Four essays in econometrics

Laurent Davezies

Institut d'Etudes Politiques de Paris

ECOLE DOCTORALE DE SCIENCES PO

Programme doctoral en économie

Département d'économie de Sciences Po

Doctorat en Sciences économiques

Four Essays in Econometrics

LAURENT DAVEZIES

Thèse dirigée par Jean-Marc Robin

Soutenue le 19 décembre 2013

Jury:

- Christian Bontemps, Rapporteur, Directeur de recherche à l'École Nationale d'Aviation Civile et à la Toulouse School of Economics
- Laurent Gobillon, Suffragant, Chargé de Recherche à l'Institut National d'Études Démographiques
- Marc Henry, Rapporteur, Professor of Economics at Pennstate University
- Koen Jochmans, Suffragant, Professeur d'Économie à l'Institut d'Études Politiques de Paris
- Jean-Marc Robin, Directeur de thèse, Professeur d'Économie à l'Institut d'Études Politiques de Paris

Je remercie en premier lieu mon directeur de thèse, Jean-Marc Robin pour avoir accepté de soutenir mon inscription en thèse malgré mon absence de diplôme de Master Recherche. Je remercie également les rapporteurs, Christian Bontemps et Marc Henry qui ont accepté de prendre du temps pour lire ce travail et rédiger un rapport ainsi les suffragants, Laurent Gobillon et Koen Jochmans qui ont également accepté de discuter ce travail.
Je remercie également les coauteurs avec qui j'ai travaillé sur les sujets présentés dans cette thèse: Romain Aeberhardt, Magali Beffy et Xavier D'Haultfœuille. Les autres coauteurs, Manon Garrouste et Léa Toulemon avec lesquels je travaille actuellement méritent également d'être remerciés puisque ces collaborations contribuent également à me faire progresser. Je remercie aussi Denis Fougère avec lequel j'ai travaillé par le passé et que j'ai toujours trouvé extrêmement stimulant. Mes pensées vont également à Delphine et Alain, tous deux disparus, qui ont fortement contribué à ma construction personnelle et à développer mon appétit pour les sciences sociales.
La possibilité offerte par l'Insee de laisser du temps à certains administrateurs de l'Insee pour se frotter aux exigences de la recherche académique me permet aujourd'hui de soutenir cette thèse après trois ans passés au CREST (Centre de Recherche en Économie et en Statistique). Ce mode de management constitue une singularité dans l'administration française. J'espère m'être montré digne de la chance qui m'a été donnée. Aujourd'hui, je suis convaincu que l'administration française gagnerait beaucoup à s'ouvrir plus largement à la recherche académique.

Contents

Chapter 1

Introduction

Cette thèse compile quatre travaux d'économétrie sur lesquels j'ai travaillé ces trois dernières années. De manière informelle l'économétrie peut se définir comme la manière dont des données empiriques sont utilisées pour dégager certaines connaissances sur des grandeurs d'intérêt pour l'économiste. En ce sens, on pourrait penser que l'économétrie ne présente pas de différence fondamentale de nature avec la statistique. Néanmoins, l'économétrie n'est pas réductible à la statistique dans le sens où les paramètres d'intérêt sont issus d'une théorie économique ou sociologique (plus ou moins implicite, plus ou moins formalisée). Une bonne part du travail en économétrie consiste donc à articuler un modèle statistique avec un cadre conceptuel issu des sciences sociales. Ce cadre conceptuel peut être par exemple une certaine forme de rationalité dans les arbitrages opérés par les agents, ou encore l'existence d'institutions telles que les marchés qui confrontent offre et demande pour allouer les ressources. En ce sens, l'économétre s'intéresse donc à des modèles statistiques spécifiques.

Au delà de cette définition rapide (et critiquable), le champ des travaux en économétrie est très varié. D'abord parce que la nature des modèles statistiques mobilisés diffère selon les problématiques étudiées. Ensuite parce que la nature des travaux peut aller de la statis-

tique théorique à la mesure empirique de l'effet de certaines politiques publiques en passant par la méthodologie computationnelle.

Cette thèse regroupe quatres travaux présentés dans un ordre quelque peu arbitraire mais que l'on pourrait classer comme allant du plus théorique et général au plus empirique et spécifique à une question donnée. Le but de cette introduction est de vulgariser en français les quatres chapitres de cette thèse. Dans cette optique les références à la littérature scientifique sont volontairement réduites. Le lecteur profane peut donc se contenter de la lecture de cette introduction alors que le lecteur averti peut directement se reporter aux chapitres rédigés en anglais.

1.1 Chapitre 1: Caractérisation des régions d'identification par les parties extrêmes d'un ensemble de distributions.

Ce chapitre a été écrit dans le cadre d'un travail en collaboration avec Xavier d'Haultfoeuille.

1.1.1 Simple exemple introductif

Ce premier chapitre s'intéresse à la caractérisation des régions d'identification d'un paramètre. La notion de région d'identification a été promue par Manski dans une série d'articles et un livre[1] au tournant du siècle. Un paramètre théorique est identifié (ponctuellement) si seulement une valeur de ce paramètre est compatible avec ce qu'observerait un économètre

[1]Partial Identification of Probability Distributions, Springer Series in Statistics, 2003

ayant un échantillon d'observations de très grande taille. Imaginons par exemple, le cas d'un économiste qui souhaite mesurer le taux de chômage au sein d'une population. Sous les hypothèses usuelles de l'économétrie, en observant le statut sur le marché du travail d'un échantillon d'individus, notre économiste pourra estimer le taux de chômage. La loi des grands nombres implique que si la taille de l'échantillon interrogé devient très grand l'estimateur sera de plus en plus proche du "vrai" taux de chômage. Avec un échantillon de taille infini le taux de chômage est donc ponctuellement identifié. Cependant, en pratique la situation peut être moins favorable : imaginons que pour 20% des personnes notre économiste n'arrive pas à observer le statut sur le marché du travail. Imaginons que parmi les individus observés le taux de chômage est de 10%, que peut-on en déduire sur la valeur du taux de chômage dans la population active ? Sans faire d'hypothèses supplémentaires, on sait simplement que le taux de chômage se situe entre 8%=10%*80%+0%*20% (cas où tous les individus non observés travaillent) et 28%=10%*80%+100%*20% (cas où tous les individus non observés sont chômeurs). Sans faire d'hypothèses supplémentaires, l'économiste ne peut exclure aucune des valeurs de l'intervalle $[8\%; 28\%]$. Le paramètre n'est donc pas ponctuellement identifié. A l'inverse, l'observation d'un grand échantillon permet à l'économiste de rejeter les valeurs inférieures à 8% ou supérieures à 28%. Les données sont donc informatives et on dit que le paramètre est partiellement identifié. On appelle région d'identification du taux de chômage l'intervalle $[8\%; 28\%]$. Des hypothèses supplémentaires peuvent permettre de réduire cette région d'identification : par exemple si l'économiste pense que les personnes en emploi sont plus difficiles à contacter que les personnes au chômage, le taux de chômage parmi les individus non observés doit être inférieur à celui des individus observés : sous cette hypothèse supplémentaire la région

d'identification se réduit à l'intervalle $[8\%; 10\%]$. Une telle hypothèse est donc très informative car elle réduit très significativement la région d'identification. Une autre hypothèse souvent adoptée dans les travaux empiriques consiste à supposer que la sélection est ignorable (c'est à dire que les individus observés et non observés sont "comparables"). Dans ce cas la région d'identification est réduite à la seule valeur de 8%, et le paramètre redevient ponctuellement identifié.

L'économiste peut également faire des hypothèses qui conduisent à une région d'identification vide, cela signifie alors que les hypothèses formulées par l'économiste sont rejetées par l'observation, car aucune valeur du taux de chômage n'est susceptible d'être compatible simultanément avec les hypothèses et l'observation.

Au vu des faits stylisés concernant les taux de réponse aux enquêtes, il est plus que probable que les femmes répondent plus facilement aux enquêtes que les hommes toutes choses égales par ailleurs. C'est donc une hypothèse que peut raisonnablement faire l'économiste. Imaginons que notre économiste fasse deux enquêtes dans les régions de Syldavie et de Bordurie. Imaginons que la structure des individus interrogés soit la suivante:

Table 1.1: Structure des individus interrogés

(a) Syldavie

	Répondants	Non répondants	Total
Homme Chômeur	5	9	50
Homme Salarié	36		
Femme Chômeuse	3	11	50
Femme Salariée	36		
Total	80	20	100

(b) Bordurie

	Répondants	Non répondants	Total
Homme Chômeur	4	9	50
Homme Salarié	37		
Femme Chômeuse	4	11	50
Femme Salariée	35		
Total	80	20	100

A première vue, deux remarques s'imposent. Premièrement, les comportements de réponse et le statut des hommes et des femmes sur le marché du travail semble proches en Bordurie et en Syldavie. Deuxièmement, le taux de réponse des femmes est inférieur à celui des hommes, or notre économiste pense qu'elles répondent plus toutes choses égales par ailleurs. Cela signifie que le statut sur le marché du travail joue sur le comportement de réponse et que les femmes répondent moins du fait de leur situation sur le marché du travail.

Déterminer la région d'identification du taux de chômage pour ces deux régions sous l'hypothèse que les femmes répondent plus toutes choses égales par ailleurs n'a rien d'évident. Un petit calcul montre alors que la région d'identification du taux de chômage de la Syldavie est [944/39;28]$\simeq$[24,21;28] alors que celle de la Bordurie est vide ! L'hypothèse que

les femmes répondent plus toute choses égales par ailleurs ne résiste pas à la confrontation avec les données pour la Bordurie alors qu'elle est soutenable pour la Syldavie.

1.1.2 Formulation générale du problème

Dans l'exemple présenté ci-dessus, la détermination de la région d'identification peut se faire par des calculs relativement élémentaires[2]. Plusieurs autres types de problème ont été résolus par des méthodes d'intuition puis de vérification : l'économiste arrive au vu du problème à avoir une intuition de la région d'identification, il vérifie ex-post que toutes les valeurs possibles du paramètre sont bien dans la région considérée et réciproquement que tout point de la région considérée est compatible simultanément avec les hypothèses et l'observation. Ce mode de preuve est donc spécifique au problème étudié, pour un nouveau problème l'économiste ne dispose pas de théorèmes généraux lui permettant de caractériser simplement la région d'identification. A notre connaissance les seules méthodes générales concerne la classe des problèmes qui s'expriment sous forme de conditions de moments, autrement dit par des conditions sur certaines moyenne. Dans ces problèmes, le paramètre est caractérisé par un nombre fini d'(in)égalités de moyennes calculables sur les données et les hypothèses retenues par l'économiste sont également caractérisées par un nombre fini ou infini d'(in)égalités de moyennes. Lorsque toutes les variables ne peuvent prendre qu'un nombre fini de valeurs et que le problème s'exprime au moyen d'un nombre fini de conditions de moments, Manski et Horowitz ont mis en avant le fait que la solution du problème pouvait se calculer numériquement par des algorithmes standards (parmi eux

[2]Mais suffisamment pénibles pour qu'on ne les détaille ici.

citons l'algorithme du simplexe). Lorsque le paramètre est un moment de la distribution jointe de deux variables mais que seules les distributions marginales sont identifiées dans les données, le problème est alors un "problème de transport optimal". Pour prendre un exemple simple, imaginons que notre économiste veuille calculer la moyenne des taux d'effort des ménages français locataires, c'est à dire la moyenne des ratio loyers/revenus. Si notre économiste ne dispose que de données concernant les loyers d'un côté et de données concernant les revenus de ménages de l'autre il n'est pas capable de calculer la moyenne des taux d'effort. Il pourra seulement déduire des bornes des données observées. Lorsque le problème peut s'exprimer sous la forme d'un problème de transport optimal, les travaux de Ekeland, Galichon et Henry ont mis en avant la dualité de Monge-Kantorovitch pour reformuler le problème initial en un problème plus simple. Enfin, Beresteanu, Molchanov et Molinari ont proposé d'utiliser les outils mathématiques des ensembles aléatoires pour traiter ce type de problèmes. Au delà de cette classe de problèmes (restrictions exprimables sous forme de conditions de moments), à notre connaissance, il n'existe pas de résultats généraux.

Nous considérons la classe des problèmes qui peuvent se formuler de la manière suivante : le paramètre d'intérêt θ_0 est caractérisé par une distribution de probabilité P_0 imparfaitement observée. Cette caractérisation s'écrit sous la forme $q(\theta_0, P_0) = 0$. Par ailleurs, l'économiste peut imposer des restrictions supplémentaires sur P_0 au moyen d'hypothèses qu'il juge crédibles. L'observation des données et les restrictions postulées impliquent que P_0 appartient à un ensemble de distributions de probabilité particulier que nous notons $\mathcal{R}$. La région d'identification de θ_0 est donc l'ensemble des valeurs θ telles qu'il existe une distribution de probabilité compatible avec les données, les hypothèses de l'économiste et

la valeur θ, mathématiquement cela signifie qu'il existe P dans $\mathcal{R}$ telle que $q(\theta, P) = 0$. Pour toute valeur θ du paramètre, nous notons $\mathcal{R}_\theta = \{P \in \mathcal{R}, q(\theta, P) = 0\}$, l'ensemble des distributions compatibles avec les données, les restrictions postulées par l'économiste et la valeur θ du paramètre. La valeur θ est dans la region d'identification du paramètre si et seulement si $\mathcal{R}_\theta$ est non vide.

Reprenons le simple exemple introductif, notons Y_i la variable qui vaut 1 si l'individu i est au chômage et 0 s'il est en emploi, notons D_i la variable qui vaut 1 si Y_i est observée et 0 sinon, et X_i la variable qui vaut 1 quand l'individu i est une femme et 0 sinon. On observe la distribution de (DY, D, X) mais le taux de chômage est une fonction de la distribution de Y donc de la distribution (Y, D, X). Dans ce cadre, $q(\theta, P) = \theta - \int y dP(y, d, x)$. Les contraintes sur P imposées par les données sont telles que

$$\mathbb{P}((DY, Y, X) \in A) = \int \mathbb{1}_{\{(dy,d,x)\in A\}} dP(y, d, x) \text{ pour tout } A \subset \{0; 1\}^3, \tag{1.1.1}$$

où le membre de gauche de l'équation précédente est observé dans les données. Si l'économiste ne souhaite pas imposer de restrictions supplémentaires, alors $\mathcal{R}$ est l'ensemble des distributions qui vérifient l'équation (1.1.1) et $\mathcal{R}_\theta$ est le sous ensemble des distributions de $\mathcal{R}$ telles que le taux de chômage vaut θ.

Si l'économiste souhaite ajouter l'hypothèse que les personnes en emploi sont plus difficiles à joindre, il ajoutera la contrainte sur P suivante :

$$\mathbb{P}(D = 1|Y = 1, X = a) \geq \mathbb{P}(D = 1|Y = 0, X = a) \text{ pour } a \in \{0; 1\},$$

autrement dit

$$\frac{\int \mathbb{1}_{\{d=1,y=1,x=a\}} dP(y,d,x)}{\int \mathbb{1}_{\{y=1,x=a\}} dP(y,d,x)} \geq \frac{\int \mathbb{1}_{\{d=0,y=1,x=a\}} dP(y,d,x)}{\int \mathbb{1}_{\{y=1,x=a\}} dP(y,d,x)}.$$

De la même manière on peut formuler l'hypothèse que les femmes répondent plus que les hommes à situation comparable sur le marché du travail:

$$\mathbb{P}(D=1|Y=a,X=1) \geq \mathbb{P}(D=1|Y=a,X=0) \text{ pour } a \in \{0;1\}.$$

Pour toute valeur θ, nous postulons également la convexité de l'ensemble $\mathcal{R}_\theta$ des distributions compatibles avec les données, les restrictions postulées par l'économiste et la valeur θ. Si P_1 et P_2 sont deux distributions différentes des variables dans la population, toutes deux compatibles avec les données, les restrictions postulées et la valeur θ, cela signifie que la population peut être décrite de deux manières différentes sans que l'on puisse distinguer l'une de l'autre sur la base des données observées, des hypothèses faites et de la valeur du paramètre d'intérêt. Une autre formulation consiste à penser qu'on a potentiellement deux populations régies par des comportements différents qui sont indiscernables compte tenu des hypothèses retenues, des observations faites et de la valeur du paramètre d'intérêt. La convexité de $\mathcal{R}_\theta$ signifie simplement que toute population obtenue en mélangeant des individus des deux populations considérées précédemment sera également compatible avec les données, les hypothèses retenues par l'économiste et la valeur θ du paramètre considérée. Cette stabilité par mélange est en fait assez naturelle. A notre connaissance, elle est vérifiée dans tous les problèmes empiriques intéressants pour un économiste.

1.1.3 Apport du chapitre 1

Au delà de l'exemple simple présenté introduction, le cadre général dans lequel nous nous plaçons couvre un grand nombre de problèmes déjà traités dans la littérature : modèles de sélection, problème de transport optimal, inégalités et/ou égalités de moments, variables instrumentales... De plus ce cadre couvre des cas intéressants qui n'ont pas été traités dans la littérature à notre connaissance. En ce sens, le cadre que nous développons permet d'unifier de nombreux problèmes et de généraliser les techniques de résolutions.

Supposons pour simplifier que $\mathcal{R}_\theta$ soit fermé pour la convergence en distribution (cette hypothèse est relâchée dans l'article). Nous montrons que dans ce cadre, les distributions dans $\mathcal{R}_\theta$ qui ne peuvent pas s'écrire comme un mélange de deux distributions de $\mathcal{R}_\theta$ jouent un rôle important. Nous notons $\text{ext}(\mathcal{R}_\theta)$ cet ensemble qui est l'ensemble des parties extrêmes de $\mathcal{R}_\theta$. Tout d'abord, nous avons vu que pour déterminer si la valeur θ était ou non dans la région d'identification, il fallait déterminer si $\mathcal{R}_\theta$ était vide ou non. Nous montrons que cela revient au même de vérifier si $\text{ext}(\mathcal{R}_\theta)$ est vide ou non. Ce résultat est trivial dès lors qu'un convexe fermé non vide admet une partie extreme. Le théorème de Krein-Milman nous assure que c'est le cas dès que $\mathcal{R}_\theta$ est compact. Bien que fermé (pour la convergence en distribution) et borné (pour la norme en variation totale), $\mathcal{R}_\theta$ n'a aucune raison d'être compact pour la topologie associée à la convergence en distribution. Par ailleurs, il existe de nombreux espaces dans lesquels cette hypothèse de compacité ne peut être relâchée dans le théorème de Krein-Milman. Nous montrons cependant que sur l'espace des distributions de probabilité, la compacité pour la convergence en distribution n'est pas nécessaire pour assurer l'existence de parties extremes des convexes fermés non

vides.

Le deuxième apport concerne l'identification des moments de P_0. Ce qui revient à se restreindre au cas où $q(\theta, P) = \theta - \int f(u)dP(u)$. La région d'identification de θ_0 est donc un intervalle d'extrémités $\inf_{P\in\mathcal{R}} \int fdP \geq 0$ et $\sup_{P\in\mathcal{R}} \int fdP \geq 0$. Or, calculer de telles bornes revient à résoudre un programme d'optimisation sur un espace de distributions $\mathcal{R}$. Nous montrons que l'optimisation sur $\mathcal{R}$ peut être remplacée par une optimisation sur un espace plus petit, à savoir $\text{ext}(\mathcal{R})$.

