L}^1(\nu), 0 \leq f \leq 1, -1 \leq g \leq 0, \\ &\quad f(y) + g(u) \leq 1_{\{u \notin \Gamma(y)\}}, f \text{ upper semicontinuous}\}.\end{aligned}$$

Now,  $(f, g)$  can be written as a convex combination of pairs  $(1_A, -1_B)$  in  $\mathcal{F}$ . Indeed,  $f = \int_0^1 1_{\{f \geq x\}} dx$  and  $g = \int_0^1 -1_{\{g \leq -x\}} dx$ , and for all  $x$ ,  $1_{\{f \geq x\}}(y) - 1_{\{g \leq -x\}}(u) \leq 1_{\{u \notin \Gamma(y)\}}$ . Since

the functional on the right-hand side of (20) is linear, the supremum is attained on such a pair  $(1_A, -1_B)$ . Hence, the right-hand side of (20) specializes to

$$\sup_{A \times B \subseteq D} (P(A) - 1 + \nu(B)). \quad (31)$$

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  $u \notin \Gamma(A)$ , 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 (21) follows immediately.

### Proof of Theorem 3a:

Let  $A_0$  be the subset of  $\mathcal{Y}$  that achieves the maximum of  $\delta(A) = P(A) - \nu(\Gamma(A))$  over  $A \in \mathcal{S} \setminus \mathcal{S}_b$ . Call  $\delta_0 = \delta(A_0)$ , and note that  $\delta_0 < 0$ . We have

$$\begin{aligned} \sqrt{n}T_{2\mathcal{Y}}(P_n, \Gamma, \nu) &= \sup_{A \in 2\mathcal{Y}} [\mathbb{G}_n(A) + \sqrt{n}(P(A) + \nu(\Gamma(A)))] \\ &= \max\left\{ \sup_{\mathcal{S}_b} \mathbb{G}_n, \sup_{A \in 2\mathcal{Y} \setminus \mathcal{S}_b} [\mathbb{G}_n(A) + \sqrt{n}(P(A) + \nu(\Gamma(A)))] \right\}. \end{aligned}$$

The second term in the maximum of the preceding display is dominated by

$$\sup_{2\mathcal{Y} \setminus \mathcal{S}_b} \mathbb{G}_n + \sqrt{n}\delta_0,$$

whose limsup is almost surely non-positive. Hence (23) follows from the convergence of the empirical process. (25) follows from the fact that, under (24), for all  $n$  sufficiently large,  $\hat{\mathcal{S}}_{b, h_n}$  is almost surely equal to  $\mathcal{S}_b$ .

### Proof of Theorem 3b:

Consider two sequences of positive numbers  $l_n$  and  $h_n$  such that they both satisfy (27),  $l_n > h_n$  and  $(l_n - h_n)^{-1} \sqrt{\frac{\ln \ln n}{n}} \rightarrow 0$ . Notice that  $\{\emptyset, \mathcal{Y}\} \subseteq \mathcal{S}_b, \mathcal{S}_{b, h}, \hat{\mathcal{S}}_{b, h}$  for any  $h > 0$ . Since  $\mathbb{G}_n(\mathcal{Y}) = 0$ , we therefore have  $\sup_{\mathcal{S}_b} \mathbb{G}_n$ ,  $\sup_{\mathcal{S}_{b, l_n}} \mathbb{G}_n$  and  $\sup_{\hat{\mathcal{S}}_{b, h_n}} \mathbb{G}_n$  non-negative. Hence, calling  $\zeta_n$  the indicator function of the event  $\sup_{\mathcal{S}} \mathbb{G}_n \leq (l_n - h_n)\sqrt{n}$ , we can write

$$\begin{aligned} \zeta_n \sup_{\mathcal{S}_b} \mathbb{G}_n &\leq \zeta_n \max \left\{ \sup_{\mathcal{S}_b} [\mathbb{G}_n + \sqrt{n}(P - \nu\Gamma)], \sup_{\mathcal{S} \setminus \mathcal{S}_b} [\mathbb{G}_n + \sqrt{n}(P - \nu\Gamma)] \right\} \\ &\leq \zeta_n \sqrt{n}T_{\mathcal{S}}(P_n, \Gamma, \nu) \\ &\leq \zeta_n \sup_{\hat{\mathcal{S}}_{b, h_n}} \mathbb{G}_n \\ &\leq \zeta_n \sup_{\mathcal{S}_{b, l_n}} \mathbb{G}_n, \end{aligned}$$

