Journal of Econometrics 177 (2013) 109–115

Contents lists available at ScienceDirect

Journal of Econometrics

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

Dilation bootstrap

Alfred Galichon [a], Marc Henry [b,*]

[a] Department of Economics, Sciences Po, 28 rue des Saints-Pères, 75007 Paris, France

[b] Université de Montréal, Département de sciences économiques, 3150, Jean-Brillant, Montréal, Québec H3C 3J7, Canada

ARTICLE INFO

Article history:
Received 29 January 2013
Received in revised form
29 January 2013
Accepted 2 July 2013
Available online 19 July 2013

JEL classification:
C15
C31

Keywords:
Partial identification
Dilation bootstrap
Quantile process
Optimal matching

ABSTRACT

We propose a methodology for combining several sources of model and data incompleteness and partial identification, which we call *Composition Theorem*. We apply this methodology to the construction of confidence regions with partially identified models of general form. The region is obtained by inverting a test of internal consistency of the econometric structure. We develop a dilation bootstrap methodology to deal with sampling uncertainty without reference to the hypothesized economic structure. It requires bootstrapping the quantile process for univariate data and a novel generalization of the latter to higher dimensions. Once the dilation is chosen to control the confidence level, the unknown true distribution of the observed data can be replaced by the known empirical distribution and confidence regions can then be obtained as in Galichon and Henry (2011) and Beresteanu et al. (2011).

0. Introduction

In several rapidly expanding areas of economic research, the identification problem is steadily becoming more acute. In policy and program evaluation (Manski, 1990) and more general contexts with censored or missing data (Shaikh and Vytlacil, 2011; Magnac and Maurin, 2008) and measurement error (Chen et al., 2005), ad hoc imputation rules lead to fragile inference. In demand estimation based on revealed preference (Blundell et al., 2008) the data is generically insufficient for identification. In the analysis of social interactions (Brock and Durlauf, 2001; Manski, 2004), complex strategies to reduce the large dimensionality of the correlation structure are needed. In the estimation of models with complex strategic interactions and multiple equilibria (Bjorn and Vuong, 1985; Tamer, 2003), assumptions on equilibrium selection mechanisms may not be available or acceptable.

More generally, in all areas of investigation with structural data insufficiencies or incompletely specified economic mechanisms, the hypothesized structure fails to identify a unique possible data generating mechanism for the data that is actually observed. In such cases, many traditional estimation and testing techniques become inapplicable and a framework for inference in incomplete models is developing, with an initial focus on estimation of the set of structural parameters compatible with true data distribution (hereafter *identified set*). A question of particular relevance in applied work is how to construct valid confidence regions for the identified set. Formal methodological proposals abound since the seminal work of Chernozhukov et al. (2007), but computational efficiency is still a major concern.

In the present work, we propose a methodology that clearly distinguishes how to deal with sampling uncertainty on the one hand, and model uncertainty on the other, so that unlike previous methodological proposals, a search in the parameter space is conducted only once, thereby greatly reducing the computational burden. The key to this separation is to deal with sampling variability without any reference to the hypothesized structure, using a methodology we call the *dilation method*. This consists in dilating each point in the space of observable variables in such a way that the empirical probability (which is known) of a dilated set dominates the true probability (which is unknown) of the original set (before dilation). The unknown true probability (i.e. the true data generating mechanism) is then removed from the analysis, and we can proceed as if the problem were purely deterministic, hence apply the methods proposed in Galichon and Henry (2011) and Beresteanu et al. (2011).

To construct confidence regions of level $1-\alpha$ for the identified set, such a dilation $y \rightrightarrows J(y)$ (where $\rightrightarrows$ denotes a one-to-many map) must satisfy $\tilde{Y}^* \in J(\tilde{Y})$ a.s. for some pair of random vectors $(\tilde{Y}^*, \tilde{Y})$,

* Corresponding author. Tel.: +1 514 343 2404.
E-mail addresses: alfred.galichon@sciences-po.fr (A. Galichon), marc.henry@umontreal.ca (M. Henry).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.07.001

with probability $1-\alpha$, where $\tilde{Y}$ is drawn from the true distribution of observable variables and $\tilde{Y}^*$ is drawn from the empirical distribution relative to the observed sample. We propose a dilation bootstrap procedure to construct J, in which bootstrap realizations Y_j^b, $j=1,\ldots,n$ are matched one-to-one with the original sample points Y_j, $j=1,\ldots,n$ so as to minimize $\eta_n^b = \max_{j=1,\ldots,n} \|Y_j^b - Y_{\sigma(j)}\|$, where the permutation σ defines the matching. The α quantile of the distribution of η_n^b then defines the radius of the dilation.

When the observable Y is a random variable, the dilation bootstrap relies on bootstrapping the quantile process, as proposed by Doss and Gill (1992). However, bootstrapping the quantile process relies on order statistics and had no higher dimensional generalization to date. This is now provided by the dilation bootstrap, which removes the constraint on dimension through the appeal to optimal matching. Although the problem of finding minimum cost matchings (called the *assignment* or *marriage problem*) is very familiar to economists, as far as we know, its application within an inference procedure is unprecedented.

The rest of the paper is organized as follows. The next section describes the econometric framework and introduces the *Composition Theorem* and the dilation method the latter justifies. The Composition Theorem is of independent interest as it provides a methodology for combining several sources of data insufficiencies and model incompleteness into a unified partial identification analysis. Section 2.1 discusses the application of the Composition Theorem to constructing confidence regions for partially identified parameters. Section 2.3 presents the bootstrap feasible dilation and its theoretical underpinnings. Section 3 presents simulation evidence on the performance of the dilation bootstrap in comparison with alternative methods. Section 4 explains how the method extends to higher dimensions and discrete choice and the last section concludes.

