A New Characterization of Identified Sets in Partially Identified Models*

Laurent Davezies† Xavier D'Haultfœuille‡

February 2, 2016

Abstract

We study the partial identification of models involving missing data in a broad sense. The framework is general enough to encompass treatment effect models, models with unobserved heterogeneity or some incomplete models. In these models, computing the identification region of the parameter of interest often requires to solve an equation defined on an infinite dimensional space, which is usually impossible in practice. We show however that the identification region can be characterized more simply under a convexity assumption, a condition satisfied in all standard settings. The identification region is then defined by extremal elements of a convex set. Exploiting such a result, we recover some previous results of the literature. We also apply our methodology to sample selection models without standard instruments, but where the probability of selection depends monotonically on the outcome, on a covariate, or on both.

Keywords: missing data, partial identification, extremal points.

JEL classification numbers: C14, C21, C26, C61.

*We would like to thank Stéphane Bonhomme and Marc Henry for their helpful remarks and suggestions. We are also grateful to Matthieu Fradelizi and Gilles Godefroy for their insights on Choquet Theorem.

†CREST. E-mail address: laurent.davezies@ensae.fr.

‡CREST. E-mail address: xavier.dhaultfoeuille@ensae.fr.

1 Introduction

In this paper, we reinvestigate partial identification with missing data, considered in a broad sense. This topic has been an active area of research, following the pioneering work of Manski (1989, 1990, 2003). While Manski initially focused on missing data specifically, his ideas have been successfully applied to limited dependent variable models (see, e.g., Chesher, 2010, Chesher et al., 2013, Bontemps et al., 2012), panel data models (see Honore & Tamer, 2006, Chernozhukov et al., 2013, Rosen, 2012) and incomplete models (see, e.g., Ciliberto & Tamer, 2009, Galichon & Henry, 2011, and Beresteanu et al., 2011), among others.

An issue that often arises in this literature is that the identification region is defined by an optimization over an infinite dimensional space, which is typically the space of a probability distribution that is at least partially unobserved. Such an optimization is often impossible to solve both in theory and computationally. For some models and parameters, closed form of the bounds of the identified set have been derived by specific methods, but general tools are still lacking. Important exceptions are the applications of random set theory, put forward by Beresteanu et al. (2011), and optimal transportation, considered by Galichon & Henry (2011) and Ekeland et al. (2010) when the identification region can be expressed only by moment conditions. Our first contribution is to propose a framework where the task of computing the identification region is much reduced. This framework encompasses standard missing data problems such as nonresponse or treatment effects models, but also models with unobserved heterogeneity, including fixed effects panel data models and incomplete models. The only substantial assumption that we consider is a convex restriction. Basically, we impose that if two at least unobserved probability distributions are consistent with the data and the model for a given value of the parameter of interest, then any mixture of these two probability distributions should also be consistent with the data and the model. We have not been able to find a natural example where this assumption would not be satisfied.

In this context, we prove that the identification region is characterized by its extreme points only. This is convenient, because in many cases the set of extreme points is small. This result may be seen as a kind of generalization, in an infinite setting, of the well known result that to optimize linear functionals on a convex, compact, finite dimensional set, one has to consider extreme points of this set only.[1] The infinite dimensional generalization is

[1]Interestingly, the finite dimensional result has already been used to derive bounds in partial identification problems, see Balke & Pearl (1997) and Freyberger & Horowitz (2012).

involved however, because the set we consider is not compact under the standard topology, and linear functionals need not be continuous. The proof of our result relies extensively on two powerful results in functional analysis: the Banach-Alaoglu theorem, which ensures, basically, that the set we consider is compact under a convenient topology, and the Choquet theorem, which gives an integral representation to every point of a given compact metrizable convex set.

In some problems, optimizing over extreme points may still be impossible. When the space is defined by an infinity number of constraints, for instance, the set of extreme points is typically infinite dimensional. Our second contribution is to give conditions under which the identification region can be well approximated by the sequence of identification regions corresponding to approximate models, that converge to the true one. In the case of a countable infinite number of constraints, such as sequence may correspond to models satisfying the first n constraints only, for instance. We also show that when the restrictions on the approximating sequence are not satisfied, convergence may not occur.

We then apply our main result to moment equality problems. The difference with standard GMM is that here moment equalities involve probability distributions of at least partially unobserved variables. Using a result of Douglas (1964), we characterize the set of extreme points in this context. We also show that it is finite dimensional when the number of equalities is finite. We obtain as corollaries recent results by Chernozhukov et al. (2013) and D'Haultfoeuille & Rathelot (2014) on the computation of bounds for average marginal effects in nonlinear panel data models and for segregation indices with small units, respectively. Using this result, we also provide another proof of Monge-Kantorovitch duality, thus making the link between our approach and optimal transportation theory.

Finally, we apply our framework to the sample selection model under monotonicity restrictions. More precisely, we suppose that the outcome, or a discrete covariate, or both, affects monotonically the probability of selection. These conditions are rather weak and likely to be plausible in many setting. Interestingly, the monotonicity on the outcome is very similar, but stronger, than the stochastic dominance condition considered for instance by Blundell et al. (2007). Our method proves very useful for deriving bounds on parameters of interest. While the bounds do not take any closed form, the bounds can be obtained by computation as the set of extreme points is finite dimensional.

The paper is organized as follows. The second section develops the general framework and presents the main results. The third section applies this result to several moment equalities problems. The fourth section applies it to the sample selection model under monotonicity restrictions. All proofs are deferred to appendix.

2 Problem and general results

2.1 Anatomy of the problem

We are interested in a parameter $\theta_0 \in \Theta$ related to a probability measure $P_0 \in \mathcal{P}$ of a random vector $U \in \mathcal{S}$, with $\mathcal{S}$ a closed subset of $\mathbb{R}^k$. More precisely, we suppose that there exists a known function q from $\mathcal{Q} \subset \Theta \times \mathcal{P}$ to $\mathbb{R}^l$ such that $q(\theta_0, P_0) = 0$. As we are mainly concerned with missing data, U is not fully observed in general, so that P_0, and hence θ_0, is not point identified in general. On the other hand, P_0 satisfies some restrictions, as it should be compatible both with the data and possible additional restrictions. We let $\mathcal{R}$ denote all these restrictions. Note that the difference between the restrictions $q(\theta_0, P_0) = 0$ and $P_0 \in \mathcal{R}$ is that the latter is independent of θ_0. We summarize our framework in the following assumption.

Assumption 1 (Framework)
The true parameter θ_0 and distribution P_0 satisfy $q(\theta_0, P_0) = 0$, where q is known, and $P_0 \in \mathcal{R}$. These restrictions exhaust the information on (θ_0, P_0).

This assumption implies in particular that the identification region of θ_0, Θ_0, is defined by[2]

$$\Theta_0 = \text{cl}\left(\{\theta \in \Theta : \exists P \in \mathcal{R} : q(\theta, P) = 0\}\right), \tag{2.1}$$

where cl (.) denotes the closure. The problem with this formulation is that it is intractable in general. Because $\mathcal{R}$ is often infinite dimensional, checking the existence of a $P \in \mathcal{R}$ satisfying $q(\theta, P) = 0$ is likely to be a very difficult task. We now impose additional restrictions in order to obtain a much more tractable form for the identification region. Namely, we assume that if two unknown distributions distributions P_1 and P_2 in $\mathcal{R}$ satisfy $q(\theta, P_1) = q(\theta, P_2) = 0$ for some θ, then every mixture P of P_1 and P_2 will also belong to $\mathcal{R}$ and satisfy $q(\theta, P) = 0$.

Assumption 2 (Convex restriction)
$\mathcal{R}_\theta = \{P \in \mathcal{R} : q(\theta, P) = 0\}$ *is convex for every* $\theta \in \Theta$.

[2]Each time we write $q(\theta, P)$, we implicitly assume that (θ, P) belongs to $\mathcal{Q}$. Hence, (2.1) should be understood as $\Theta_0 = \text{cl}\left(\{\theta \in \Theta : \exists P \in \mathcal{R} : (\theta, P) \in \mathcal{Q} \text{ and } q(\theta, P) = 0\}\right)$. $(\theta, P) \in \mathcal{Q}$ simply means that θ is well defined for $P_0 = P$. For instance if $q(\theta, P) = \int m(u, \theta) dP(u)$ for a known function m on $\mathcal{S} \times \Theta$, $\mathcal{Q}$ is the set of (θ, P) such that $\int |m(u, \theta)| dP(u) < \infty$.

This restriction actually holds in many missing data problems, as shown in the examples below. In the following we will give some results under Assumption 1 and 2. We also provide more precise results when Assumption 2 is replaced by the following condition.

Assumption 3 (Convex restriction and linear parameter)
$\mathcal{R}$ is convex and closed for the weak convergence. Moreover, $q(\theta, P) = \theta - \int f(u)dP(u)$ with f a known (or identifiable) real function satisfying $\int |f(u)|dP_0(u) < \infty$.

The function q is defined on $\mathbb{R} \times \mathcal{I}(f)$ with $\mathcal{I}(f) = \{P \in \mathcal{P} : \int |f|dP < \infty\}$. The restriction $\int |f(u)|dP_0(u) < \infty$ thus ensures that the true parameter is well-defined. Under Assumption 3, and because $P \mapsto \int f dP$ is linear, Θ_0 is an interval of $\mathbb{R}$, $\Theta_0 = [\underline{\theta}, \overline{\theta}]$. It suffices therefore to compute

$$\underline{\theta} = \inf_{P \in \mathcal{R} \cap \mathcal{I}(f)} \int f dP \text{ and } \overline{\theta} = \sup_{P \in \mathcal{R} \cap \mathcal{I}(f)} \int f dP. \tag{2.2}$$

However, even in this simpler case, the computation of the bounds requires an infinite dimensional optimization, which is not tractable in practice. We show in the following subsection how to reduce this computational task.

For the sake of simplicity, we consider in Assumption 3 that f is a real function, but the generalization to vector-valued functions ($f \in \mathbb{R}^p$) can be handled by using support functions of convex sets. Θ_0 is indeed convex whether f is real or not. It is therefore characterized by its support function (see, e.g., Hiriart-Urruty & Lemaréchal, 2001, p. 134) defined by

$$S(\lambda) = \sup_{P \in \mathcal{R} \cap \mathcal{I}(f)} \lambda' \int f dP,$$

for all λ belonging to the unit sphere of $\mathbb{R}^p$. In other words, instead of focusing on $\overline{\theta}$ and $\underline{\theta}$, we should focus on $\overline{\theta}_\lambda = \sup_{P \in \mathcal{R} \cap \mathcal{I}(f)} \int [\lambda' f] dP$, for all λ in the unit sphere of $\mathbb{R}^p$. Another generalization is the case where $q(\theta, P) = \int m(u, \theta) dP(u)$ with range of m in $\mathbb{R}^p$. For instance, θ can be the coefficient of a linear regression, and m represent the moment derived from orthogonality conditions between residuals and instruments. In this case, $\theta \in \Theta_0$ if and only if $\sup_{P \in \mathcal{R} \cap \mathcal{I}(f)} \int \lambda' m dP$ and $\inf_{P \in \mathcal{R} \cap \mathcal{I}(f)} \int \lambda' m dP$ have not the same sign for every λ in the unit sphere of $\mathbb{R}^p$.

2.2 Examples

2.2.1 Missing data with a known link.

In this example, observed data O are related with partially unobserved variable U through a known link: $O = s(U)$ where s is a known function. s is non-injective in general, so that

we cannot recover U given O. The parameter of interest θ_0 depends on the probability distribution of U, so here $P_0 = P^U$. A first example of this framework is unit nonresponse, where $O = (D, DY, DX)$ and $U = (D, Y, X)$, D being the dummy of response, Y the outcome and X are covariates. A second is the sample selection model (see, e.g., Heckman, 1974), where $O = (D, DY, X)$ and $U = (D, Y, X)$. A third is nonresponse on covariates, with $O = (D, Y, DX)$ and $U = (D, Y, X)$. Finally, this model also encompasses treatment effects, where $O = (T, Y_T, X)$ and $U = (T, (Y_t)_{t\in\mathcal{T}}, X)$. Here $T \in \mathcal{T}$ denotes the treatment and Y_t denotes the potential outcome corresponding to a treatment equal to t.

In this general missing data framework, Assumption 2 is satisfied if θ_0 is defined by moment equalities, so that $q(\theta, P) = \int m(\theta, u)dP(u)$, and under many different sets of additional restrictions. The first case is when there is actually no additional restriction. Then $\mathcal{R} = g^{-1}(\{P_0\})$, where g is a linear mapping from $\mathcal{P}$ to $\mathcal{P}$ defined by

$$g(P)(A) = \int \mathbb{1}\{s(u) \in A\}dP(u).$$

$\mathcal{R} = g^{-1}(\{P_0\})$ simply means that $P_0 = P^U$ should be compatible with the data and the link function s. Then Assumption 2 is satisfied because $\mathcal{R}_\theta = \mathcal{R} \cap \{P : \int m(u, \theta)dP(u) = 0\}$, and both sets are convex.

Assumption 2 also holds in the sample selection model if one of the covariates, say X_1, satisfies the exclusion restriction $Y \perp\!\!\!\perp X_1|X_2$, with $X = (X_1, X_2)$. Here, X_1 is a variable affecting D but not Y directly. These restrictions have been studied , either together with functional form restrictions (Heckman, 1974, Gronau, 1974), or alone (see Manski, 2003, chapter 2). In this last case, $\mathcal{R}$ is the set of all probability distributions in $g^{-1}(\{P_0\}) = \{P \in \mathcal{P} : \forall A, \int \mathbb{1}_{\{(d,dy,x)\in A\}}dP(d, y, x) = \int \mathbb{1}_{\{(d,dy,x)\in A\}}dP_0(d, y, x)\}$ satisfying this conditional independence restriction. This example is less trivial because the conditional independence restriction alone is not preserved by convex combinations. However, $\mathcal{R}$, and thus also $\mathcal{R}_\theta$, is still convex for all $\theta \in \Theta$.[3] The same result applies to the treatment effect example, with $(Y_t) \perp\!\!\!\perp X_1|X_2$.

In the sample selection literature, we often focus on coefficients of regression. In this case $q(\theta, P) = \int(y - x_2\theta)x_2'dP(y, d, x_1, x_2)$. Such parameter is not point-identified in general: we often use the restriction $Y \perp\!\!\!\perp X_1$ with shape restriction on $\mathbb{E}(Y|X_1, X_2, D = 1)$ to

[3]To see this, take $(P_1^{D,Y,X_1,X_2}, P_2^{D,Y,X_1,X_2}) \in g^{-1}(\{P_0\})^2$ and satisfying the conditional independence restriction. Let $P^{D,Y,X_1,X_2} = \lambda P_1^{D,Y,X_1,X_2} + (1-\lambda)P_2^{D,Y,X_1,X_2}$, with $\lambda \in [0, 1]$, then $P^{Y,X_1,X_2} = \lambda P_1^{Y,X_1,X_2} + (1-\lambda)P_2^{Y,X_1,X_2}$. The data restrictions impose $P_1^{D,X_1,X_2} = P_2^{D,X_1,X_2} = P^{D,X_1,X_2}$. Thus, $P^{Y|X_1,X_2}$ satisfy $P^{Y|X_1,X_2} = \lambda P_1^{Y|X_1,X_2} + (1-\lambda)P_2^{Y|X_1,X_2} = \lambda P_1^{Y|X_2} + (1-\lambda)P_2^{Y|X_2} = P^{Y|X_2}$. And so, P^{D,Y,X_1,X_2} satisfies the conditional independence restriction.

ensure point-identification (Heckman (1974)). Manski (2003) and Kitagawa (2010) relaxe such assumptions to characterize Θ_0. D'Haultfœuille (2010) also discuss identification of θ under restrictions that $D \perp\!\!\!\perp X_1|Y$ (see also Ramalho & Smith (2011)). More generally we can focus on the identification of the full joint distribution of (Y, X) and consequently on every parameter that depends on this distribution (moment, inequality index, quantile,...). In the last Section of this paper, we apply our result to characterize Θ_0 when the selection is monotonous in X and/or in Y.

In treatment effect literature, we often focus on average treatment effect, i.e.

$$q(\theta, P) = \theta - \int f(y,t)dP(y,t),$$

with f the identifiable function: $f(y,t) = y\left(\frac{\mathbb{1}_{\{t=1\}}}{\int \mathbb{1}_{\{t=1\}}dP_0^T(t)} - \frac{\mathbb{1}_{\{t=0\}}}{\int \mathbb{1}_{\{t=0\}}dP_0^T(t)}\right)$ and $\mathcal{R}$ is the set of distributions of (Y_0, Y_1, T) such that $P^{Y_0(1-T)+Y_1T,T} = P_0^{Y_0(1-T)+Y_1T,T}$. Apart in case of randomized experiment with perfect compliance, this parameter is generally not identified. Huge literature about this type of model focus on various parameters: local average treatment effects (Imbens & Angrist, 1994, Angrist et al., 1996), quantile treatment effects (Doksum, 1974, Chernozhukov & Hansen, 2005, Abadie et al., 2002, Firpo, 2007), values of counterfactual distributions (Abadie, 2002) etc... All these examples are embedded in our framework.

2.2.2 Unobserved heterogeneity

In this example, we suppose that the probability distribution of an observed variable O conditional on an (at least partially) unobserved heterogeneity U is a known function of θ_0. θ_0 may also satisfy moment restrictions $\int g(u, \theta_0)dP^U(u) = 0$. In this example, $P_0 = P^U$, $\mathcal{R} = \mathcal{P}$ and

$$q(\theta, P) = \max\left(\sup_{A \text{ measurable set}} \left|\int P^{O|U}(A|u,\theta)dP(u) - P^O(A)\right|, \left\|\int g(u,\theta)dP(u)\right\|\right).$$

Note that q is known since $P^{O|U}(A|u,\theta)$ is known. For each θ, $\mathcal{R}_\theta$ is convex since $\mathcal{P}$ is convex and the maps $P \mapsto \int P^{O|U}(A|u,\theta)dP(u)$ and $P \mapsto \int g(u,\theta)dP(u)$ are linear.