where the first inequality holds because the left-hand side is equal to the first term in the right-hand side, the second inequality holds trivially as an equality since  $\mathcal{S} = \mathcal{S}_b \cup \mathcal{S} \setminus \mathcal{S}_b$ , the third inequality holds because on  $\mathcal{S} \setminus \hat{\mathcal{S}}_{b,h_n}$ , we have by definition  $\mathbb{G}_n + \sqrt{n}(P - \nu\Gamma) = \sqrt{n}(P_n - \nu\Gamma) \leq -h_n \leq 0$ , and the last inequality holds because on  $\{\zeta_n = 1\}$ , we have that  $A \in \hat{\mathcal{S}}_{b,h_n}$  implies  $\nu\Gamma(A) \leq P_n(A) + h_n = P(A) + (P_n - P)(A) + h_n \leq P(A) + \sup_{\mathcal{S}} \mathbb{G}_n / \sqrt{n} + h_n \leq P(A) + l_n - h_n + h_n = P(A) + l_n$ , which implies that  $A \in \mathcal{S}_{b,l_n}$ .

By Lemma 3a and Appendix A1, we have that both  $\sup_{\mathcal{S}_b} \mathbb{G}_n$  and  $\sup_{\mathcal{S}_{b,l_n}} \mathbb{G}_n$  converge weakly to  $\sup_{\mathcal{S}_b} \mathbb{G}$ . It is shown below that  $\zeta_n \rightarrow_p 1$ , so that Slutsky's lemma (lemma 2.8 page 11 of van der Vaart (1998)) yields the weak convergence of  $\zeta_n \sup_{\mathcal{S}_b} \mathbb{G}_n$  and  $\zeta_n \sup_{\mathcal{S}_{b,l_n}} \mathbb{G}_n$  to the same limit, and hence that of  $\zeta_n T_{\mathcal{S}}(P_n, \Gamma, \nu)$  and  $\zeta_n \sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G}_n$ . It follows from Slutsky's lemma again that

$$\sqrt{n}T_{\mathcal{S}}(P_n, \Gamma, \nu) \rightsquigarrow \sup_{\mathcal{S}} \mathbb{G} \quad \text{and} \quad \sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G}_n \rightsquigarrow \sup_{\mathcal{S}_b} \mathbb{G},$$

which proves (23).

We now prove that  $\zeta_n \rightarrow_p 1$ . Indeed, for any  $\epsilon > 0$ ,  $P(|\zeta_n - 1| > \epsilon) = P(\zeta_n = 0) = P(\sup_{\mathcal{S}} \mathbb{G}_n > (l_n - h_n)\sqrt{n}) \rightarrow 0$  by the Law of Iterated Logarithm, since  $(l_n - h_n)\sqrt{n} \gg \sqrt{\ln \ln n}$  by assumption.

There remains to show (25). Defining  $\xi_n$  as the indicator of the set

$$\{-h_n\sqrt{n} \leq \sup_{\mathcal{S}} \mathbb{G}_n \leq (l_n - h_n)\sqrt{n}\},$$

we have the inequalities

$$\xi_n \sup_{\mathcal{S}_b} \mathbb{G} \leq \xi_n \sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G} \leq \xi_n \sup_{\mathcal{S}_{b,l_n}} \mathbb{G}.$$

Indeed, the first inequality holds because  $\sup_{\mathcal{S}} \mathbb{G}_n \geq -h_n\sqrt{n}$  implies that  $P_n(A) \geq P(A) - h_n$  for all  $A$ , hence that  $\mathcal{S}_b \subseteq \hat{\mathcal{S}}_{b,h_n}$ ; and the second inequality holds because on  $\{\xi_n = 1\}$ , we have that  $A \in \hat{\mathcal{S}}_{b,h_n}$  implies  $\nu\Gamma(A) \leq P_n(A) + h_n = P(A) + (P_n - P)(A) + h_n \leq P(A) + \sup_{\mathcal{S}} \mathbb{G}_n / \sqrt{n} + h_n \leq P(A) + l_n - h_n + h_n = P(A) + l_n$ , which implies that  $A \in \mathcal{S}_{b,l_n}$ .