Nous étudions également ce qui se passe lorsque l'ensemble $\mathcal{R}$ peut s'écrire comme une intersection dénombrable d'ensembles $\cap_{n\in\mathbb{N}}\mathcal{R}_n$. Cela présente un intérêt lorsque le problème étudié est tel que $\text{ext}(\mathcal{R}_\theta)$ reste difficile à caractériser mais que les parties extrêmes de $\mathcal{R}_{\theta,n} = \{P \in \mathcal{R}_n : q(\theta, P) = 0\}$ sont faciles à caractériser pour tout n. Dans ce cas, nous explicitons des conditions suffisantes (et "presque nécessaires") sous lesquelles la région d'identification de θ_0 se déduit des régions d'identification correspondant aux problèmes où les restrictions $\mathcal{R}$ ont été remplacées par les restrictions $\mathcal{R}_n$. Reprenons notre exemple, imaginons que notre économiste observe les impôts payés des Syldaves et des Bordures interrogés (par exemple parce qu'il échantillonne les individus de son enquête dans les fichiers fiscaux). Imaginons par exemple que notre économiste est persuadé que la probabilité de répondre à l'enquête décroisse avec le revenu conditionnellement à la situation sur le marché du travail. Dans ce cas, si X représente les impôts payés par l'individu, cela signifie que $\mathbb{P}(D = 1|Y = a, X = x) \geq \mathbb{P}(D = 1|Y = a, X = x')$ pour $a \in \{0; 1\}$ et $x' \geq x$. Conformément à ce qui précède, notons $\mathcal{R}$ l'ensemble des probabilités qui vérifient une telle propriété et $\mathcal{R}_\theta$ l'ensemble des distributions dans $\mathcal{R}$ qui sont en plus compatibles avec un taux de chômage de valeur θ. Une telle restriction conduit à une region d'identification

particulièrement difficile à caractériser directement (car la caractérisation de ext($\mathcal{R}_\theta$) est compliquée). Cependant si σ est une bijection strictement croissante de $[0;1]$ dans $[0;+\infty]$, alors les *a priori* de notre économiste impliquent que pour $n \in \mathbb{N}$:

$$\forall k = 1, ..., n-1, \forall a \in \{0;1\},$$
$$\mathbb{P}\left(D=1|Y=a, X \in \left[\sigma\left(\tfrac{k-1}{n}\right); \sigma\left(\tfrac{k}{n}\right)\right[\right) \geq \mathbb{P}\left(D=1|Y=a, X \in \left[\sigma\left(\tfrac{k}{n}\right); \sigma\left(\tfrac{k+1}{n}\right)\right[\right).$$

Notons $\mathcal{R}_n$ l'ensemble des probabilités qui vérifient la restriction précédente, et $\mathcal{R}_{\theta,n}$ le sous ensemble des distributions qui sont en plus compatibles avec un taux de chômage de valeur θ. On a alors $\mathcal{R} = \cap_{n\in\mathbb{N}}\mathcal{R}_n$ et $\mathcal{R}_\theta = \cap_{n\in\mathbb{N}}\mathcal{R}_{\theta,n}$. D'autre part nous sommes capables de caractériser ext($\mathcal{R}_n$) et ext($\mathcal{R}_{\theta,n}$) et donc de caractériser les régions d'identification du paramètre dans le cas où notre économiste postule la restriction $\mathcal{R}_n$ au lieu de $\mathcal{R}$. Nous donnons des conditions précises sous lesquelles on peut alors retrouver la région d'identification du paramètre sous $\mathcal{R}$ à partir des régions d'identification sous $\mathcal{R}_n$. Nous donnons également des contre-exemples dans le cas où les conditions exhibées ne sont pas vérifiées.

Après avoir exprimé la région d'identification de θ_0 comme fonctions des parties extrêmes de $\mathcal{R}_\theta$, nous appliquons cette méthodologie pour retrouver de nombreux résultats de la littérature : identification partielle de l'effet marginal moyen dans un modèle non linéaire sur données de panel, problème des moments (classique ou ses extensions), identification partielle d'un paramètre défini par un nombre fini d'égalités ou d'inégalité de moments, dualité de Monge-Kantorovitch en transport optimal, identification partielle de paramètres structurels dans des jeux avec équilibres multiples avec ou sans possibilité de jouer en stratégies mixtes...

Enfin dans une dernière partie nous utilisons les outils développés pour résoudre un prob-

lème non traité dans la littérature et qui n'est qu'une généralisation du problème du taux de chômage en Borduro-Syldavie présenté précédemment au cas où les variables intervenant dans le problème peuvent prendre des valeurs réelles quelconques.

1.2 Chapitre 2: Méthode pour corriger une certaine forme d'attrition endogène dans les panels.

Ce chapitre a été écrit dans le cadre d'un travail en collaboration avec Xavier d'Haultfoeuille.

La non réponse dans les enquêtes conduit souvent l'économètre à travailler sur un échantillon sélectionné (i.e. non représentatif dans un langage profane). Or les grandeurs qui intéressent au final l'économiste, le sociologue, le décideur public sont définies sur la population d'intérêt et pas sur la population des répondants aux enquêtes. La non-réponse est d'autant plus problématique que le comportement de réponse des individus est lié aux variables d'intérêt. Les données de panel, en suivant les individus dans le temps, permettent d'identifier des covariations de variables pour un même individu au cours du temps. Pour de telles données, on n'effectue donc pas des comparaisons entre individus mais des comparaisons entre différents moments pour un même individu. Cela présente l'avantage de contrôler l'effet des variables constantes au cours du temps, qu'elles soient observables ou inobservables. Cependant sur ces données, il existe presque toujours un phénomène de non réponse spécifique : l'attrition, c'est à dire le fait que le suivi des individus s'interrompe. Dans ce cas, quelque chose a donc rendu impossible la collecte d'une information qui avait pu être collectée à une date précédente. Si une variable d'intérêt n'est observée qu'à

une date initiale, mais n'a pas pu l'être à une date ultérieure, c'est que des changements sont survenus, ces changements concernant le comportement de réponse peuvent être (et sont certainement !) corrélés aux changements concernant les variables d'intérêt. Or avec des données de panel, on utilise principalement les variations des variables observées aux cours du temps pour identifier les paramètres d'intérêt, si ces variations sont corrélées avec l'attrition, cela introduit une sélection de l'échantillon. Le problème de sélection dû à l'attrition est généralement considéré comme susceptible d'introduire plus de biais que la non réponse totale (i.e. à toutes les dates) car la non réponse totale peut-être rationalisée par l'effet de facteurs inobservables constants au cours du temps.

Les méthodes de correction de la sélection due à l'attrition dans les panels s'appuient généralement sur des hypothèses d'attrition ignorable : en contrôlant des variables observées à la première interrogation on suppose que l'on contrôle tous les facteurs communs susceptibles d'expliquer simultanément l'attrition et les variations futures des variables d'intérêt. Sous cette hypothèse, les variations futures des variables d'intérêt sont donc décorrelées de l'attrition. Par exemple, dans une enquête en panel pour mesurer le statut sur le marché du travail, on suppose que la probabilité de sortir du panel entre la date 1 et la date 2 dépend uniquement du statut sur le marché du travail à la date 1 mais pas du statut sur le marché du travail à la date 2. Dans ce dernier cas, il apparaît clairement qu'une telle hypothèse est peu crédible, les individus pouvant être plus ou moins enclins à répondre à l'enquête en fonction de leur situation à la date d'interrogation.

Une approche alternative (et plus originale) mise en oeuvre par Hausman et Wise en 1979 consiste à supposer au contraire que l'attrition dépend seulement des valeurs contemporaines des variables d'intérêt mais pas des valeurs passées. Pour reprendre l'exemple précé-

dent, on suppose donc que le fait de ne pas répondre à l'enquête en deuxième interrogation dépend de la situation sur le marché du travail de l'individu au moment de la deuxième interrogation mais pas de son statut au moment de la première. Cependant, il arrive que dans certains cas, l'attrition dépende simultanément des valeurs de la variable d'intérêt aux deux dates. C'est de manière assez évidente le cas pour l'enquête emploi française : la collecte des données a lieu en enquêtant de manière répétée les ménages d'un échantillon de logements. Les ménages qui déménagent ne peuvent donc être enquêtés après leur déménagement. Or il est raisonnable de penser que les transitions professionnelles sont partiellement concomitantes avec des déménagements. Dans un tel cas, la probabilité d'attrition entre deux dates d'enquête dépend mécaniquement du statut sur le marché du travail aux deux dates.

Pour prendre en compte un phénomène d'attrition aussi complexe, Hirano, Imbens, Ridder et Rubin (2001) ont généralisé les deux approches précédentes en autorisant la probabilité d'attrition à dépendre du statut sur le marché du travail aux deux dates. Cependant, ils imposent des restrictions fonctionnelles sur la manière dont la probabilité d'attrition dépend des variables d'intérêt. Ces restrictions ne sont pas soutenables dans le cas de l'enquête emploi française car elles ne sont pas compatibles avec le fait que les déménagements surviennent plus fréquemment lorsque les individus changent de statut sur le marché du travail. De plus, Hirano, Imbens, Ridder et Rubin (2001) ont besoin d'un échantillon de rafraîchissement. Dans notre exemple, cela signifie qu'ils identifient au moyen de données auxiliaires la distribution marginale des statuts sur le marché du travail en deuxième date.

Nous proposons une méthode alternative, qui repose sur d'autres hypothèses. L'avantage de notre approche est qu'elle ne nécessite pas de disposer d'un échantillon de rafraîchisse-

ment et qu'elle n'impose aucune restriction fonctionnelle sur la probabilité d'attrition. En revanche, nous devons disposer d'une variable qui est corrélée au fait de répondre en deuxième date conditionnellement au statut sur le marché du travail aux deux dates. Sous cette hypothèse, nous explicitons les conditions sous lesquelles un moment des variables d'intérêt peut être identifié. Nous proposons et dérivons le comportement asymptotique d'un estimateur dans le cas où le support des variables d'intérêt est fini, nous dérivons également des tests. Et nous appliquons notre méthode au calcul d'une matrice de transition entre emploi, chômage et inactivité à partir des données de l'enquête emploi de l'INSEE. Dans le cas où la variable d'intérêt a un support infini, nous dérivons la borne d'efficacité semi-paramétrique du moment considéré et nous exhibons une condition nécessaire pour qu'il existe un estimateur semi-paramétrique $\sqrt{n}$ convergent.

La dérivation d'un estimateur ayant de bonnes propriétés dans le cas d'une variable d'intérêt dont le support est infini reste à traiter.

1.3 Chapitre 3: Méthode pour estimer simplement les paramètres d'un modèle logistique avec dépendance d'état sur données de panels.

Ce chapitre a été écrit dans le cadre d'un travail en collaboration avec Romain Aeberhardt.

L'économétrie des panels s'est fortement développée depuis une vingtaine d'années. Les données de panels présentent de nombreux avantages mais aussi des difficultés de traitement économétrique spécifiques, notamment dans le cas de modèles non linéaires. Disposer de

données de panels permet en effet de contrôler dans les estimations un effet des variables inobservées constantes au cours du temps. Une telle entreprise peut néanmoins poser des difficultés : il faut souvent arbitrer entre possibilité d'identification et facilité d'estimation d'un coté et crédibilité de la modélisation de l'hétérogénéité inobservée de l'autre. Cet arbitrage est particulièrement crucial dans les modèles non linéaires pour lesquels une modélisation à "effets aléatoires" (c'est à dire une modélisation qui s'appuie sur de fortes hypothèses concernant la distribution de l'hétérogénéité inobservée) est assez simple à implémenter et pour lesquels une modélisation à "effets fixes" (c'est à dire une modélisation relativement agnostique concernant la distribution de l'hétérogénéité inobservée) pose des problèmes théoriques et pratiques importants.

Une autre spécificité importante des données de panels est qu'elle invite naturellement l'économiste à modéliser une dépendance d'état : c'est à dire le fait que les valeurs actuelles des variables d'intérêt puissent dépendre des valeurs passées de ces mêmes variables. Dans le cadre d'un modèle non linéaire, il devient alors assez difficile de concilier une approche agnostique concernant la distribution de l'hétérogénité inobservée constante au cours du temps et l'estimation d'un modèle avec dépendance d'état. Pour des modèles à variables qualitatives avec dépendance d'état et effets fixes, Bo Honoré et Ekaterini Kyriazidou ont proposé un estimateur basé sur des comparaisons de paires de périodes par individu. Dans le cas où on suppose en plus que l'hétérogénéité inobservée variable au cours du temps suit une distribution logistique, ils dérivent également les propriétés asymptotiques de leur estimateur dont la vitesse de convergence n'est pas en $\sqrt{n}$ (comme souvent dans le cas d'estimateurs semi-paramétriques).

L'estimateur proposé par Honoré et Kyriazidou n'est que la solution de la maximisation

d'une fonction objectif calculée sur les données. En toute généralité, il faut donc programmer la maximisation d'une telle fonction ce qui peut sembler un peu fastidieux à un économètre appliqué. D'autre part, il faut également programmer le calcul des écarts-types associés à l'estimation ce qui rend la méthode peu attractive en pratique par rapport à des modélisations concurrentes pour lesquelles des routines programmées existent déjà. Dans ce travail, nous montrons que cet estimateur présente un avantage peu mis en avant par les auteurs : il est facilement calculable (ainsi que les écart-types estimés) au moyen d'une simple régression logistique ! Il suffit simplement de réordonner les données selon une procédure que nous décrivons, de calculer des pondérations ad-hoc et d'estimer un modèle de régression logistique avec un calcul robuste des écarts-types. L'estimateur ainsi calculé aussi bien que la p-value associée à la statistique de Student sont asymptotiquement sans biais. A distance finie, nous montrons également au moyen d'une simulation Monte-Carlo, que la prise en compte d'une possible dépendance intra-individuelle des observations permet d'obtenir des tests ayant de meilleures propriétés à distance finie que ceux proposés originellement par Honoré et Kyriazidou.

1.4 Chapitre 4: Evaluation de la politique des Réseaux Ambition Réussite (RAR) par régression par discontinuité.

Ce chapitre a été écrit dans le cadre d'un travail en collaboration avec Magali Beffy.

Ce travail cherche à évaluer la politique des Réseaux Ambition Réussite mis en place à la rentrée de septembre 2006. Cette politique consiste à sélectionner un petit nombre de

collèges pour lesquels des moyens spécifiques sont mobilisés. Cette réforme de la politique d'éducation prioritaire était motivée à l'époque par la volonté de recentrer la politique sur un petit nombre d'établissements et d'éviter de saupoudrer les moyens mobilisés. Les politiques publiques lorsqu'elles sont très ciblées sont par nature difficiles à évaluer. Les établissements sont sélectionnés par la politique parce qu'ils présentent des caractéristiques particulières, comparer naïvement les établissements sélectionnés et les établissements non sélectionnés pour évaluer l'effet de la politique semble alors une mauvaise idée.

Une partie de la littérature empirique s'est attaquée à l'évaluation des politiques d'éducation prioritaire en utilisant la méthode de "différences de différences". C'est en particulier la stratégie suivie par Benabou, Kramarz et Prost. Dans cette méthode on fait l'hypothèse que les établissements sélectionnées peuvent avoir des caractéristiques particulières mais que la différence entre établissements sélectionnés et établissements non sélectionnés serait restée stable dans le temps si la politique n'avait pas été mise en place. Ainsi, si les différences entre établissements sélectionnés et non sélectionnés évoluent après la mise en place de la politique, on peut attribuer ces évolutions à la politique. Dans ces méthodes d'estimation tout repose sur une hypothèse de différence stable dans le temps. Or il existe de nombreuses raisons de penser que les différences entre établissements scolaires ne sont pas stables dans le temps. Par exemple, certains établissements font face à des difficultés croissantes, parce que certains quartiers s'appauvrissent alors que d'autres voient leur missions facilitées par un embourgeoisement local. Si la politique sélectionne des établissements faisant face à des difficultés croissantes, alors la méthode des "différences de différences" ne fournie qu'une estimation biaisée de l'effet de la politique publique. Un autre problème avec les "différences de différences" peut survenir si la sélection des étab-

lissements se fait sur la base des valeurs passées des variables d'intérêt. Imaginons par exemple que l'on décide de mettre en place une politique visant à réduire les inégalités de réussite au brevet des collèges entre établissements. Imaginons que les établissements sélectionnés dans le traitement soient ceux qui aient eu les moins bons résultats au brevet des collèges. Même si la politique n'a aucun effet, l'année suivante la différence entre établissements traités et non-traités au brevet des collèges se réduira par un phénomène de regression vers la moyenne (phénomène qui traduit simplement le fait que les plus mauvais établissements une année donnée ne seront pas systématiquement aussi mal classés l'année suivante). Dans ce cas, la méthode de différences de différences conduit à une évaluation trop optimiste des effets de la politique.

Dans ce travail, nous proposons une évaluation de la politique des RAR qui repose sur une autre méthode d'estimation et d'identification: la regression par discontinuité. La sélection des établissements en RAR s'est faite en partie sur la base de la proportion d'enfants issus de catégories sociales défavorisées et de la proportion d'élèves en retard de deux ans à l'entrée en 6ème en 2004. Nous mettons en évidence l'existence d'une forte discontinuité dans la probabilité d'affectation des établissements dans le dispositif (RAR) selon ces deux critères. La probabilité estimée pour un établissement d'être classé en RAR augmente fortement au delà du seuil de 10 % d'élèves en retard de deux ans dans l'établissement et au delà du seuil de 67 % d'élèves issus de CSP défavorisées dans l'établissement. L'exploitation de ces deux discontinuités peut permettre d'identifier l'effet pour les établissements d'être classé en RAR. A partir de différentes données administratives de l'Education Nationale, nous construisons un panel de collèges et nous comparons des collèges proches des seuils de discontinuité mis en évidence. L'hypothèse identifiante est la suivante : certains collèges

ne sont pas classés en RAR car ils se situent juste en dessous des seuils mais ils l'auraient été s'ils s'étaient situés juste au dessus de ces seuils. De même certains collèges sont classés en RAR car ils se situent juste au dessus des seuils mais ne l'auraient pas été s'ils s'étaient situés juste en dessous. Intuitivement, les seuils produisent donc une quasi-expérience aléatoire. Une telle méthode offre donc l'avantage de faire des hypothèses assez faibles concernant les mécanismes de sélection dans le traitement. L'inconvénient est que les estimations ne peuvent pas s'extrapoler aux établissements loin des seuils de discontinuité sans hypothèse forte sur l'homogénéité des effets.

Les résultats de ce travail suggèrent que l'effet de la politique est hétérogène (en général, les valeurs obtenues diffèrent selon la discontinuité mobilisée). Cependant, certains faits stylisés se dégagent de l'analyse : tout d'abord, il semble que les annonces ministérielles n'aient pas été entièrement respectées pour les établissements près des seuils de sélection. Il nous est difficile de retrouver dans les données les enseignants supplémentaires promis pour les collèges traités juste aux dessus des seuils par rapport au collèges non traités juste en dessous. Deuxièmement, la réforme a changé la structure des enseignants affectés dans les établissements traités : nous observons une augmentation de la proportion de professeurs âgés et de professeurs n'ayant pas les qualifications habituelles en collège (agrégation, certification,...). Une explication probable de ce phénomène est que des instituteurs ont été affectés dans les collèges pour améliorer la transition "école primaire-collège". Troisièmement, la réforme a accru les phénomènes de ségrégations scolaires entre établissements lorsque celle-ci est mesurée par la profession des parents. Quatrièmement nous observons une augmentation des écarts de résultats au brevet des collèges entre établissements traités et non traités. Cela peut s'expliquer par une augmentation de la ségrégation scolaire des

élèves entre établissements ou par des effets contre-productifs de la politique sur les élèves scolarisés dans les établissements en RAR.

Au final, le bilan de cette politique apparaît assez sombre, même si des études supplémentaires sont encore nécessaires pour comprendre pourquoi les résultats des élèves au brevet des collèges diminuent dans les établissements traités par rapport aux établissements non traités. Cette dernière question ne peut pas se traiter avec les seuls données mobilisées dans ce travail.

Chapter 2

Partial identification with Missing Data

2.1 Introduction

In this chapter, 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., 2011, Bontemps et al., 2012), panel data models (see Honore & Tamer, 2006, Chernozhukov et al., 2012, 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 contexte, 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 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. (2012) and D'Haultfoeuille & Rathelot (2011) 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 re-

[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).

strictions. 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.

2.2 Problem and general results

2.2.1 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]

$$\Theta_0 = \mathrm{cl}\left(\{\theta \in \Theta : \exists P \in \mathcal{R} : q(\theta, P) = 0\}\right), \tag{2.2.1}$$

[2]Each time we write $q(\theta, P)$, we implicitly assume that (θ, P) belongs to $\mathcal{Q}$. Hence, (2.2.1) should be understood as $\Theta_0 = \mathrm{cl}\left(\{\theta \in \Theta : \exists P \in \mathcal{R} : (\theta, P) \in \mathcal{Q} \text{ and } q(\theta, P) = 0\}\right)$. $(\theta, P) \in \mathcal{Q}$ 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 $\int |m(u, \theta)| dP(u) < \infty$.

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* $\theta \in \Theta$.