This framework includes the example of panel data model, with $O = ((Y_t)_{t=1...T}, (X_t)_{t=1...T})$ and $U = ((X_t)_{t=1...T}, \alpha)$, Y_t denoting the outcome at date t, X_t covariates at t and α an unobserved fixed effect. If we consider a parametric panel data model, distribution of $Y = (Y_t)_{t=1...T}$ conditional on $X = (X_t)_{t=1...T}$, α and θ_0 is known. This is the case if $Y_t = g(X_t, \alpha, \varepsilon_t, \beta_0)$ where the $(\varepsilon_t)_{t=1...T}$ are i.i.d., independent of (X, α) and with a known distribution and β_0 is a subvector of θ_0. Dynamic Markov models are also allowed for, by

simply adding the first period outcomes to U. In this example, if we are interested in β_0, $\theta_0 = \beta_0$ and $g(u, \theta) = 0$, namely, there is no additional restriction on θ_0. But we may also be interested in the average effect Δ_0 of a binary covariate X_{1t}, defined by

$$\Delta_0 = \int \left[E\left(Y_t | X_{1t} = 1, X_{2t} = x_2, \alpha = a, \beta_0\right) - E\left(Y_t | X_{1t} = 0, X_{2t} = x_2, \alpha = a, \beta_0\right)\right] dP^{X_{2t},\alpha}(x_2, a),$$

where $X_t = (X_{1t}, X_{2t})$. In this case, $\theta_0 = (\beta_0, \Delta_0)$, and

$$g(x_1, x_2, a, \beta, \Delta) = E\left(Y_t | X_{1t} = 1, X_{2t} = x_2, \alpha = a, \beta\right) - E\left(Y_t | X_{1t} = 0, X_{2t} = x_2, \alpha = a, \beta\right) - \Delta.$$

2.2.3 Models with multiple equilibria

In this example, we consider a simple entry game with two players studied, among others, by Bresnahan REF, Tamer (2003), Ciliberto & Tamer (2009), Ekeland et al. (2010), Galichon & Henry (2011), Beresteanu et al. (2011). The payoffs of the two players are given by the following matrix:

	2 enters	2 does not enter
1 enters	$(\theta_1 + \varepsilon_1, \theta_2 + \varepsilon_2)$	$(\varepsilon_1, 0)$
1 does not enter	$(0, \varepsilon_2)$	$(0, 0)$

Figure 1: Payoffs of entry game

Here θ_1 and θ_2 are nonpositive and $(\varepsilon_1, \varepsilon_2)$ are supposed to be observed by both players. When $(\varepsilon_1, \varepsilon_2) \in [0; -\theta_1] \times [0; -\theta_2]$, the game has two Nash equilibria in pure strategy (1 enters, 2 does not or 2 enters, 1 does not) and one in mixed strategy (1 enters with probability $-\varepsilon_2/\theta_2$ and 2 enters with probability $-\varepsilon_1/\theta_1$). For other values of $(\varepsilon_1, \varepsilon_2)$, the game has a unique equilibrium.

The parameters (θ_1, θ_2) and the payoffs shifters ε_1 and ε_2 are known by the players but not by the econometrician, who only observes the actions of the players. Then, letting Y_k be the dummy of entry of player $k \in \{1, 2\}$, the quantities $\pi_{ij} = \mathbb{P}(Y_1 = i, Y_2 = j)$ are identified for $(i, j) \in \{0; 1\}^2$. We also suppose that the distribution of $(\varepsilon_1, \varepsilon_2)$ is known up to some parameters. Hereafter, we suppose as an example that it is bivariate normal, with $E(\varepsilon_k) = \alpha_k$, $V(\varepsilon_k) = 1$ $(k \in \{1, 2\})$ and $\text{Cov}(\varepsilon_1, \varepsilon_2) = \rho$. In this example P_0 is the set of distributions of $(Y_1, Y_2, \varepsilon_1, \varepsilon_2)$ and $\theta_0 = (\alpha_1, \alpha_2, \rho, \theta_1, \theta_2)$.

There is not a unique way to define $\mathcal{R}_\theta$ and q and we shall present only one possibility. Let $S_{00} =]-\infty; 0] \times]-\infty; 0]$, $S_{11} = [-\theta_1; +\infty[\times [-\theta_2; +\infty[$, $S_{01} =]-\infty; 0[\times [0; +\infty[\cup]-\infty; -\theta_1] \times$

$[-\theta_2; +\infty[$, $S_{10} = [0; +\infty[\times]-\infty; 0[\cup [-\theta_1; +\infty[\times]-\infty; -\theta_2]$ and $S_{..} = [0, -\theta_1] \times [0, -\theta_2]$. For any $w = (w_{01}, w_{10}, w_m)$, let $q_w = (q_{00w}, q_{01w}, q_{10w}, q_{11w})'$ be defined as

$$
\begin{aligned}
q_{00w}(\varepsilon_1, \varepsilon_2) &= \mathbb{1}_{S_{00}}(\varepsilon_1, \varepsilon_2) + \frac{(\theta_1 + \varepsilon_1)(\theta_2 + \varepsilon_2)}{\theta_1 \theta_2} w_m(\varepsilon_1, \varepsilon_2) \mathbb{1}_{\{(\varepsilon_1, \varepsilon_2) \in S_{..}\}}, \\
q_{01w}(\varepsilon_1, \varepsilon_2) &= \mathbb{1}_{S_{01}}(\varepsilon_1, \varepsilon_2) + \left[w_{01}(\varepsilon_1, \varepsilon_2) - \frac{(\theta_1 + \varepsilon_1)\varepsilon_2}{\theta_1 \theta_2} w_m(\varepsilon_1, \varepsilon_2) \right] \mathbb{1}_{S_{..}}(\varepsilon_1, \varepsilon_2), \\
q_{10w}(\varepsilon_1, \varepsilon_2) &= \mathbb{1}_{S_{10}}(\varepsilon_1, \varepsilon_2) + \left[w_{10}(\varepsilon_1, \varepsilon_2) - \frac{\varepsilon_1(\theta_2 + \varepsilon_2)}{\theta_1 \theta_2} w_m(\varepsilon_1, \varepsilon_2) \right] \mathbb{1}_{S_{..}}(\varepsilon_1, \varepsilon_2), \\
q_{11w}(\varepsilon_1, \varepsilon_2) &= \mathbb{1}_{S_{11}}(\varepsilon_1, \varepsilon_2) + \frac{\varepsilon_1 \varepsilon_2}{\theta_1 \theta_2} w_m(\varepsilon_1, \varepsilon_2) \mathbb{1}_{S_{..}}(\varepsilon_1, \varepsilon_2).
\end{aligned}
$$

$w_{10}(\varepsilon_1, \varepsilon_2)$ (resp. $w_{01}(\varepsilon_1, \varepsilon_2)$) represents the probability that the players chooses the pure strategy that 1 enters while 2 does not (resp. 2 enters while 1 does not), while $w_m(\varepsilon_1, \varepsilon_2)$ corresponds to the probability that a couple of players chooses the mixed strategy. Finally, let $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})'$. Then $q(\theta, P) = \mathbb{1}_{^cC(\theta,P)}$ with

$$
\begin{aligned}
C(\theta, P) = \Big\{ & \exists (w_{01}, w_{10}, w_m) \text{ functions from } [0, -\theta_1] \times [0, -\theta_2] \text{ into } [0,1] \text{ such that } w_{01} + w_{10} + w_m = 1, \\
& \int q_w(\varepsilon_1, \varepsilon_2) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = \pi \text{ and } \forall (u_1, u_2) \in \mathbb{R}^2, \\
& \int \mathbb{1}\{\varepsilon_1 \leq u_1; \varepsilon_2 \leq u_2\} dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = \Phi_2(u_1 - \alpha_1, u_2 - \alpha_2, \rho) \Big\}.
\end{aligned}
$$

With this definition of q, $\mathcal{R}$ corresponds to the set of probability distributions on $\{0;1\}^2 \times \mathbb{R}^2$ such that

$$
\forall (i,j) \in \{0;1\}^2, \quad \int \mathbb{1}_{\{i,j\}}(y_1, y_2) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = \pi_{ij}.
$$

2.3 Main theoretical results

Our main result, Theorem 2.1 below, is that Θ_0 can be characterized by extreme points of $\mathcal{R}_\theta$. In the separable and linear case, the bounds of Θ_0 can be obtained by an optimization on a smaller set than $\mathcal{R} \cap \mathcal{I}(f)$. Let ext $\left(\overline{\mathcal{R}}_\theta\right)$ denote the set of extreme points of the closure for weak convergence of $\mathcal{R}_\theta$, and ext$(\mathcal{R})$ denote the set of extreme points of $\mathcal{R}$.

Theorem 2.1 (Main result)

1. *Under Assumptions 1 and 2,* $\Theta_0 = \{\theta \in \Theta : ext\left(\overline{\mathcal{R}}_\theta\right) \neq \emptyset\}$.

2. *If Assumption 3 also holds, then*

$$
\underline{\theta} = \inf_{P \in ext(\mathcal{R}) \cap \mathcal{I}(f)} \int f(u) dP(u) \text{ and } \overline{\theta} = \sup_{P \in ext(\mathcal{R}) \cap \mathcal{I}(f)} \int f(u) dP(u).
$$

This theorem shows the existence of extreme points of $\mathcal{R}_\theta$ (when $\mathcal{R}_\theta$ is not empty) and that the identification region is completely characterized by ext $\left(\overline{\mathcal{R}}_\theta\right)$. Similarly, for a linear parameter, the bounds of the identification region can be attained by considering only extreme distributions of $\mathcal{R}$. Our main result is particularly helpful if ext $\left(\overline{\mathcal{R}}_\theta\right)$ (respectively ext$(\mathcal{R})$) is easily characterizable and finite dimensional, because optimization on ext $\left(\overline{\mathcal{R}}_\theta\right)$ is then doable in practice. We provide an example below, and consider a larger class in Subsection 3 below. Theorem 2.1 is also useful if ext $\left(\overline{\mathcal{R}}_\theta\right)$ is complicated but there exists a simpler set A such that ext $\left(\overline{\mathcal{R}}_\theta\right) \subset A \subset \mathcal{R}_\theta$.

Example (continued): models with multiple equilibria.

By what precedes, $\theta \in \Theta_0$ if and only if for the corresponding $(\alpha_1, \alpha_2, \rho)$, there exists $w = (w_{01}, w_{10}, w_m)$ with values in $[0,1]$ such that $w_{01} + w_{10} + w_m = 1$ and

$$\int q_w(\varepsilon_1, \varepsilon_2) dF_\theta(\varepsilon_1, \varepsilon_2) = \pi,$$

where F_θ is the cumulative distribution function (cdf) of a bivariate normal with expectation (α_1, α_2), unit variance and correlation ρ.

The set ext $\left(\overline{\mathcal{R}}_\theta\right)$ corresponds to the case where w_{01}, w_{10} and w_m take their values in $\{0;1\}$. To see this, suppose that there exists $(i,j) \in \{01;10;m\}$ such that $(w_i, w_j) \in]0,1[^2$. Then consider $w_i^1 = w_i + \delta$, $w_j^1 = w_j - \delta$, $w_i^2 = w_i - \delta$, $w_j^2 = w_j + \delta$ with δ a function bounded by $\min(w_i, w_j, 1 - w_i, 1 - w_j)$ and positive on a set $A \in \mathbb{R}^2$ satisfying $\mathbb{P}((\varepsilon_1, \varepsilon_2) \in A) > 0$. Because $(w_i, w_j) = 1/2[(w_i^1, w_j^1) + (w_i^2, w_j^2)]$, the corresponding distribution cannot be an extreme point of cl $(\mathcal{R}_\theta)$. Let $u = (u_{00}, u_{01}, u_{10}, u_{11})$ be in the unit sphere $\mathbb{S}_4$ of $\mathbb{R}^4$. Then $\theta \in \Theta_0$, if and only if:

$$\forall u \in \mathbb{S}_4 : \sup_{P \in \mathcal{R}_\theta} u' \int q_w(\varepsilon_1, \varepsilon_2) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) \geq u'\pi,$$

Now, we can show that $\mathcal{R}_\theta$ is closed. DETAILS? Then, by Theorem 2.1.2, $\theta \in \Theta_0$ if and only if

$$\forall u \in \mathbb{S}_4 : \sup_{(w_{01}, w_{10}, w_m) \in \{(0;0;1),(0;1;0),(1;0;0)\}^{\mathbb{R}^2}} \int u' q_w(\varepsilon_1, \varepsilon_2) dF_\theta(\varepsilon_1, \varepsilon_2) \geq u'\pi.$$

The maximization can be done pointwise inside the integral and then:

$$\begin{aligned} &\sup_{(w_{01}, w_{10}, w_m) \in \{(0;0;1),(0;1;0),(1;0;0)\}^{\mathbb{R}^2}} u' q_w(\varepsilon_1, \varepsilon_2) \\ &= \sum_{i,j \in \{0;1\}} u_{ij} \mathbb{1}_{S_{ij}}(\varepsilon_1, \varepsilon_2) + \max\left(u_{00} \frac{(\theta_1 + \epsilon_1)(\theta_2 + \epsilon_2)}{\theta_1 \theta_2} - u_{01} \frac{(\theta_1 + \varepsilon_1)\varepsilon_2}{\theta_1 \theta_2} - u_{10} \frac{\varepsilon_1(\theta_2 + \varepsilon_2)}{\theta_1 \theta_2} + u_{11} \frac{\varepsilon_1 \varepsilon_2}{\theta_1 \theta_2}, u_{01}, u_{10} \right) \\ &\qquad \times \mathbb{1}_{S_{\cdot\cdot}}(\varepsilon_1, \varepsilon_2). \end{aligned}$$

This reasoning applies to every games with unobserved heterogeneity belonging to a parametric family and with a finite number of strategies. This is precisely the result obtained by Beresteanu et al. (2011) using results from the random set theory □

Theorem 2.1.2 is well known when $\mathcal{R}$ is finite-dimensional. This case occurs in parametric models or when $\mathcal{S}$ is finite, since $\mathcal{R}$ is a subset of probability distributions with support included in $\mathcal{S}$. Let us recall the argument in this latter case. Without loss of generality, let $\mathcal{S} = \{1, ..., I\}$. Any $P \in \mathcal{R}$ is characterized by $P(\{i\})$ for $i = 1, ..., I$. We can therefore assimilate $\text{cl}(\mathcal{R}_\theta)$ with the compact subset of $[0, 1]^I$ of all $\lambda = (P(\{1\}), ..., P(\{I\}))'$ corresponding to a measure such that $q(P, \theta) = 0$. When, as here, $\text{cl}(\mathcal{R}_\theta)$ is a finite-dimensional, compact and convex set, $\text{ext}\left(\overline{\mathcal{R}_\theta}\right)$ is nonempty (see, e.g., Proposition 2.3.3 in Hiriart-Urruty & Lemaréchal, 2001) as soon as $\mathcal{R}_\theta$ is nonempty. In the linear case, a similar result holds for $\mathcal{R}$, which is also a compact set. Let $a = (f(1), ..., f(I))'$, we then get $\int f(u)dP(u) = a'\lambda$. Hence,

$$\overline{\theta} = \max_{\lambda \in \mathcal{R}} a'\lambda,$$

and similarly for the lower bound. Moreover, by Minkowski Theorem (see, e.g., Hiriart-Urruty & Lemaréchal, 2001, Theorems 2.3.4),

$$\mathcal{R} = \text{co}(\text{ext}(\mathcal{R})),$$

where $\text{co}(A)$ denotes the convex hull of a set A. As a result, any $\lambda \in \mathcal{R}$ can be written as $\lambda = \sum_{k=1}^{K} \alpha_k \lambda_k$, with $\lambda_k \in \text{ext}(\mathcal{R})$, $\alpha_k \geq 0$ and $\sum_{i=k}^{K} \alpha_k = 1$. This implies that $a'\lambda \leq \max_{k=1...K} a'\lambda_k$, and therefore

$$\max_{\lambda \in \mathcal{R}} a'\lambda = \max_{\lambda \in \text{ext}(\mathcal{R})} a'\lambda. \tag{2.3}$$

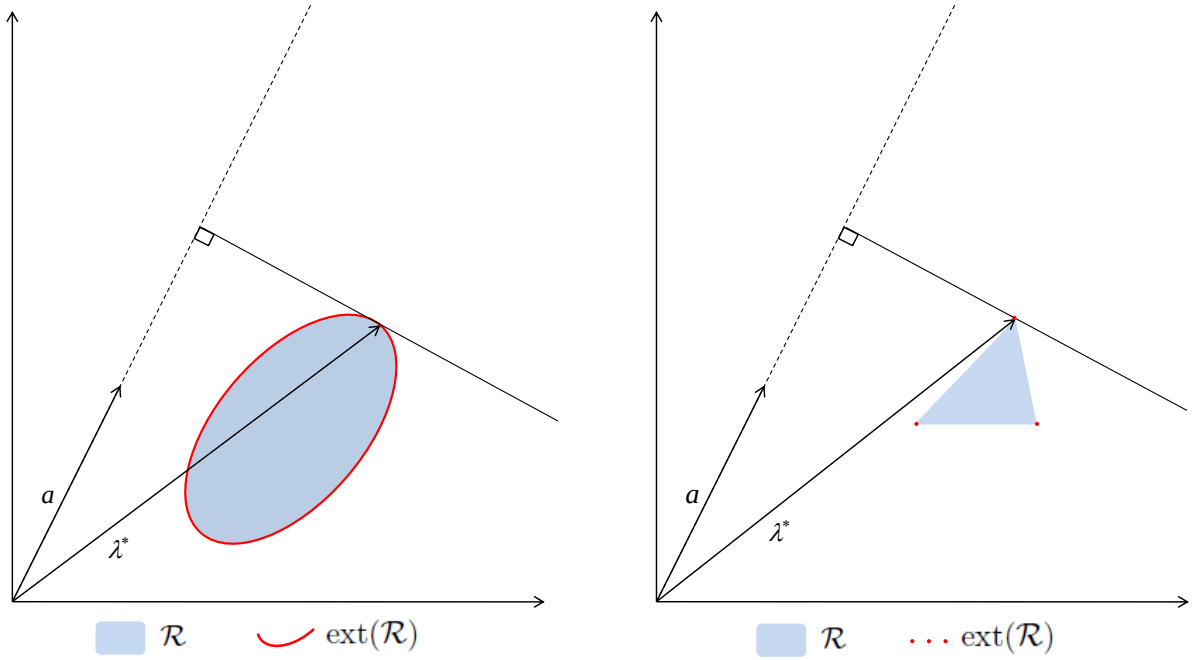
Figures 2a and 2b display two examples of extremal sets of a compact convex set, illustrate Minkowski Theorem and Equality (2.3). We are looking for the vector $\lambda \in \mathcal{R}$ that maximizes the (oriented) norm of its projection on the line generated by a. In both cases the maximum is reached on an extremal element of $\mathcal{R}$. In the first example, $\text{ext}(\mathcal{R})$ has an infinite number of points but is a one-dimensional manifold, whereas $\text{ext}(\mathcal{R})$ consists of only three points in the second example. This case corresponds to a standard linear programming problem, where optimization is conducted on a polyhedron. In such a case, a possibility is simply to compare $a'\lambda$ on each of these values.[4]

[4]This solution is inefficient, though, as the number of vertices can be very large. Simplex or interior point algorithms are much more efficient.