By Lemma 3a suitably modified to apply to the oscillations of  $\mathbb{G}$  instead of the oscillations of  $\mathbb{G}_n$ , we have that  $\sup_{\mathcal{S}_{b,l_n}} \mathbb{G}$  converges weakly to  $\sup_{\mathcal{S}_b} \mathbb{G}$ . It is shown below that  $\xi_n \rightarrow_p 1$ , so that Slutsky's lemma yields the weak convergence of  $\xi_n \sup_{\mathcal{S}_b} \mathbb{G}_n$  and  $\xi_n \sup_{\mathcal{S}_{b,l_n}} \mathbb{G}$  to the same limit, and hence that of  $\xi_n \sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G}$ . It follows from Slutsky's lemma again that

$$\sup_{\hat{\mathcal{S}}_{b,h_n}} \mathbb{G} \rightsquigarrow \sup_{\mathcal{S}_b} \mathbb{G},$$

which proves (25).

We now prove that  $\xi_n \rightarrow_p 1$ . Indeed, for any  $\epsilon > 0$ ,  $P(|\xi_n - 1| > \epsilon) = P(\zeta_n = 0) = P(\sup_{\mathcal{S}} \mathbb{G}_n > (l_n - h_n)\sqrt{n} \text{ or } \sup_{\mathcal{S}} \mathbb{G}_n < -h_n\sqrt{n}) \rightarrow 0$  by the Law of Iterated Logarithm, since  $(l_n - h_n)\sqrt{n} \gg \sqrt{\ln \ln n}$  and  $h_n\sqrt{n} \gg \sqrt{\ln \ln n}$  by assumption.

### Proof of Lemma 3a:

Take a bandwidth sequence  $l_n$  that satisfies (27), and take  $\mathcal{S}_{b,l_n}$  as in definition 3.3. Under assumption FS, take  $A \in \mathcal{S}_{b,l_n}$  and an  $A_0 \in \mathcal{S}_b$  such that  $d_H(A, A_0) \leq \zeta_n = Kl_n^\eta$  (we suppress the dependence of  $A_b$  on  $A$  for ease of notation). As  $\mathcal{S}_b \subseteq \mathcal{S}_{b,l_n}$ , one has

$$\sup_{A \in \mathcal{S}_b} \mathbb{G}_n(A) \leq \sup_{B \in \mathcal{S}_{b,l_n}} \mathbb{G}_n(A) \quad (32)$$

Second, since  $A_b \subseteq A$ , one has

$$\begin{aligned} \sup_{A \in \mathcal{S}_{b,l_n}} \mathbb{G}_n(A) &= \sup_{A \in \mathcal{S}_{b,l_n}} [\mathbb{G}_n(A_b) + \mathbb{G}_n(A \setminus A_b)] \\ &\leq \sup_{A \in \mathcal{S}_{b,l_n}} [\mathbb{G}_n(A_b)] + \sup_{A \in \mathcal{S}_{b,l_n}} [\mathbb{G}_n(A \setminus A_b)]. \end{aligned}$$

If we have that

$$\sup_{A \in \mathcal{S}_{b,l_n}} |\mathbb{G}_n(A \setminus A_b)| = O_{\text{a.s.}} \left( \sqrt{\zeta_n \ln \ln n} \right),$$

then

$$\sup_{A \in \mathcal{S}_{b,l_n}} \mathbb{G}_n(A) = \sup_{A \in \mathcal{S}_{b,l_n}} [\mathbb{G}_n(A_b)] + O_{\text{a.s.}} \left( \sqrt{\zeta_n \ln \ln n} \right) \quad (33)$$

noting the dependence of  $A_b$  on  $A$  in the expression above. But since  $A_b \in \mathcal{S}_b$ , one has  $\sup_{A \in \mathcal{S}_{b,l_n}} [\mathbb{G}_n(A_b)] \leq \sup_{A \in \mathcal{S}_b} \mathbb{G}_n(A)$ . This fact, along with (32) and (33), yields the result.