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, $\mathcal{R}$ is convex. Because $P \mapsto \int f dP$ is linear, Θ_0 is then an interval of $\mathbb{R}$, $\Theta_0 = [\underline{\theta}, \overline{\theta}]$. We are thus reduced to compute

$$\underline{\theta} = \inf_{P \in \mathcal{R} \cap \mathcal{I}(f)} h(P) \quad \text{and} \quad \overline{\theta} = \sup_{P \in \mathcal{R} \cap \mathcal{I}(f)} h(P). \tag{2.2.2}$$

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 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 $\mathbb{R}^p$ with $p = dim(Im(f))$. 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 concernes 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 $\mathbb{R}^p$.

2.2.2 Examples

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 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*a*)). More generally we can focus on the identification of the full jointe 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.

[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 $P^{Y|X_1,X_2} = \lambda P_1^{Y|X_1,X_2} + (1-\lambda) P_2^{Y|X_1,X_2} = \lambda P_1^{Y|X_2} + (1-\lambda) P_2^{Y|X_2} = P^{Y|X_2}$ ($P_1^{X_1,X_2} = P_2^{X_1,X_2} = P^{X_1,X_2}$ and $P_1^{X_2} = P_2^{X_2} = P^{X_2}$ imply the first and third equalities, the second equality is implied by the conditional independence restriction of P_1 and P_2). And so, P^{D,Y,X_1,X_2} satisfies the conditional independence restriction.

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 $O = ((Y_t)_{t=1\dots T}, (X_t)_{t=1\dots T})$ and $U = ((X_t)_{t=1\dots T}, \alpha)$, 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 $Y = (Y_t)_{t=1\dots T}$ conditional on $X = (X_t)_{t=1\dots T}$, α and θ_0 is known. This is the case if $Y_t = g(X_t, \alpha, \varepsilon_t, \beta_0)$ where the $(\varepsilon_t)_{t=1\dots T}$ 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

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

Models with multiple equilibria

In this example, we consider a simple entry game with two players originally studied in Tamer (2003) and after that in 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 2.1: Payoffs of entry game

The payoffs shifters ε_1 and ε_2 are observed by players but not by econometrician. θ_1 and θ_2 are unknown non positive parameters. When ε_1 (respectively ε_2) is greater than $-\theta_1$ (respectively $-\theta_2$), the player 1 (respectively the player 2) always enters. On the other side, when ε_1 (respectively ε_2) is lower than 0, the player 1 (respectively the player 2) never enters. 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 $-\varepsilon_1/\theta_1$). Let Y_1 (respectively Y_2) the dummy variable coding the entry of player 1 (respectively of player 2). Econometrician observes a large number of independent realizations of $(Y_1, Y_2) \in \{0; 1\}^2$, and then knows the quantity $\mathbb{P}((Y_1, Y_2) = (y_1, y_2))$ for $(y_1, y_2) \in \{0; 1\}^2$. She/he also knows the functional form of the payoffs but does not know the realization of $(\varepsilon_1, \varepsilon_2)$ nor the value of θ_1 and θ_2. Econometrician assumes that the distribution of $(\varepsilon_1, \varepsilon_2)$ belongs to a set of probability $\mathcal{E}$, for instance the normale bivariate distribution with $\mathbb{E}(\varepsilon_1, \varepsilon_2) = (\alpha_1, \alpha_2)$, $\mathbb{C}ov(\varepsilon_1, \varepsilon_2) = \rho$ and $\mathbb{V}(\varepsilon_1) = \mathbb{V}(\varepsilon_2) = 1$. In this example P_0 is the set of distribution of $(Y_1, Y_2, \varepsilon_1, \varepsilon_2)$, $\theta_0 = (\alpha_1, \alpha_2, \rho, \theta_1, \theta_2)$.

For given shifters $(\varepsilon_1, \varepsilon_2) \in [0; -\theta_1] \times [0; -\theta_2]$, let $w_{10}(\varepsilon_1, \varepsilon_2)$ the probability that a couple of players chooses the pure strategy such that 1 enters and 2 does not. Similarly, let $w_{01}(\varepsilon_1, \varepsilon_2)$ the probability that a couple of players chooses the pure strategy such that 2 enters and 1 does not, and $w_m(\varepsilon_1, \varepsilon_2)$ the probability that a couple of players chooses the mixed strategy. Let

$$
\begin{aligned}
&S_{00} =]-\infty; 0] \times]-\infty; 0], \\
&S_{11} = [-\theta_1; +\infty[\times [-\theta_2; +\infty[, \\
&S_{01} =]-\infty; 0[\times [0; +\infty[\cup]-\infty; -\theta_1] \times [-\theta_2; +\infty[, \\
&S_{10} = [0; +\infty[\times]-\infty; 0[\cup [-\theta_1; +\infty[\times]-\infty; -\theta_2] \\
&\text{and } S_{..} = [0, -\theta_1] \times [0, -\theta_2].
\end{aligned}
$$

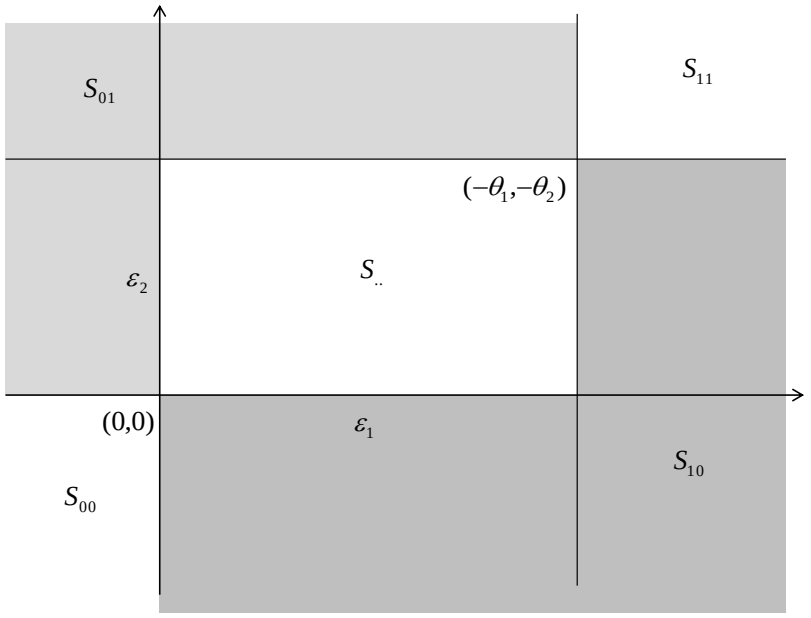


Figure 2.2: Partition of Type Spaces of Entry Game

There is not only one way to define $\mathcal{R}$ and q for a such example, we only give here one possibility. For instance, we can define $q(\theta, P) = \mathbb{1}_{\{C \text{ is false}\}}$ where C is equivalent to the following condition:

$\exists w_{01}, w_{10}, w_m$ measurable functions from $[0, -\theta_1] \times [0, -\theta_2]$ into $[0, 1]$ such that:

$$
\begin{aligned}
&w_{01} + w_{10} + w_m = 1, w_{01} \geq 0, w_{10} \geq 0 \text{ and } w_m \geq 0 \\
&\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{11}\}} + \frac{\varepsilon_1\varepsilon_2}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} - \mathbb{1}_{\{(y_1,y_2)=(1,1)\}} \right) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = 0 \\
&\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{00}\}} + \frac{(\theta_1+\varepsilon_1)(\theta_2+\varepsilon_2)}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} - \mathbb{1}_{\{(y_1,y_2)=(0,0)\}} \right) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = 0 \\
&\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{10}\}} + \left[w_{10}(\varepsilon_1,\varepsilon_2) - \frac{\varepsilon_1(\theta_2+\varepsilon_2)}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \right] \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} - \mathbb{1}_{\{(y_1,y_2)=(1,0)\}} \right) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = 0 \\
&\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{01}\}} + \left[w_{01}(\varepsilon_1,\varepsilon_2) - \frac{(\theta_1+\varepsilon_1)\varepsilon_2}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \right] \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} - \mathbb{1}_{\{(y_1,y_2)=(0,1)\}} \right) dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = 0 \\
&\forall (u_1, u_2) \in \mathbb{R}^2, \int \mathbb{1}\{\varepsilon_1 \leq u_1; \varepsilon_2 \leq u_2\} dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = \Phi_2(u_1 - \alpha_1, u_2 - \alpha_2, \rho)
\end{aligned}
$$

And with this definition of q, $\mathcal{R}$ corresponds to the set of probability distribution concen-

trated on $\{0;1\}^2 \times \mathbb{R}^2$ such that:

$$\forall (y_1, y_2) \in \{0;1\}^2, \quad \int \mathbb{1}_{\{(Y_1,Y_2)=(y_1,y_2)\}} dP(y_1, y_2, \varepsilon_1, \varepsilon_2) = \mathbb{P}((Y_1, Y_2) = (y_1, y_2)).$$

Remember that the right hand side of the previous equation is identified by the data.

2.2.3 Main theoretical results

Our main result, Theorem 2.2.1 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 $\text{ext}(\overline{\mathcal{R}_\theta})$ denote the set of extreme points of the closure for weak convergence of $\mathcal{R}_\theta$, and $\text{ext}(\mathcal{R})$ denote the set of extreme points of $\mathcal{R}$.

Theorem 2.2.1 (Main result)

1. *Under Assumptions 1 and 2,*

$$\Theta_0 = \{\theta \in \Theta : ext(\overline{\mathcal{R}_\theta}) \neq \emptyset\}.$$

2. *Moreover if Assumption 3 also holds, then:*

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

This theorem shows existence of extreme points of $\mathcal{R}_\theta$ (when $\mathcal{R}_\theta$ is not empty) and that identification region is completely characterized by $\text{ext}(\overline{\mathcal{R}_\theta})$. For a linear parameter, the unobserved distribution that achieved the bounds of the identification region can be obtained by considering only extreme distributions of $\mathcal{R}$. Our main result is particularly helpful if $\text{ext}(\overline{\mathcal{R}_\theta})$ (respectively $\text{ext}(\mathcal{R})$) is easily characterizable and finite-dimensional, because optimization on $\text{ext}(\overline{\mathcal{R}_\theta})$ is then doable in practice. We provide an important class of such examples in Subsection 2.3 below. Theorem 2.2.1 is also useful if $\text{ext}(\overline{\mathcal{R}_\theta})$ is complicated but there exists a simpler set A such that $\text{ext}(\overline{\mathcal{R}_\theta}) \subset A \subset \mathcal{R}_\theta$.

Theorem 2.2.1.2 is well-known when $\mathcal{R}$ is finite-dimensional. This can occurs in parametric models or when $\mathcal{R} = \mathcal{P}$, with $\mathcal{S}$ finite[4]. Let us recall the usual argument in this case. First,

[4]Remember that $\mathcal{P}$ is the set of probability distribution on $\mathcal{S}$ with $\mathcal{S}$ that does not depend on θ.

we can assume without loss of generality, that the support of distributions is included in $\{1, ..., I\}$. In this case any $P \in \mathcal{R}$ is characterized by $P(\{i\})$ for $i = 1, ..., I$. We can therefore assimilate $\overline{\mathcal{R}_\theta}$ 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, $\overline{\mathcal{R}_\theta}$ is a finite-dimensional, compact and convex set, $\text{ext}(\overline{\mathcal{R}_\theta})$ 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, similar results 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),

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

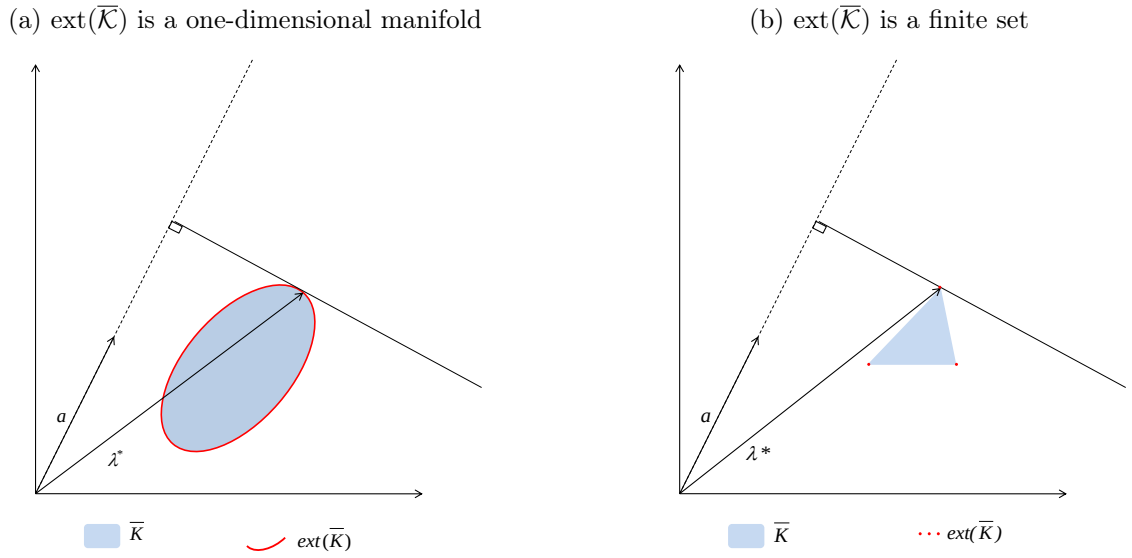
where $\text{co}(A)$ denotes the convex hull of a set A. As a result, any $\lambda \in \mathcal{R}$ can be written as $\lambda = \sum_{k=1}^{K} \alpha_k \lambda_k$, with $\lambda_k \in \text{ext}(\mathcal{R})$, $\alpha_k \geq 0$ and $\sum_{i=k}^{K} \alpha_k = 1$. This implies that $a'\lambda \leq \max_{k=1...K}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

$$\max_{\lambda \in \mathcal{R}} a'\lambda = \max_{\lambda \in \text{ext}(\mathcal{R})} a'\lambda. \tag{2.2.3}$$

Figures 2.3a and 2.3b display two examples of extremal sets of a compact convex set, illustrate Minkowski Theorem and Equality (2.2.3). 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.[5]

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

Figure 2.3: Linear optimization on compact convex sets of $\mathbb{R}^2$



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

Extending the results to the case where $\mathcal{R}$ is infinite dimensional space, on the other hand, is far from straightforward. 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, $\mathcal{R}_\theta$ is only bounded here, and so $\overline{\mathcal{R}}_\theta$ is not necessarily compact. Indeed, closed and bounded sets need not be compact in infinite dimension. 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 has extreme points, such convex is not equal to the closure of the convex hull of his 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 ext($\mathcal{K}$) is empty □

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

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

Thus, cl (co(ext($\mathcal{K}$))) is the set of constant functions from $[-1;1]$ to itself, and cl (co(ext($\mathcal{K}$))) $\neq \mathcal{K}$. It follows that optimization of linear forms on $\mathcal{K}$ does not reduce to the optimization on 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 ensures 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

$$\sup_{P\in \text{cl}(\text{co}(\text{ext}(\mathcal{R})))\cap\mathcal{I}(f)} \int f dP = \sup_{P\in \text{co}(\text{ext}(\mathcal{R}))\cap\mathcal{I}(f)} \int f dP. \tag{2.2.4}$$

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 ext(A). Using this integral representation, we are able to show directly Theorem 2.2.1, without having to prove (2.2.4).

2.2.4 Converging outer bounds

It may happen that the set ext ($\mathcal{R}$) is difficult to characterize or is 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 $\text{ext}(\overline{\mathcal{R}}_{\theta,n})$ in such case. We state the following Assumption which mimics Assumptions 2 and 3.

Assumption 4 (Intersection of decreasing convex sets)

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

2. *$\mathcal{R}_n$ 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 $q(\theta, P) = \theta - \int f dP$.*

When $\theta \mapsto \text{ext}(\overline{\mathcal{R}}_\theta)$ 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 $\Theta_{0n} = \{\theta : \overline{\mathcal{R}}_{\theta,n} \neq \emptyset\}$ 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 2.2.2 (Converging outer regions)
Under Assumptions 1 and 4.1, if for every θ it exists $\varepsilon > 0$ and $n_0 \in \mathbb{N}$ such that:

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

Then $\Theta_0 = \cap_{n\in\mathbb{N}}\Theta_{0n}$ 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 Assumption 4.2, 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 2.2.3 (Converging outer bounds for linear parameter)
Under Assumptions 1 and 4.2, if
(i) f is continuous and
(ii) 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 $\underline{\theta}_n \to \underline{\theta}$ and $\overline{\theta}_n \to \overline{\theta}$.

If we are interested only by the result on the upper bounds of θ, only the lower semi-continuity of f is needed. And for the lower bound of θ, we only need the upper semi-continuity.

We now show that Conditions (i) and (ii) in the previous corollary cannot be weakened easily. First, the result does not hold in general if f is not continuous, as the following counterexample shows.

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 condition (ii) of Corollary 2.2.3 hold. 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, $\sup_{P\in\mathcal{R}_n} \int f dP \geq 3/4$.

We also prove with the following counterexample that the result does not hold in general if Condition (ii) of Corollary 2.2.3 is not satisfied.

Counterexample 4: restriction on $\mathcal{R}_n$.
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 Assumptions 4.1 and Condition (i) of Corollary 2.2.3 hold but not Condition (ii). Finally, $\overline{\theta}_n = +\infty > \overline{\theta} = 0$.

2.3 Application to problems with moment equalities

Extremal points are the keystone of our strategy to characterize Θ_0. So an important point is to have some useful results to characterize the extreme parts. A huge literature have focus 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 the inequalities of moments and when we consider closure for the weak convergence of distributions.

Theorem 2.3.1 (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 $P \in ext(\overline{\mathcal{K}})$ only if $span(\mathcal{G}, 1)$ is dense in $L_1(P)$.
If $\mathcal{L}$ is not empty, $P \in ext(\mathcal{L})$ (respectively $P \in ext(\overline{\mathcal{L}})$) only if $span(\mathcal{G}, 1)$ is dense in $L_1(P)$.

2.3.1 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 finit 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 2.3.2 *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

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

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 $\mathcal{R}_n = \{P \in \mathcal{P} : \int |g_k| < \infty$ and $\int g_k dP = 0$ for $k \leq n\}$.

Example 1: Average effects in nonlinear panel data models.

Chernozhukov et al. (2012) 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. (2012) 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 2.3.2 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 (2011) consider the issue of measuring segregation on small units of size K, such as small firms or classrooms.
For each unit i $(i = 1, ..., N)$, there exists a theoretical probability p_i that an individual belongs to the minority. If all the p_i are the same between firms or classrooms there is no segregation. So segregation is measured by an inequality index, such as the Duncan or Theil indices, on the distribution of p. However, because firms or classrooms are small one only observes a proportion $\widehat{p}_i$ that is measured with error on p_i. The distribution of $\widehat{p}$ is over-dispersed compared to the one of p. In such a model, only the K first moments of p are identifiable (see D'Haultfoeuille & Rathelot, 2011). Theorem 2.3.2 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.