Figure 2: Linear optimization on compact convex sets of $\mathbb{R}^2$

(a) ext($\mathcal{R}$) is a one-dimensional manifold

(b) ext($\mathcal{R}$) is a finite set



$$\lambda^* = \arg\max_{\lambda \in \mathcal{R}} a'\lambda = \arg\max_{\lambda \in \text{ext}(\mathcal{R})} a'\lambda$$

Extending the results to the case where $\mathcal{R}$ is infinite dimensional space, on the other hand, is challenging. The Krein-Milman Theorem, which is the usual generalization of the Minkowski Theorem in infinite dimension, states that any compact convex set is the closure of the convex hull of its extreme points. However, cl $(\mathcal{R}_\theta)$ is only closed and bounded here. Because it may be infinite dimensional, it is not necessarily compact. In general, the lack of compactness of $\mathcal{R}$ and $\mathcal{R}_\theta$ can have severe consequences as the following counterexamples show. The first shows that a closed and convex subset of a Banach space needs not have extreme points. The second proves that even if a closed and bounded convex set has extreme points, it may not be equal to the closure of the convex hull of its extreme points.

Counterexample 1: Existence of extreme points.
Let $\mathcal{K}$ denote the set of real valued continuous functions f from $[0;1]$ such that $\sup_{x\in[0;1]} |f(x)| \leq 1$ and $f(0) = 0$. $\mathcal{K}$ is a bounded, closed and convex set for the supremum norm in the Banach space of continuous functions from $[0;1]$ to $\mathbb{R}$. However, it is easy to see that it has no extreme points $\square$

Counterexample 2: Convex hull of extreme points.
Let $\mathcal{K}$ be the set of real valued continuous functions f from $[-1;1]$ such that $\sup_{x\in[-1;1]} |f(x)| \leq 1$. $\mathcal{K}$ is a bounded, closed and convex set of a Banach space, and the set of its extreme points $\text{ext}(\mathcal{K})$ satisfies

$$\text{ext}(\mathcal{K}) = \{f : f(x) = 1 \text{ for } x \in [-1;1] \text{ or } f(x) = -1 \text{ for } x \in [-1;1]\}.$$

Thus, $\text{cl}\,(\text{co}(\text{ext}(\mathcal{K})))$ is the set of constant functions from $[-1;1]$ to itself, and $\text{cl}\,(\text{co}(\text{ext}(\mathcal{K}))) \neq \mathcal{K}$. It follows that optimization of linear forms on $\mathcal{K}$ does not reduce to the optimization on $\text{ext}(\mathcal{K})$. Consider for instance h the linear form defined by $h(f) = \int xf(x)dx$. In this case,

$$\sup_{f\in\text{ext}(\mathcal{K})} h(f) = 0 < 1 = \sup_{f\in\mathcal{K}} h(f) \ \square$$

In the linear case, to ensure compactness of $\mathcal{R}$ and thus use the Krein-Milman Theorem, a possibility would be to choose a convenient topology, the weak-* topology for instance. This would ensure that

$$\sup_{P\in\mathcal{R}\cap\mathcal{I}(f)} \int f dP = \sup_{P\in\text{cl}(\text{co}(\text{ext}(\mathcal{R})))\cap\mathcal{I}(f)} \int f dP.$$

Moreover, we can easily prove, as we did before, that

$$\sup_{P\in\text{co}(\text{ext}(\mathcal{R}))\cap\mathcal{I}(f)} \int f dP = \sup_{P\in\text{ext}(\mathcal{R})\cap\mathcal{I}(f)} \int f dP.$$

An issue arises, however, at this stage. It is not straightforward that

$$\sup_{P\in\text{cl}(\text{co}(\text{ext}(\mathcal{R})))\cap\mathcal{I}(f)} \int f dP = \sup_{P\in\text{co}(\text{ext}(\mathcal{R}))\cap\mathcal{I}(f)} \int f dP. \tag{2.4}$$

This holds if $P \mapsto \int f dP$ is continuous, but in our infinite-dimensional setting this is a restrictive condition. Also, the choice of the topology matters there. Under the weak-* topology, continuity of such map holds only if f is continuous and vanishes at infinity. These restrictions do not hold for standard choices of f such as $f(u) = u$ (if support of U is unbounded) or $f(u) = \mathbb{1}\{u \leq t\}$. To be able to drop these restrictions, we rely on an extension of the Krein-Milman Theorem, namely the Choquet Theorem. Basically, this result provides a representation of any element of a compact, convex set A by an integral over $\text{ext}(A)$. Using this integral representation, we are able to show directly Theorem 2.1, without having to prove (2.4).

2.4 Converging outer bounds

It may happen that the set $\text{ext}(\mathcal{R})$ is difficult to characterize or too large to yield a tractable optimization algorithm. In such circumstances, we may still be able to compute outer bounds arbitrarily close to the true ones, if $\mathcal{R}$ can be written as the intersection of a decreasing sequence $(\mathcal{R}_n)_{n\in\mathbb{N}}$. In this case, $\mathcal{R}_\theta = \cap_{n\in\mathbb{N}}\mathcal{R}_{\theta,n}$ with $\mathcal{R}_{\theta,n} = \{P \in \mathcal{R}_n : q(\theta, P) = 0\}$. We discuss the characterization of Θ_0 by $\text{ext}(\text{cl}(\mathcal{R}_{\theta,n}))$ in such cases. The following assumption corresponds , which mimics Assumptions 2 and 3.

Assumption 4 (Intersection of decreasing convex sets) *(i)* $\mathcal{R}_{\theta,n} = \{P \in \mathcal{R}_n : q(\theta, P) = 0\}$ *is a decreasing sequence of closed and convex subsets of* $\mathcal{P}$ *such that* $\mathcal{R}_\theta = \cap_{n\in\mathbb{N}}\mathcal{R}_{\theta,n}$.

(ii) For every θ*, there exists* $\varepsilon > 0$ *and* $n_0 \in \mathbb{N}$ *such that:*

$$\sup_{P\in ext(\mathcal{R}_{n_0,\theta})} \int ||u||^\varepsilon dP(u) < \infty.$$

When $\theta \mapsto \text{ext}(\text{cl}(\mathcal{R}_\theta))$ is difficult to characterize then Θ_0 remains difficult to characterize. On the other hand, under the previous assumption, if we are able to compute $\Theta_{0n} = \{\theta : \text{cl}(\mathcal{R}_{\theta,n}) \neq \emptyset\}$ for every n, this will give us a sequence of decreasing outer regions because $\Theta_0 \subset \Theta_{0n}$. Such a sequence Θ_{0n} will converge to the identification region only if $\Theta_0 = \lim_{n\to+\infty} \downarrow \Theta_{0n} = \cap_{n\in\mathbb{N}}\Theta_{0n}$. The following theorem gives technical conditions under which such convergence holds.

Theorem 2.2 (Converging outer regions)
Under Assumptions 1 and 4, $\Theta_0 = \cap_{n\in\mathbb{N}}\Theta_{0n}$ *with* $\Theta_{0n} = \{\theta : ext(\mathcal{R}_{\theta,n}) \neq \emptyset\}$.

The previous theorem may be adapted to the special case where the parameter is a moment of P, under the following assumption.

Assumption 5 (Intersection of decreasing convex sets, separable case) *(i)* $\mathcal{R}_n$ *is a decreasing sequence of closed and convex subsets of* $\mathcal{P}$ *such that* $\mathcal{R} = \cap_{n\in\mathbb{N}}\mathcal{R}_n$ *and* $q(\theta, P) = \theta - \int f dP$.

(ii) f *is continuous.*

(iii) There exists $\varepsilon > 0$ *and* n_0 *such that:*

$$\sup_{P\in ext(\mathcal{R}_n)} \int |f(u)|^{1+\varepsilon} \vee |u|^\varepsilon dP < +\infty,$$

Then let $\underline{\theta}_n = \inf_{P \in \text{ext}(\mathcal{R}_n) \cap \mathcal{I}(f)} \int f(u)dP(u)$ and $\overline{\theta}_n = \sup_{P \in \text{ext}(\mathcal{R}_n) \cap \mathcal{I}(f)} \int f(u)dP(u)$.

Corollary 2.3 (Converging outer bounds for linear parameters)
Under Assumptions 1 and 5, $\underline{\theta}_n \to \underline{\theta}$ and $\overline{\theta}_n \to \overline{\theta}$.

Note that if we are interested only by the result on the upper bound (resp. lower bound) of θ, only the lower (resp. upper) semi-continuity of f is needed. On the other hand, the result may fail to hold if we weaken further Conditions (ii) and (iii) of Assumption 5, as the following counterexamples show.

Counterexample 3: continuity.
Let $\mathcal{S} = [-1; 1]$, and $f(x) = \mathbb{1}_{\{x>0\}}$. Suppose that $\mathcal{R}$ is defined by the following moments conditions:

$$\mathcal{R} = \left\{ P : \int_{-1}^{1} x^k dP(x) = 0 \text{ if } k \text{ is odd and } \int_{-1}^{1} x^k dP(x) = 1/(2(k+1)) \text{ if } k \text{ is even} \right\}.$$

Let $\mathcal{R}_n$ be the set of distributions corresponding to the n first moments conditions in $\mathcal{R}$. We show in the appendix that Assumption 4 and conditions (i) and (iii) of Assumption 5 hold, but not condition (ii). We also establish that $\mathcal{R}$ is reduced to the singleton $1/2\mathcal{U}_{[-1;1]} + 1/2\delta_0$, so that $\sup_{P \in \mathcal{R}} \int f dP = 1/4$. On the other hand, $\sup_{P \in \mathcal{R}_n} \int f dP \geq 3/4$.

Counterexample 4: uniform integrability condition.
Let $\mathcal{S} = \mathbb{R}$, $f(x) = x$ and consider the functions

$$g(x) = q \max_{j=1...k} (1 - p_j |x - s_j|)^+ ,$$

where $k \in \mathbb{N}$, $(q, s_1, ..., s_k) \in \mathbb{Q}^{k+1}$, $(p_1, ..., p_k) \in \mathbb{Q}_+^k$, $|q| \leq 1$ and $q \max_{i=1...k} p_i \leq 1$. Because the class $\mathcal{G}$ of such functions is countable, we can write $\mathcal{G} = \{(g_i)_{i \in \mathbb{N}}\}$. Let $Z \sim N(0, 1)$ and let

$$\begin{aligned} \mathcal{R} &= \left\{ P \in \mathcal{P} : \int g_i(x) dP(x) = \mathbb{E}(g_i(Z)),\ \forall\, i \in \mathbb{N} \right\}, \\ \mathcal{R}_n &= \left\{ P \in \mathcal{P} : \int g_i(x) dP(x) = \mathbb{E}(g_i(Z)),\ i \in \{1, ..., n\} \right\}. \end{aligned}$$

We show in the appendix that Assumption 4 and conditions (i) and (ii) of Corollary 2.3 hold, but not condition (iii). Finally, $\overline{\theta}_n = +\infty > \overline{\theta} = 0$.

3 Application to problems with moment equalities

As extremal points are the keystone of our strategy to characterize Θ_0, it is important to be able to characterize them. A large literature has focused on the extreme parts

of distributions in various contexts. In many of cases, $\mathcal{R}_\theta$ can be expressed as a set of probability distributions that verify a set of moments as in a GMM estimation or in optimal transportation problem (Ekeland et al. (2010), Galichon & Henry (2011), Chiappori et al. (2010)). For optimal transportation problem a significant literature has focus on this problem (see for instance Ahmad et al. (2011) for a recent work on this topics). We will give a useful result to characterize the extreme part of $\mathcal{R}_\theta$ in such a case.
An important result have been given by Douglas (1964) in case of equality of moments. We extend his result to inequalities of moments.

Theorem 3.1 (Extension of Douglas (1964))
Let $\mathcal{G}$ a family of real valued functions and let

$$\mathcal{K} = \left\{ P \in \mathcal{P},\ s.t.\ \ \forall g \in \mathcal{G} \quad \int |g| dP < +\infty \quad and \quad \int g dP = 0 \right\},$$

$$\mathcal{L} = \left\{ P \in \mathcal{P},\ s.t.\ \ \forall g \in \mathcal{G} \quad \int |g| dP < +\infty \quad and \quad \int g dP \geq 0 \right\}.$$

If $\mathcal{K}$ is not empty, $P \in ext(\mathcal{K})$ if only if $span(\mathcal{G}, 1)$ is dense in $L_1(P)$ and $P \in ext(cl(\mathcal{K}))$ only if $span(\mathcal{G}, 1)$ is dense in $L_1(P)$.
If $\mathcal{L}$ is not empty, $P \in ext(\mathcal{L})$ (respectively $P \in ext(cl(\mathcal{L}))$) only if $span(\mathcal{G}, 1)$ is dense in $L_1(P)$.

3.1 Finite number of moments equalities and/or inequalities

We derive from the result of Douglas an interesting result when the parameter θ and the restrictions are defined by a finite number of (in)equalities of moments. In this case, optimization on the possibly infinite dimensional set $\mathcal{R}_\theta$ can be reduced to a finite dimensional problem. In fact optimization can be done only on distributions that have a limited number of points in their support. Let $\mathcal{P}_j$ the subset of $\mathcal{P}$ consisting of distributions that have at most j support points.

Theorem 3.2 *If $q(P, \theta) = \int m(U, \theta) dP$ with $m = (m_1, ...m_l)$ and $m_i(., \theta)$ continuous and bounded real valued functions on $\mathcal{S}$. If $g_1, ..., g_k$ are continuous and bounded functions on $\mathcal{S}$ such that*

$$\mathcal{R} = \left\{ P \in \mathcal{P},\ s.t.\ \ \forall j = 1, ..., k, \int g_j dP = 0 \right\},$$

then

$$\Theta_0 \ = \ \left\{ \theta : min_{P \in \mathcal{P}_{k+l+1}} \left(\sum_{i=1}^{l} \left|\left| \int m_i(u, \theta) dP(u) \right|\right| + \sum_{j=1}^{k} \left|\left| \int g_j(u) dP(u) \right|\right| \right) = 0 \right\}.$$

ADD THE SEPARABLE CASE

The previous theorem shows that infinite dimensional optimization can be replaced by an optimization on a finite dimensional space. Moreover, the previous theorem is stated with moment equalities, but can also be easily adapted when for some g_i, we only have the inequality condition $\int g_i dP \geq 0$. In this case one need to replace $\left|\left|\int g_j(u)dP(u)\right|\right|$ by $\left(\int g_j(u)dP(u)\right)^-$ in the characterization of Θ_0 for the corresponding g_i (where $(a)^- = -a$ if $a < 0$ and 0 otherwise).

When conditions of moments are given by a countable linearly independent family of continuous and bounded functions $\mathcal{G} = \{g_1, g_2, ...\}$, we can use the results of previous Section with the sequence $\mathcal{R}_n = \{P \in \mathcal{P} : \int |g_k| < \infty \text{ and } \int g_k dP = 0 \text{ for } k \leq n\}$.

Example 1: Average effects in nonlinear panel data models.

Chernozhukov et al. (2013) derive bounds of average and quantile effect in nonseparable panel models. To simplify consider the average treatment effects on a simple non parametric binary panel model with a binary regressor. Let $Y_{it} \in \{0;1\}$, $X_{it} \in \{0;1\}$ and $Y_i = (Y_{i1}, ..., Y_{iT})$, $X_i = (X_{i1}, ..., X_{iT})$. Chernozhukov et al. (2013) assume that it exists $\alpha_i \in \mathbb{R}^k$ and $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT}) \in \mathbb{R}^T$ such that $Y_{it} = g_0(X_{it}, \alpha_i, \varepsilon_{it})$. The Average Treatment Effect is given by:

$$\begin{aligned} ATE &= \int \left[g_0(1,a,e) - g_0(0,a,e)\right] dF_{\alpha_i,\varepsilon_{i1}}(a,e) \\ &= \int \left[g_0(1,a,e) - g_0(0,a,e)\right] dF_{X_i,\alpha_i,\varepsilon_{i1}}(x,a,e) \end{aligned}$$

For every $(y,x) \in \{0;1\}^{2T}$, identification based on the data of $\mathbb{P}(Y = y, X = x)$ gives a constraint of moment on $F_{X_i,\alpha_i,\varepsilon_i}$:

$$\int \mathbb{1}\{g_0(u,a,e) = y, u = x\} dF_{X_i,\alpha_i,\varepsilon_i}(u,a,e) = \mathbb{E}\left(\mathbb{1}\{Y_i = y, X_i = x\}\right).$$

Without supplementary assumptions, Theorem 3.2 ensures that extremal points of set of distributions $(X_i, \alpha, \varepsilon_i)$ compatible with the data are mixture of Dirac distribution with at most 2^{2T} support points in $\{0;1\}^T \times \mathbb{R}^k \times \mathbb{R}$. Bounds on ATE are given by optimization on 2^{2T} values of $(\alpha_i, \varepsilon_i)$ such that both individuals having same trajectory (Y_i, X_i) share the same value of $(\alpha_i, \varepsilon_i)$.

Example 2: measuring segregation in small units.

Cortese et al. (1978), Winship (1977), Carrington & Troske (1997), Allen et al. (2009), Rathelot (2011), D'Haultfoeuille & Rathelot (2014) consider the issue of measuring segregation on small units such as small firms or classrooms. For simplicity, suppose that these units have constant size equal to K. For each unit i ($i = 1, ..., N$), let us denote by p_i the probability that an individual belongs to the minority. Note that p_i differs from the actual

proportion $\widehat{p}_i$. If all the p_i are the same between units, there is no segregation. More generally, segregation is measured by an inequality index, such as the Duncan or Theil indices, on the distribution of the p_i's. It is not straightforward to make some inference on such inequality indices, because the distribution of the observed $\widehat{p}_i$'s is not equal to the one of p. In such a model, only the K first moments of p are identifiable (D'Haultfoeuille & Rathelot, 2014). Theorem 3.2 shows that sharp bounds for $D(F_p)$ and $T(F_p)$ can be computed by maximization (respectively minimization) on distributions that have $K+1$ support points.