We now show that we have indeed that

$$\sup_{A \in \mathcal{S}_{b,l_n}} |\mathbb{G}_n(A \setminus A_b)| = O_{\text{a.s.}} \left( \sqrt{\zeta_n \ln \ln n} \right).$$

This relies on the construction of a local empirical process relative to the thin regions  $A \setminus A_b$ . First consider such a region. If  $A \in \mathcal{S}_b$ , the result holds trivially, so that we may assume that  $A \in \mathcal{S}_{b,l_n} \setminus \mathcal{S}_b$ , so that  $A \setminus A_b$  is not empty. We distinguish the case where  $A$  is a bounded rectangle, and the cases where  $A$  is unbounded.

- (i)  $A$  is a bounded rectangle, i.e. of the form  $(y_1, z_1) \times \dots \times (y_{d_y}, z_{d_y})$ , with  $y_1, \dots, y_{d_y}, z_1, \dots, z_{d_y}$  real. Then, since  $d_H(A, A_b) \leq \zeta_n$ ,  $A_b$  is also a bounded rectangle, and the  $A \setminus A_b$  is the union of at least one (since  $A$  and  $A_b$  are distinct) and at most  $f(d_y)$  (the number of faces of a rectangle in  $\mathbb{R}^{d_y}$ ) rectangles with at least one dimension bounded by  $\zeta_n$ .
- (ii)  $A$  is an unbounded rectangle, i.e. of the same form as above, except that some of the edges are  $+\infty$  or  $-\infty$ . Then  $A_b$  is also an unbounded rectangle, and  $A \setminus A_b$  is also the union of a finite number of rectangles with one dimension bounded by  $\zeta_n$ .

In both cases (i), and (ii),  $A \setminus A_b$  is the union of a finite number of rectangles with at least one dimension bounded by  $\zeta_n$ . Hence if we control the supremum of the empirical process on one of these thin rectangles, when  $A$  ranges over  $\mathcal{S}_{b,l_n}$ , we can control it on  $A \setminus A_b$ .

Hence, it suffices to prove that

$$\sup_{A \in \mathcal{S}_{b,l_n}} |\mathbb{G}_n(\varphi_n(A))| = O_{\text{a.s.}} \left( \sqrt{\zeta_n \ln \ln n} \right),$$

where  $\varphi_n$  is the homothety that carries  $A$  into one of the thin rectangles described above.

As an homothety,  $\varphi_n$  is invertible and bi-measurable, and since  $\varphi_n(A)$  has at least one dimension bounded by  $\zeta_n$ , and  $P$  is absolutely continuous with respect to Lebesgue measure,  $P(\varphi_n(A)) = O(\zeta_n)$  uniformly when  $A$  ranges over  $\mathcal{S}_{b,l_n}$ . Now, for any  $A \in \mathcal{S}_{b,l_n}$ , we have

$$\begin{aligned} \mathbb{G}_n(\varphi_n(A)) &= \sqrt{n} [P_n(\varphi_n(A)) - P(\varphi_n(A))] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (1_{\{\varphi_n(A)\}}(Y_i) - \mathbb{E}_P(1_{\{\varphi_n(A)\}}(Y))) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (1_A(\varphi_n^{-1}(Y_i)) - \mathbb{E}_P(1_A(\varphi_n^{-1}(Y)))) \\ &:= \sqrt{\zeta_n} L_n(1_A, \varphi_n), \end{aligned}$$

where  $L_n(1_A, \varphi_n)$  is defined as

$$\frac{1}{\sqrt{n\zeta_n}} \sum_{i=1}^n (1_A(\varphi_n^{-1}(Y_i)) - \mathbb{E}_P(1_A(\varphi_n^{-1}(Y))))$$

to conform with the notation of Einmahl and Mason (1997).