1. Dilation method and Composition Theorem

We consider the problem of inference on the structural parameters of an economic model, when the latter are (possibly) only partially identified. The economic structure is defined as in Jovanovic (1989), which generalizes (Koopmans and Reiersøl, 1950). Variables under consideration are divided into two groups. Latent variables U capture unobserved heterogeneity in the model. They are typically not observed by the analyst, but some of their components may be observed by the economic actors. Observable variables Y include outcome variables and other observable heterogeneity. They are observed by the analyst and the economic actors. We call *observable distribution* P the true probability distribution generating the observable variables, and denote by ν the probability distribution that generated the latent variables U. The econometric structure under consideration is given by a binary relation between observable and latent variables, i.e. a subset of $\mathcal{Y}\times\mathcal{U}$, which can be written without loss of generality as a correspondence from $\mathcal{U}$ to $\mathcal{Y}$. The pair of random vectors (Y,U) involved in the model is generated by a probability distribution, that we denote π. Since the vector U is unobservable, the probability distribution π is not directly identifiable from the data. However, the econometric model imposes restrictions on π. The distribution of its component Y is the observable distribution P. The distribution of its component U is the hypothesized probability distribution $\nu(\cdot|\theta)$. Finally, the joint distribution is further restricted by the fact that it gives probability 0 to the event $\{Y \notin G(U|\theta)\}$. This leads to the following:

Assumption 1 (*Econometric Specification*)**.**

(i) Observable variables Y, with realizations $y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$ and latent variables U, with realizations $u \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$, are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and satisfy the relation $Y \in G(U) \subseteq \mathcal{Y}$ almost surely.

(ii) The correspondence $G : \mathcal{U} \to 2^{\mathcal{Y}}$ is known by the analyst up to a finite dimensional vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$. It is denoted $G(\cdot;\theta)$. For all $\theta \in \Theta$, $G(\cdot;\theta)$ is measurable (i.e. the set $\{u : G(u;\theta) \cap A \neq \varnothing\}$ is measurable for each open subset A of $\mathcal{Y}$) and has non empty and closed values.

(iii) The distribution ν of the unobservable variables U is assumed to belong to a parametric family $\nu(\cdot|\theta)$, $\theta \in \Theta$. The same notation is used for the parameters of ν and G to highlight the fact that they may have components in common.

In more realistic models, computing $G(\cdot;\theta)$ can be challenging as $G(u,\theta)$ is a set, so the combinatorial complexity of the problem may be severe. Note that the measurability and closed values assumptions are very mild conditions. The assumption that the correspondence is non-empty, however, may be restrictive. In the revealed preferences example, we require that the demand correspondence be non empty. In the games example, we require existence of an equilibrium.

Example 1 (*Revealed Preferences*)**.** This approach is particularly well suited to revealed preference analysis. Suppose Y is the vector of observed choices made by an agent, possibly over several periods. Suppose the agent maximized a utility $u(Y,U|\theta)$ under constraints $g(Y,U|\theta) \leq 0$ (budget constraints, etc...), where θ is a vector of structural parameters (including elasticities, risk aversion, etc...) and U a random vector describing unobserved heterogeneity. Call $G(U|\theta)$ the demand correspondence, i.e. the set of utility maximizing choices. Then $G(U|\theta)$ exhausts all the information embodied in the utility maximization model.

Example 2 (*Games*)**.** Another family of examples of our framework arises with parametric games. Let N players with observable characteristics $X=(X_1,\ldots,X_N)$ and unobservable characteristics $U=(U_1,\ldots,U_N)$ have strategies $Z=(Z_1,\ldots,Z_N)$ and payoffs parameterized by X, U, Z and θ. For a given choice of equilibrium concept in pure strategies, call $\mathcal{C}(X,U,\theta)$ the equilibrium correspondence, i.e. the set of pure strategy equilibrium profiles. Then the empirical content of the game is characterized by $Z \in \mathcal{C}(X,U,\theta)$, which can be equivalently rewritten $Y \in G(U;\theta)$ with $Y=(Z,X)$.

For any given value of the structural parameter vector θ, a joint distribution satisfying all these restrictions may or may not exist. If it does, it is generally non unique. The identified set Θ_I is the collection of values of the structural parameter vector θ for which such a joint probability distribution does indeed exist. If $\Theta_I = \varnothing$, the model is rejected; on the contrary, if Θ_I is nonempty, then it may contain one point, in which case the parameter vector θ is point identified, or several, in which case the parameter θ is set identified. The identified Θ_I, first formalized in this way in Galichon and Henry (2006) is sometimes called "sharp identification region" to emphasize the fact that it exhausts all the information on the parameter available in the model. We can characterize it in the following way, which we take as our formal definition.

Definition 1 (*Identified Set*)**.**

$$\Theta_I = \left\{\theta \in \Theta \mid \exists \tilde{Y} \sim P,\ \tilde{U} \sim \nu(\cdot|\theta) : \mathbb{P}(\tilde{Y} \notin G(\tilde{U}|\theta)) = 0\right\}.$$

Our inference method on the identified set will be based on a general way of combining sources of uncertainty (sampling uncertainty or data incompleteness) by composition of correspondences. Suppose the probability measure Q on $\mathcal{Y}$ is the known distribution of a random vector Z and that it is related to the true unknown distribution P of the observed variables Y by the following relation:

Assumption 2 (*Dilation*)**.** There exists a correspondence $J : \mathcal{Y} \rightrightarrows \mathcal{Y}$ such that $\mathbb{P}(\tilde{Z} \notin J(\tilde{Y})) \leq \beta$ for some $\tilde{Z} \sim Q$, $\tilde{Y} \sim P$ and $0 \leq \beta < 1$.