3.2 Optimal transportation problem

Theorem 3.1 associated with our main result allows to recover the Monge-Kantorovitch duality used in optimal transportation with other techniques of proofs than these used by, e.g., Villani (2003, 2009).

The Monge-Kantorovitch duality is used to maximize $\int f(u_1,u_2)dP(u_1,u_2)$, a moment that depends on two sets of variables U_1 and U_2, when we only known the marginal distributions P_1 and P_2 of U_1 and U_2 (but not the joint distribution of (U_1,U_2)). In this case, moments that depend only from marginal distributions are known. So the programm can be written as $\sup_{P\in\mathcal{R}} \int f(u,v)dP(u,v)$, with

$$\mathcal{R} = \left\{ \begin{array}{ll} P : \forall (g,h) \in L_1(P_1)\times L_1(P_2), & \int g(u_1)dP(u_1,u_2) = \int g(u_1)dP_1(u_1) \\ & \text{and } \int h(u_2)dP(u_1,u_2) = \int h(u_2)dP_2(u_2) \end{array} \right\}.$$

Theorem 2.1 ensures that maximization can be done only on $\text{ext}(\mathcal{R})$ instead of $\mathcal{R}$ and Theorem 3.1 ensures that for every $P \in \text{ext}(\mathcal{R})$, $f(u_1,u_2)$ can be written as $g(u_1)+h(u_2)$ with $(g,h) \in L_1(P_1)\times L_1(P_2)$. So we recover a deep result simply using our main result and some classical characterization of extreme parts. Moreover one can easily give a more general result when we have more than two marginal distributions.

Let $U = (U_1, U_2, ..., U_n)$ a random vector in $\mathbb{R}^{d_1} \times ... \times \mathbb{R}^{d_n}$. The random sub-vectors U_i $(i = 1, ..., n)$ are distinct but can overlap, i.e. $U_i = (U_{i1}, ..., U_{id_i})$ can have common component with $U_{i'} = (U_{i'1}, ..., U_{i'd_{i'}})$. Let P_i $(i = 1...n)$ the probability distributions of U_i supported by $\mathcal{S}_i \subset \mathbb{R}^{d_i}$. We assume that each P_i is identified (by the data or by additional restrictions) but not the full distribution of U. The parameter of interest is a moment of U, $\theta_0 = \int f(u)dP(u)$. So the set of restrictions compatible with the data and the restrictions can be expressed as an infinite number of moments:

$$\mathcal{R} = \{P \in \mathcal{P}, \forall g \in L_1(P_i), \int g(u_i)dP(u_1,...,u_n) = \int g(u_i)dP_i(u_i)\}.$$

Theorem 3.3 (Monge-Kantorovitch duality)
Let f a function, we have:

$$\sup_{P\in\mathcal{R}\cap\mathcal{I}(f)} \int f dP \;=\; \inf_{g_i\in L_1(P_i)\sum g_i\geq f} \sum_{i=1}^n \int g_i(u_i) dP_i(u_i),$$

and

$$\inf_{P\in\mathcal{R}\cap\mathcal{I}(f)} \int f dP \;=\; \sup_{g_i\in L_1(P_i)\sum g_i\leq f} \sum_{i=1}^n \int g_i(u_i) dP_i(u_i),$$

where

$$\mathcal{R} = \{P\in\mathcal{P}, \forall g\in L_1(P_i), \int g(u_i)dP(u_1,...,u_n) = \int g(u_i)dP_i(u_i)\}.$$

For $n = 2$, this results have been used by Ekeland et al. (2010) and Galichon & Henry (2011) to characterize the identification regions in various models.

4 Application to the sample selection model

4.1 The setting

The sample selection model is an important particular case of our general framework with $U = (D, Y, X)$ and $O = (DY, D, X)$ and $\text{Supp}(D) = \{0; 1\}$. It has been widely used in parametric framework and under the usual restriction $Y \perp\!\!\!\perp X$ since Heckman (1974). More recently, Manski (2003) and Kitagawa (2010) have discuss issue of identification in a nonparametric framework. However, the existence of a variable X satisfying $X \perp\!\!\!\perp Y$ is often questionable. However, one can work under alternative restrictions.

Assumption 6 (Sample Selection)
We observe an iid sample of (D, DY, X) with Supp$(D) = \{0; 1\}$.

In line with the previous sections, $P_0^{D,DY,X}$, $P_0^{Y|D=1,X=x}$ and $P_0^{Y|D=1}$ denote respectively the identified distributions of (D, DY, X), $Y|D = 1, X = x$ and $Y|D = 1$.
In this section, we derive the extreme points of joint distribution (D, Y, X) under monotonicity conditions on the selection. This ensures identification of Θ_0 when θ is defined by moments conditions $\mathbb{E}(m(D, Y, X, \theta)) = 0$. We consider hereafter the two following conditions.

Assumption 7 (Monotonicity in X)
$x \mapsto \mathbb{E}(D|Y, X = x)$ *is increasing almost surely.*

Assumption 8 (Monotonicity in Y)
$y \mapsto \mathbb{E}(D|Y = y, X)$ *is increasing almost surely.*

Note that the two previous assumptions can not be expressed as moment inequalities. So we are not in the framework detailed in the previous section. This shows that our result also apply to setup not treated in the literature.
The two previous assumptions are credible in some situations. Consider for instance the female labour supply model of Gronau (1974). In this model, individuals self-select into the labour market if their potential wage Y is larger than their reservation wage W^*: $D = \mathbb{1}\{Y \geq W^*\}$. Suppose also that $W^* = g(X) + \xi$, where $\xi \perp\!\!\!\perp (X, Y)$ (cf. Equation 15 of Gronau (1974)). In this case,

$$\mathbb{P}(D = 1|X, Y) = F_\xi(Y - g(X)),$$

where F_ξ denotes the cdf of ξ. In this framework, the missing at random assumption is never satisfied. On the other hand, in this framework[5] $\mathbb{P}(D = 1|X, Y = y)$ is an increasing function of y. Similarly, in some cases, it might be reasonable to impose monotonicity conditions on g and in this case $\mathbb{P}(D = 1|X = x, Y)$ is a monotone function of x. For instance, it seems reasonable to assume that $x \mapsto g(x)$ is increasing when considering X as the number of children.
Moreover, to ensure a sufficient regularity to the model we assume supplementary technical conditions.

Assumption 9 (Support Condition) *1.* $P(D = 1|X) > 0$.

2. The support of $Y|X, D = 0$ *is included in the support of* $Y|X, D = 1$, X *almost-surely.*

3. $\mathcal{Y} = Supp(Y|X, D = 1)$, X *almost-surely.*

4. $\mathcal{X} = Supp(X) = \{x_1, x_2, ..., x_J\}$ *with* $x_1 < ... < x_J$.

Under Assumption 9.1 and 9.2, the support of (Y, X) is equal to the support of (DY, X). If relaxed, auxiliary information would be needed to identify this support of (Y, X), a necessary step for obtaining informative bounds on many parameters of interest. Assumption 9.3 is essentially made for sake of simplicity, and the following reasoning can be easily adapted when the support of $Y|D = 1, X = x$ depends on x.

[5]The model of Gronau is restrictive because $\xi \perp\!\!\!\perp (X, Y)$ contrary to the most popular assumptions (Heckman (1974)). We can also consider more general frameworks for instance, $W^* = g(X, Y) + \xi$, with $F_{\xi|Y,X}(Y - g(Y, X))$ monotone in Y. Such assumption is made for instance by Blundell et al. (2007).

Note also that under Assumption 9, the inverse of probability of selection $\rho(y,x) = 1/\mathbb{P}(D=1|Y=y,X=x)$ is defined almost surely. In the following, $p(x)$ denotes $\mathbb{P}(D=1|X=x)$.
Under Assumption 9, $P_0^{Y|D=1}$ is a dominant measure of $P_0^{Y|D=1,X=x}$, so we can define $f_{Y|D=1,X=x}$ as the density of $P_0^{Y|D=1,X=x}$ with respect to the distribution of $P_0^{Y|D=1}$ for every $x \in \mathcal{X}$.

Assumption 10 (Continuity of density ratios) *$f_{Y|D=1,X=x_j}/f_{Y|D=1,X=x_i}$ is continuous for every $(x_i,x_j) \in \mathcal{X}^2$.*

4.2 Characterization of the bounds

Rather than considering, $\mathcal{K}$, the set of distributions of (D,Y,X) compatible with the data and the model, we focus hereafter for convenience on $\mathcal{C}$, the set of distributions of $(Y|D=0,X=x_1,Y|D=0,X=x_2,...,Y|D=0,X=x_J)$. Because $\mathcal{K}$ and $\mathcal{C}$ are related through a one-to-one mapping, characterizing $\mathcal{C}$ is equivalent to characterizing the set of distributions rationalized by the data and the model. The following lemma also proves that Assumption **??**old in this set-up, so that we can apply Theorem 2.1 and focus on ext($\mathcal{C}$) only.

Lemma 4.1 *Suppose that Assumptions 6, 9-10 and either 7, 8 or both hold. Then Assumptions 1 and 2 are satisfied.*

This lemma, together with Theorem **??**, ensures that we can focus on ext($\mathcal{C}$) only. We first characterize this set under Assumption 7. For $(x_i,x_j) \in \mathcal{X}^2$, let $r_{i,j} = p(x_j)(1-p(x_i))f_{Y|D=1,X=x_j}/(p(x_i)(1-p(x_j))f_{Y|D=1,X=x_i})$.

Proposition 4.1 (Monotone Selection in X with discrete X)
Suppose that Assumptions 6, 7, 9 and 10 hold.

$$ext(\mathcal{C}) \subset \left\{ \begin{array}{l} \sum_{j=1}^{J}\left[\sum_{k=\max(i,j)}^{J} w_{kj} r_{ki}(y_j)\right]\delta_{y_j},\ (y_1,...,y_J,w_{11},...,w_{JJ}) \in \mathcal{Y}^J \times [0,1]^{J(J+1)/2}, \\ \sum_{j=1}^{J}\left[\sum_{k=\max(i,j)}^{J} w_{kj} r_{ki}(y_j)\right] = 1 \end{array} \right\},$$

Assumption 7 is rejected if and only if the following set is empty:

$$H = \left\{ (y_1,...,y_J,w_{11},...,w_{JJ}) \in \mathcal{Y}^J \times [0,1]^{J(J+1)/2},\ \sum_{j=1}^{J}\left[\sum_{k=\max(i,j)}^{J} w_{kj} r_{ki}(y_j)\right] = 1 \right\}.$$

The bound of $\mathbb{E}(h(Y))$ *are given by:*

$$\mathbb{E}(h(Y)D) + \sup_{(y_1,...,w_{JJ})\in H} \sum_{i=1}^{J} \mathbb{P}(D=0 \cap X_i = x_i) \sum_{j=1}^{J} \left[\sum_{k=\max(i,j)}^{J} w_{kj} r_{ki}(y_j) \right] h(y_j)$$

$$\mathbb{E}(h(Y)D) + \inf_{(y_1,...,w_{JJ})\in H} \sum_{i=1}^{J} \mathbb{P}(D=0 \cap X_i = x_i) \sum_{j=1}^{J} \left[\sum_{k=\max(i,j)}^{J} w_{kj} r_{ki}(y_j) \right] h(y_j)$$

If $\mathcal{Y}$ *is compact and* h *continuous the upper bound (respectively lower bound) is trivial i.e. equal to* $\mathbb{E}(h(Y)D) + \max h(\mathcal{Y})$ *(respectively* $\mathbb{E}(h(Y)D) + \min h(\mathcal{Y})$*) if and only if it exists* $\overline{y}$ *(respectively* $\underline{y}$*) such that* $h(\overline{y}) = \max h(\mathcal{Y})$ *(respectively* $h(\underline{y}) = \min h(\mathcal{Y})$*) and* $r_{i+1,i}(\overline{y}) \leq 1$ *(respectively* $r_{i+1,i}(\underline{y}) \leq 1$*).*

The last part of the previous proposition shows that Assumption 7 is often informative when h is continuous and $\mathcal{Y}$ is compact. This is also the case when $\mathcal{Y}$ is not compact but in this case it is more tricky to derive a necessary and sufficient condition.
The previous proposition can be used to derive bounds with multiple discrete variable X and when we assume that $\mathbb{E}(D|Y, X_1 = x_1, X_2 = x_2, ..., X_k = x_k)$ is increasing in x_i for every i. In this case $\mathcal{C}$ is a vector of $J_1 \times ... \times J_k$ probabilities corresponding to the distributions of $Y|D = 0, X_1 = x_1, ..., X_k = x_k$ where x_i vary among J_i values. For each i and each value of $(x_j)_{j\neq i}$ the vector of distribution $P_{i,(x_j)_{i\neq j}}$ corresponds to the distribution vector of distributions of $(Y|D = 0, (X_j)_{j\neq i} = (x_j)_{j\neq i}, X_i = x_{il})$ when x_{il} vary from x_{i1} to x_{iJ_i} and belongs to a set of distributions described in the previous Proposition (replacing J by J_i). Combination of such Assumptions of componentwise monotonicity concerning variables X may decrease drastically the set of identification of the joint distribution $(Y, D, X_1, ..., X_k)$, and in some case the set of identification is empty.
We can also derive bounds under monotonicity Assumption in Y. And in this case, set of extreme parts of $\mathcal{C}$ takes a particularly simple expression.

Proposition 4.2 (Monotone Selection in Y **with discrete** X**)**
Suppose that Assumptions 6, 8, 9 and 10 hold.

$$ext(\mathcal{C}) = \left\{ (P_1, ..., P_J) : \forall j \in J, \exists y_j \in \mathcal{Y} \;\; s.t. \;\; P_j = P_0^{Y|D=1, X=x_j, Y\leq y_j} \right\}.$$

Assumption 8 of monotonicity can not be rejected. The bounds of $\mathbb{E}(h(Y))$ *are given by:*

$$\mathbb{E}(h(Y)D) + \sum_{j=1}^{J} \mathbb{P}(D=0 \cap X = x_j) \sup_{y_j\in\mathcal{Y}} \mathbb{E}(h(Y)|Y \leq y_j, D=1, X=x_j)$$

$$\mathbb{E}(h(Y)D) + \sum_{j=1}^{J} \mathbb{P}(D=0 \cap X=x_j) \inf_{y_j \in \mathcal{Y}} \mathbb{E}(h(Y)|Y \leq y_j, D=1, X=x_j)$$

Moreover if h is increasing (respectively decreasing), the upper bound (respectively the lower bound) corresponds to the estimation under the MAR Assumption

$$\sum_{j=1}^{J} \mathbb{P}(X=x_j)\mathbb{E}(h(Y)|D=1, X=x_j),$$

and the lower bound (respectively the upper bound) corresponds to the trivial bound

$$\mathbb{E}(h(Y)D) + \mathbb{P}(D=0)h(\inf(\mathcal{Y}))$$

$$\textit{respectively } \mathbb{E}(h(Y)D) + \mathbb{P}(D=0)h(\sup(\mathcal{Y}))$$

A natural extension consists to combine both assumptions of monotonicity in X and in Y.

Proposition 4.3 (Monotone Selection in Y and X with discrete X)
Suppose that Assumptions 6, 7, 8, 9 and 10 hold.

$$ext(\mathcal{C}) \subset \left\{ \begin{array}{l} (P_1, ..., P_J) : \exists (\alpha_1, ..., \alpha_J) \in \mathbb{R}^{+J}, \exists (y_{11}, ..., y_{JJ}) \in \mathcal{Y}^{J^2} \\ \forall j, P_j(]-\infty; y]) = \frac{p(x_j)}{1-p(x_j)} \sum_{i=1}^{J} \alpha_i \mathbb{P}(Y \leq y_{ij} \wedge y | D=1, X=x_j) \\ y_{i,j} \leq y_{i,j+1} \wedge y_{i+1,j} \end{array} \right\}.$$

Assumptions 7 and 8 of monotonicity are jointly rejected if and only if the following set is empty:

$$H = \left\{ \begin{array}{l} (\alpha_1, ..., \alpha_J, y_{11}, ..., y_{JJ}) \in \mathbb{R}^{+J} \times \mathcal{Y}^{J^2} \textit{ such that} \\ \forall j, \frac{p(x_j)}{1-p(x_j)} \sum_{i=1}^{J} \alpha_i \mathbb{P}(Y \leq y_{ij} | D=1, X=x_j) = 1 \\ \forall (i,j), y_{i,j} \leq y_{i,j+1} \wedge y_{i+1,j} \end{array} \right\}$$

And the bounds of $\mathbb{E}(h(Y))$ are given by:

$$\mathbb{E}(h(Y)D) + \sup_{(\alpha_1,...,\alpha_j,y_{11},...,y_{JJ}) \in H} \sum_{j=1}^{J} \sum_{i=1}^{J} \alpha_i \mathbb{E}(D \mathbb{1}_{\{X_j=x_j\}} h(Y) \mathbb{1}_{\{Y \leq y_{i,j}\}})$$

$$\mathbb{E}(h(Y)D) + \inf_{(\alpha_1,...,\alpha_j,y_{11},...,y_{JJ}) \in H} \sum_{j=1}^{J} \sum_{i=1}^{J} \alpha_i \mathbb{E}(D \mathbb{1}_{\{X_j=x_j\}} h(Y) \mathbb{1}_{\{Y \leq y_{i,j}\}})$$

This result can be used to derive sharp bounds with several discrete X and when assumption of monotonicity of selection is made for each variable. In this case ext$(\mathcal{C})$ is a set of $J_1 \times ... \times J_k$ distributions. The distributions corresponding to $Y|D=0, X_1 = x_1, ..., X_l = x_j, ..., X_k = x_k$ when x_l varies from x_{l1} to x_{lJ_l}, depends on at most J_l^2 values

of Y $(y_{ij}(x_1, ..., x_{l-1}, x_{l+1}, ..., x_J))$and on J_l positives values $\alpha(x_1, ..., x_{l-1}, x_{l+1}, ..., x_k)$ that verify the relations in the right hand side equation in the previous proposition.
When the support of Y is bounded, the previous result can be used to construct converging outer bounds when X is a continuous variable. Indeed, the range of X can be partitioned in a n subintervals. Note that if Assumptions of monotonicity holds for the continuous X, they hold also for discretized variables. So using the previous result, we get outer bounds. When the length of subintervals tends to zero (and then n tends to infinity) we can apply Theorem 2.2, with $q(\theta, P) = \theta - P$ and with $\mathcal{R}_n$ derived by Assumption of monotonicity for discretized variables.