Example 3: models with multiple equilibria.
Following Tamer (2003), consider the model presented in section 2.2.2. First, note that $\mathcal{R}_\theta$ is closed[6]. $\theta \in \Theta_0$ if and only if for the corresponding $(\alpha_1, \alpha_2, \rho)$ it exists w_{01}, w_{10} and w_m with values in $[0, 1]$ such that following moments conditions holds:

$$\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{00}\}} + \frac{(\theta_1+\varepsilon_1)(\theta_2+\varepsilon_2)}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} \right) dF(\varepsilon_1,\varepsilon_2) = \mathbb{P}((Y_1,Y_2)=(0,0))$$
$$\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{01}\}} + \left[w_{01}(\varepsilon_1,\varepsilon_2) - \frac{(\theta_1+\varepsilon_1)\varepsilon_2}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \right] \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} \right) dF(\varepsilon_1,\varepsilon_2) = \mathbb{P}((Y_1,Y_2)=(0,1))$$
$$\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{10}\}} + \left[w_{10}(\varepsilon_1,\varepsilon_2) - \frac{\varepsilon_1(\theta_2+\varepsilon_2)}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \right] \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} \right) dF(\varepsilon_1,\varepsilon_2) = \mathbb{P}((Y_1,Y_2)=(1,0))$$
$$\int \left(\mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{11}\}} + \frac{\varepsilon_1\varepsilon_2}{\theta_1\theta_2} w_m(\varepsilon_1,\varepsilon_2) \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} \right) dF(\varepsilon_1,\varepsilon_2) = \mathbb{P}((Y_1,Y_2)=(1,1))$$

With F the cumulative function of a bivariate normale with expectation (α_1, α_2) and correlation ρ and where the left hand side of the previous equations are identified in the data. The extreme parts of $\overline{\mathcal{R}}_\theta$ correspond to the case where[7] w_{01}, w_{10} and w_m takes their values in $\{0; 1\}$. Let $u = (u_{00}, u_{01}, u_{10}, u_{11})$ in the unit sphere $\mathbb{S}_4$ of $\mathbb{R}^4$, let π the vector of $\mathbb{R}^4$ obtained by concatenation of the right hand side of previous equations and let $m(.,.)$

[6]Indeed, $\mathcal{R}_\theta$ is characterized by moment (in)equalities of continuous and bounded functions on $\{0;1\}^2 \times \mathbb{R}^2$: $\int f(\varepsilon_1,\varepsilon_2) dP(y_1,y_2,\varepsilon_1,\varepsilon_2) = \int f(\varepsilon_1,\varepsilon_2) dF(\varepsilon_1,\varepsilon_2)$ for all f continuous and bounded,
$\int \left(\mathbb{1}_{\{\{(0,1);(1,0)\}\cap A\neq\emptyset\}} \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{..}\}} + \sum_{(d_1,d_2)\in A} \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{d_1 d_2}\}} \right) dF(\varepsilon_1,\varepsilon_2)$
$\geq \int \mathbb{1}_{\{(y_1,y_2)\in A\}} dP(y_1,y_2,\varepsilon_1,\varepsilon_2) \geq \int \sum_{(d_1,d_2)\in A} \mathbb{1}_{\{(\varepsilon_1,\varepsilon_2)\in S_{d_1 d_2}\}} dF(\varepsilon_1,\varepsilon_2)$
and $\int \mathbb{1}_{\{(y_1,y_2)\in A\}} dP(y_1,y_2,\varepsilon_1,\varepsilon_2) = \mathbb{P}((Y_1,Y_2)\in A)$ for every $A \subset \{(0,0);(0,1);(1,0);(1,1)\}$.

[7]If not, it exists a measurable set $A \in \mathbb{R}^2$ such that $\mathbb{P}((\varepsilon_1,\varepsilon_2)\in A) > 0$ and it exists $(i,j) \in \{01;10;m\}$ such that $(w_i, w_j) \in]0,1[^2$, considers $w_i^1 = w_i + \delta$, $w_j^1 = w_j - \delta$, $w_i^2 = w_i - \delta$, $w_j^2 = w_j + \delta$ with δ a positive function on A bounded by $min(w_i, w_j, 1-w_i, 1-w_j)$. Because $(w_i, w_j) = 1/2[(w_i^1, w_j^1) + (w_i^2, w_j^2)]$, the corresponding distribution can not be an extreme point of $\overline{\mathcal{R}}_\theta$.

the function from $\mathbb{R}^2$ to $\mathbb{R}^4$ obtained by concatenation of the integrands of the left hand side of the previous equations. So $\theta \in \Theta_0$, if and only if:

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

or equivalently, using the fact that $\mathcal{R}_\theta$ is closed and Theorem 2.2.1.2:

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

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

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

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 of random sets theory.

2.3.2 Optimal transportation problem

Theorem 2.3.1 associated with our main result allows to recover the Monge-Kantorovitch duality used in optimal transportation with other technics of proofs than these used in Villani (2003) or in Villani (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 2.2.1 ensures that maximization can be done only on $\text{ext}(\mathcal{R})$ instead of $\mathcal{R}$ and Theorem 2.3.1 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 2.3.3 (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 identify identification regions in various models.

2.4 Application to the sample selection model

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, existence of X such that $X \perp\!\!\!\perp Y$ is often questionable. However, one can work under alternative restrictions.

Assumption 5 (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 6 (Monotonicity in X)
$x \mapsto \mathbb{E}(D|Y, X = x)$ *is increasing almost surely.*

Assumption 7 (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[8] $\mathbb{P}(D = 1|X, Y = y)$ 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 8 (Support Condition)

[8]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).

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 8.1 and 8.2, 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 8.3 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.
Note also that under Assumption 8, 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 8, $P_0^{Y|D=1}$ 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 9 (Continuity of density ratio)
$f_{Y|D=1,X=x_j}/f_{Y|D=1,X=x_i}$ *is continuous for every* $(x_i,x_j)\in\mathcal{X}^2$.

If $\mathcal{K}$ denotes the set of distributions of (D,Y,X) compatible with the data and assumptions made by econometrician and $\mathcal{C}$ denotes the set of distributions of
$(Y|D=0,X=x_1,Y|D=0,X=x_2,...,Y|D=0,X=x_J)$, Assumption 9 is a simple sufficient condition to ensure that $\mathcal{K}$ and $\mathcal{C}$ are closed for the weak convergence.
Note that distributions of (D,Y,X) is in one to one affine mapping with distributions of $(Y|D=0,X=x_1,Y|D=0,X=x_2,...,Y|D=0,X=x_J)$:

$$\begin{array}{l}\forall P\in\mathcal{K},\exists(P_1,...,P_J)\in\mathcal{C},\\ \int f(d,y,x)dP(d,y,x) \quad = \quad \mathbb{P}(D=1)\int f(1,y,x)dP_0(1,y,x)\\ \qquad\qquad\qquad\qquad +\sum_j(1-p(x_j))\mathbb{P}(X=x_j)\int f(0,y,x_j)dP_j(y)\end{array}$$

So, to characterize $\mathcal{K}$ and ext$(\mathcal{K})$ we can characterize $\mathcal{C}$ and ext$(\mathcal{C})$.
We first characterize ext$(\mathcal{C})$ under Assumption 6 and/or Assumption 7. For $(x_i,x_j)\in\mathcal{X}^2$, let $r_{x_i,x_j}=p(x_j)(1-p(x_i))f_{Y|D=1,X=x_j}/(p(x_i)(1-p(x_j))f_{Y|D=1,X=x_i})$.

Proposition 2.4.1 (Monotone Selection in X with discrete X)
Suppose that Assumptions 5, 6, 8 and 9 hold.

$$ext(\mathcal{C}) = \left\{ M_{y_1,...,y_J} \begin{pmatrix} \delta_{y_1} \\ \vdots \\ \delta_{y_J} \end{pmatrix} (y_1, ..., y_J) \in \mathcal{Y}^J : diag(M_{y_1,...,y_J}) \geq 0 \right\},$$

where $M_{y_1,...,y_J}$ is the upper triangular matrix of size J with (i,j) component $\mathbb{1}_{\{i \leq j\}} r_{x_i,x_j}(y_j) \left(1 + \sum_{p=2}^{J-j+1} \sum_{j=l_1<...<l_p \leq J} (-1)^{p-1} \prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)$, and where $diag(M_{y_1,...,y_J}) \geq 0$ should be understood as positivity of diagonal component of $M_{y_1,...,y_J}$. Assumption 6 of monotonicity is rejected if and only if there exists no $(y_1, ..., y_J) \in \mathcal{Y}^J$ such that $diag(M_{y_1,...,y_J}) \geq 0$.

The previous result can be used to derive bounds with multiple discret 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 2.4.2 (Monotone Selection in Y with discrete X)
Suppose that Assumptions 5, 7, 8 and 9 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 7 of monotonicity can not be rejected.

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

Proposition 2.4.3 (Monotone Selection in Y and X with discrete X)
Suppose that Assumptions 5, 6, 7, 8 and 9 hold.

$$ext(\mathcal{C}) \subset \left\{ \begin{array}{l} (P_1, ..., P_J) : \exists(\alpha_1, ..., \alpha_J) \in \mathbb{R}^{+J}, \exists(y_{11}, ..., y_{JJ}) \in \mathcal{Y}^{J^2} \\ \forall j \in J, P_j = \sum_{i=1}^{J} \alpha_i \frac{p(x_j)}{1-p(x_j)} P_0^{Y|D=1, Y \leq y_{ij}} \\ \sum_{i=1}^{J} \alpha_i \mathbb{1}_{\{y \leq y_{i,j+1}\}} \leq \sum_{i=1}^{J} \alpha_i \mathbb{1}_{\{y \leq y_{i,j}\}} \end{array} \right\}.$$

Assumptions 6 and 7 of monotonicity are jointly rejected if and only if the set in the right hand side is empty.

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 2.2.2, with $q(\theta, P) = \theta - P$ and with $\mathcal{R}_n$ derived by Assumption of monotonicity for discretized variables.

2.5 Proofs

2.5.1 Proof of Theorem 2.2.1

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 du 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 $\overline{\mathcal{R}}$ (respectively $\widetilde{\mathcal{R}}$) the τ-closure (respectively the $\tau^\star$-closure) of $\mathcal{R}$:

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

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

Similarly, let $\overline{\mathcal{R}}_\theta$ (respectively $\widetilde{\mathcal{R}}_\theta$) 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):

$$\begin{array}{ccccccc}
\mathcal{R}_\theta & \subset & \overline{\mathcal{R}}_\theta & \subset & \widetilde{\mathcal{R}}_\theta & & \\
\cap & & \cap & & \cap & & \\
\mathcal{R} & \subset & \overline{\mathcal{R}} & \subset & \widetilde{\mathcal{R}} & & \\
 & & \cap & & \cap & & \\
\mathcal{I}(f) & \subset & \mathcal{P} & \subset & \mathcal{B} & \subset & \mathcal{M}.
\end{array}$$

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 $\widetilde{\mathcal{R}}_\theta$ 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[9] probability measure μ_P on $\widetilde{\mathcal{R}}_\theta$ supported by $\text{ext}(\widetilde{\mathcal{R}}_\theta)$ such that

$$P = \int_{\text{ext}(\widetilde{R}_\theta)} R d\mu_P(R)$$

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

$$\int_{\mathcal{S}} f dP = \int_{\text{ext}(\widetilde{R}_\theta)} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R). \tag{2.5.1}$$

4. To prove the first part of the Theorem, we need to extend the Equality 2.5.1 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 2.5.1 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 2.5.1 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

[9]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.

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 2.5.1 holds for every continuous and bounded f. By dominated convergence theorem, Borel-measurability of $R \mapsto \int f dR$ and Equality 2.5.1 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 $R \in \widetilde{\mathcal{R}}_\theta$. 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 2.5.1 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 2.5.1 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 $\text{ext}(\widetilde{\mathcal{R}}_\theta)$ by $\text{ext}(\widetilde{\mathcal{R}}_\theta) \cap \mathcal{P}$ in Equation 2.5.1.

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

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:
$$\forall P \in \mathcal{R}, \quad \theta = \int_{\text{ext}(\widetilde{\mathcal{R}}) \cap \mathcal{P}} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R). \tag{2.5.2}$$
   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:

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

The previous Equality ensures that $\mu_P(\mathcal{I}(f)) = 1$, so we can replace $\text{ext}(\widetilde{\mathcal{R}}) \cap \mathcal{P}$ by $\text{ext}(\widetilde{\mathcal{R}}) \cap \mathcal{P} \cap \mathcal{I}(f)$ in Equation 2.5.2.

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 2.5.1 holds for every f Lebesgue-measurable and $P \in \mathcal{I}(f)$:

$$\int_{\mathcal{S}} f dP = \int_{\text{ext}(\widetilde{\mathcal{R}})\cap\mathcal{P}\cap\mathcal{I}(f)} \left(\int_{\mathcal{S}} f dR \right) d\mu_P(R). \tag{2.5.3}$$

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

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.

2.5.2 Proof of Theorem 2.2.2

We have: $\Theta_0 = \{\theta : \mathcal{R}_\theta \neq \emptyset\} = \{\theta : \cap_{n\in\mathbb{N}} \overline{\mathcal{R}}_{n,\theta} \neq \emptyset\} \subset \cap_{n\in\mathbb{N}} \Theta_{0,n}$.
To prove the reverse inclusion, consider $\theta \in \cap_{n\in\mathbb{N}} \Theta_{0,n}$. Because $\sup_{P\in\text{ext}(\mathcal{R}_{n_0,\theta})} \int ||u||^\varepsilon dP(u)$ is finite, Theorem 2.2.1.1 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 2.2.1.1 ensures that $\Theta_{0,n} = \{\theta : \text{ext}\,(\mathcal{R}_{n,\theta}) \neq \emptyset\}$.

2.5.3 Proof of Theorem 2.2.3

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 $\overline{\theta}_n \geq \overline{\theta}$.

2.5.4 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:

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

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

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

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

$\mathcal{P}_n^\varepsilon$ 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 $P_n^\varepsilon \in \mathcal{R}_n$.

2.5.5 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 4.1 holds. Condition (i) of Corollary 2.2.3 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 $\overline{\theta}_n = +\infty$.

2.5.6 Proof of Theorems 2.3.1, 2.3.2 and 2.3.3

Theorem 2.3.1

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 $Q \in \text{ext}(\overline{\mathcal{J}})$, 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 $\overline{\mathcal{J}}$ 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 $L_1(Q)$.

Theorem 2.3.2

To prove Theorem 2.3.2, note that $\mathcal{R}_\theta = \overline{\mathcal{R}}_\theta$ and apply Theorem 2.3.1.
We have $dim\,(\text{span}(g_1, ...g_k, m_1, ..., m_l, 1)) = r \leq l + k + 1$. Let $P \in \overline{\mathcal{R}}_\theta$ 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 $\overline{\mathcal{R}}_\theta$. We deduce that extreme point of $\overline{\mathcal{R}}_\theta$ are supported by at most r points in $\mathcal{S}$. This achieves the proof of Theorem 2.3.2.

Theorem 2.3.3

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 2.3.1 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.

2.5.7 Proof of Propositions 2.4.1, 2.4.2 and 2.4.3

Bounds under Assumption 6

Under Assumption 8, 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 6. 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[10] such that

[10]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 $\mathcal{C}$.

$\left(P^{Y|D=0,X=x_j}\right)_{j=1,...,J} \in \mathcal{C}$ 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 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 ν_{x_i,x_j} the endomorphism on the space of positive measure concentrated on $\mathcal{Y}$ defined by:

$$\nu_{x_i,x_j}(Q)(B) = \int_B r_{x_i,x_j}(y) dQ(y)$$

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

Note that $\nu_{x_i,x_i} = Id$, $\nu_{x_i,x_j}(\delta_y) = r_{x_i,x_j}(y)\delta_y$ and $\nu_{x_i,x_j} \circ \nu_{x_j,x_k} = \nu_{x_i,x_k}$.

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 6, $s(y, x_i, x_j)$ 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 8 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 $x_j \leq x_i$, $P_0^{Y|D=0,X=x_j}(B) = \int_B s(y, x_i, x_j) d\nu_{x_i,x_j}(P_0^{Y|D=0,X=x_i})(y)$.

Under Assumption 6, for every $i = 1, ..., J - 1$ we have

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

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 $P^{Y|X=x_i,D=0} = \nu_{x_{i+1},x_i}(P^{Y|X=x_{i+1},D=0}) + R_i$, 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 6 holds.

We deduce that:

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

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

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

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

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 9 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 $j = 2, ..., J$:

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

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

Step 3: Elements of ext($\mathcal{C}$) have finite support

Let $(P_i)_{i=1,...,J}$ an extreme point of $\mathcal{C}$ and let $R_i = P_i - \nu(P_{i+1})_{x_{i+1},x_i}$. Assume that P_J is not an extreme point in the space of probability distributions, then it exists $\lambda \in]0; 1[$, P_J^1 and P_J^2 two probability distributions such that $P_J = \lambda P_J^1 + (1-\lambda)P_J^2$. For $k = 1, 2$, let P_i^k and R_i^k defined recursively by $R_i^k = (1 - \nu_{x_{i+1},x_i}(P_{i+1}^k)(\mathcal{Y}))\frac{R_i}{R_i(\mathcal{Y})}$ and $P_i^k = \nu_{x_{i+1},x_i}(P_{i+1}^k) + R_i^k$.

Because $\nu_{x_J,x_{J-1}}(P_J) = \lambda\nu_{x_J,x_{J-1}}(P_J^1) + (1-\lambda)\nu_{x_J,x_{J-1}}(P_J^2)$, a decreasing recurrence on i shows that $P_i = \lambda P_i^1 + (1-\lambda)P_i^2$ for every $i = 1, ..., J$. It follows that P_J is a Dirac distribution. Now assume that $R_{J-1} = P_{J-1} - \nu_{x_J,x_{J-1}}(P_J)$ is not an extreme point in the space of positive measure of mass $1 - \nu_{x_J,x_{J-1}}(P_J)(\mathcal{Y})$. Then it exists $\lambda, R_{J-1}^1, R_{J-1}^2$ such that $R_{J-1} = \lambda R_{J-1}^1 + (1-\lambda)R_{J-1}^2$. For $k = 1, 2$, let $P_J^k = P_J$ and $P_i^k = \nu_{x_{i+1},x_i}(P_{i+1}^k) + R_i^k$ and $R_i^k = (1 - \nu_{x_{i+1},x_i}(P_{i+1}^k)(\mathcal{Y}))\frac{R_i}{R_i(\mathcal{Y})}$. We have $P_i = \lambda P_i^1 + (1-\lambda)P_i^2$ for every $i = 1, ..., J$. And then the support of R_{J-1} is a point of $\mathcal{Y}$. Similar reasoning shows that every for every i, $\text{Supp}(R_i)$ is reduced to a point of $\mathcal{Y}$.

It follows that it exists $(y_1, y_2, ..., y_J) \in \text{Supp}(Y|X = x_1, D = 1) \times ... \times \text{Supp}(Y|X = x_k, D = 1)$ and $(w_1, w_2, ..., w_{J-1}) \in [0; 1]^{J-1}$ such that

$$\begin{pmatrix} w_1\delta_{y_1} \\ \vdots \\ w_{J-1}\delta_{y_{J-1}} \\ \delta_{y_J} \end{pmatrix} = \begin{pmatrix} R_1 \\ \vdots \\ R_{J-1} \\ P_J \end{pmatrix}.$$

Step 4: Characterization of ext($\mathcal{C}$)
Because $(P_i)_{i=1,...,J}$ is a linear transformation of $(R_i)_{i=1,...,J}$, it exists a square matrix $M_{y_1,y_2,...y_J}$ such that:

$$M_{y_1,...,y_J}\begin{pmatrix} \delta_{y_1} \\ \vdots \\ \delta_{y_J} \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_J \end{pmatrix}.$$

Let $m_{i,j}$ the component of $M_{y_1,...,y_J}$ on the row i and column j.

We have $m_{J,j} = \mathbb{1}_{\{j=J\}}$, note that
$r_{x_i,x_j}(y_j)\left(1 + \sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right) = 1$ for $i = j = J$ (with the usual convention $\sum_{p=2}^{1} a_p = 0$).

Now for $i = 2, ..., J$ assume that $m_{i,j}$ are such that:

$$m_{i,j} = \mathbb{1}_{\{i\leq j\}} r_{x_i,x_j}(y_j)\left(1 + \sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right),$$

this means that:

$$P_i = \sum_{j=i}^{J} r_{x_i,x_j}(y_j)\left(1 + \sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\delta_{y_j}.$$

Because $P_{i-1} = \nu_{x_{i-1},x_i}(P_i) + \left(1 - \nu_{x_{i-1},x_i}(P_i)(\mathcal{Y})\right)\delta_{y_{i-1}}$, we have:

$$\begin{array}{l} P_{i-1} = \sum_{j=i}^{J} r_{x_{i-1},x_i}(y_j)\left[r_{x_i,x_j}(y_j)\left(1+\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\right]\delta_{y_j} \\ + \left(1 - \sum_{j=i}^{J} r_{x_{i-1},x_i}(y_j)\left[r_{x_i,x_j}(y_j)\left(1+\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\right]\right)\delta_{y_{i-1}}. \end{array}$$

Because $r_{x_{i-1},x_i}(y_j) r_{x_i,x_j}(y_j) = r_{x_{i-1},x_j}(y_j)$, we have:

$$\begin{array}{l} P_{i-1} = \sum_{j=i}^{J}\left[r_{x_{i-1},x_j}(y_j)\left(1+\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\right]\delta_{y_j} \\ + \left(1 - \sum_{j=i}^{J} r_{x_{i-1},x_j}(y_j) + \sum_{j=i}^{J}\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p} r_{x_{i-1},x_j}(y_j)\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\delta_{y_{i-1}}. \end{array}$$

We can rewrite the factor of $\delta_{y_{i-1}}$:

$$\begin{array}{rcl} P_{i-1} & = & \sum_{j=i}^{J}\left[r_{x_{i-1},x_j}(y_j)\left(1+\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\right]\delta_{y_j} \\ & & \left(1+\sum_{p=2}^{J-j+1}\sum_{i-1=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\delta_{y_{i-1}}. \end{array}$$

Lastly, because $r_{x_{i-1},x_j}(y_j) = 1$ for $j = i-1$, we have:

$$P_{i-1} = \sum_{j=i-1}^{J}\left[r_{x_{i-1},x_j}(y_j)\left(1+\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right)\right]\delta_{y_j} .$$

By decreasing recurrence, it follows that for every (i,j):

$$m_{i,j} = \mathbb{1}_{\{i\leq j\}} r_{x_i,x_j}(y_j)\left(1+\sum_{p=2}^{J-j+1}\sum_{j=l_1<...<l_p\leq J}(-1)^{p-1}\prod_{k=1}^{p-1} r_{x_{l_k},x_{l_{k+1}}}(y_{l_{k+1}})\right).$$

Note that $\sum_j m_{i,j} = 1$ and that $m_{i,j} = r_{x_i,x_{i+1}}(y_j) m_{i+1,j}$ for every $j > 1$. It follows that $(P_1, ...P_J)$ is a vector of probability if and only if $m_{i,i} \geq 0$ for every i.