Conditions A(i)-A(iv) of the latter hold for  $a_n = b_n = l_n$  and  $a = 0$  under (27), and conditions S(i)-S(iii) and F(ii) and F(iv)-F(viii) hold because  $\mathcal{F}$  is here the class of indicator functions of  $\mathcal{S}_{b,l_n}$  which, as a subclass of  $\mathcal{S}$ , is a Vapnik-Červonenkis class of sets. Hence Theorem 1.2 of Einmahl and Mason (1997) holds, and

$$\sup_{A \in \mathcal{S}_{b,l_n}} |L_n(1_A, \varphi_n)| = O_{\text{a.s.}} \left( \sqrt{\ln \ln n} \right)$$

so that the desired result holds.

### Proof of Lemma 3b:

Consider  $\mathcal{S} = \{(y, z) : (y, z) \in \mathbb{R}^{2d_y}\}$ . It is a Vapnik-Červonenkis class. Indeed, if  $d_y = 1$ , its Vapnik-Červonenkis 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 Vapnik-Červonenkis index of  $\mathcal{S}$  is  $2d_y + 1$  (see Example 2.6.1 page 135 of van der Vaart and Wellner (1996)). Hence the class  $\mathcal{S}_K$  is also Vapnik-Červonenkis. The latter 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  $\mathcal{S} \sqcup \dots \sqcup \mathcal{S}$ , where the “square union” of two classes of sets  $\mathcal{S}_1$  and  $\mathcal{S}_2$  is defined by  $\mathcal{S}_1 \sqcup \mathcal{S}_2 = \{A_1 \cup A_2 : A_1 \in \mathcal{S}_1, A_2 \in \mathcal{S}_2\}$ .

### Proof of Theorem 3c:

From Fact 2, we know that we can restrict attention to closed subsets of  $\mathcal{Y}$ . Take  $F$  one such subset. By the outer regularity of Borel probability measures, for all  $n$  there is an open set  $\mathcal{O}'_n$  such that  $F \subseteq \mathcal{O}'_n$  and  $P(\mathcal{O}'_n) \leq P(F) + 1/n$ . Since  $\mathcal{O}'_n$  is open, for each  $y \in F$ , there exists  $r_y > 0$  such that the open ball  $B(y, r_y)$  centered at  $y$  with radius  $r_y$  is included in  $\mathcal{O}'_n$ , and by construction, the open set  $\tilde{\mathcal{O}}'_n = \bigcup_{y \in F} B(y, \min(r_y, 1/n^2))$  covers  $F$ . As a closed subset of a compact set,  $F$  is compact. Hence we can call  $\mathcal{O}_n$  the finite sub-covering of  $F$  extracted from  $\tilde{\mathcal{O}}'_n$ .  $\mathcal{O}_n$  is therefore a finite union of open balls with positive radii, i.e. it belongs to  $\tilde{\mathcal{S}}_{\text{SW}}$ . By construction of  $\mathcal{O}_n$ , we have  $d_H(\mathcal{O}_n, F) \leq 1/n^2$ , and we know that  $\Gamma(F) \subseteq \Gamma(\mathcal{O}_n)$ , and we shall now show that  $\nu(\Gamma(\mathcal{O}_n))$  converges to  $\nu(\Gamma(F))$  to yield the result that  $\tilde{\mathcal{S}}_{\text{SW}}$  is core determining.

Consider the following partition  $\mathcal{Y} = \mathcal{Y}_I \cup \mathcal{Y}_n^- \cup \mathcal{Y}_n^+$  with:

$$\begin{aligned} \mathcal{Y}_I &= \{y \in \mathcal{Y} : \nu(\Gamma(y)) = 0\}, \\ \mathcal{Y}_n^- &= \{y \in \mathcal{Y} : 0 < \nu(\Gamma(y)) < 1/n\}, \\ \mathcal{Y}_n^+ &= \{y \in \mathcal{Y} : \nu(\Gamma(y)) \geq 1/n\}. \end{aligned}$$

Define  $F_I = F \cap \mathcal{Y}_I$ ,  $F_n^- = F \cap \mathcal{Y}_n^-$  and  $F_n^+ = F \cap \mathcal{Y}_n^+$ , and similarly for  $\mathcal{O}_n$ , with  $\mathcal{O}_n^I$  denoting  $\mathcal{O}_n \cap \mathcal{Y}_I$ .