4.3 Illustration

Finally, we show through a particular example, that the monotonicity condition, though rather weak, can deliver tight bounds, especially compared to the bounds obtained without any assumption (referred to as the Manski bounds hereafter). We consider X^* and Y, two uniform variables on $[0, 1]$ with copula

$$C(x, y; \theta) = (1 - \theta)xy + \theta \min(x, y).$$

This copula is a mixture between the independent copula and the Fréchet copula corresponding to perfect, positive dependence. We also suppose the following linear probability model

$$P(D = 1|X^*, Y) = \alpha + \beta X^* + \gamma Y + \delta X^* Y.$$

Finally, we consider X to be a discretization of X^*: $X = j \in \{1, ..., J\}$ if $X^* \in [(j - 1)/J, j/J)$. One can show that in this case, both Assumptions 7 and 8 hold as soon as XX. We introduce X^* and its discretized version to illustrate the convergence of the sequence of outer bounds considered in Subsection **??**.

References

Abadie, A. (2002), 'Bootstrap tests for distributional treatment effects in instrumental variable models', *Journal of the American Statistical Association* **97**(457), 284–292.

Abadie, A., Angrist, J. & Imbens, G. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earning', *Econometrica* **70**(1), 91–117.

Ahmad, N., Kim, H. & McCann, R. (2011), 'Optimal transportation, topology and uniqueness', *Bulletin of Mathematical Sciences* **1**(1), 13–32.

Allen, R., Burgess, S. & Windmeijer, F. (2009), More reliable inference for segregation indices. University of Bristol Working Paper No 09/216.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**, 444–455.

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

Beresteanu, A., Molchanov, I. & Molinari, F. (2011), 'Sharp identification regions in models with convex moment predictions', *Econometrica* **79**(6), 1785–1821.

Billingsley, P. (1995), *Probability and Measure, Thrid Edition*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons.

Blundell, R., Gosling, A., Ichimura, H. & Meghir, C. (2007), 'Changes in the distribution of male and female wages accounting for employment composition using bounds', *Econometrica* **75**, 323–363.

Bontemps, C., Magnac, T. & Maurin, E. (2012), 'Set identified linear models', *Econometrica* **80**(3), 1129–1155.

Carrington, W. J. & Troske, K. R. (1997), 'On measuring segregation in samples with small units', *Journal of Business & Economic Statistics* **15**(4), 402–09.

Chernozhukov, V., Fernández-Val, I., Hahn, J. & Newey, W. (2013), 'Average and quantile effects in nonseparable panel models', *Econometrica* **81**(2), 535–580.

Chernozhukov, V. & Hansen, C. (2005), 'An iv model of quantile treatment effects', *Econometrica* **73**(1), 245–261.

Chesher, A. (2010), 'Instrumental variable models for discrete outcomes', *Econometrica* **78**, 575–601.

Chesher, A., Rosen, A. & Smolinski, K. (2013), 'An instrumental variable model of multiple discrete choice', *Quantitative Economics* **4**, 157–196.

Chiappori, P.-A., Nesheim, L. & McCann, R. J. (2010), 'Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness', *Economics Theory* **42**, 317–35.

Ciliberto, F. & Tamer, E. (2009), 'Market structure and multiple equilibria in airline markets', *Econometrica* **77**(6), 1791–1828.

Cortese, C., Falk, F. & Cohen, J. (1978), 'Understanding the standardized index of dissimilarity: Reply to massey', *American Sociological Review* **43**(4), 590–592.

D'Haultfœuille, X. (2010), 'A new instrumental method for dealing with endogenous selection', *Journal of Econometrics* **154**, 1–15.

D'Haultfoeuille, X. & Rathelot, R. (2014), Measuring segregation on small units : A partial identification analysis. Crest Working Paper.

Doksum, K. (1974), 'Empirical probability plots and statistical inference for nonlinear models in the two-sample case', *Annals of Statistics* **2**(2), 267–277.

Douglas, R. (1964), 'On extremal measures and subspace density', *Michigan Mathematical Journal* **11**(3), 243–246.

Ekeland, I., Galichon, A. & Henry, M. (2010), 'Optimal transportation and the falsifiability of incompletely specified economic models', *Economic Theory* **42**, 355–374.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.

Freyberger, J. & Horowitz, J. (2012), Identification and shape restrictions in nonparametric instrumental variables estimation. CEMMAP Working Paper CWP15/12.

Frontini, M. & Tagliani, A. (2011), 'Hausdorff moment problem and maximum entropy: On the existence conditions', *Applied Mathematics and Computation* **218**, 430–433.

Galichon, A. & Henry, M. (2011), 'Set identification in models with multiple equilibria', *Review of Economic Studies* **78**(4), 1264–1298.

Gronau, R. (1974), 'Wage comparisons - a selectivity bias', *Journal of Political Economy* **82**, 119–1143.

Gut, A. (2005), *Probability: A Graduate Course*, Springer-Verlag, New York.

Heckman, J. J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica* **42**, 679–694.

Hiriart-Urruty, J.-B. & Lemaréchal, C. (2001), *Fundamentals of Convex Analysis*, Springer.

Honore, B. & Tamer, E. (2006), 'Bounds on parameters in panel dynamic discrete choice models', *Econometrica* **74**(3), 611–629.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–75.

Kitagawa, T. (2010), Testing for instrument independence in the selection model. Unpublished working paper, http://www.homepages.ucl.ac.uk/ uctptk0/Research/TestER.pdf.

Klopotowski, A., Nadkarni, M. G. & Bhaskara-Rao, K. P. S. (2003), 'When is $f(x_1, x_2, ..., x_n) = u_1(x_1) + ... + u_n(x_n)$?', *Proceedings of The Indian Academy of Sciences - Mathematical Sciences* **113**(1), 77–86.

Lebesgue, H. (1905), 'Sur les fonctions représentables analytiquement', *Journal de mathématiques pures et appliquées* **6**, 139–216.

Leoni, G. (2009), *A First Course in Sobolev Space*, Vol. 105 of *Graduate Studies in Mathematics*, American Mathematical Society.

Manski, C. F. (1989), 'Anatomy of the selection problem', *Journal of Human Resources* **24**, 343–360.

Manski, C. F. (1990), 'Nonparametric bounds on treatment effects', *The American Economic Review, Papers and Proceedings* **80**, 319–323.

Manski, C. F. (2003), *Partial Identification of Probability Distribution*, Springer-Verlag.

Phelps, R. R. (2001), *Lectures on Choquet's Theorem, Second Edition*, Vol. 1757 of *Lecture Notes in Mathematics*, Springer.

Ramalho, E. A. & Smith, R. J. (2011), 'Discrete choice nonresponse', *Review of Economic Studies* .

Rathelot, R. (2011), Measuring segregation when units are small: a parametric approach. Crest Working Paper.

Rosen, A. (2012), 'Set identification via quantile restrictions in short panels', *Journal of Econometrics* **166**(1), 127–137.

Rudin, W. (1987), *Real and Complex Analysis. Thrid Edition.*, McGraw-Hill.

Tamer, E. (2003), 'Incomplete simultaneous discrete response model with multiple equilibria', *Review of Economic Studies* **70**(1), 147–165.

Tao, T. (2010), *An epsilon of Room, I: Real Analysis, pages from year three of a mathematical blog*, Vol. 58 of *Graduate Studies in Mathematics*, American Mathematical Society.

Van Der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge Unversity Press.

van der Vaart, A. & Wellner, J. (1996), *Weak convergence and Empirical Processes*, Springer.

Villani, C. (2003), *Topics in Optimal Transportation*, Vol. 58 of *Graduate Studies in Mathematics*, American Mathematical Society.

Villani, C. (2009), *Optimal Transport: Old and New*, Springer.

Winship, C. (1977), 'A revaluation of indexes of residential segregation', *Social Forces* **55**(4), 1058–1066.

Appendix: proofs

Proof of Theorem 2.1

Before to detail the proofs, we will fix some notations.
We consider the set $\mathcal{M}$ of signed measures concentrated on $\mathcal{S}$ equipped with the norm of total variation $|.|_{TV}$. Let $\mathcal{B}$ the unit ball of $(\mathcal{M}, |.|_{TV})$. Remember that $\mathcal{P}$ is a subset of probability measures in $\mathcal{B}$.
Because we do not assume that the unknown probability distribution P_0 is supported by a compact of $\mathbb{R}^k$, some sequences of probability measure in $\mathcal{K}$ can send some mass to infinity (think to the sequence of Dirac δ_n for $n \in \mathbb{N}$). To control some disagreements of this property, a weakest topology than the topology of the weak convergence is used in some steps of the proof. For $E \subset \mathbb{R}^k$, let $\mathcal{C}(E, \mathbb{R})$ the set of continuous function from E to $\mathbb{R}$ and $\mathcal{C}_c(E, \mathbb{R})$ the set of continuous function from E to $\mathbb{R}$ with a compact support (the topology on E is the usual subspace topology). This functional spaces can be equipped with the supremum norm $||.||_\infty$.
Let τ denotes the weak topology, i.e. the topology associated to the weak convergence of measures, i.e. the weakest topology such that for every $f \in \mathcal{C}(\mathcal{S}, \mathbb{R})$:

$$P \mapsto \int_{\mathcal{S}} f dP \text{ is continuous}$$

Let $\tau^\star$ denotes the weak-$\star$ topology, i.e. the weakest topology such that for every $f \in \mathcal{C}_c(\mathcal{S}, \mathbb{R})$:

$$P \mapsto \int_{\mathcal{S}} f dP \text{ is continuous}$$

Some authors define the weak-$\star$ topology with the class of function that vanish to infinity (as Rudin (1987), in Theorem 6.19). In our context the two definitions are equivalent (see for instance Rudin (1987), Paragraph 6.18). Other authors defined such topology as the vague topology (see for instance Tao (2010), page 166).
$\tau^\star$ is weakest than τ in the sens that $\tau^\star \subset \tau$ (a $\tau^\star$-open set is a τ-open set and then a τ-compact set is a $\tau^\star$-compact set).
Let $\text{cl}(\mathcal{R})$ (respectively $\text{cl}^\star(\mathcal{R})$) the τ-closure (respectively the $\tau^\star$-closure) of $\mathcal{R}$:

$$\text{cl}^\star(\mathcal{R}) = \left\{ \mu \in \mathcal{M} \text{ s.t. } \exists P_n \in \mathcal{R} \text{ s.t. } \forall f \in \mathcal{C}_c(\mathcal{S}, \mathbb{R}) \quad \int_{\mathcal{S}} f dP_n \to \int_{\mathcal{S}} f d\mu \right\}$$

$$\text{cl}(\mathcal{R}) = \left\{ \mu \in \mathcal{M} \text{ s.t. } \exists P_n \in \mathcal{R} \text{ s.t. } \forall f \in \mathcal{C}(\mathcal{S}, \mathbb{R}) \quad \int_{\mathcal{S}} f dP_n \to \int_{\mathcal{S}} f d\mu \right\}$$

Similarly, let $\mathrm{cl}\left(\mathcal{R}_\theta\right)$ (respectively $\mathrm{cl}^\star\left(\mathcal{R}\right)_\theta$) the τ-closure (respectively the $\tau^\star$-closure) of $\mathcal{R}_\theta$.
The different sets of measures are included as follow (for every measurable function f):

$$\begin{array}{ccccccc}
\mathcal{R}_\theta & \subset & \mathrm{cl}\left(\mathcal{R}_\theta\right) & \subset & \mathrm{cl}^\star\left(\mathcal{R}_\theta\right) & & \\
\cap & & \cap & & \cap & & \\
\mathcal{R} & \subset & \mathrm{cl}\left(\mathcal{R}\right) & \subset & \mathrm{cl}^\star\left(\mathcal{R}\right) & & \\
 & & \cap & & \cap & & \\
\mathcal{I}(f) & \subset & \mathcal{P} & \subset & \mathcal{B} & \subset & \mathcal{M}.
\end{array}$$

The proof is divided in 8 steps. Steps 1 to 5 concern the first part of the Theorem, and steps 6 to 8 concern the specific case where θ is a moment of U. Steps 1, 2 and 3 rely on deep but usual Theorems of functional analysis and are compactly exposed. The step 4 and 6 need more extended developments and rely on usual tools of integration theory (monotone convergence theorem, dominated convergence theorem, completion of the measure space...). The Steps 5, 7 and 8 do not present conceptual difficulties and are quite short.

1. Riesz theorem for bounded linear form (for instance Rudin (1987), Theorem 6.19) implies that the space $(\mathcal{M}, |.|_{TV})$ is the dual of $(\mathcal{C}_c(\mathcal{S}, \mathbb{R}), ||.||_\infty)$.

2. By Banach-Alaoglu Theorem, $\mathcal{B}$ is $\tau^\star$-compact. For every K compact of $\mathbb{R}^k$, $(\mathcal{C}_c(K, \mathbb{R}), ||.||_\infty)$ is a separable normed vector space as subspace of $(\mathcal{C}(K, \mathbb{R}), ||.||_\infty)$ separable normed vector space. Let K_n a sequence of compact such that $\cup_{n\in\mathbb{N}} K_n = \mathbb{R}^k$. Because $\mathcal{S}$ is closed then $K_n \cap \mathcal{S}$ is compact and $\mathcal{S} = \cup_{n\in\mathbb{N}}(K_n \cap \mathcal{S})$. It follows that $(\mathcal{C}_c(\mathcal{S}, \mathbb{R}), ||.||_\infty)$ is a separable normed vector space as countable union of separable space. It follows that every $\tau^\star$-closed ball is metrizable (see for instance Theorems A.48 in Leoni (2009)). Then $\mathrm{cl}^\star\left(\mathcal{R}_\theta\right)$ is $\tau^\star$-compact and metrizable.

3. The Choquet theorem (Phelps (2001), Chapter 3) ensures that for every P in $\mathcal{R}_\theta$, it exists a Radon[6] probability measure μ_P on $\mathrm{cl}^\star\left(\mathcal{R}_\theta\right)$ supported by $\mathrm{ext}(\mathrm{cl}^\star\left(\mathcal{R}_\theta\right))$ such that

$$P = \int_{\mathrm{ext}(\mathrm{cl}^\star(R)_\theta)} R d\mu_P(R)$$

which means that: for all $f \in \mathcal{C}_c(\mathcal{S}, \mathbb{R})$, we have:

$$\int_{\mathcal{S}} f dP = \int_{\mathrm{ext}(\mathrm{cl}^\star(R)_\theta)} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R). \tag{.1}$$

[6]For the definition of Radon measure, see for instance Tao (2010), Definition 1.10.2.. A Radon probability measure is a positive Radon measure with total mass 1.

4. To prove the first part of the Theorem, we need to extend the Equality .1 to the set of functions f that are continuous and bounded by 1. So we have to prove two results. The first one is that the right hand sign of Equation .1 is defined when f is continuous and bounded, this is equivalent to show that $R \mapsto \int f dR$ is Lebesgue measurable with respect to μ_P. The second one is that the both sides of Equation .1 are equal when f is bounded and continuous. To prove this two results we use standard technics of integration on topological spaces. Let ψ_n a continuous function with values between 0 and 1 and such that $\psi_n = 1$ on $[-n; n]^k \cap \mathcal{S}$ with support equal to $[-n-1; n+1]^k \cap \mathcal{S}$. $\psi_n f$ is continuous with compact support and converges pointwise to f and is dominated by 1, so the dominated convergence applied to measures P and R (for R in $\text{Supp}(\mu_P)$) and to measure μ_P implies that $R \mapsto \int f dR$ is Borel-measurable with respect to μ_P and Equality .1 holds for every continuous and bounded f. By dominated convergence theorem, Borel-measurability of $R \mapsto \int f dR$ and Equality .1 can be proved for the Baire class 1 functions, i.e. the bounded functions that are pointwise limit of continuous functions. By induction, the result also hold for every class of the Baire functions. Because the Baire functions are the Borel measurable functions (see Lebesgue (1905)), Equality holds for bounded and Borel measurable functions. Next, for C a set such that $C \subset B$ with B a P-negligible Borel set, $h_B : R \mapsto R(B)$ is a positive and Borel measurable function and by the previous reasoning we have $\int h_B(R) d\mu_P(R) = \int R(B) d\mu_P(R) = P(B) = 0$. Let $h_C : R \mapsto R(C)$, we have $0 \leq h_C(R) \leq h_B(R)$ for every $R \in \text{cl}^\star(\mathcal{R}_\theta)$. It follows that h_C is Lebesgue measurable and such that $\int h_C(R) d\mu_P(R) = \int R(C) d\mu_P(R) = 0$. Then Lebesgue measurability for $R \mapsto \int f dR$ with respect to μ_P and Equality .1 holds for every f indicatrice of null-set C and then for every indicatrice of $B \cup C$, with B Borel set and C null-set. It follows that $R \mapsto \int f dR$ is Lebesgue measurable with respect to μ_P as soon as f is Lebesgue measurable with respect to P. It follows that if f is continuous and bounded $R \mapsto \int f dR$ is Lebesgue measurable with respect to μ_P and Equation .1 holds.

   Note that for $f = 1$, we have $\int R(\mathcal{S}) d\mu_P(R) = 1$, so $\mu_P(\mathcal{P}) = 1$. And then we can replace $\text{ext}(\text{cl}^\star(\mathcal{R})_\theta)$ by $\text{ext}(\text{cl}^\star(\mathcal{R})_\theta) \cap \mathcal{P}$ in Equation .1.

5. To achieved the proof for the first part of the Theorem, we have to show that $\mathcal{R}_\theta \neq \emptyset \Rightarrow \text{ext}(\text{cl}(\mathcal{R}_\theta)) \neq \emptyset$.
   We have already proved that $\mathcal{R}_\theta \neq \emptyset \Rightarrow \text{ext}(\text{cl}^\star(\mathcal{R}_\theta)) \cap \mathcal{P} \neq \emptyset$.
   Because $\text{cl}(\mathcal{R}_\theta) = \text{cl}^\star(\mathcal{R}_\theta) \cap \mathcal{P}$ (see for instance, Billingsley (1995) on the vague convergence), we have $\text{ext}(\text{cl}^\star(\mathcal{R}_\theta)) \cap \mathcal{P} \subset \text{ext}(\text{cl}(\mathcal{R}_\theta))$.