Bounds under Assumption 7

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[11].

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

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:

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

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 a cumulative distribution function 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 7.

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 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 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 6 and 7

$\mathcal{C}$ is the intersection of distributions compatible with the data and assumptions 6 and 7. Because the sets of distribution that verify Assumption 6 or Assumption 7 are close for the weak convergence (cf. the two previous demonstration). $\mathcal{C}$ is closed as the intersection of closed sets.

We denotes by G_0 the cumulative distribution function 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 8). In this case if $F = (F_1, ..., F_J)$ is the cumulative distributions 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 polynôme 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 $\# \cup_{j=1,...,J} \{g_j(y) : y \in \mathcal{Y}, g_j(y) > 0\} \leq J$.

Chapter 3

Endogeneous Attrition in Panels

3.1 Introduction

Panel data are very useful to distinguish between state dependence and unobserved heterogeneity (see, e.g., Heckman, 2001), to analyze the dynamics of variables such as income (see, e.g., Hall & Mishkin, 1982) or spells in duration analysis (see, e.g., Lancaster, 1990). However, these advantages may be counterbalanced by attrition, which can be especially severe when units are observed over a long period of time. Besides, attrition is often considered more problematic than standard nonresponse, because the reasons of attrition are often related to the outcomes of interest, or variations in these outcomes. Several solutions have been considered in the literature to handle this issue. A first is to suppose that attrition is exogenous, i.e. depends on lagged values that are observed by the econometrician (see, e.g., Little & Rubin, 1987). This, however, rules out a dependence between attrition and current outcomes, and is thus likely to fail in many cases. A second model takes the opposite point of view by assuming attrition to depend on contemporaneous values only (see Hausman & Wise, 1979). To handle more complex attrition patterns, Hirano et al. (2001) generalize the two previous models by allowing attrition to depend both on contemporaneous and lagged values. This generalization is made possible when a refreshment sample, i.e. a sample of new units surveyed at each period, is available. Hirano et al. (2001) also impose that the probability of attrition depends on past and current outcomes through a binary model excluding any interaction between these two variables.

In this paper, we consider still another approach, based on instruments. Contrary to Hirano et al. (2001), we do not impose any functional restrictions on the probability of attrition conditional on lagged and contemporaneous values. Refreshment sample are not needed either. On the other hand, an instrument independent of attrition conditional

on past and contemporaneous outcomes is supposed to be available. A rank condition between the instrument and the contemporaneous outcome, which can be stated in terms of completeness, is also needed. Hence, the instrument is typically a lagged variable that affects the contemporaneous outcome but not directly attrition. We can use for instance past outcomes obtained from a retrospective questionnaire. We show that under a nonlinear fixed effect model, such a variable is likely to meet the nonparametric rank condition, and satisfies also the conditional independence condition if attrition only depends on transitions on the outcome.

An advantage of our method is that even if no more instruments than outcomes are available, we can test for implications of the conditional independence assumption. Another way of testing this assumption is to use refreshment samples, even though they are unnecessary in our setting. With such samples, the marginal distribution of the contemporaneous outcome is directly identified. We can then compare this distribution with the one obtained under our identifying restriction.

We also conduct inference under such an attrition process. In the case of discrete outcomes and instruments, the model is parametric and a straightforward constrained maximum likelihood estimation procedure is proposed. In the continuous case, the model is semiparametric and estimation is more involved. We show that our setting is closely related to the one of additive, nonparametric, instrumental variable models. Similarly to Severini & Tripathi (2011), we provide a necessary and sufficient condition for the semiparametric efficiency bound to be finite, and derive the bound in this case. We also adapt, under this condition, an estimator recently proposed by Santos (2011) for nonparametric, instrumental variable models.

Finally, we apply our results to study transitions on the French labor market, using the labor force survey of the French national institute of statistics (INSEE). This survey, which interviews people in the same housings during eighteen months, is one of the most important one of INSEE. An important issue however is that the survey does not follow individuals but housings. Thus, attrition is closely related to moving of individuals. We provide evidence that these movings are themselves related to transitions on the labor market in a way that violates the additive restriction considered by Hirano et al. (2001). With either the test described above or the refreshment sample, we do not reject the conditional independence assumption with past employment status used as an instrument. Our estimates confirm that attrition is highly related to transitions in the labor market. We show that this has important implications for the estimation of the probabilities of transition on the labor market.

The paper is organized as follows. In the second section, we study identification and testability under endogenous attrition, and compare our model with the existing literature. In the third section, we develop inference for both discrete or continuous outcomes. The fourth section is devoted to our application. Finally, the fifth section concludes. All proofs are gathered in the appendix.

3.2 Identification

3.2.1 The setting and main result

For simplicity, we consider a panel dataset with two dates $t = 1, 2$, and also suppose that there is no or ignorable nonresponse at date 1. We let $D = 1$ if the unit is observed at date 2, $D = 0$ otherwise. We let Y_t denote the outcome at t and $Y = (Y_1, Y_2)$. We also consider an instrument Z_1 whose role will be explained below, and let $Z = (Y_1, Z_1)$. For the sake of simplicity, we do not introduce covariates here, though the extension with covariates would be straightforward. We focus hereafter on the identification of either the joint distribution of (D, Y, Z) or on a parameter $\beta_0 = E(g(Y, Z))$. Our first assumption states the observational problem.

Assumption 10 *The distribution of* (D, DY_2, Z) *is identified.*

To satisfy this requirement, Z_1 can be observed at the first period, or at the second period if some information on nonrespondents is available at the second period. It also holds if Z_1 (together with Y) is observed only when $D = 1$, provided that the distribution of (Z_1, Y_1) is identified for instance through another dataset. Of course, to achieve full identification of the distribution of (D, Y, Z), restrictions are needed. If attrition directly depends on the outcome Y, the usual assumption of exogenous selection fails, and it may be difficult to find an instrument that affects the selection variable but not the outcome. On the other hand, a variable Z_1 related to Y but not directly to D may be available in this case. We thus assume the following:

Assumption 11 $D \perp\!\!\!\perp Z_1 | Y$.

This assumption is identical to the one considered by D'Haultfœuille (2010) in the case of endogenous selection. It was also considered by Chen (2001), Tang et al. (2003) and Ramalho & Smith (2011*b*) in a nonresponse framework. Intuitively, it states that the

attrition equation depends on Y_1 and Y_2 but not on Z_1. If Y_2 was endogenous (but always observed) in this equation, we could instrument it by Z_1 to identify the causal effect of Y_2 on D. Here the problem is actually slightly different: Y_2 is observed only when $D = 1$. The identification strategy is similar, however, as we use the instrument to recover the conditional distribution of attrition.

Let $P(Y) = \Pr(D = 1|Y)$. Because identification is based on inverse probability weighted moment conditions, we assume the following:

Assumption 12 *$P(Y) > 0$ almost surely.*

This assumption is similar to the common support condition in the treatment effects literature. It does not hold if D is a deterministic function of Y, as in simple truncation models where $D = \mathbb{1}\{g(Y) \geq y_0\}$, y_0 denoting a fixed threshold.

Before stating our main result, let us introduce some notations. For any random variable U and $p > 0$, let $L^p(U)$ (respectively $L^p(U|D = 1)$) denote the space of functions q satisfying $E(|q(U)|^p) < +\infty$ (respectively $E(|q(U)|^p|D = 1) < +\infty$). Note that $1/P \in L^1(Y|D = 1)$ because $E(1/P(Y)|D = 1) = 1/E(D)$. For any set $A \subset L^1(U|D = 1)$, let also

$$A^{\perp} = \{q \in L^1(U|D = 1) : \forall a \in A, E(|q(U)a(U)||D = 1) < \infty,\ E(q(U)a(U)|D = 1) = 0\}.$$

The following operator will be important for identification issues:

$$\begin{array}{rccl} T: & L^1(Y|D=1) & \to & L^1(Z|D=1) \\ & q & \mapsto & (z \mapsto E(q(Y)|D=1, Z=z)). \end{array} \tag{3.2.1}$$

Because Y is observed when $D = 1$, T is identified. Besides, and as indicated previously, identification hinges upon dependence conditions between Y_2 and Z, which are actually related to the null space $\text{Ker}\,(T)$ of T. Let $\mathcal{F} = \{q \in L^1(Y|D = 1) : q(Y) \geq 1 - 1/P(Y) \text{ a.s.}\}$ and for $f \in L^1(Y, Z)$,

$$\mathcal{F}_f = \left\{q \in L^1(Y|D = 1) : q(Y) \geq 1 - 1/P(Y) \text{ a.s. and } E(|q(Y)f(Y, Z)||D = 1) < \infty\right\}.$$

Finally, in the case where $g \in L^1(Y, Z)$ we denote $\beta(Y) = E[g(Y, Z)|Y]$. Our main result is the following.

Theorem 3.2.1 *If assumptions 10-12 hold, then:*

1. *The distribution of (D, Y, Z) is identified if and only if $Ker\,(T) \cap \mathcal{F} = \{0\}$.*

Moreover, if $g \in L^1(Y, Z)$,

2. *The set of identification of* β_0 *is* $\{\beta_0 + E(D)E\left[\beta(Y)h(Y)|D=1\right] : h \in Ker\,(T) \cap \mathcal{F}_g\}$.

3. β_0 *is identified if and only if* $\beta(.) \in (Ker\,(T) \cap \mathcal{F}_g)^{\perp}$.

Let us provide the intuition for the easiest result, i.e. the "if" part of the first statement. We rely on the fact that under Assumptions 11 and 12, it is sufficient to identify $P(Y)$ to recover the whole distribution of (D, Y, Z). Besides, we show that this function satisfies

$$T\left(\frac{1}{P}\right) = w, \tag{3.2.2}$$

where $w(Z) = 1/\Pr(D=1|Z)$. Because T and w are identified, P is identified if there is a unique solution in $(0, 1]$ of this equation. This uniqueness can be established if $\mathrm{Ker}\,(T) \cap \mathcal{F} = \{0\}$.

The identifying condition $\mathrm{Ker}\,(T) \cap \mathcal{F} = \{0\}$ is related to various completeness conditions considered in the literature (see, e.g., Newey & Powell, 2003, Severini & Tripathi, 2006, Blundell et al., 2007, D'Haultfœuille, 2011, Andrews, 2011 and Hu & Shiu, 2013). Our condition is intermediate between the stronger "standard" completeness condition $\mathrm{Ker}\,(T) = \{0\}$ and the bounded completeness condition $\mathrm{Ker}\,(T) \cap \mathcal{B} = \{0\}$, where $\mathcal{B}$ is the set of bounded functions. When Y and Z have a finite support (respectively by $(1, ..., I)$ and $(1, ..., J)$), this assumption is satisfied if $\mathrm{rank}(M) = I$, where M is the matrix of typical element $\Pr(Y = i|D = 1, Z = j)$ (see Newey & Powell (2003)).[1] Hence, the support of Z must be at least as rich as the one of Y $(J \geq I)$ and the dependence between the two variables must be strong enough for I linearly independent conditional distributions to exist. Because the matrix M is identified, it is straightforward to test for this condition, using for instance the determinant of MM' (see Subsection 3.1 below). When Y and Z are continuous, it is far more difficult to characterize them. Conditions have been provided by Newey & Powell (2003), D'Haultfœuille (2011), Andrews (2011) and Hu & Shiu (2013). We consider below another example, related to our panel framework, where the restriction $\mathrm{Ker}\,(T) \cap \mathcal{F} = \{0\}$ is satisfied.

The third statement of Theorem 3.2.1 shows that when we consider only one parameter rather than on the full distribution of (D, Y, Z), identification is achieved under weaker restrictions. To see this, note that $\mathcal{F}_g \subset \mathcal{F}$ and then $(\mathrm{Ker}\,(T) \cap \mathcal{F})^{\perp} \subset (\mathrm{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$.

[1] It is not equivalent to this full rank conditions because of the inequality constraints induced by $\mathcal{F}$. One can show however that both are equivalent when $P(Y) < 1$.

Thus, $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$ implies that $\beta(.) \in (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$. On the other hand, we may have $\beta(.) \in (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$ and $\text{Ker}\,(T) \cap \mathcal{F} \neq \{0\}$. This result is closely related to Lemma 2.1 of Severini & Tripathi (2011), who consider identification of linear functionals related to a nonparametric instrumental regression. Finally, the second statement of Theorem 3.2.1 describes the identification set of β_0 in general.

As an illustration of Theorem 3.2.1 with continuous outcomes, suppose that we observe at the first date a past outcome Y_0, thanks to a retrospective questionnaire. This will be the case in the application considered in Section 4. Suppose also that the outcomes satisfy the following nonlinear fixed effect model:

$$\Lambda(Y_t) = U + \varepsilon_t, \tag{3.2.3}$$

where $\Lambda(.)$ is a strictly increasing real function and $(U, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ are independent. Such a model generalizes standard linear fixed effect model $Y_t = U + \varepsilon_t$ and is close to the accelerated failure time model in duration analysis. Note that we do not introduce covariates here for simplicity, but our result can be extended to the more realistic model considered by Evdokimov (2011), namely $\Lambda(Y_t, X_t) = \psi(U, X_t) + \varepsilon_t$ with Λ strictly increasing in Y_t, provided that the covariates X_t are always observed at each period. We also suppose that attrition only depends on current outcomes and transitions:

$$D = g(Y_1, Y_2, \eta), \quad \eta \perp\!\!\!\perp (Y_0, Y_1, Y_2). \tag{3.2.4}$$

Finally, we impose the following technical restriction on U, ε_0 and ε_2. For any random variable V, we let Ψ_V denote its characteristic function.

Assumption 13 *U admits a density with respect to the Lebesgue measure, whose support is the real line. Ψ_{ε_0} vanishes only on isolated points. The distribution of ε_2 admits a continuous density f_{ε_2} with respect to the Lebesgue measure. Moreover, $f_{\varepsilon_2}(0) > 0$ and there exists $\alpha > 2$ such that $t \mapsto t^{\alpha} f_{\varepsilon_2}(t)$ is bounded. Lastly, Ψ_{ε_2} does not vanish and is infinitely often differentiable in $\mathbb{R} \backslash A$ for some finite set A.*

The assumption imposed on the characteristic function of ε_0 is very mild and satisfied by all standard distributions. The conditions on ε_2 are more restrictive but hold for many distributions such as the normal, the Student with degrees of freedom greater than one[2] and the stable distributions with characteristic exponent greater than one. The following

[2]See e.g. Mattner (1992) for a proof that the conditions on the characteristic function of Student distributions are indeed satisfied.

proposition shows that under these conditions, the model is fully identified using Y_0 as the instrument.

Proposition 3.2.2 *Let $Z = (Y_0, Y_1)$, and suppose that Assumptions 12, 13, Equations (3.2.3) and (3.2.4) hold. Then Assumption 11 holds and $Ker(T) \cap \mathcal{F} = \{0\}$. Thus, the distribution of (D, Y, Z) is identified.*

3.2.2 Partial identification and testability

Apart from point identification under various completeness conditions, our attrition model displays two interesting features. First, Assumption 11 is refutable, contrary to the ignorable attrition assumption $D \perp\!\!\!\perp Y_2 | Y_1$ discussed below. Second, we can obtain bounds on parameters of interest when the model is underidentified, i.e. when the above completeness condition fails to hold. Both are due to the fact that solutions to Equation (3.2.2) must lie in $[0, 1]$. These inequality constraints can be used both for testing and bounding parameters of interest.

To see this, consider the case where (Y, Z) has a finite support. If Y and Z take respectively I and J distinct values, then (3.2.2) can be written as a linear system of J equations with I unknown parameters and the inequality constraints:

$$\Pr(D = 0, Z = j) = \sum_{i=1}^{I} b_i \Pr(D = 1, Y = i, Z = j),\ b_i \geq 0.$$

Of course, the model is overidentified and thus testable when $I > J$, but we can also test for the inequality constraints when $I \leq J$. We derive a formal statistical test of this condition in Subsection 3.1 below. We can also partially identify parameters of interest in the underidentified case $I < J$, still using the fact that the $(b_i)_{i=1...I}$ are positive.

Finally, a stronger test of the conditional independence assumption can be derived if a refreshment sample is available, as in Hirano et al. (2001). In this case, the marginal distribution of Y_2 is identified. Then we can reject the conditional independence assumption if for all Q satisfying $T(1/Q) = w$, there exists t such that

$$E\left[\frac{D\mathbb{1}\{Y_2 \leq t\}}{Q(Y)}\right] \neq \Pr(Y_2 \leq t).$$

3.2.3 Comparison with the literature

We compare our approach with the most usual models of attrition.

Missing at random attrition

This model, which has been considered by, e.g., Rubin (1976) and Abowd et al. (1999), posits that D only depends on Y_1:

$$D \perp\!\!\!\perp Y_2 | Y_1. \tag{3.2.5}$$

Identification of the joint distribution of (Y_1, Y_2) follows directly from the fact that, letting f_{D,Y_1,Y_2} denote the density of (D, Y_1, Y_2) with respect to an appropriate measure,

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{f_{D,Y_1,Y_2}(1, y_1, y_2)}{\Pr(D = 1 | Y_1 = y_1)}.$$

Condition (3.2.5) is the equivalent, in a panel setting, of the so-called missing at random assumption (see, e.g., Little & Rubin, 1987) or the unconfoundedness assumption in the treatment effect literature (see for instance Imbens, 2004). Because it rules out any dependence between attrition and current outcomes, it is likely to fail in many cases. In a labor force survey, for instance, house moving is a common source of attrition, and is itself related to changes in employment and/or earnings.

Dependence on current values

Compared to the first, the logic of this model is the opposite, as attrition is related to current values only:

$$D \perp\!\!\!\perp Y_1 | Y_2. \tag{3.2.6}$$

This assumption has been considered by Hausman & Wise (1979) in a parametric model. This assumption takes into account nonignorable attrition, but in a special way. It rules out in particular the possibility that transitions (namely, functions of Y_1 and Y_2) explain attrition. Abstracting from the parametric restrictions of Hausman & Wise (1979), identification can be proved along the same lines as previously. It suffices to solve in g the functional equation

$$E\left[g(Y_2) | D = 1, Y_1\right] = 1/\Pr(D = 1 | Y_1).$$

Under completeness conditions similar to the one above, this equation admits a unique solution in g, namely $1/\Pr(D = 1 | Y_2 = .)$.

Standard instrumental strategy

Attrition can be considered a particular selection problem, and thus be treated using the same tools. A classical solution (see, e.g., Heckman, 1974, Angrist et al., 1996, or Heckman & Vytlacil, 2005) is to use an instrument Z that affects attrition but not directly the current outcome:

$$Y_2 \perp\!\!\!\perp Z | Y_1.$$

Such an exclusion restriction may be credible if for some exogenous reasons, some individuals were less likely to be interviewed at the second period. However, as pointed out by Manski (2003), an important drawback of this assumption is that it is not sufficient in general to point identify the distribution of Y_2. Basically, this can be achieved only if there exists some z such that the probability of attrition $\Pr(D = 1|Z = z, Y_1 = y_1)$ is equal to zero, or is arbitrarily close to zero under continuity conditions. With limited variations in this probability, the distribution of Y_2 can only be set identified.