Consider first  $\mathcal{O}_n^I \setminus F_I$ . Assumption (CD3) yields immediately that  $\nu(\Gamma(\mathcal{O}_n^I \setminus F_I)) \downarrow 0$ .

Consider now  $\mathcal{O}_n^- \setminus F_n^-$ . Under assumption (CD6),  $\nu(\Gamma(\mathcal{Y}_n^-)) \downarrow 0$ , hence  $\nu(\Gamma(\mathcal{O}_n^- \setminus F_n^-)) \downarrow 0$ .

Consider now  $\mathcal{O}_n^+ \setminus F_n^+$ . Consider the disjoint connected components of  $\Gamma(\mathcal{O}_n^+)$ . Their  $\nu$  measure is at least  $1/n$  by construction, hence by the compactness of  $\mathcal{U}$ , the number  $J_n$  of disjoint connected components of  $\Gamma(\mathcal{O}_n^+)$  is no greater than  $n$ . We have shown above that  $d_H(\mathcal{O}_n, F) < 1/n^2$ , hence we have  $d_H(\mathcal{O}_n^+, F_n^+) < 1/n^2$ . By assumption (CD5), this implies that  $d_H(\Gamma(\mathcal{O}_n^+), \Gamma(F_n^+)) = O(1/n^2)$ . Hence for  $n$  sufficiently large, all the disjoint connected components of  $\Gamma(\mathcal{O}_n^+)$  intersect

$\Gamma(F_n^+)$ . Call  $(C_j)_{j=1}^{J_n}$  the disjoint connected components of  $\Gamma(\mathcal{O}_n^+)$ . We have

$$\nu(\Gamma(\mathcal{O}_n^+)) = \sum_{j=1}^{J_n} \nu(\Gamma(C_j)) = \sum_{j=1}^{J_n} (\nu(\Gamma(C_j)) + O(1/n^2)) = \nu(\Gamma(F_n^+)) + O(1/n),$$

where the second equality holds under assumption (CD2). Since  $F_n^+ \subseteq \mathcal{O}_n^+$ , we therefore have the desired result  $\nu(\Gamma(\mathcal{O}_n^+ \setminus F_n^+)) \downarrow 0$ , which completes the proof.

### Proof of Theorem 3d:

From fact 2, we can restrict attention to closed subsets of  $\mathcal{Y} = \mathbb{R}$ . Call  $\mathcal{Y}_I$  the subset of  $\mathcal{Y}$  defined by  $u(y) = l(y)$   $P$ -almost surely (and therefore everywhere since  $u$  and  $l$  are increasing). Note that the restriction of  $\nu\Gamma$  to  $\mathcal{Y}_I$  is a probability measure. Consider a closed subset  $F$  of  $\mathcal{Y}$ . Call  $F_I = F \cap \mathcal{Y}_I$  (resp.  $F_U = F \setminus F_I$ ) the intersection of  $F$  with  $\mathcal{Y}_I$  (resp. its complementary). Because of the monotonicity of the envelopes,  $\nu(\Gamma(F)) = \nu(\Gamma(F_I)) + \nu(\Gamma(F_U))$ , hence we only need to prove the result for closed subsets of  $\mathcal{Y}_I$  and for closed subsets of  $\mathcal{Y} \setminus \mathcal{Y}_I$ .