6. We have to prove the second part of the Theorem, so hereafter we assume that Assumption 3 holds. If f is a bounded function, by a reasoning similar to the previous step, we have:

$$\forall P \in \mathcal{R}, \quad \theta = \int_{\text{ext}(\text{cl}^\star(\mathcal{R}))\cap\mathcal{P}} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R). \tag{.2}$$

If f is Lebesgue measurable but unbounded and $P \in \mathcal{I}(f)$, consider $g_n = |f| \wedge n$. For every n, $R \in \mathcal{P} \mapsto \int g_n dR$ is Lebesgue measurable and integrable with respect to μ_P. Because $g_n \uparrow |f|$, the monotone convergence theorem (with respect to R) implies that $\int_{\mathcal{S}} g_n dR \uparrow \int_{\mathcal{S}} |f| dR$. The monotone convergence theorem (with respect to the measure μ_P) implies that $R \mapsto \int_{\mathcal{S}} |f| dR$ is Lebesgue-measurable and integrable with respect to μ_P and then:

$$\int_{\text{ext}(\text{cl}^\star(\mathcal{R}))\cap\mathcal{P}} \left(\int_{\mathcal{S}} |f| dR \right) d\mu_P(R) = \int_{\mathcal{S}} |f| dP.$$

The previous Equality ensures that $\mu_P(\mathcal{I}(f)) = 1$, so we can replace $\text{ext}(\text{cl}^\star(\mathcal{R})) \cap \mathcal{P}$ by $\text{ext}(\text{cl}^\star(\mathcal{R})) \cap \mathcal{P} \cap \mathcal{I}(f)$ in Equation .2.

Now let $e_n = (f \wedge n) \vee (-n)$, for every $R \in \mathcal{P} \cap \mathcal{I}(f)$, the dominated convergence theorem (with respect to R) implies that $\int e_n dR \to \int f dR$. Because $R \mapsto \int_{\mathcal{S}} |f| \, dR$ is integrable (with respect to the measure μ_P) and $\left| \int_{\mathcal{S}} e_n dR \right| \leq \int_{\mathcal{S}} |e_n| \, dR \leq \int_{\mathcal{S}} |f| \, dR$, the dominated convergence theorem (with respect to μ_P) implies that Equation .1 holds for every f Lebesgue-measurable and $P \in \mathcal{I}(f)$:

$$\int_{\mathcal{S}} f dP = \int_{\text{ext}(\text{cl}^\star(\mathcal{R}))\cap\mathcal{P}\cap\mathcal{I}(f)} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R). \tag{.3}$$

7. Because $\mathcal{R} = \text{cl}^\star(\mathcal{R}) \cap \mathcal{P}$ and because $\mu(\mathcal{S}) \leq 1$ for $\mu \in \text{cl}^\star(\mathcal{R})$, we have $\text{ext}(\mathcal{R}) = \text{ext}(\text{cl}^\star(\mathcal{R})) \cap \mathcal{P}$.

8. Because μ_P is a probability measure, we have for every f Lebesgue-measurable and $P \in \mathcal{R} \cap \mathcal{I}(f)$:

$$\begin{array}{rcl} \inf_{R\in\text{ext}(\mathcal{R})\cap\mathcal{I}(f)} \int_{\mathcal{S}} f dR & = & \int_{\text{ext}(\mathcal{R})\cap\mathcal{I}(f)} \left(\inf_{R\in\text{ext}(\mathcal{R})\cap\mathcal{I}(f)} \int_{\mathcal{S}} f dR \right) d\mu_P(S) \\ & \leq & \int_{\text{ext}(\mathcal{R})\cap\mathcal{I}(f)} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R) \\ & = & \int_{\mathcal{S}} f dP \end{array}$$

And then $\inf_{R\in\text{ext}(\mathcal{R})\cap\mathcal{I}(f)} \int_{\mathcal{S}} f dR \leq \inf_{R\in\mathcal{R}\cap\mathcal{I}(f)} \int_{\mathcal{S}} f dR = \underline{\theta}$.
Because $\text{ext}(\mathcal{R}) \subset \mathcal{R}$, we have the reverse inequality.
Similar reasoning holds for the upper bound.

Proof of Theorem 2.2

We have: $\Theta_0 = \{\theta : \mathcal{R}_\theta \neq \emptyset\} = \{\theta : \cap_{n\in\mathbb{N}} \mathrm{cl}\,(\mathcal{R})_{n,\theta} \neq \emptyset\} \subset \cap_{n\in\mathbb{N}}\Theta_{0,n}$.
To prove the reverse inclusion, consider $\theta \in \cap_{n\in\mathbb{N}}\Theta_{0,n}$. Because $\sup_{P\in \mathrm{ext}(\mathcal{R}_{n_0,\theta})} \int ||u||^\varepsilon dP(u)$ is finite, Theorem 2.1.1 ensures that this is also the case when the sup is taken over $\mathcal{R}_{n_0,\theta}$. Markov's inequality ensures that $\mathcal{R}_{n_0,\theta}$ is uniformly tight (cf. Van Der Vaart (1998) for a definition of uniform tightness). Consider a sequence $P_n \in \mathcal{R}_{n,\theta}$, for $n \geq n_0$, $P_n \in \mathcal{R}_{n_0,\theta}$, the Prohorov's Theorem (cf. Van Der Vaart (1998)) ensures that it exist a subsequence $P_{\sigma(n)}$ and a distribution $P^\star$ such that $P_{\sigma(n)}$ converges weakly to $P^\star$. Because the $(\mathcal{R}_{n,\theta})_{n\in\mathbb{N}}$ is a decreasing sequence of closed sets, $P^\star \in \cap_{n\in\mathbb{N}}\mathcal{R}_{\sigma(n),\theta} \subset \cap_{n\in\mathbb{N}}\mathcal{R}_{n,\theta}$, and so $\cap_{n\in\mathbb{N}}\mathcal{R}_{n,\theta} \neq \emptyset$.
To achieved the proof, note that Theorem 2.1.1 ensures that $\Theta_{0,n} = \{\theta : \mathrm{ext}\,(\mathcal{R}_{n,\theta}) \neq \emptyset\}$.

Proof of Theorem 2.3

We make the proof for the upper bound only, the reasoning being similar for the lower bound.
By a similar reasoning to the previous Proof we know that if P_n is a sequence of distribution in $\mathcal{R}_n$, we have a subsequence $P_{\sigma(n)}$ that converge weakly to $P^\star \in \cap_{n\in\mathbb{N}}\mathcal{R}_n = \mathcal{R}$.
Now,

$$\int |f(u)|\mathbb{1}\{|f(u)| \geq x\}dP_{\sigma(n)}(u) \leq \frac{1}{x^\varepsilon}\int |f(u)|^{1+\varepsilon}dP_n(u) \leq \frac{M}{x^\varepsilon}.$$

Thus,

$$\lim_{x\to\infty} \limsup_{n\in\mathbb{N}} \int |f(u)|\mathbb{1}\{|f(u)| \geq x\}dP_{\sigma(n)}(u) = 0.$$

This, combined with the weak convergence of $(P_{\sigma(n)})_{n\in\mathbb{N}}$ and the continuity of f, implies (see, e.g., Van Der Vaart, 1998, Theorem 2.20)

$$\int f dP_{\sigma(n)} \to \int f dP \leq \overline{\theta}.$$

Now, by definition of P_n, $\lim_n \overline{\theta}_n = \lim_n \overline{\theta}_{\sigma(n)} = \lim_n \int f dP_{\sigma(n)}$. Therefore,

$$\lim_n \overline{\theta}_n \leq \overline{\theta}.$$

This implies the result because $\overline{\theta}_n \geq \overline{\theta}$.

Proof of Counterexample 3

Because $x \mapsto x^k$ is continuous and bounded on $\mathcal{S}$ for every $k \in \mathbb{N}$, $\mathcal{R}$ and $\mathcal{R}_n$ are closed for the weak convergence.
A simple calculation shows that $P^\star = \frac{1}{2}\mathcal{U}_{[-1;1]} + \frac{1}{2}\delta_0$ belongs to $\mathcal{R}$. Because $\mathcal{S}$ is bounded, $P^\star$ is defined by its moments (see, e.g. Gut, 2005, Theorems 8.1 and 8.3) and then $\mathcal{R} = \{P^\star\}$.
Now for $n \in \mathbb{N}^\star$, $\varepsilon > 0$ consider distribution P_n^ε such that $P_n^\varepsilon = \frac{1}{4}\mathcal{U}_{[-1;0]} + \frac{1}{2}\delta_\varepsilon + \frac{1}{4}Q_n^\varepsilon$, with Q_n^ε a probability distribution.
$P_n^\varepsilon \in \mathcal{R}_n$ if and only if

$$Q_n^\varepsilon \in \mathcal{P}_n^\varepsilon = \{Q \in \mathcal{P} :\ Q([0;1]) = 1 \text{ and } \int x^k dQ(x) = m_k(\varepsilon) = 1/(k+1) - 2\varepsilon^k \text{ for } k = 1, ..., n\}.$$

Let $m_0(\varepsilon) = 1$ and $M_{1,n}(\varepsilon)$ and $M_{2,n}(\varepsilon)$ the Hankel matrices defined by:

$$M_{1,n}(\varepsilon) = \begin{cases} \begin{bmatrix} m_0(\varepsilon) & m_1(\varepsilon) & ... & m_{n/2}(\varepsilon) \\ m_1(\varepsilon) & m_2(\varepsilon) & ... & m_{n/2+1}(\varepsilon) \\ \vdots & \vdots & ... & \vdots \\ m_{n/2}(\varepsilon) & \dots & \dots & m_n(\varepsilon) \end{bmatrix} & \text{if } n \text{ is even} \\ \begin{bmatrix} m_1(\varepsilon) & m_2(\varepsilon) & ... & m_{(n+1)/2}(\varepsilon) \\ m_2(\varepsilon) & m_3(\varepsilon) & ... & m_{(n+1)/2+1}(\varepsilon) \\ \vdots & \vdots & ... & \vdots \\ m_{(n+1)/2}(\varepsilon) & \dots & \dots & m_n(\varepsilon) \end{bmatrix} & \text{if } n \text{ is odd} \end{cases}$$

$$M_{2,n}(\varepsilon) = \begin{cases} \begin{bmatrix} m_1(\varepsilon) - m_2(\varepsilon) & ... & m_{n/2}(\varepsilon) - m_{n/2+1}(\varepsilon) \\ \vdots & & \vdots \\ m_{n/2}(\varepsilon) - m_{n/2+1}(\varepsilon) & \dots & m_{n-1}(\varepsilon) - m_n(\varepsilon) \end{bmatrix} & \text{if } n \text{ is even} \\ \begin{bmatrix} m_0(\varepsilon) - m_1(\varepsilon) & ... & m_{(n-1)/2}(\varepsilon) - m_{(n-1)/2+1}(\varepsilon) \\ \vdots & & \vdots \\ m_{(n-1)/2}(\varepsilon) - m_{(n-1)/2+1}(\varepsilon) & \dots & m_{n-1}(\varepsilon) - m_n(\varepsilon) \end{bmatrix} & \text{if } n \text{ is odd} \end{cases}$$

$\mathcal{P}_n^\varepsilon$ contains a continuous probability measure if and only if $\det(M_{1,n}(\varepsilon)) \geq 0$ and $\det(M_{2,n}(\varepsilon)) \geq 0$ (Frontini & Tagliani (2011)). For n even, $M_{1,n}(0)$ is the Hilbert matrix of size $n/2+1$, and then $\det(M_{1,n}(0)) > 0$. For n odd, $M_{1,n}(0)$ is a principal submatrix of the Hilbert matrix $M_{1,n+1}(0)$ and we also have $\det(M_{1,n}(0)) > 0$.

Now consider $M_{2,n}(0)$. For n odd, $M_{2,n}(0) = M_{1,n-1}(0) \circ M_{1,n}(0)$, where $\circ$ is the Hadamard product. The Oppenheim's inequality then implies that $\det(M_{2,n}(0)) > 0$. For n even,

$M_{2,n}(0) = A \circ B$, with A and B principal submatrix of the Hilbert matrix $M_{1,n}$ and then by similar argument $\det(M_{2,n}(0)) > 0$.

Because $\varepsilon \mapsto (\det(M_{1,n}(\varepsilon)), \det(M_{2,n}(\varepsilon)))$ is continuous, $\det(M_{1,n}(\varepsilon))$ and $\det(M_{2,n}(\varepsilon))$ are positive for sufficiently small ε. Then for sufficiently small ε, $\mathcal{P}_n^\varepsilon$ contains a probability distribution Q_n^ε dominated by the Lebesgue measure on $[0;1]$. For $P_n^\varepsilon = \frac{1}{4}\mathcal{U}_{[-1;0]} + \frac{1}{2}\delta_\varepsilon + \frac{1}{4}Q_n^\varepsilon$, we have $\int f(x)dP_n^\varepsilon(x) = 3/4$ with $P_n^\varepsilon \in \mathcal{R}_n$.

.1 Proof of Counterexample 4

By construction, $\mathcal{R}_n$ is a decreasing sequence of convex sets. It is also closed for the weak convergence because the functions $(g_k)_{k\in\mathbb{N}}$ are continuous and bounded and $\mathcal{R} = \cap_{n\in\mathbb{N}}\mathcal{R}_n$. Thus Assumption 4.**??** holds. Condition (i) of Corollary 2.3 trivially holds. By Theorem 1.12.2 of van der Vaart & Wellner (1996), the class $\mathcal{G}$ is convergence-determining. As a consequence, the moments $(E(g_k(X)))_{k\in\mathbb{N}}$ determine the distribution of X, implying that $\mathcal{R} = \{N(0,1)\}$. Hence, $\sup_{P\in\mathcal{R}} \int f dP = 0$.

We now prove that for all n, $\overline{\theta}_n = +\infty$. Remark that the functions $(g_k)_{k\in\mathbb{N}}$ are compactly supported. Let $x_n = \inf\{x \geq 2 : \ g_k(x) = 0 \ \forall\, k \in \{1, ..., n\}\}$. Let Φ denote the cdf of Z and for any $y \geq 1$, let

$$F_y(x) = \Phi(x)\mathbb{1}_{\{x\leq x_n\}} + \Phi(x_n)\mathbb{1}_{\left\{x_n \leq x < \frac{y}{1-\Phi(x_n)}\right\}} + \mathbb{1}_{\left\{\frac{y}{1-\Phi(x_n)} \leq x\right\}}.$$

Remark that this definition is valid since $x < y/(1-\Phi(x))$ for all $x \geq 2$ and $y \geq 1$. By construction, F_y is a cdf such that the corresponding probability measure P_y satisfies $P_y \in \mathcal{R}_n$. Moreover,

$$\int |x| dP_y(x) = \int_{-\infty}^{x_n} |x| d\Phi(x) + (1-\Phi(x_n)) \times \frac{y}{1-\Phi(x_n)} < +\infty,$$

so that $P_y \in \mathcal{I}(f)$. Finally,

$$\int x dP_y(x) = E\left(Z\mathbb{1}\{Z \leq x_n\}\right) + y.$$

Because y was arbitrary, we obtain $\overline{\theta}_n = +\infty$.

Proof of Theorems 3.1, 3.2 and 3.3

.1.1 Theorem 3.1

Let $\mathcal{J} = \mathcal{K}$ (or respectively $\mathcal{L}$). Let $Q \in \text{ext}(\mathcal{J})$ and assume that $\text{span}(\mathcal{G}, 1)$ is not dense in $L_1(Q)$. Then it follows by the Hahn-Banach Theorem that it exists a non null and

continuous linear form on $L_1(Q)$ that vanishes on $\text{span}(\mathcal{G}, 1)$. The identification $L_1^\star(Q) = L_\infty(Q)$ ensures that it exists non null $h \in L_\infty(Q)$ such that $\int ghdQ = 0$ for every $g \in \text{span}(\mathcal{G}, 1)$. For any measurable set A, let $Q_1(A) = Q(A) + \frac{1}{||h||_\infty}\int_A hdQ$ and $Q_2(A) = Q(A) - \frac{1}{||h||_\infty}\int_A hdQ$, Q_1 and Q_2 are positive measures. Then Q_1 and Q_2 are in $\mathcal{J}$ and $(Q_1 + Q_2)/2 = Q$. This is absurde.

Let $Q \in \text{ext}(\text{cl}\,(\mathcal{J}))$, and assume that $\text{span}(\mathcal{G}, 1)$ is not dense in $L_1(Q)$. It exists non null $h \in L_\infty(Q)$ such that $\int ghdQ = 0$ for every $g \in \text{span}(\mathcal{G}, 1)$. It exists $Q_n \in \mathcal{J}$ such that Q_n converges weakly to Q. Let $Q_1 = Q + \frac{1}{||h||_\infty}\int hdQ$, $Q_2 = Q - \frac{1}{||h||_\infty}\int hdQ$, $Q_{1n} = Q_n + \frac{1}{||h||_\infty}\int hdQ$ and $Q_{2n} = Q_n - \frac{1}{||h||_\infty}\int hdQ$. Q_{in} is a sequence in $\mathcal{J}$ weakly converging to Q_i. So Q_1 and Q_2 are in $\text{cl}\,(\mathcal{J})$ and $Q = (Q_1 + Q_2)/2$. This is absurde.

We have proved the "only if" parts of Theorem. We now turn to the "if" part of the Theorem.

Let $Q \in \mathcal{K} \setminus \text{ext}(\mathcal{K})$, it exists Q_1 and Q_2 in $\mathcal{K}$ such that $Q = (Q_1 + Q_2)/2$ with $Q_1 \neq Q_2$. It follows that $2Q(A) \geq Q_1(A) \geq 0$ for every measurable set $A \in \mathcal{S}$. It follows by the Radon-Nikodym Theorem that it exists $h \in L_\infty(Q)$ such that $dQ_1 = hdQ$. Because $Q_1 \neq Q_2$, it exists a measurable set A such that $Q(A) \neq Q_1(A)$. Then for every $g \in \text{span}(\mathcal{G}, 1)$:

$$|Q(A) - Q_1(A)| = \left|\int \mathbb{1}_A(1-h)dQ\right| = \left|\int (\mathbb{1}_A - g)(1-h)dQ\right| \leq ||1-h||_\infty \int |\mathbb{1}_A - g|dQ,$$

and thus $\text{span}(\mathcal{G}, 1)$ is not dense in $L_1(Q)$.

.1.2 Theorem 3.2

To prove Theorem 3.2, note that $\mathcal{R}_\theta = \text{cl}\,(\mathcal{R}_\theta)$ and apply Theorem 3.1.