Additive restriction on the probability of attrition

Hirano et al. (2001) propose a two-period framework that generalize both previous examples in the sense that D may depend on both Y_1 and Y_2. This generalization is possible when a refreshment sample, which allows one to identify directly the distribution of Y_2, is available. Note that because, the distribution of Y_1 is also identified from the panel at date 1, the problem reduces to recover the copula of (Y_1, Y_2). For that purpose, Hirano et al. (2001) also suppose that

$$1/\Pr(D = 1|Y_1, Y_2) = g(k_1(Y_1) + k_2(Y_2)), \tag{3.2.7}$$

where g is a known function while $k_1(.)$ and $k_2(.)$ are unknown. They show that $k_1(.)$ and $k_2(.)$ are identified by the knowledge of the marginal distributions of Y_1 and Y_2. This allows them to recover the joint distribution of (Y_1, Y_2), since, by Bayes' rule,

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1,Y_2|D=1}(y_1, y_2) \Pr(D = 1) g(k_1(y_1) + k_2(y_2)).$$

Compared to our approach, Hirano et al. (2001) do not rely on any exclusion restriction. This comes at the cost of imposing the additive restriction on $\Pr(D = 1|Y_1, Y_2)$, which may be restrictive (see below), and having a refreshment sample, which is not needed in our case.

Though the identification proof of Hirano et al. (2001) is much different from ours, the two frameworks are actually related. As shown by Bhattacharya (2008), identification in this

additive model can be directly obtained from the functional equations

$$E\left[g(k_1(Y_1)+k_2(Y_2))|D=1,Y_i\right]=1/\Pr(D=1|Y_i).$$

Thus, identification is actually achieved along similar lines as in our case, the instrument Z being equal to (Y_1, Y_2). The difference here is that only the marginal distributions of the instrument is identified. This is the reason why they have to impose Model (3.2.7) to the attrition process. Such a restriction is not innocuous. If attrition depends on transitions, then their restriction is likely fails to hold. If, as in our application, attrition occurs for individuals who move, and that moving itself occurs with a large probability when employment status changes, then $\Pr(D=1|Y_1,Y_2)$ depends on $\mathbb{1}\{Y_1=Y_2\}$. Model (3.2.7) cannot handle such attrition processes.

Attrition with unobserved heterogeneity

Finally, Sasaki (2012) proposes a very different approach where attrition at time t depends on Y_t and a constant unobserved heterogeneity term U that also affects the dynamics of Y_t. Such a model is attractive if individual fixed effects affect both the dynamics of Y_t and the decision to respond to the panel. He shows that the dynamics of Y_t, the attrition rule and the initial conditions (the joint distribution of Y_1 and U) are identified under, basically, four restrictions. First, Y_t should follow a Markov model of order 1, conditional on U. Second, attrition at date t should be independent of past outcomes, conditional on (U, Y_t). Third, both the law of dynamics and the attrition rule should be time invariant. Fourth, the number of periods of observations should be at least three, and a proxy of U, independent of other variables conditional on U, should be available. If such a proxy does not exist, the length of the panel should be at least six.

Even if attractive, his approach is more demanding than ours in terms of data, since it requires at least three periods. Besides, he also relies on exclusion restrictions, and contrary to our approach, it is not clear whether these restrictions are testable or not.

3.3 Estimation

We now turn to inference within our framework of endogenous attrition. As previously, we focus on the estimation of the distribution of (D, Y, Z), but also on the parameter $\beta_0 = E(g(Y, Z))$, which can be estimated under restrictions detailed before. We first posit an i.i.d. sample of n observations.

Assumption 14 *We observe an iid sample of size n of* (D, DY_2, Z).

We consider two cases subsequently. The first one, in line with our application, assume that the support of (Y, Z) is finite. In this setting, we derive a simple and efficient estimator and a test of the rank condition and exclusion restriction. We then turn to the continuous case, where we investigate conditions for root-n estimability of β_0, derive the semiparametric efficiency bound when it exists and propose an estimator under this condition.

3.3.1 The discrete case

We denote the support of Y_t and Z_1 by respectively $\{1, ..., I\}$ and $\{1, ..., J\}$, with $I \leq J$. In this case, the data (D, DY_2, Z) are distributed according to a multinomial distribution. To obtain asymptotically efficient estimators, we consider constrained maximum likelihood estimation hereafter.

For a fixed y, let $p_{1ij} = \Pr(D = 1, Y_2 = i, Z_1 = j|Y_1 = y)$ and $p_{0.j} = \Pr(D = 0, Z_1 = j|Y_1 = y)$ denote the probabilities corresponding to the observations, and define $p_1 = (p_{111}, ..., p_{1IJ})$, $p_{0.} = (p_{0.1}, ..., p_{0.J})$ and $p = (p_1, p_{0.})$. Note that we let the dependence in y implicit hereafter. p is the natural parameter of the statistical model here, as it fully describes the distribution of (D, DY_2, Z_1) conditional on Y_1. However, it does not directly allow us to recover the whole distribution of (D, Y_2, Z_1). This is why we also introduce $p_{0ij} = \Pr(D = 0, Y_2 = i, Z_1 = j|Y_1 = y)$, and $p_0 = (p_{011}, ..., p_{0IJ})$ as p_1. Then any parameter θ_0 of the distribution of (D, Y_2, Z_1) is a function of (p_0, p_1), and we write $\theta_0 = g(p_0, p_1)$. We thus consider here implicitly parameters that depend on the distribution of (D, Y_2, Z_1) conditional on Y_1. Unconditional parameters depend on all the different (p_0, p_1) corresponding to different values of Y_1, and on the marginal distribution of Y_1. We can estimate them similarly, using the empirical distribution of Y_1. Because Assumption 11 does not impose any restriction on the distribution of Y_1, such estimators are also asymptotically efficient.

Finally, we adopt the same notations for the constrained maximum likelihood estimator $\widehat{p}$ as for p. We let $n_{1ij} = \sum_{k:Y_{1k}=y} D_k \mathbb{1}\{Y_{2k} = i\}\mathbb{1}\{Z_{1k} = j\}$ and $n_{0.j} = \sum_{k:Y_{1k}=y}(1-D_k)\mathbb{1}\{Z_{1k} = j\}$. The following proposition shows how to compute $\widehat{p}$ and an efficient estimator of θ_0 in our attrition model.

Proposition 3.3.1 *Suppose that Assumptions 10-12 hold. Then the maximum likelihood*

estimator $\widehat{p}$ satisfies

$$(\widehat{p},\widehat{b}) = \arg \max_{(q,b)\in[0,1]^{(I+1)J}\times\mathbb{R}^I} \sum_{j=1}^{J}\left[n_{0.j}\ln q_{0.j} + \sum_{i=1}^{I} n_{1ij}\ln q_{1ij}\right]$$
$$s.t. \left| \begin{array}{ll} \sum_{j=1}^{J}\left[q_{0.j} + \sum_{i=1}^{I} q_{1ij}\right] = 1, & \\ b_i \geq 0 & i = 1,...,I, \\ \sum_{i=1}^{I} q_{1ij} b_i = q_{0.j} & j = 1,...,J. \end{array} \right. \quad (C)$$

Suppose moreover that the matrix P_1 of typical element p_{1ij} has rank I, $(P_1P_1')^{-1}P_1p_0 > 0$ (where the inequality should be understood componentwise) and g is differentiable. Then θ_0 is identifiable and $\widehat{\theta} = g(\widehat{p}_0,\widehat{p}_1)$, with $\widehat{p}_0 = (\widehat{p}_{011},...,\widehat{p}_{0IJ})$ and for all (i,j), $\widehat{p}_{0ij} = \widehat{b}_i\widehat{p}_{1ij}$, is asymptotically normal and efficient.

Proposition 3.3.1 establishes that the maximum likelihood of p can be obtained by a constrained maximization with quite simple (although nonlinear) constraints. It also shows how to compute an asymptotically efficient estimator of θ_0. The idea behind the introduction of the $(b_i)_{1\leq i\leq I}$ is that, by Bayes' rule and Assumption 11,

$$p_{0ij} = \frac{\Pr(D=0|Y_1=y,Y_2=i)}{\Pr(D=1|Y_1=y,Y_2=i)}p_{1ij},$$

and b_i represents the odds $\Pr(D=0|Y_1=y,Y_2=i)/\Pr(D=1|Y_1=y,Y_2=i)$. The inequality constraints $b_i \geq 0$ then ensure that $\Pr(D=1|Y_1=y,Y_2=i)$ is indeed a probability, while the equality constraints are a rewriting of Equation (3.2.2) in this discrete context (see the proof of Proposition 3.3.1 in the appendix).

The condition $\text{rank}(P_1) = I$ implies $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$, and is thus sufficient for the identification of θ_0 by Theorem 3.2.1. It can be easily tested in the data because under the null hypothesis that $\text{rank}(P_1) < I$, we have $\mu_0 \equiv \det(P_1P_1') = 0$. Then, letting $\widehat{\mu} = \det(\widehat{P}_1\widehat{P}_1')$, $\sqrt{n}\widehat{\mu}$ tends to a zero mean normal variable under the null by the delta method. We use this result to test for the rank condition in our application (see Section 4 below).

$\widehat{\theta}$ is asymptotically normal and efficient when $(P_1P_1')^{-1}P_1p_0 > 0$. When $(P_1P_1')^{-1}P_1p_0 = 0$, the true parameters lie at the boundary of the parameter space. $\widehat{\theta}$ is still a root-n consistent estimator in this case. However, it is not asymptotically normal anymore (see, e.g., Andrews, 1999, for a thorough study of such cases). Moreover, the standard bootstrap typically fails to be valid (see Andrews, 2000, for an illustration). Subsampling remains

valid, on the other hand. We use it in the application when the estimator is at the boundary or close to it.

Finally, as noted before, we can test Assumption 11 by two ways. The first and standard one is that the equality constraints in (C) may not hold when $J > I$, because there is no $(b_i)_{1\leq i\leq I}$ such that $\sum_{i=1}^{I} b_i p_{1ij} = p_{0.j}$. Basically, this arises when the different values of Z are not "compatible", as with the Sargan test in linear IV models. The second is that the $(b_i)_{1\leq i\leq I}$ satisfying these equality constraints must be nonnegative. This may not hold in general, even when $I = J$. To test for both conditions simultaneously, we use the same Wald statistic as the one considered by Kodde & Palm (1986). In our framework, the unconstrained model where Assumption 11 does not necessarily hold is simply the multinomial model on (D, DY_2, Z) parameterized by p, and the maximum likelihood estimator $\widehat{p}^U$ simply corresponds to the sample proportions. The constraints (C) corresponding to Assumption 11 hold if and only if there exists $b \geq 0$ (understood componentwise) such that $P_1' b = p_{0.}$. If P_1 is full rank, the latter equation has a unique solution, the least square solution $(P_1 P_1')^{-1} P_1 p_{0.}$. Therefore, if $\text{rank}(P_1) = I$, Assumption 11 is equivalent to

$$\left[P_1'(P_1P_1')^{-1}P_1 - \text{Id}\right] p_{0.} = 0, \quad (P_1P_1')^{-1}P_1 p_{0.} \geq 0, \tag{3.3.1}$$

where Id is the identity matrix. The idea, therefore, is to see whether $\left[P_1'(\widehat{P}_1^U P_1')^{-1}\widehat{P}_1^U - \text{Id}\right]\widehat{p}_{0.}^U$ is close to zero and $(\widehat{P}_1^U P_1')^{-1}\widehat{P}_1^U \widehat{p}_{0.}^U$ is positive componentwise, where $\widehat{P}_1^U$ and $\widehat{p}_{0.}^U$ are the estimators of P_1 and $p_{0.}$ obtained from $\widehat{p}^U$.

Let us rewrite the two constraints of (3.3.1) as $h_1(p) = 0$ and $h_2(p) \geq 0$, and let $h(p) = (h_1(p), h_2(p))$. Let also $\mathcal{H}_0 = \{0\}^J \times \mathbb{R}^{+I}$ denote the set of $h = (h_1, h_2)$ satisfying these constraints. Denote by Σ_{ii} (resp. Σ_{12}) the asymptotic variance of $\widehat{h}_i \equiv h_i(\widehat{p}^U)$ (resp. covariance of $h_1(\widehat{p}^U)$ and $h_2(\widehat{p}^U)$), and by Σ the asymptotic variance of $\widehat{h} \equiv h(\widehat{p}^U)$. Finally, let $\widehat{\Sigma}$ denote a consistent estimator of Σ. The test statistic W_n is then defined as

$$W_n = n \min_{h\in\mathcal{H}_0} \left(h - \widehat{h}\right)' \widehat{\Sigma}^- \left(h - \widehat{h}\right), \tag{3.3.2}$$

where $\widehat{\Sigma}^-$ denotes the Moore-Penrose inverse of $\widehat{\Sigma}$. $\widehat{\Sigma}$ is not full rank because the rank of Σ_{11} is $J - I$, while $h_1(p) \in \mathbb{R}^J$. This is logical, since we only have $J - I$ overidentifying equality constraints here. Computing W_n is straightforward as it corresponds to a quadratic programming problem.

We now indicate how to compute critical values that are asymptotically valid under the null. We do not rely on the asymptotic result of Kodde & Palm (1986) here as they only

compute the critical value corresponding to the least favorable case of the null hypothesis. Namely, they compute c such that $\sup_{h\in\mathcal{H}_0}\lim_{n\to\infty}\Pr_h(W_n \geq c) = \alpha$. This leads to a conservative test, and therefore to low power if the null hypothesis is violated. By contrast, we compute here a critical value corresponding to the most plausible DGP satisfying the null hypothesis, given the data. Therefore, our test is not conservative for a whole range of DGP satisfying the null hypothesis (see Proposition 3.3.2 below).

Let $(h_{10}, h_{20}) = h_0 = h(p_1, p_{0.})$ denote the true parameter. The asymptotic distribution of W_n depends on whether the components $(h_{20i})_{1\leq i\leq I}$ are equal to zero or not. Let $\mathcal{R}_j$ be equal to $\mathbb{R}^+$ if $h_{20j} = 0$, and to $\mathbb{R}$ otherwise. Then let

$$\mathcal{H}(h_0) = \{0\}^J \times \mathcal{R}_1 \times ... \times \mathcal{R}_I.$$

We show in the proof of Proposition 3.3.2 below that

$$\lim_{n\to\infty}\Pr(W_n > w) = \Pr\left(\min_{h\in\mathcal{H}(h_0)}(h-U)'\,\Sigma^-\,(h-U) > w\right) \tag{3.3.3}$$

where $U \sim \mathcal{N}(0,\Sigma)$. To compute the level of the test based on this asymptotic distribution, we need to estimate $\mathcal{H}(h_0)$. Following, e.g., Rosen (2008) or Andrews & Soares (2010), we consider a sequence $(c_n)_{n\in\mathbb{N}}$ such that $c_n \to \infty$ and $c_n/\sqrt{n} \to 0$. We let $\widehat{\mathcal{R}}_j$ be equal to $\mathbb{R}^+$ if $\widehat{h}_{2j} \leq c_n/\sqrt{n}$, and to $\mathbb{R}$ otherwise, and

$$\widehat{\mathcal{H}(h_0)} = \{0\}^J \times \widehat{\mathcal{R}}_1 \times ... \times \widehat{\mathcal{R}}_I.$$

Finally, let $\widehat{c}_\alpha$ satisfy

$$\widehat{c}_\alpha = \inf\left\{c > 0 : \Pr\left(\min_{h\in\widehat{\mathcal{H}(h_0)}}\left(h-\widehat{U}\right)'\widehat{\Sigma}^-\left(h-\widehat{U}\right) > c\right) \leq \alpha\right\}, \tag{3.3.4}$$

where $\widehat{U} \sim \mathcal{N}(0,\widehat{\Sigma})$. $\widehat{c}_\alpha$ or, similarly, the p-value of the test, can be obtained easily by simulations.

Proposition 3.3.2 *For any $\alpha \leq 1/2$, the test defined by the critical region $\{W_n > \widehat{c}_\alpha\}$ is consistent. Its asymptotic level is α if $J > I$ or $\mathcal{R}_i = \mathbb{R}^+$ for some $i \in \{1,...,I\}$, and 0 otherwise.*

Note that the asymptotic distribution of W_n is degenerated when $I = J$ and $\mathcal{R}_i = \mathbb{R}$ for all i, which is logical since there is no overidentifying equality constraints and the inequality constraints are not binding. In this case, the asymptotic level of the test will be 0 rather

than α, as could be expected. In all other cases, the test has a non-degenerated distribution and its asymptotic level is exactly α. Following the analysis of Kodde & Palm (1986), it is also possible to express this asymptotic distribution as a mixture of chi-square. The corresponding weights, however, do not have a closed form in general, so that it is actually easier to approximate the asymptotic distribution using (3.3.3). We use such simulations to compute our p-values in the application below.

3.3.2 The continuous case

The situation is more involved when (Y, Z) is continuous, because the distribution of (Y, Z) depends on the nonparametric function $P(.)$ that is identified through an integral equation. We mostly focus on the estimation of $\beta_0 = E[g(Y, Z)]$ here. The key insight is that this problem is closely related to the estimation of linear functionals in additive, nonparametric instrumental variables (IV) models. Recall that such models satisfy

$$Y = m(X) + \varepsilon,\ E(\varepsilon|Z) = 0.$$

These models have been investigated by, among others, Newey & Powell (2003), Hall & Horowitz (2005), Santos (2011) and Severini & Tripathi (2011). m is identified through the integral equation $E(Y - m(Z)|X) = 0$. This identifying equation is similar to ours, namely $E(1 - D/P(Y)|Z) = 0$. Rather than m itself, one may be interested in linear functionals of m, $\theta_0 = E[\phi(X)m(X)]$, where ϕ is known. In our context, we also have to estimate a linear functional of $1/P$, since $\beta_0 = E[D\beta(Y)/P(Y)]$. Given these analogies, it is not surprising that a similar methodology can be applied to our setting. An overview of the relationship between the two problems is given by Table 3.1.

Table 3.1: Analogy with additive nonparametric IV problems

	Endogeneous Attrition	Additive nonparametric IV
Observed variables	(D, DY, Z)	(Y, X, Z)
Unknown function	$1/P(.)$	$m(.)$
Exclusion restriction	$E\left(1 - \frac{D}{P(Y)}|Z\right) = 0$	$E(Y - m(X)|Z) = 0$
Operator T	$T(f) = E(f(Y)|D = 1, Z = .)$	$T(f) = E(f(X)|Z = .)$
Operator $T^\star$	$T^\star(f) = E(f(Z)|D = 1, Y = .)$	$T^\star(f) = E(f(Z)|X = .)$
Parameter of interest	$\beta_0 = E\left(\frac{Dg(Y)}{P(Y)}\right)$	$\theta_0 = E\left(\phi(X)m(X)\right)$
Root-n estimability condition: $\exists q \in L^2(Z)$ s.t.	$T^\star(q) = \beta(.)$	$T^\star(q) = \phi(.)$
Estimating equation	$\beta_0 = E(q(Z))$	$\theta_0 = E(Yq(Z))$
Estimator	$\widehat{\beta} = \widehat{E}\left(\widehat{q}(Z)\right)$	$\widehat{\theta} = \widehat{E}(Y\widehat{q}(Z))$

The first issue we investigate is the root-n estimability of β_0, that is to say, the existence of regular estimators converging at the root-n rate to β_0 (see, e.g., van der Vaart, 2000, Chapter 25). Our results are closely related to those of Severini & Tripathi (2011) in the classical IV framework. Let T^* be the adjoint operator of T, defined in (3.2.1):[3]

$$\begin{array}{rccl} T^*: & L^2(Z|D=1) & \to & L^2(Y|D=1) \\ & q & \mapsto & (y \mapsto E(q(Z)|D=1, Y=y)). \end{array}$$

Actually, we only need considering the restriction $T^*_{\mathcal{Y}_0}(q)$ of $T^*(q)$ on $\mathcal{Y}_0 = \text{Supp}(Y|D=0)$, which is included in $\text{Supp}(Y|D=1)$ under Assumption 12. By Assumptions 11 and 12, $E(q(Z)|D=1, Y) = E(q(Z)|D=0, Y)$ $P^{Y|D=0}$- almost surely. This allows us to extend $T^*_{\mathcal{Y}_0}(q)$ on $L^2(Z|D=0)$. By a slight abuse of notation, this extension, as well as the restriction of $\beta(.)$ on $\mathcal{Y}_0$, are also denoted by T^* and $\beta(.)$. The condition for root-n estimability is the following.