Take  $F$  a subset of  $\mathcal{Y}_I$ . The restriction  $\nu\Gamma|_{\mathcal{Y}_I}$  of  $\nu\Gamma$  to  $\mathcal{Y}_I$  is a probability measure, and the class of sets  $\mathcal{C}_I$  defined by  $\mathcal{C}_I = \{A \in \mathcal{Y} : A = \tilde{A} \cap \mathcal{Y}_I, \tilde{A} \in \mathcal{C}\}$  is value determining for  $\nu\Gamma|_{\mathcal{Y}_I}$ . By the monotonicity of the envelopes, we have  $\nu(\Gamma(\tilde{A})) = \nu(\Gamma(A)) + \nu(\Gamma(\tilde{A} \setminus A))$  (with the notation of the definition of  $\mathcal{C}_I$  above). Hence, if  $\nu(\Gamma(A)) \geq P(A)$  for all  $A \in \mathcal{C}$ , then  $\nu(\Gamma(A)) \geq P(A)$  for all  $A \subseteq \mathcal{Y}_I$ .

We can now restrict attention to the case where the upper and lower envelopes are distinct, in which case, for a closed set  $F$ ,  $\Gamma(F)$  has at most a countable number of connected parts, which we denote  $C_n$ ,  $n \in \mathbb{Z}$ , ordered in the sense that  $\inf C_n > \sup C_{n-1}$ . By construction, each  $C_n$  is the image by  $\Gamma$  of a subset  $F_n$  of  $F$ .  $\Gamma$  being convex-valued, the monotonicity of the envelopes  $u$  and  $l$  implies upper-semicontinuity of  $l$  and lower-semicontinuity of  $u$ . Therefore,  $C_n = \Gamma(F_n) = \Gamma([\inf F_n, \sup F_n])$ , and we deduce that  $\nu\Gamma(F) = \nu\Gamma(\bigcup_n I_n)$  where  $(I_n)_{n \in \mathbb{Z}}$  is a countable collection of disjoint closed intervals in  $\mathbb{R}$ . Hence if we show that  $\nu\Gamma(I) \geq P(I)$  for any interval  $I$ , then we have  $\nu\Gamma(F) = \sum_n \nu\Gamma(I_n) \geq \sum_n P(I_n) \geq P(F)$ , and the inequality holds for  $F$ .

Now, for any  $y_1 < y_2 \in \mathbb{R}$  we have  $P(y_1, y_2] = P(y_1, +\infty) + P(-\infty, y_2] - 1 \leq \nu\Gamma(y_1, +\infty) + \nu\Gamma(-\infty, y_2] - 1 = \nu(u(y_2) - l(y_1)) = \nu\Gamma(y_1, y_2]$  where  $u$  (resp.  $l$ ) is the upper (resp. lower) envelope, and the result follows.

## References

- ANDREWS, D., S. BERRY, and P. JIA (2003): “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,” Cemmap Working Papers, CWP10/06.
- BLUNDELL, R., M. BROWNING, and I. CRAWFORD (2005): “Best nonparametric bounds on demand responses,” 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. (1954): “Theory of capacities,” *Annales de l’Institut Fourier*, 5, 131–295.
- DEMPSTER, A. P. (1967): “Upper and lower probabilities induced by a multi-valued mapping,” *Annals of Mathematical Statistics*, 38, 325–339.
- DUDLEY, R. (2003): *Real Analysis and Probability*. Cambridge University Press.
- EINMAHL, U., and D. MASON (1997): “Gaussian approximation of local empirical processes indexed by functions,” *Probability Theory and Related Fields*, 107, 283–311.
- GALICHON, A., and M. HENRY (2006): “A duality approach to inference in models defined by moment inequalities,” unpublished manuscript.
- 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.
- JOVANOVIĆ, 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.
- MAGNAC, T., and E. MAURIN (2005): “Partial identification in monotone binary models: discrete regressors and interval data,” unpublished manuscript.
- 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.
- 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): “Threshold crossing models and bounds on treatment effects: a nonparametric analysis,” NBER Technical Working Paper 0307.
- STRASSEN, V. (1965): “The existence of probability measures with given marginals,” *Journal of Mathematical Statistics*, 36, 423–439.
- VAN DER VAART, A. (1998): *Asymptotic Statistics*. Cambridge University Press.
- VAN DER VAART, A., and J. WELLNER (1996): *Weak Convergence and Empirical Processes*. New York: Springer.
- WASSERMAN, L. (1990): “Prior envelopes based on belief functions,” *Annals of Statistics*, 18, 454–464.