We have $dim\,(\text{span}(g_1, ...g_k, m_1, ..., m_l, 1)) = r \leq l + k + 1$. Let $P \in \text{cl}\,(\mathcal{R}_\theta)$ such that it exists $A_1, ..., A_{r+1}$ disjoint subsets of $\mathcal{S}$ such that $P(A_i) > 0$ then

$\mathcal{F} = \{f : \mathcal{S} \mapsto \mathbb{R} : \exists(\alpha_1, \alpha_2, ..., \alpha_{r+1}) \in \mathbb{R}^{r+1}, f(y) = \sum_{i=1}^{r+1} \alpha_i \mathbb{1}_{\{y \in A_i\}}\}$.

$\mathcal{F} \subset L^1(P)$ and $dim(\mathcal{F}) = r + 1 > dim\,(\text{span}(g_1, ..., g_k, m_1, ..., m_l, 1))$, then

$\text{span}(g_1, ..., g_k, m_1, ..., m_l, 1)$ is not dense in $L^1(P)$. It follows from previous Lemma, that P is not an extreme point of $\text{cl}\,(\mathcal{R}_\theta)$. We deduce that extreme point of $\text{cl}\,(\mathcal{R}_\theta)$ are supported by at most r points in $\mathcal{S}$. This achieves the proof of Theorem 3.2.

.1.3 Theorem 3.3

Note that $\mathcal{R}$ is closed for the weak convergence because:

$$\{P \in \mathcal{P}, \forall g \in \mathcal{C}_b(\text{Supp}(P_i)), \int g(u_i)dP(u_1, ..., u_n) = \int g(u_i)dP_i(u_i)\}$$
$$=$$
$$\{P \in \mathcal{P}, \forall g \in L_1(P_i), \int g(u_i)dP(u_1, ..., u_n) = \int g(u_i)dP_i(u_i)\}.$$

Let A_k a countable basis of open sets of $\mathbb{R}^{d_1} \times ... \times \mathbb{R}^{d_n}$ and let $P \in \text{ext}(\mathcal{R}) \cap \mathcal{I}(f)$. Theorem 3.1 implies that every function $u = (u_1, ..., u_n) \mapsto \mathbb{1}_{\{u \in A_k\}}$ is such that:

$$\mathbb{1}_{\{u \in A_k\}} = \lim_l \sum_{i=1}^n g_{i,l}(u_i) \quad - \quad P \ a.s.$$

It follows (see Klopotowski et al. (2003), Theorem 3.1) that P has a support $\mathcal{S} \subset \prod_i \text{Supp}(P_i)$ such that for every function $f \in L_1(P)$ it exists n functions g_i $(i = 1, ..., n)$ such that:

$$\forall u = (u_1, ..., u_n) \in \mathcal{S}, \quad f(u) = \sum_i g_i(u_i)$$

Let $\pi_{-j}\mathcal{S} = \{(u_1, ..., u_{j-1}, u_{j+1}, ..., u_n) \in \prod_{i \neq j} \text{Supp}(P_i) \quad s.t. \quad \exists u_j \quad s.t. \quad (u_1, ...u_n) \in \mathcal{S}\}$. Now, for every $u_1 \in \text{Supp}(P_1)$ let $\widetilde{g}_1(u_1) = \sup_{(u_2, u_3, ..., u_n) \in \pi_{-1}\mathcal{S}} f(u_1, ..., u_n) - \sum_{i=2}^n g_i(u_i)$ and for $j \geq 2$, $\widetilde{g}_j(u_j) = \sup_{(u_1, ..., u_{j-1}, u_{j+1}, ..., u_n) \in \pi_{-j}\mathcal{S}} f(u_1, ..., u_n) - \sum_{i=1}^j \widetilde{g}_i(u_i) - \sum_{i=j+1}^n g_i(u_i)$. We have $f(u_1, ..., u_n) \leq \sum_i \widetilde{g}_i(u_i)$ on $\prod_{i=1,...,n} \text{Supp}(P_i)$ and the equality holds on $\mathcal{S}$. Note that $\int \widetilde{g}_i(u_i) dP_i(u_i) = \int \widetilde{g}_i(u_i) dP(u_1, ..., u_n) = \int g_i(u_i) dP(u_1, ..., u_n) = \int g_i(u_i) dP_1(u_1)$, then $\widetilde{g}_i \in L_1(P_i)$. It follows that $\int f dP = \sum_i \int \widetilde{g}_i dP_i$ and our main result ensures:

$$\sup_{P \in \mathcal{R} \cap \mathcal{I}(f)} \int f dP = \sup_{P \in \text{ext}(\mathcal{R}) \cap \mathcal{I}(f)} \int f dP \geq \inf_{g_i \in L_1(P_i) \sum g_i \geq f} \sum_{i=1}^n \int g_i(u_i) dP_i(u_i).$$

The reverse inequality is obvious. Changing f by $-f$ gives the result on the lower bound.

Proof of Propositions 4.1, 4.2 and 4.3

Bounds under Assumption 7

Under Assumption 9, the support of (Y, X) is included in the support of $DY, X|D = 1$ which is identified in the data. Let $\mathcal{S}$ (respectively $\mathcal{S}_1$, $\mathcal{Y}$ and $\mathcal{X} = \{x_1, ..., x_J\}$) the supports of (D, DY, X) (respectively (Y, X), Y and X). For every distribution Q concentrated on $\{0; 1\} \times \mathcal{S}_1$, $Q^{Y|D=d, X=x}$, $Q^{Y|D=d}$, $Q^{D,X}$, Q^Y, $Q^{D,DY,X}$ denote the conditional, the marginal or the joint distributions derived from Q and E_Q denotes the expectation operator with respect to the measure Q. Let $\mathcal{K}$ the set of distributions of (D, Y, X) compatible with the data and the Assumption 7. For every $P \in \mathcal{K}$, $P^{D,X} = P_0^{D,X}$ and $P^{Y|D=1,X=x} = P_0^{Y|D=1,X=x}$ then P is characterized by $\left(P^{Y|D=0,X=x_j}\right)_{j=1,...,J}$. Let $\mathcal{C}$ the set of vector of distribution[7] such that $\left(P^{Y|D=0,X=x_j}\right)_{j=1,...,J} \in \mathcal{C}$ if and only if $P = \sum_x P_0^{Y|D=1,X=x} P_0^{D,X} \delta_1^D + P^{Y|D=0,X=x} P_0^{D,X} \delta_0^D$ is

[7]We assume for sake of simplicity that $E_{P_0}(D|X = x_j) < 1$ for every j. If this is not the case the present demonstration can be easily adapted after exclusion of the corresponding component in the vector $\mathcal{C}$.

in $\mathcal{K}$. Extreme points of $\mathcal{K}$ are one-to-one with extreme points of $\mathcal{C}$, so we will characterize $\mathcal{C}$ and his extreme points.

We will proceed in five steps: first we characterize $\mathcal{C}$, then we show that $\mathcal{C}$ is closed for the weak convergence, third we show that the component j of an element of $\text{ext}(\mathcal{C})$ has at most $J-j+1$ point of support, forth we fully characterize $\text{ext}(\mathcal{C})$ and lastly we give a necessary and sufficient condition under which $\text{ext}(\mathcal{C})$ is empty.

Step 1: Characterization of $\mathcal{C}$

$P_0^{Y|D=1}(A)=0$ implies that $P_0^{Y|D=1,X=x_j}(A)=0$ for every $j=1,...,J$, so we can define $f_{Y|D=1,X=x_j}$ as the density of $P_0^{Y|D=1,X=x_j}$ with respect to the distribution of $P_0^{Y|D=1}$.

For any $(x_i,x_j)\in\mathcal{X}\times\mathcal{X}$, let $\nu_{i,j}$ the endomorphism on the space of positive measure concentrated on $\mathcal{Y}$ defined by:

$$\nu_{i,j}(Q)(B)=\int_B r_{i,j}(y)dQ(y)$$

$$\text{with}\quad r_{i,j}=p(x_j)(1-p(x_i))f_{Y|D=1,X=x_j}/(p(x_i)(1-p(x_j))f_{Y|D=1,X=x_i}).$$

Note that $\nu_{i,i}=Id$, $\nu_{i,j}(\delta_y)=r_{i,j}(y)\delta_y$ and $\nu_{i,j}\circ\nu_{j,k}=\nu_{i,k}$.

Let $\left(P^{Y|X=x_j,D=0}\right)_{j=1,...,J}\in\mathcal{C}$.

For $(y,x_i,x_j)\in\mathcal{Y}\times\mathcal{X}\times\mathcal{X}$ such that

$(E_{P_0}(D|Y=y,X=x_i),E_{P_0}(D|Y=y,X=x_j))\notin\{(0,0);(1,1)\}$,

let $s(y,x_i,x_j)=\left(\frac{1}{E_{P_0}(D|Y=y,X=x_j)}-1\right)/\left(\frac{1}{E_{P_0}(D|Y=y,X=x_i)}-1\right)$.

Under Assumption 7, $s(y,x_i,x_j)$ is greater than one if and only if $x_j<x_i$. Note that $1>E_{P_0}(D|Y,X=x_i)$, $P_0^{Y|D=0,X=x_i}$-almost-surely and that Assumption 9 ensures also that $E_{P_0}(D|Y,X=x_i)>0$, $P_0^{Y|D=0,X=x_i}$-almost-surely. So, $y\mapsto s(y,x_i,x_j)$ is positive and mesurable function with respect to the Lebesgue σ-algebra on $\text{Supp}(P_0^{Y|D=0,X=x_i})$.

The Bayes formula implies that:

$$\begin{aligned}dP_0^{Y|D=0,X=x} &= \left(\frac{1}{E_{P_0}(D=1|Y=y,X=x)}-1\right)\frac{p(x)}{1-p(x)}dQ_0^{Y|D=1,X=x}\\ &= \left(\frac{1}{E_{P_0}(D=1|Y=y,X=x)}-1\right)\frac{p(x)}{1-p(x)}f_{Y|D=1,X=x}dP_0^{Y|D=1}\end{aligned}$$

Then for all $x_j\leq x_i$, $P_0^{Y|D=0,X=x_j}(B)=\int_B s(y,x_i,x_j)d\nu_{i,j}(P_0^{Y|D=0,X=x_i})(y)$.

Under Assumption 7, for every $i=1,...,J-1$ we have

$P_0^{Y|D=0,X=x_i}(B)\geq\nu_{i+1,i}(P_0^{Y|D=0,X=x_{i+1}})(B)$ then if it exists a positive measure R_i dominated by the distribution of $Y|X=x_i,D=0$ such that $P_0^{Y|X=x_i,D=0}=\nu_{i+1,i}(P_0^{Y|X=x_{i+1},D=0})+R_i$.

Reciprocally, if $\left(P^{Y|X=x_j,D=0}\right)_{j=1,...,J}$ is a vector of distribution such that for every $i=1,...,J-1$ it exists R_i a positive measure dominated by the distribution of $Y|X=x_i,D=0$

such that $P^{Y|X=x_i,D=0} = \nu_{i+1,i}(P^{Y|X=x_{i+1},D=0}) + R_i$, then the selection mechanism defined by

$$E_P(D=1|Y \in B, X=x) = \frac{p(x_1)P_0^{Y|D=1,X=x}(B)}{p(x)P_0^{Y|D=1,X=x}(B)+(1-p(x))P^{Y|D=0,X=x}(B)}$$

rationalizes the data, and the positivity of R_i for $i = 1, ..., J-1$ ensures that Assumption 7 holds.

We deduce that:

$$\mathcal{C} = \left\{ \begin{array}{ll} (P_j)_{j=1,...,J}: & \text{Supp}(P_j) \subset \mathcal{Y}, \\ & \exists (R_j)_{j=1,...,J-1} \text{ positive measures s.t. } P_j = \nu_{j+1,j}(P_{j+1}) + R_j \end{array} \right\}$$

Because $\nu_{j+1,j} \circ \nu_{j,j-1} = \nu_{j+1,j-1}$ and $\nu_{j,j} = Id$, this also means:

$$\mathcal{C} = \left\{ (P_j)_{j=1,...,J}: \ \exists (R_j)_{j=1,...,J} \text{ positive measures on } \mathcal{Y} \text{ s.t. } P_j = \sum_{k=j}^{J} \nu_{k,j}(R_k) \right\}$$

Step 2: $\mathcal{C}$ is closed and convex

The linearity of $\nu_{j+1,j}$ implies that $\mathcal{C}$ is convex (and so is $\mathcal{K}$). To prove that $\mathcal{C}$ is closed remark that:

$$\mathcal{C} = \left\{ \begin{array}{ll} (P_j)_{j=1,...,J}: & \text{Supp}(P_j) \subset \text{Supp}(Y|D=1, X=x_j), \\ & \forall g \text{ positive, bounded and continuous,} \\ & \int g(y)dP_j(y) \geq \int g(y) r_{j+1,j}(y) dP_{j+1}(y) \end{array} \right\}.$$

Let $(P_{1,n}, P_{2,n}, ..., P_{J,n})$ a sequence in $\mathcal{C}$ such that $P_{j,n}$ weakly converges to P_j for all j. Let g a continuous and bounded function from $\mathcal{Y}$ to $\mathbb{R}^+$ and $g_k = \left(\frac{k}{f} \wedge 1\right) g$. Assumption 10 implies that g_k and $g_k f$ are continuous and bounded, moreover $g_k f \uparrow gf$ and $g_k \uparrow g$ when $k \to +\infty$. It follows that $j = 2, ..., J$:

$$\begin{array}{rcl} \int g f dP_j & = & \lim_k \int g_k f dP_j \\ & = & \lim_k \lim_n \int g_k f dP_{j,n} \\ & \leq & \lim_k \lim_n \int g_k dP_{j-1,n} \\ & = & \lim_k \int g_k dP_{j-1} \\ & = & \int g dP_{j-1} \end{array}$$

Then $(P_1, P_2, ..., P_J) \in \mathcal{C}$, so $\mathcal{C}$ and $\mathcal{K}$ are closed for the weak convergence.

Step 3: If $(P_1, ..., P_J) \in \text{ext}(\mathcal{C})$, then $(P_1, ..., P_J) = A(R_1, ..., R_J)$ with $\#\cup_{k=1}^{j} \text{Supp}(R_k) \leq j$

Let $\mathcal{P}_\mathcal{Y}$ (respectively $\mathcal{M}_\mathcal{Y}^+$ and $\mathcal{M}_\mathcal{Y}$) the set of probability measures concentrated on $\mathcal{Y}$ (respectively the set of positive measures and the set of signed measures concentrated on $\mathcal{Y}$). $\mathcal{P}_\mathcal{Y}^J$ (respectively $\mathcal{M}_\mathcal{Y}^{+J}$ and $\mathcal{M}_\mathcal{Y}^J$) denotes a vector of size J with component in $\mathcal{P}_\mathcal{Y}$ (respectively $\mathcal{M}_\mathcal{Y}^+$ and $\mathcal{M}_\mathcal{Y}$).

Because $\nu_{k,j} \circ \nu_{j,k} = Id$, $\nu_{k,j}$ is a linear isomorphism on $\mathcal{M}_{\mathcal{Y}}$ and because $r_{k,j}(y)$ are non-negative functions of y, $\nu_{k,j}$ is moreover a positive endomorphism in the sense that $\nu_{k,j}(\mathcal{M}_{\mathcal{Y}}^{+}) \subset \mathcal{M}_{\mathcal{Y}}^{+}$. Let A the endomorphism on $\mathcal{M}_{\mathcal{Y}}^{J}$ defined by the matrix of component $\nu_{j,i}\mathbb{1}_{\{i \leq j\}}$ at row i and column j. A is a triangular matrix of endomorphisms with isomorphisms on the diagonal, A is an linear isomorphism on $\mathcal{M}_{\mathcal{Y}}^{J}$ and A^{-1} denotes its inverse. The characterization that we obtain in step 1 of the proof ensures that $\mathcal{C} = \mathcal{P}_{\mathcal{Y}}^{J} \cap A\left(\mathcal{M}_{\mathcal{Y}}^{+J}\right)$ or equivalently that $\mathcal{C} = A\left(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^{J}\right)\right)$.
Let $P \in \text{ext}(\mathcal{C})$, it exists $R \in \mathcal{M}_{\mathcal{Y}}^{+J}$ such that $P = A(R)$. Suppose that $R \notin \text{ext}(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^{J}\right)$ then it exists $(R^1, R^2) \in \left(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^{J}\right)\right)^2$ and $\lambda \in]0,1[$ such that $R^1 \neq R^2$ and $R = \lambda R^1 + (1-\lambda)R^2$. It follows that $P = \lambda A(R^1) + (1-\lambda)A(R^2)$ with $A(R^1) \neq A(R^2)$, which is absurd. Then:

$$\text{ext}(\mathcal{C}) \subset A\left(\text{ext}\left(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^{J}\right)\right)\right).$$

The reverse inclusion also holds : indeed for $R \in \text{ext}\left(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^{J}\right)\right)$, if we assume that $A(R) = P \notin \text{ext}(\mathcal{C})$ then it exists λ, P^1 and P^2 such that $R = \lambda A^{-1}(P^1) + (1-\lambda)A^{-1}(P^2)$, which is absurd. So we are lead back to characterize ext $\left(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^{J}\right)\right)$, ie the extremal parts of:

$$\left\{ \; R \in \mathcal{M}_{\mathcal{Y}}^{+J} \text{ such that } \forall j = 1, ..., J, \sum_{k=j}^{J} \int r_{k,j}(y) dR_k(y) = 1 \; \right\}$$

Let $\overline{R}_L$ the sum of the L first components of R, $\overline{R}_L = \sum_{l=1}^{L} R_l$. For $k \leq L \leq J$, $\frac{dR_k}{d\overline{R}_L}$ denotes the Radon Nikodym density of R_k with respect to $\overline{R}_L$. For any $L \leq J$:

$$\begin{array}{rcl} \sum_{k=j}^{J} \int r_{k,j}(y) dR_k(y) & = & \int \sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k}{d\overline{R}_L}(y) d\overline{R}_L(y) \\ & & + \sum_{k=L+1}^{J} \int r_{k,j}(y) dR_k(y) \end{array}$$

For any $L = 1, ..., J$, assume that $\text{span}\left(\left(\sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k}{d\overline{R}_L}(y) \right)_{j=1,...,L} \right)$ is not dense in $L^1(\overline{R}_L)$. The Hahn-Banach Theorem and the identification $L^{1\star}(\mu) = L^{\infty}(\mu)$ ensure that it exists h such that

$$\int \sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k}{d\overline{R}_L}(y) h(y) d\overline{R}_L(y) = 0.$$

Let R^+ and R^- such that $dR_j^{\pm} = \left(1 \pm \frac{h}{||h||_{\infty}}\right) dR_j$ for $j \leq L$ and $R_j^{\pm} = R_j$ for $j > L$. By construction $R = R^+/2 + R^-/2$ with $(R^+, R^-) = \left(\mathcal{M}_{\mathcal{Y}}^{+J}\right)^2$.