Assumption 15 *$g \in L^2(Y, Z)$ and there exists $q \in L^2(Z|D=0)$ such that $T^*(q) = \beta(.)$ and*

$$E\left[\frac{1 - P(Y)}{P(Y)}\left(q(Z) - g(Y, Z)\right)\left(q(Z) - g(Y, Z)\right)'\right] < \infty.$$

The condition $g \in L^2(Y, Z)$ is standard to derive the asymptotic distribution of $\frac{1}{n}\sum_{i=1}^n g(Y_i, Z_i)$, even when Y is always observed. The second condition is similar to the one considered by Severini & Tripathi (2011), namely the existence of q satisfying $T^\star(q) = \phi$ in

[3]We define here our operators on L^2 rather than on L^1, as in Section 2. This is not really a restriction since square integrability is required for root-n consistency in the first place.

their context. If the standard completeness condition holds, then $\text{Ker}\,(T) = \{0\}$ and $\overline{\mathcal{R}\,(T^*)} = L^2(Y|D=1)$, where $\mathcal{R}\,(T^*)$ denotes the range of T^* and $\overline{A}$ denotes the closure of A. As a consequence if $g \in L^2(Y,Z)$, then $\beta(.)$ lies in $\overline{\mathcal{R}\,(T^*)}$. However, when Y is continuous, $\mathcal{R}\,(T^*)$ is not closed in general, so that even if the standard completeness holds, it may happen that $\beta(.) \notin \mathcal{R}\,(T^*)$. In such a case, the following theorem states that β_0 can not be consistently estimated at the root-n rate, as in the additive nonparametric IV problem. We also provide the semiparametric efficiency bound under Assumption 15.

Theorem 3.3.3 *Suppose that Assumptions 10-12 hold, and $\beta(.) \in (Ker\,(T) \cap \mathcal{F}_g)^{\perp}$. Then a regular root-n estimator of β_0 exists only if Assumption 15 holds and in this case the semi-parametric efficiency bound of θ_0 is:*

$$V^* = V\left(g(Y,Z)\right) + \min_{q(.)\in T^{*-1}(\{\beta(.)\})} E\left[\frac{1-P(Y)}{P(Y)}\left(q(Z)-g(Y,Z)\right)\left(q(Z)-g(Y,Z)\right)'\right]. \tag{3.3.5}$$

The second part of the theorem shows that the asymptotic efficiency bound comprises two terms. The first corresponds to the standard estimation of β_0 without any attrition, i.e. when $D=1$. The second accounts for attrition, and is indeed, loosely speaking, increasing with $P(Y)$. It is also related to the quality of the approximation of $g(Y,Z)$ by functions of Z. If g only depends on Z, this term disappears, which makes sense because we can estimate directly β_0 by the sample average of $g(Z)$. On the other hand, if $g(Y,Z)$ only depends on Y, the expectation on the right-hand side of (3.3.5) can be rewritten as

$$E\left[\frac{1-P(Y)}{P(Y)}V\left(q(Z)|Y\right)\right].$$

Hence, if there is a strong dependence between Y and Z, we may expect this second term to be small.

Turning to inference, a key observation for estimating β_0 under Assumption 15 is that

$$\beta_0 = E\left[\beta(Y)\right] = E\left[E[q(Z)|D=1,Y]\right] = E\left[E[q(Z)|Y]\right] = E[q(Z)],$$

where the third equality follows by conditional independence. Once more, a similar estimating equation arises in nonparametric IV models, since we have $\theta_0 = E[Yq(Z)]$. In a similar way as Santos (2011), the idea is to estimate q first, and then estimate β_0 by taking the sample average of the $(\widehat{q}(Z_i))_{i=1\ldots n}$. A difficulty is that the q satisfying Equation 15

may not be unique. Santos (2011) proposes to choose the one with the smallest norm. Adapting his idea to our context, we consider

$$\widehat{q} = \arg\min_{q \in \Theta_n} \sum_{i=1}^{n} q(Z_i)^2 \quad \text{s.t.} \quad \frac{a_n}{n} \sum_{i=1}^{n} \widehat{E}\left([\beta(Y) - q(Z)] \widehat{f}_{Y|D=1}(Y) | Y = y_i, D = 1 \right)^2 \leq b_n.$$

Θ_n denotes a sieve space, i.e. a subset of $L^2(Z)$ such that $\Theta_n \subset \Theta_{n+1}$ and $q \in \overline{\cup \Theta_n}$. $\widehat{f}_X$ is a kernel estimator of f_X and $\widehat{E}[U|V=v]$ is a linear sieve estimator of $E[U|V=v]$, for any random variables (U, V).[4] The constraint of the program defines the set of functions $q \in \Theta_n$ that approximately satisfy $E(q(Z)|Y) = \beta(Y)$. Among those functions, $\widehat{q}$ is the one with the smallest norm. Santos (2011) shows that under technical conditions and with appropriate smoothing parameters, the corresponding estimator of θ_0 is root-n consistent and asymptotically normal. It is unclear, on the other hand, whether this estimator reaches the semiparametric efficiency bound.

3.4 Application

3.4.1 Introduction

In this section, we apply the previous results to estimate transitions on employment status in the French labor market. Beyond the unemployment rate, measuring such transitions is important to assess, for instance, the importance of short and long-term unemployment. We use for that purpose the Labor Force Survey (LFS) conducted by the French national institute of statistics (INSEE). This survey is probably the best tool to measure such transitions in France. Compared to administrative data or other surveys, it properly measures unemployment with respect to the standard ILO definition, has a comprehensive coverage of the population and has a large sample size. Since 2003, the French LFS is a rotating panel with approximately 5,900 new households each quarter. Each household is interviewed during six waves. On the first and sixth wave, interviews are face to face, while on the others they are conducted by telephone. It has been argued that the use of phone may introduce specific measurement errors (see, e.g., Biemer, 2001), so we focus on the first and last interrogations hereafter. We also restrict ourselves to people between 15 and 65 and pool together all labor force surveys on the period 2003-2005.

[4] Here, we have supposed that $\beta(Y)$ is known, which is the case in the common situation where $g(Y, Z)$ does not depend on Z. Otherwise, $\beta(Y)$ should also be estimated with a nonparametric estimator of $E(g(Y, Z)|D = 1, Y)$.

Table 3.2: Summary statistics on the French LFS.

Statistics	All	Men	Women
Main sample:			
Number of individuals	107,031	52,245	54,786
Attrition rate on last waves	21.78%	22.26%	21.31%
Participation rate on first waves	68.17%	73.91%	62.69%
(Uncorrected) participation rate on last waves	67.38%	72.75%	62.32%
Unemployment rate on first waves	9.68%	9.05%	10.39%
(Uncorrected) unemployment rate on last waves	8.02%	7.22%	8.90%
Refreshment sample for last waves:			
Number of observations	109,404	53,337	56,067
Participation rate on the refreshment sample	67.92%	73.31%	62.78%
Unemployment rate on the refreshment sample	9.97%	9.43%	10.57%

Sources: French LFS, first waves between 2003 and 2005, individuals between 15 and 65 year old.

Table 3.2 provides some summary statistics on our dataset, which emphasize that attrition may be problematic in the LFS survey. This is especially striking when we compare the (uncorrected) participation and unemployment rate on last waves and the one on the refreshment sample, which corresponds to entrants interviewed at the same time. We observe differences around 1.5 percent points on participation rates, and around 2 percent points on unemployment rates. To understand these differences, recall that in the French LFS, moving households are not followed by interviewers, who stick instead on housings which were selected in the first waves. This is likely to affect activity rates and transition estimates on the labor market, because transitions are very different for moving and non-moving households.

This latter fact can be illustrated using the French sample of the European Survey on Income Living Conditions (SILC). Contrary to the LFS, this panel follows individuals even if they move. It is therefore possible to estimate the difference in the transition matrix for those who have moved and the others. Note, on the other hand, that it is difficult to use its results as a benchmark, for several reasons. First, and most importantly, the status on the labor market is not obtained with the same questions as in the LFS, and it is well-known that this matters much for defining in particular unemployment (for evidence on this issue in France, see, e.g., Guillemot, 1996, and Gonzalez-Demichel & Nauze-Fichet, 2003). Second, still around 40% of the individuals in the French sample of SILC that move from one year to another are lost, so the bias stemming from such nonrespondents may still be substantial. That said, Table 3.3 shows that the difference in the transitions on

the labor market between individuals who have moved and the others are substantial. In particular, the diagonal of the transition matrix is much smaller for individuals who move. The difference reaches around 30 percentage points for inactive people. This suggests that the MAR and HIRR methods may overestimate the diagonal of the transition matrix.

Table 3.3: Comparison moving and non moving people in SILC

	Non moving			Moving		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV-MAR						
Y_1 = Empl.	92.86	3.50	3.64	90.53	3.99	5.48
	(0.34)	(0.27)	(0.23)	(1.34)	(0.95)	(0.99)
Y_1 = Unempl.	30.48	51.31	18.21	38.32	41.55	20.13
	(1.84)	(2.05)	(1.60)	(6.33)	(6.60)	(0.52)
Y_1 = Out L.F.	7.40	5.83	86.77	34.64	8.05	57.30
	(0.48)	(0.45)	(0.64)	(3.67)	(2.13)	(3.77)

Sources: French sample of SILC 2004/2005, individuals between 15 and 65.

Notes: standard error in parentheses.

As suggested in Section 2, we propose to correct for potentially endogenous attrition by using past employment status, measured by a retrospective question asked on the first waves. The underlying assumption is that attrition depends on the current transition on this outcome, but not on previous ones. This assumption is plausible if most of the endogeneity in attrition stems from the moving of households. The instrument Z we use is employment status six months before the first wave. We choose to divide this variable in three categories (unemployed, employed, and out of labour force), in the same way as our outcome, which is contemporary employment status.

3.4.2 The results

We first check the rank condition between Z_1 and Y_2 conditional on gender and Y_1, relying on the determinant test proposed in Subsection 3.1. Results are displayed in Table 3.4. The p-value of the rank test associated to any state Y_1 are always smaller than 10% for both men and women. We also implement the test developed in the Proposition 3.3.2, using $c_n = \ln(n)$. Though some inequality constraints are binding with Y_1 = Unempl., we do not reject the independence assumption $Z \perp\!\!\!\perp D|Y_1, Y_2$ here, the p-value being close to 0.50. The p-values equal to one that we obtain correspond to situations where the inequality constraints are not binding. In such a case, $W_n = 0$ and we accept the null hypothesis at any level.

Table 3.4: Rank test between Z and Y_2 conditional on gender and Y_1.

	P-value (Men)	P-value (Women)
$Y_1 =$ Empl.	0.004	0.001
$Y_1 =$ Unempl.	0.077	0.057
$Y_1 =$ Out L.F.	0.059	0.091

Sources: French LFS (2003-2005).
Notes: the p-values are obtained by bootstrap with 1,000 bootstrap samples.

Table 3.5: Test of $Z \perp\!\!\!\perp D|Y_1, Y_2$ by gender.

	P-value (Men)	P-value (Women)
$Y_1 =$ Empl.	1	1
$Y_1 =$ Unempl.	0.491	0.488
$Y_1 =$ Out L.F.	1	1

Sources: French LFS (2003-2005).
Notes: we use the test based on W_n and $\widehat{c}_\alpha$ defined in (3.3.2) and (3.3.4).

Second, we estimate the probabilities of attrition (or non-attrition) conditional on (Y_1, Y_2). Our results, displayed in Table 3.6, confirm that attrition is related to transitions on employment status. People who remain stable on the labor market have always a significant larger probability to respond in the second wave than people who change. In particular, we observe a large attrition for those who move from employment to unemployment or inactivity whereas attrition seems negligible for those who remain unemployed at both periods. As suggested above, such transitions are likely to be related to house movings. For instance, transitions from inactivity to employment or unemployment mostly correspond to students who enter the labor market and move at the same time. Such features cannot be captured under the missing at random (MAR) scheme $D \perp\!\!\!\perp Y_2|Y_1$, or the additive model of Hirano et al. (2001). In particular, they tend to underestimate the probability of attrition for people whose status change on the labor market, and to overestimate them for stable trajectories (see Table 3.7 for the tests on the difference between our IV models and the two others). Note also that we estimate the probability of attrition to be zero for people who remain unemployed. This indicates that for those people, the inequality constraint $b_i \geq 0$ is binding. This could suggest that the exclusion restriction is violated. However, the test

conducted previously shows that the unconstrained estimator, if negative, is actually close to zero, and we cannot reject at standard levels that the true value is actually positive. That $\widehat{b_i} = 0$ for individuals initially unemployed also indicates that the true value of b_i may be equal to zero, in which case the estimator is not asymptotically normal. We therefore use subsampling rather than the bootstrap or the normal approximation for inference on this subpopulation.

Table 3.6: Estimation of $P(D = 1|Y_1, Y_2)$ by gender under various assumptions.

	Men			Women		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV						
Y_1 = Empl.	84.33 (0.83)	34.33 (3.93)	46.31 (7.11)	85.72 (0.92)	46.45 (6.17)	44.57 (4.42)
Y_1 = Unempl.	55.56 (4.73)	100 (4.6)	51.01 (7.13)	52.46 (4.88)	100 (4.45)	76.38 (6.15)
Y_1 = Out L.F.	54.83 (10.91)	55.85 (11.15)	85.72 (2.01)	56.43 (11.88)	67.1 (17.22)	84.34 (1.11)
MAR						
Y_1 = Empl.	78.22 (0.22)	78.22 (0.22)	78.22 (0.22)	79.00 (0.22)	79.00 (0.22)	79.00 (0.22)
Y_1 = Empl.	65.9 (0.81)	65.9 (0.81)	65.9 (0.81)	69.77 (0.78)	69.77 (0.78)	69.77 (0.78)
Y_1 = Empl.	79.52 (0.35)	79.52 (0.35)	79.52 (0.35)	79.77 (0.28)	79.77 (0.28)	79.77 (0.28)
HIRR						
Y_1 = Empl.	79.01 (0.29)	59.98 (2.13)	76.44 (2.17)	79.57 (0.3)	66.59 (2.18)	77.57 (2.01)
Y_1 = Empl.	75.84 (1.4)	55.55 (1.25)	73.01 (2.06)	76.04 (1.45)	61.89 (1.37)	73.81 (1.75)
Y_1 = Empl.	82.41 (1.68)	65.09 (2.33)	80.15 (0.45)	82.01 (1.68)	69.99 (2.01)	80.18 (0.37)

Sources: French LFS (2003-2005).

Notes: the standard errors, in parentheses, are computed with the bootstrap except for Y_1 = Unempl. in the IV case, where we use subsampling.

Table 3.7: Comparison between our method and other ones on $\widehat{P}(D = 1|Y_1, Y_2)$

	Men			Women		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV-MAR						
Y_1 = Empl.	6.11 (<0.001)	−43.89 (<0.001)	−31.91 (<0.001)	6.72 (<0.001)	−32.55 (<0.001)	−34.43 (<0.001)
Y_1 = Unempl.	−10.34 (0.018)	34.1 (<0.001)	−14.9 (0.009)	−17.31 (<0.001)	30.23 (<0.001)	6.61 (0.322)
Y_1 = Out L.F.	−24.69 (0.024)	−23.67 (0.034)	6.2 (0.002)	−23.33 (0.049)	−12.66 (0.462)	4.58 (<0.001)
IV-HIRR						
Y_1 = Empl.	5.32 (<0.001)	−25.64 (<0.001)	−30.13 (<0.001)	6.15 (<0.001)	−20.14 (0.002)	−33 (<0.001)
Y_1 = Unempl.	−20.28 (<0.001)	44.45 (<0.001)	−22.01 (<0.001)	−23.58 (<0.001)	38.11 (<0.001)	2.57 (0.862)
Y_1 = Out L.F.	−27.58 (0.012)	−9.24 (0.415)	5.57 (0.005)	−25.57 (0.033)	−2.89 (0.867)	4.16 (<0.001)

Sources: French LFS (2003-2005).

Notes: the p-values, in parentheses, are computed with the bootstrap (Y_1 = Empl. and Out L.F.) and subsampling (Y_1 = Unempl.).

Before presenting our results on transitions, we estimate the distribution of Y_2 with our IV method and compare it with the one of the refreshment sample. We also estimate this distribution supposing that data are missing at random (MAR), i.e. $D \perp\!\!\!\perp Y_2|Y_1$. Table 3.8 shows that on the five statistics related to the distribution of Y_2, our estimator is close, and not statistically significant at usual levels, to the one based on the refreshment sample. Those based on the MAR assumptions, on the other hand, do differ significantly for several features of Y_2. In other words, we can reject, using the refreshment sample, the hypothesis that attrition only depends on past outcomes, while our independence condition is not rejected in the data. Note that we cannot use the refreshment sample to properly compare our method with the one of Hirano et al. (2001) because by construction, their estimator exactly matches the distribution of Y_2 on the refreshment sample.

Table 3.8: Comparison of the methods with the refreshment sample

	Men			Women		
	REF.	MAR	IV	REF.	MAR	IV
$P(Y_2 = \text{Empl.})$	66.4	67.47 (<0.0001)	64.59 (0.055)	56.15	56.81 (0.0054)	55.07 (0.159)
$P(Y_2 = \text{Unempl.})$	6.92	5.62 (<0.0001)	7.53 (0.235)	6.63	5.78 (<0.0001)	6.51 (0.801)
$P(Y_2 = \text{Out L.F.})$	26.69	26.92 (0.2641)	27.88 (0.159)	37.22	37.4 (0.4243)	38.42 (0.127)
Participation rate	73.31	73.08 (0.2641)	72.12 (0.159)	62.78	62.6 (0.4243)	61.58 (0.127)
Unemployment rate	9.43	7.68 (<0.0001)	10.44 (0.146)	10.57	9.24 (<0.0001)	10.58 (0.982)

Sources: French LFS (2003-2005).

Notes: the p-values of the difference with the refreshment sample, in parentheses, are obtained using the bootstrap (MAR) and subsampling (IV).

Finally, we compute transitions on the labor market using our IV method, the MAR assumption and the additive method of Hirano et al. (2001) (see Table 3.9). Not surprisingly given the discrepancies on the probabilities of attrition, our results differ significantly from those obtained by the other methods. Other methods lead in particular to a higher stability on the labor market. This is not surprising, given the assumptions underlying these methods. Table 3.3 suggests that there is a specific effect of being in the diagonal on the transition matrix on attrition, but neither the MAR assumption nor the additivity condition of Hirano et al. (2001) can incorporate such effects. The final results suggest that this could lead to important biases on the estimation of transitions.

Table 3.9: Estimated probability of transitions by gender under various assumptions

	Men			Women		
	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.	Y_2 = Empl.	Y_2 = Unempl.	Y_2 = Out L.F.
IV						
Y_1 = Empl.	85.86 (0.83)	6.12 (0.69)	8.02 (1.11)	83.46 (0.9)	4.73 (0.62)	11.81 (1.15)
Y_1 = Unempl.	49 (2.97)	25.85 (2.33)	25.15 (2.86)	52.61 (2.9)	25.3 (2.25)	22.08 (2.6)
Y_1 = Out L.F.	13.81 (2.59)	6.48 (1.15)	79.72 (1.81)	12.74 (2.24)	5.92 (1.49)	81.34 (1.07)
MAR						
Y_1 = Empl.	92.56 (0.16)	2.69 (0.1)	4.75 (0.13)	90.56 (0.18)	2.78 (0.1)	6.66 (0.16)
Y_1 = Empl.	41.31 (1.03)	39.23 (0.97)	19.46 (0.82)	39.56 (0.99)	36.27 (0.94)	24.18 (0.86)
Y_1 = Empl.	9.52 (0.28)	4.55 (0.21)	85.93 (0.34)	9.02 (0.22)	4.98 (0.17)	86 (0.27)
HIRR						
Y_1 = Empl.	91.64 (0.2)	3.5 (0.12)	4.86 (0.14)	89.92 (0.22)	3.3 (0.12)	6.79 (0.18)
Y_1 = Empl.	35.9 (0.93)	46.54 (0.92)	17.57 (0.79)	36.29 (0.94)	40.87 (0.88)	22.85 (0.84)
Y_1 = Empl.	9.19 (0.28)	5.56 (0.23)	85.26 (0.36)	8.77 (0.22)	5.68 (0.17)	85.55 (0.29)

Sources: French LFS (2003-2005).