For $k \leq L$, note that $\frac{dR_k}{d\overline{R}_L} = \frac{dR_k^+}{d\overline{R}_L^+}$, and then:

$$\begin{aligned}
\sum_{k=j}^{J} \int r_{k,j}(y) dR_k(y) &= \int \sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k^+}{d\overline{R}_L^+}(y) d\overline{R}_L(y) \\
&\quad + \sum_{k=L+1}^{J} \int r_{k,j}(y) dR_k(y) \\
&= \int \sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k^+}{d\overline{R}_L^+}(y) d\overline{R}_L(y) \\
&\quad + \frac{1}{||h||_\infty} \int \sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k}{d\overline{R}_L}(y) h(y) d\overline{R}_L(y) \\
&\quad + \sum_{k=L+1}^{J} \int r_{k,j}(y) dR_k^+(y) \\
&= \int \sum_{k=j}^{L} r_{k,j}(y) \frac{dR_k^+}{d\overline{R}_L^+}(y) d\overline{R}_L^+(y) \\
&\quad + \sum_{k=L+1}^{J} \int r_{k,j}(y) dR_k^+(y) \\
&= \sum_{k=j}^{J} \int r_{k,j}(y) dR_k^+(y)
\end{aligned}$$

And then $R^+ \in \mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^J\right)$. A similar reasoning ensures also that $R^- \in \mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^J\right)$. So R is not an extreme part of $\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^J\right)$.

It follows that if $R \in \text{ext}\left(\mathcal{M}_{\mathcal{Y}}^{+J} \cap A^{-1}\left(\mathcal{P}_{\mathcal{Y}}^J\right)\right)$ then $\#\text{Supp}(\sum_{l=1}^{L} R_l) \leq L$. Because $\text{Supp}(\sum_{l=1}^{L} R_l) \subset \text{Supp}(\sum_{l=1}^{L+1} R_l)$, we deduce that $R = T^- \delta$ where T^- is a lower triangular matrix with non negative component and δ is a column vector of J Dirac measures.

$T^-\delta$ is an element of $\mathcal{M}_{\mathcal{Y}}^{+J}$ with components of the form $\left(\sum_{j=1}^{k} w_{k,j}\delta_{y_j}\right)_{k=1,\dots,J}$. It follows that $A(T^-\delta)$ is a vector with component of the form:

$$\sum_{j=1}^{J} \left[\sum_{k=\max(i,j)}^{J} w_{kj} r_{ki}(y_j) \right] \delta_{y_j}.$$

Because $r_{ii}(y_j) = 1$ and we deduce that $\sum_{j=1}^{J} w_{ij} \leq 1$ and next that $w_{ij} \in [0;1]$.

Bounds under Assumption 8

Note that the distribution of X is free in this case, so $\mathcal{C} = \prod_{j=1}^{J} \mathcal{C}_j$, with $\mathcal{C}_j$ the set of distributions of $Y|D=0, X=x_j$ for $j = 1, ..., J$ compatible with the data and assumptions. $\mathcal{C}$ is closed for the weak convergence if and only if $\mathcal{C}_j$ is closed for the weak convergence for every $j = 1, ..., J$, and in this case we have $\text{ext}(\mathcal{C}) = \prod_{j=1}^{J} \text{ext}(\mathcal{C}_j)$. And then to characterize $\mathcal{K}$ we only have to prove that $\mathcal{C}_j$ is closed and to characterize $\text{ext}(\mathcal{C}_j)$ for every $j = 1, ..., J$.

Step 1: $\mathcal{C}_j$ is closed.

Let F_0 and G_0 the cumulative distribution functions of $P_0^{Y|D=0,X=x_j}$ and $P_0^{Y|D=1}$. Let $S_0 = 1 - F_0$ and $\widetilde{\mathcal{Y}} = \{y \in \mathbb{R} : G_0(y) > G_0(\inf(\mathcal{Y}))\}$. We can exclude the case where $\widetilde{\mathcal{Y}}$ is empty[8].

For every $y \in \widetilde{\mathcal{Y}} \cap \mathcal{Y}$, $E_{P_0}(D|Y=y, X=x_j) > 0$ and the Bayes formula implies that:

$$dF_0 = \left(1/E_{P_0}(D|Y=y, X=x_j) - 1\right) \frac{p(x_j)}{1-p(x_j)} dG_0.$$

[8]In such case, the distribution of Y is a Dirac and the conclusion holds.

And then $S_0(y) = \int_{]y;+\infty]} f dG_0$, with f a non increasing function from $\widetilde{\mathcal{Y}}$ to $\mathbb{R}^+$.
It follows that for all $(y_0, y) \in \widetilde{\mathcal{Y}}^2$ such that $G_0(y) \neq G_0(y_0)$,

$$\frac{F_0(y) - F_0(y_0)}{G_0(y) - G_0(y_0)} = \mathbb{E}\left(f(Y)|Y \in]y_0; y] \cup]y; y_0], D = 1\right),$$

with f a non increasing function from $\widetilde{\mathcal{Y}}$ to $\mathbb{R}^+$.
It follows that the function

$$y \mapsto \frac{F_0(y) - F_0(y_0)}{G_0(y) - G_0(y_0)}$$

is non increasing on $\widetilde{\mathcal{Y}} \backslash G_0^{-1}(\{G_0(y_0)\})$ for every $y_0 \in \widetilde{\mathcal{Y}}$.
Reciprocally, let F be a cdf concentrated on $\mathcal{Y}$ such that

$$y \mapsto \frac{F(y) - F(y_0)}{G_0(y) - G_0(y_0)}$$

is non increasing on $\widetilde{\mathcal{Y}} \backslash G_0^{-1}(\{G_0(y_0)\})$ for every $y_0 \in \widetilde{\mathcal{Y}}$.
We can define the left limit on y_0 of such function:

$$g_l(y_0) = \lim_{y \to y_0^-} \frac{F(y) - F(y_0)}{G_0(y) - G_0(y_0)}.$$

g_l is a non increasing and left continuous function from $\widetilde{\mathcal{Y}}$ to $\mathbb{R}^+$.
Because g_l is non increasing, g_l is Riemann-Stieltjes integrable, and then for $y \in]\inf \mathcal{Y}; y_0]$:

$$\begin{array}{c} \sum_{n=0}^{N-1} g_l(y + \frac{y_0 - y}{N}(n+1)) \left[G_0(y + \frac{y_0 - y}{N}(n+1)) - G_0(y + \frac{y_0 - y}{N} n)\right] \\ \leq \sum_{n=0}^{N-1} F(y + \frac{y_0 - y}{N}(n+1)) - F(y + \frac{y_0 - y}{N} n) = F(y_0) - F(y) \leq \\ \sum_{n=0}^{N-1} g_l(y + \frac{y_0 - y}{N} n) \left(G_0(y + \frac{y_0 - y}{N}(n+1)) - G_0(y + \frac{y_0 - y}{N} n)\right) \end{array}$$

When N tends to infinity, we have:

$$F(y_0) - F(y) = \int_{]y;y_0]} g_l(y) dG_0(y)$$

For $y_0 \to +\infty$, we deduce that:

$$1 - F(y) = \int_{]y;+\infty[} g_l(y) dG_0(y).$$

For every $y \in \widetilde{\mathcal{Y}}$, let $E(D|Y = y, X = x_j) = \frac{1 - p(x_j)}{(1 - p(x_j)) g_l(y) + p(x_j)}$. This is a increasing function in y. When $\inf \mathcal{Y} > -\infty$, such function could be extended at $\inf \mathcal{Y}$ by $E(D|Y = \inf \mathcal{Y}, X = x_j) = \inf_{y \in \mathcal{Y}} \frac{1 - p(x_j)}{(1 - p(x_j)) g_l(y) + p(x_j)}$. Such mechanism of selection rationalize the data and the Assumption 8.

So we have proved that $\mathcal{C}_j$, is the set of probability distributions concentrated on $\mathcal{Y}$ with associated cdf F such that :

$$y \mapsto \frac{F(y) - F(y_0)}{G_0(y) - G_0(y_0)}$$

is non increasing on $\widetilde{\mathcal{Y}} \backslash G_0^{-1}(\{G_0(y_0)\})$ for every $y_0 > \widetilde{\mathcal{Y}}$.
For every F cdf supported on $\mathcal{Y}$, let $c(F) = \{y \in \widetilde{\mathcal{Y}} : F(y^-) = F(y)\}$ and
$d(F) = \{y \in \widetilde{\mathcal{Y}} : F(y^-) < F(y)\}$.
Let F_n a sequence of cdf of distribution in $\mathcal{C}_j$, such that F_n converge to F at every point of continuity of a cdf F. For every $y_0 \in c(F)$ and for every $y \in c(F) \backslash G_0^{-1}(\{G_0(y_0)\})$, $\frac{F_n(y)-F_n(y_0)}{G_0(y)-G_0(y_0)}$ converges to $\frac{F(y)-F(y_0)}{G_0(y)-G_0(y_0)}$.
For every $y_0 \in c(F)$ and for every $(y_1, y_2) \in \left(\widetilde{\mathcal{Y}} \backslash G_0^{-1}(G_0(y_0))\right)^2$ such that $y_1 < y_2$, it exists sequences y_{1n} and y_{2n} in $c(F)$ such that $y_{1n} \to y_1^+$, $y_{2n} \to y_2^+$, $G_0(y_{1n}) \neq G_0(y_0)$ and $G_0(y_{2n}) \neq G_0(y_0)$. Then we deduce that we have:

$$\frac{F(y_1) - F(y_0)}{G_0(y_1) - G_0(y_0)} \geq \frac{F(y_2) - F(y_0)}{G_0(y_2) - G_0(y_0)}.$$

Similarly, for every $(y_1, y_2) \in \widetilde{\mathcal{Y}}^2$ such that $y_1 < y_2$, if $y_0 \in d(F)$ such that
$G_0(y_0) \notin \{G_0(y_1); G_0(y_2)\}$ it exists $y_{0n} \in c(F)$ decreasing sequence converging to y_0 such that $G_0(y_{0n}) \notin \{G_0(y_1); G_0(y_2)\}$. Because G_0 is right continuous, we have :

$$\frac{F(y_1) - F(y_0)}{G_0(y_1) - G_0(y_0)} \geq \frac{F(y_2) - F(y_0)}{G_0(y_2) - G_0(y_0)}.$$

It follows that $\mathcal{C}_j$ is closed.

Step 2: Characterization of $\text{ext}(\mathcal{C}_j)$.
Let F a cdf of an element of $\mathcal{C}_j$. If $1 > F(\inf \mathcal{Y}) > 0$ then
$F(y) = F(\inf \mathcal{Y}) \mathbb{1}_{y \geq \inf \mathcal{Y}} + (1 - F(\inf \mathcal{Y})) \frac{(F(y)-F(\inf \mathcal{Y}) \sup)^+}{(1-F(\inf \mathcal{Y}))}$,
and $\frac{(F(y)-F(\inf \mathcal{Y}) \sup)^+}{(1-F(\inf \mathcal{Y}))}$ is a cdf of an element of $\mathcal{C}$. So if F is the cdf of an element of $\text{ext}(\mathcal{C}_j)$ then $F(\inf \mathcal{Y}) \in \{0; 1\}$.
Now assume that $F(\inf \mathcal{Y}) = 0$, then $F(y) = \int_{]-\infty; y] \cap \mathcal{Y}} g(u) dG_0(u)$ with g decreasing function from $\mathcal{Y}$ to $\mathbb{R}^+$.
Let $a < b$ two values in $\mathcal{Y}$. Assume that $\int_{]a;b] \cap \mathcal{Y}} (g(a) - g(u))(g(u) - g(b)) dG_0(u) > 0$, in this case let $P(u) = (g(a) - u)(u - g(b))(u - \mu g(a) - (1 - \mu) g(b))$ with μ such that $\mu = \frac{\int_{]a;b] \cap \mathcal{Y}} (g(a)-g(u))(g(u)-g(b))^2 dG_0(u)}{(g(a)-g(b)) \int_{]a;b] \cap \mathcal{Y}} (g(a)-g(u))(g(u)-g(b)) dG_0(u)}$. Because the derivatives of $P(u)$ are bounded on $[g(b); g(a)]$, it exists $\lambda \neq 0$ such that $g(u) - \lambda P(g(u))$ and $g(u) + \lambda P(g(u))$ are decreasing on $]a; b]$.
Moreover we have $\int_{]a;b] \cap \mathcal{Y}} P(g(u)) dG_0(u) = 0$. So F is the cdf of an element of $\text{ext}(\mathcal{C}_j)$ only if $P(g(u)) = 0$ $P_0^{Y|D=1, X=x_j, Y \in]a;b]}$-almost-surely. This means that $g(u)$ takes at most three

values on $]a;b]$: $g(a)$, $\mu g(a)+(1-\mu)g(b)$ and $g(b)$. Because such result hods for every a and b, g takes at most three values on $\mathcal{Y}$, $P_0^{Y|D=1,X=x_j}$-almost-surely. If there is two values v_1 and v_2 such that $v_1 > v_2 > 0$, we can easily find ε and ε' positive numbers sufficiently small such that $v_1 - \varepsilon = v_2 + \varepsilon'$, $v_2 - \varepsilon' > 0$ and $\varepsilon Q_0^{Y|D=1,X=x_j}(g^{-1}(v_1)) - \varepsilon' Q_0^{Y|D=1,X=x_j}(g^{-1}(v_2)) = 0$. So g takes at most one non null value, $P_0^{Y|D=1,X=x_j}$-almost-surely.

Bounds under Assumptions 7 and 8

$\mathcal{C}$ is the intersection of distributions compatible with the data and assumptions 7 and 8. Because the sets of distribution that verify Assumption 7 or Assumption 8 are close for the weak convergence (cf. the two previous demonstration). $\mathcal{C}$ is closed as the intersection of closed sets.

We denote by G_0 the cdf of $P_0^{Y|D=1}$ and we first assume that $G_0(\inf \mathcal{Y}) = 0$. Let f_j a Radon Nikodym derivative of $P_0^{Y|D=1,X=x_j}$ with respect to $P_0^{Y|D=1}$ (such derivative exists by Assumption 9). In this case if $F = (F_1, ..., F_J)$ is the cdf of $P = (P_1, ..., P_J)$ in $\mathcal{C}$, then using some results of the two previous proofs, it exists $(g_1, ..., g_J)$ in decreasing functions such that:

$$g_{j+1} \leq g_j, P_0^{Y|D=1} - a.s.$$
$$F_j(y) = \int_{]-\infty;y]} g_j(z) f_j(z) \frac{p(x_j)}{1-p(x_j)} dG_0(z)$$

Let a and b two elements of $\mathcal{Y}$ such that $a < b$.

Let $R(x) = \prod_{j=1}^{J}(g_j(a)-x)(x-g_j(b))$ and S a polynomial such that $S(x) = R(x)\left(\sum_{k=0}^{J} \alpha_k x^k\right)$. It exists $(\alpha_0, ..., \alpha_J)$ not identically equal to 0 such that

$$\int_{]a;b]} S(g_j(y)) f_j(y) \frac{p(x_j)}{1-p(x_j)} dG_0(y) = \sum_{k=0}^{J} \alpha_k \int_{]a;b]} g_j(y)^k f_j(y) \frac{p(x_j)}{1-p(x_j)} dG_0(y) = 0,$$

for every $j = 1, ..., J$. Such S has bounded derivatives on $[\min_{j=1,...,J} g_j(b); \max_{j=1,...,J} g_j(a)]$. So it exists $\lambda > 0$ and sufficiently small such that $x+\lambda S(x)$ and $x-\lambda S(x)$ are non decreasing on $[\min_{j=1,...,J} g_j(b); \max_{j=1,...,J} g_j(a)]$.

Let $u_j(y) = g_j(y) + \mathbb{1}_{\{y\in[a;b]\}} \lambda S(g_j(y))$ and $v_j(y) = g_j(y) - \mathbb{1}_{\{y\in[a;b]\}} \lambda S(g_j(y))$.

We can define the cdf $U_j(y) = \int_{]-\infty;y]} u_j(z) f_j(z) \frac{p(x_j)}{1-p(x_j)} dG_0(z)$ and
$V_j(y) = \int_{]-\infty;y]} v_j(z) f_j(z) \frac{p(x_j)}{1-p(x_j)} dG_0(z)$. $U = (U_1, ..., U_J)$ and $V = (V_1, ..., V_J)$ are in $\mathcal{C}$ and $F = (U + V)/2$. It follows that $P \in \text{ext}(\mathcal{C})$ if and only if $U = V = F$, and so if and only if $S \circ g_j = 0$ $P_0^{Y|D=1,Y\in]a;b]}$-a.s. Because S has at most $3J$ roots, $\cup_{j=1,...,J} g_j(\mathcal{Y})$ has at most $3J$ elements.

Now remark that it exists $(\beta_1, ..., \beta_{J+1}) \neq (0, ..., 0)$ such that $\sum_{k=1}^{J+1} \beta_k \int g_j(z)^k f_j(z) dG_0(z) = 0$. Because the g_j have a finite number of values, $\min_{x:g_j(x)>0} g_j(x)$ and $\max_{x:g_j(x)>0} g_j(x)$ are well defined and non negative. Let $H(x) = \sum_{k=1}^{J+1} \beta_k x^k$, H has bounded derivatives on the compact $I = [1/2 \times \min_{x:g_j(x)>0} g_j(x); 2 \times \max_{x:g_j(x)>0} g_j(x)]$. So it exists $\lambda > 0$

and sufficiently small such that $x + \lambda H(x)$ and $x - \lambda H(x)$ are non decreasing and non negative on I. Considering $t_j(x) = g_j(y) + \lambda H(g_j(y))$ and $w_j(x) = g_j(y) - \lambda H(g_j(y))$ the fact that $g_j(y) \in I \cup 0$, we deduce that $P \in \text{ext}(\mathcal{C})$, $H \circ g_j = 0$ $P_0^{Y|D=1}$-a.s. Because H has at most J non null roots, it follows that $\# \cup_{j=1,\dots,J} \{g_j(y) : y \in \mathcal{Y}, g_j(y) > 0\} \leq J$ then $g_j(y) = \sum_{i=1}^{J} a_i \mathbb{1}_{\{y \leq y_{ij}\}}$ and the result follow.