Notes: the standard errors, in parentheses, are computed with the bootstrap except for Y_1 = Unempl. in the IV case, where we use subsampling.

3.5 Conclusion

In this paper, we develop an alternative method to correct for endogenous attrition in panel. We allow for both dependence on current and past outcomes and, thanks to the availability of an instrument, do not need to impose functional restrictions on the probability of attrition. The application suggests that our method may do a good job for handling attrition processes which mostly depend on transitions.

The paper raises challenging issues, related to our main conditional independence assumption. The first is whether the refreshment sample could be used to weaken this assumption, rather than to test for it. This may be useful in settings where this condition is considered too stringent. The second is whether one can build bounds on parameters of interest if the conditional independence assumption is replaced by weaker conditions such as monotonicity ones. Finally, an issue that also arises for nonparametric additive IV models would be

to obtain efficient estimators for linear functionals under Assumption 15, and consistent estimators without such an assumption.

3.6 Appendix: proofs

Theorem 3.2.1

The distribution of (Y, Z, D) is identified if and only if the distribution of $Y|Z, D = 0$ is identified. We have:

$$\begin{aligned} f_{Y|Z=z,D=0}(y) &= \frac{f_{D,Y,Z}(0,y,z)}{f_{D,Z}(0,z)} \\ &= \frac{1}{f_{D,Z}(0,z)} P(D=0|Y=y,Z=z) f_{Y,Z}(y,z) \\ &= \frac{1}{f_{D,Z}(0,z)} \frac{P(D=0|Y=y,Z=z)}{P(D=1|Y=y,Z=z)} f_{Y,Z|D=1}(y,z) \\ &= \frac{1}{f_{D,Z}(0,z)} \frac{1-P(y)}{P(y)} f_{Y,Z|D=1}(y,z). \end{aligned}$$

Then we deduce that $Y|Z, D = 0$ is identified if and only if P is identified. Under assumptions 11 and 12 (i), the function P is such that $T(1/P) = w$ and $1/P \geq 1$. Reciprocally, let Q a function such that $1/Q \in L^1(Y|D = 1)$, $T(1/Q) = w$ and $1/Q \geq 1$. If the unobserved distribution of $Y|Z, D = 0$ is such that

$$f_{Y|Z=z,D=0}(y) = \frac{1}{f_{D,Z}(0,z)} \frac{1-Q(y)}{Q(y)} f_{Y,Z|D=1}(y,z),$$

we have $P(D = 1|Y) = Q(Y)$ and $D \perp\!\!\!\perp Z|Y$. So the set of identification of P is

$$\{Q : 1/Q \in L_1(Y|D = 1), T(1/Q) = w, 1/Q \geq 1\},$$

which is reduced to a point if and only if $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$. This proves the first point of Theorem 3.2.1.

For the second and the third points, let Q be such that $T(1/Q) = w$, $1/Q \geq 1$ and $E(|g(Y,Z)|/Q(Y)\ |D = 1) < \infty$. Choosing $f_{Y|Z,D=0}$ as above, we can rationalize that $P(D = 1|Y) = Q(Y)$, $D \perp\!\!\!\perp Z|Y$ and $g \in L_1(Y, Z)$. So the set of identification of $1/P$ is

$$\left\{1/Q : 1/Q \in L^1(Y|D = 1), T(1/Q) = w, 1/Q \geq 1, E(|g(Y,Z)|/Q(Y)|D = 1) < \infty\right\}$$

or, equivalently,

$$\{1/P + h : h \in \text{Ker}\,(T) \cap \mathcal{F}, E(|gh||D = 1) < \infty\} = 1/P + \text{Ker}\,(T) \cap \mathcal{F}_g.$$

For all $h \in \mathcal{F}_g$ the quantities $E(|g(Y,Z)h(Y)||D=1)$, $E(g(Y,Z)h(Y)|D=1)$, $E(|\beta(Y)h(Y)||D=1)$, $E(\beta(Y)h(Y)|D=1)$ are well defined and finite. Then the set identification of β_0 is

$$\{\beta_0 + E(\beta(Y)h(Y)|D=1)E(D) : h \in \text{Ker}\,(T) \cap \mathcal{F}_g\},$$

which reduces to a point if and only if $E(\beta(Y)h(Y)|D=1) = 0$ for every $h \in \text{Ker}\,(T) \cap \mathcal{F}_g$. Hence, β_0 is identified if and only if $\beta(.) \in (\text{Ker}\,(T) \cap \mathcal{F}_g)^{\perp}$ □

Proposition 3.2.2

First, remark that if $\eta \perp\!\!\!\perp (Y_0, Y_1, Y_2)$, then $\eta \perp\!\!\!\perp Y_0|Y_1, Y_2$. As a result, $D \perp\!\!\!\perp Y_0|Y$ and Assumption 11 holds with $Z_1 = Y_0$. Now, suppose that $T(h) = 0$ for $h \in \mathcal{F}$, and let us prove that $h = 0$. First, $T(h) = 0$ rewrites as

$$0 = E(Dh(Y_1, Y_2)|Y_0, Y_1) = E(\widetilde{h}(Y_1, \widetilde{Y}_2)|Y_0, Y_1),$$

with $\widetilde{Y}_t = \Lambda(Y_t)$ and $\widetilde{h}(y_1, y_2) = h(y_1, \Lambda(y_2)) \times P(y_1, y_2)$. As a result, for all $t \in \mathbb{R}$,

$$E(\widetilde{h}(Y_1, \widetilde{Y}_2)e^{it\widetilde{Y}_0}|Y_1) = 0$$

Because $\varepsilon_0 \perp\!\!\!\perp (U, Y_1, Y_2)$,

$$E(\widetilde{h}(Y_1, \widetilde{Y}_2)e^{itU}|Y_1)\Psi_{\varepsilon_0}(t) = 0.$$

Thus, by assumption, $t \mapsto E(\widetilde{h}(Y_1, \widetilde{Y}_2)e^{itU}|Y_1)$ is equal to zero except perhaps on a set of isolated points. Because this function is continuous by dominated convergence, it is actually equal to zero on the whole line. This implies (see e. g. Bierens, 1982, Theorem 1)

$$E(\widetilde{h}(Y_1, \widetilde{Y}_2)|Y_1, U) = 0.$$

Now, ε_2 is independent of (Y_1, U) and U admits a density with respect to the Lebesgue measure. Thus, for almost all y_1 and almost every u,

$$\int \widetilde{h}(y_1, u - v) f_{-\varepsilon_2}(v) dv = 0. \tag{3.6.1}$$

Fix y_1 so that Equation (3.6.1) holds for almost every u. Because $h \geq 1 - 1/P$ by assumption, $\widetilde{h}$ is bounded below by -1. Letting $g = \widetilde{h}(y_1, .)$ and $\star$ denote the convolution product, we have $(g + 1) \star f_{-\varepsilon_2} = 1$ almost everywhere. Besides, by the first step of the proof of

Proposition 2.2 of D'Haultfœuille (2010), there exist positive c_1, c_2 and $0 < \alpha' < \alpha - 2$ such that

$$c_1 \leq (f_{\varepsilon_2} \star f_{\alpha'})(x) \times (1 + |x|)^{\alpha'+1} \leq c_2, \tag{3.6.2}$$

where $f_{\alpha'}$ denotes the density of an α'-stable distribution of characteristic function $\exp(-|t|^{\alpha'})$. Moreover, because $g + 1, f_{-\varepsilon_2}$ and $f_{\alpha'}$ are nonnegative, we can apply Fubini's theorem, so that

$$(g + 1) \star (f_{-\varepsilon_2} \star f_{\alpha'}) = ((g + 1) \star f_{-\varepsilon_2}) \star f_{\alpha'} = 1 \star f_{\alpha'} = 1.$$

Thus, $g \star \phi = 0$, with $\phi = f_{-\varepsilon_2} \star f_{\alpha'}$. In other words, we have a similar result as (A.5) in the proof of D'Haultfœuille (2010). Applying the third step of this proof shows that the location family generated by ϕ is complete. Thus $g = 0$ almost everywhere. Because this reasoning holds for almost all y_1, $h(Y_1, Y_2) = 0$ almost surely. Hence, $\text{Ker}\,(T) \cap \mathcal{F} = \{0\}$, and the result follows by Theorem 3.2.1 □

Proposition 3.3.1

For simplicity, we keep hereafter the conditioning on $Y_1 = y$ implicit. We first prove that $D \perp\!\!\!\perp Z_1 | Y_2$ is equivalent to the existence of $b_i \geq 0$, for $i = 1, ..., I$ such that $\forall (i, j), p_{0ij} = b_i p_{1ij}$. For $i \in \{1, ...I\}$, let $A(i)$ the set of j such that $\Pr(Y_2 = i, Z_1 = j) > 0$. Suppose first that $Z_1 \perp\!\!\!\perp D | Y_2$. Then for all i and all $j \in A(i)$,

$$[1 - \Pr(D = 1 | Y_2 = i, Z_1 = j)] / \Pr(D = 1 | Y_2 = i, Z_1 = j)$$

does not depend on j. Thus, there exists $b_i \geq 0$ such that for all $j \in A(i)$,

$$\Pr(D = 0 | Y_2 = i, Z_1 = j) = b_i \Pr(D = 1 | Y_2 = i, Z_1 = j).$$

Multiplying both sides by $\Pr(Y_2 = i, Z_1 = j)$ and remarking that both sides are equal to 0 when $j \notin A(i)$, we get, for all j, $p_{0ij} = b_i p_{1ij}$. This proves the "only if" part.

Conversely, suppose that there exists $b_i \geq 0$ such that $p_{0ij} = b_i p_{1ij}$. Because $\sum_j p_{1ij} = P(D = 1, Y_2 = i) > 0$ by Assumption 12, $p_{0ij} = b_i p_{1ij}$ implies that

$$p_{0ij} = \frac{\sum_j p_{0ij}}{\sum_j p_{1ij}} p_{1ij} = \frac{\Pr(D = 0, Y_2 = i)}{\Pr(D = 1, Y_2 = i)} p_{1ij}.$$

In other words, $P(Z_1 = j | D = 0, Y_2 = i) = P(Z_1 = j | D = 1, Y_2 = i)$ for all j, implying that $Z_1 \perp\!\!\!\perp D | Y_2 = i$.

We deduce that the distribution of (D, DY_2, Z_1) is compatible with $D \perp\!\!\!\perp Z_1 | Y_2$ if and only if

$$\forall (i,j), \exists p_{0ij} \geq 0, \exists b_i \geq 0 : \sum_i p_{0ij} = p_{0.j} \text{ and } p_{0ij} = b_i p_{1ij}.$$

This condition is also equivalent to the existence, for all i, of $b_i \geq 0$ satisfying $p_{0.j} = \sum_i b_i p_{1ij}$ for all j. If such $(b_i)_{i=1,...,I}$ exist, indeed, p_{0ij} can always be defined by $p_{0ij} = b_i p_{1ij}$.

As a result, the maximum likelihood estimator $\widehat{p}$ of p under Assumptions 11, 12 and 14 is defined by:

$$(\widehat{p}, \widehat{b}) = \arg \max_{(q,b) \in [0,1]^{(I+1)J} \times \mathbb{R}^{+I}} \sum_{j=1}^{J} \left[n_{0.j} \ln q_{0.j} + \sum_{i=1}^{I} n_{1ij} \ln q_{1ij} \right]$$
$$\text{s.t.} \left| \begin{array}{l} \sum_{j=1}^{J} \left[q_{0.j} + \sum_{i=1}^{I} q_{1ij} \right] = 1, \\ \sum_{i=1}^{I} q_{1ij} b_i = q_{0.j} \ j = 1, ..., J. \end{array} \right.$$

To complete the proposition, remark that $\widehat{p}$ is an asymptotically efficient estimator of p. The Fisher information matrix of p is singular because $\sum_{j,i} p_{1ij} + \sum_j p_{0.j} = 1$. However the parameter u defined by the $IJ + J - 1$ first components of p has a nonsingular Fisher information matrix. If matrix P_1 has rank I then $p_0 = P_1'(P_1 P_1')^{-1} P_1 p_{0.}$, and then $p_0 = l(u)$ and $\theta = g(p_0, p_1) = k(u)$ with l and k being differentiable at u. Because $\widehat{P}_1$ has rank i with probability tending to one, $\widehat{b}$ is equal to $(\widehat{P}_1 \widehat{P}_1')^{-1} \widehat{P}_1 \widehat{p}_{0.}$ with probability tending to one and then $\widehat{b} = m(\widehat{u})$ with m differentiable with probability tending to one. It follows that $\widehat{b}$, $\widehat{p_0}$ and $\widehat{\theta}$ are efficient estimators of $(P_1 P_1')^{-1} P_1 p_{0.}$, p_0 and θ (see for instance van der Vaart, 2000, Section 8.9) □

Proposition 3.3.2

The proof proceeds in three steps. We first establish that the set $\widehat{\mathcal{H}(h_0)}$ approximates well the set $\mathcal{H}(h_0)$. Then we establish the asymptotic distribution of W_n under the null. Thirdly, we compute the asymptotic level of the test, and show that it is consistent.

Step 1. $\Pr\left(\widehat{\mathcal{H}(h_0)} = \mathcal{H}(h_0)\right) \to 1$**.**

It suffices to show that

$$\Pr\left(\widehat{h}_{2j} \leq c_n/\sqrt{n}\right) \to 1 \text{ when } h_{20j} = 0 \tag{3.6.3}$$
$$\Pr\left(\widehat{h}_{2j} > c_n/\sqrt{n}\right) \to 1 \text{ when } h_{20j} > 0 \tag{3.6.4}$$

We have

$$\sqrt{n}\left(\widehat{h}_{2j} - h_{20j}\right) \to U_j \sim \mathcal{N}(0, \sigma_j^2).$$

Fix $\varepsilon > 0$, and let c_j be such that $\Pr(U_j \leq c_j) = \Pr(U_j > -c_j) > 1 - \varepsilon$. Because $c_n \to \infty$, we have $c_n \geq c_j$ for n large enough. Thus, when $h_{20j} = 0$,

$$\Pr(\widehat{h}_{2j} \leq c_n/\sqrt{n}) \geq \Pr(\sqrt{n}\widehat{h}_{2j} \leq c_j) \to \Pr(U_j \leq c_j) > 1 - \varepsilon.$$

This establishes (3.6.3). When $h_{20j} > 0$, we have $c_n - \sqrt{n}h_{20j} \leq -c_j$ for n large enough because $c_n/\sqrt{n} \to 0$. Thus,

$$\begin{aligned}\Pr\left(\widehat{h}_{2j} > c_n/\sqrt{n}\right) &= \Pr\left(\sqrt{n}(\widehat{h}_{2j} - h_{20j}) > c_n - \sqrt{n}h_{20j}\right) \\ &\geq \Pr\left(\sqrt{n}(\widehat{h}_{2j} - h_{20j}) > -c_j\right) \to \Pr\left(U_j > -c_j\right) > 1 - \varepsilon.\end{aligned}$$

Hence, (3.6.4) also holds, ending the first step.

Step 2. Asymptotic distribution of W_n under the null.

By Step 1, we can suppose without loss of generality that $\widehat{\mathcal{H}(h_0)} = \mathcal{H}(h_0)$. We show here that (3.3.3) holds under the null hypothesis. Let $\widetilde{h} = \widehat{h}_2 - \widehat{\Sigma}_{12}'\widehat{\Sigma}_{11}^{-}\widehat{h}_1$, $U_{2n} = \sqrt{n}(\widetilde{h} - h_{20})$, $V = \Sigma_{22} - \Sigma_{12}'\Sigma_{11}^{-}\Sigma_{12}$ and $\widehat{V} = \widehat{\Sigma}_{22} - \widehat{\Sigma}_{12}'\widehat{\Sigma}_{11}^{-}\widehat{\Sigma}_{12}$. Straightforward computations show that $U_{2n} \to U_2 \sim \mathcal{N}(0, V)$. Besides, we have, following Kodde & Palm (1986),

$$\begin{aligned}W_n &= n\widehat{h}_1'\widehat{\Sigma}_{11}^{-}\widehat{h}_1 + n\min_{x \geq 0}\left(x - \widetilde{h}\right)'\widehat{V}^{-1}\left(x - \widetilde{h}\right) \\ &= n\widehat{h}_1'\widehat{\Sigma}_{11}^{-}\widehat{h}_1 + \min_{x \geq 0}\left(\sqrt{n}(x - h_{20}) - \sqrt{n}(\widetilde{h} - h_{20})\right)'\widehat{V}^{-1}\left(\sqrt{n}(x - h_{20}) - \sqrt{n}(\widetilde{h} - h_{20})\right) \\ &= n\widehat{h}_1'\widehat{\Sigma}_{11}^{-}\widehat{h}_1 + \min_{t \geq -\sqrt{n}h_{20}}(t - U_{2n})'\widehat{V}^{-1}(t - U_{2n}).\end{aligned}$$

Let $\mathcal{H}_2(h_0) = \mathcal{R}_1 \times ... \times \mathcal{R}_I$ and define

$$\widetilde{W_n} = n\widehat{h}_1'\widehat{\Sigma}_{11}^{-}\widehat{h}_1 + \min_{t \in \mathcal{H}_2(h_0)}(t - U_{2n})'\widehat{V}^{-1}(t - U_{2n}).$$

For a given ε, there exists a compact set K such that $\Pr((U_{2n}, \widehat{V}) \in K) \geq 1 - \varepsilon$ for all n large enough. Let $\pi(u, V) = \arg\min_{t \in \mathcal{H}_2(h_0)}(t - u)'V^{-1}(t - u)$. Because $\mathcal{H}_2(h_0)$ is convex, π is a function rather than simply a correspondence. Moreover, it is continuous by Berge maximum theorem (see, e.g., Carter, 2001, Theorem 2.3). Thus $\pi(K)$ is compact. As a result, for n large enough, $\pi(K)$ is included in $[-\sqrt{n}h_{201}, +\infty[\times ... \times [-\sqrt{n}h_{201}, +\infty[$. In

other words, for n large enough,

$$\Pr(W_n = \widetilde{W}_n) \geq \Pr((U_{2n}, \widehat{V}) \in K) \geq 1 - \varepsilon. \tag{3.6.5}$$

Besides, the application $\Xi \mapsto \Xi^-$ is continuous once restricted to matrices or rank J (see, e.g., Stewart, 1969). By continuity of π and the continuous mapping theorem, we have, under the null hypothesis,

$$\widetilde{W}_n \xrightarrow{\mathcal{L}} U_1 \Sigma_{11}^- U_1 + \min_{t \in \mathcal{H}_2(h_0)} (t - U_2)' V^{-1} (t - U_2) = \min_{t \in \mathcal{H}(h_0)} (t - U)' \Sigma^- (t - U), \tag{3.6.6}$$

where $U_1 \sim \mathcal{N}(0, \Sigma_{11})$ and $U = (U_1, U_2 - \Sigma_{12}' \Sigma_{11}^- U_1)$. Note that $U \sim \mathcal{N}(0, \Sigma)$. Besides, (3.6.5) implies that W_n converges to the same distribution as $\widetilde{W}_n$. Hence, (3.6.6) implies that (3.3.3) holds.

Step 3. Consistency and asymptotic level of the test.

Let us define, for any positive matrix Ξ of rank J

$$g(u, \Xi) = \min_{t \in \mathcal{H}(h_0)} (t - u)' \Xi^- (t - u).$$

Let $\widehat{U}$ be a random normal variable satisfying $\widehat{U} | \widehat{\Sigma} \sim \mathcal{N}(0, \widehat{\Sigma})$. Because $\widehat{\Sigma} \xrightarrow{P} \Sigma$, we have $(\widehat{U}, \widehat{\Sigma}) \xrightarrow{\mathcal{L}} (U, \Sigma)$. Thus, by Berge maximum theorem once more, g is continuous. As a result, by the continuous mapping theorem, $g(\widehat{U}, \widehat{\Sigma}) \xrightarrow{\mathcal{L}} g(U, \Sigma)$.

Now, suppose first that $J > I$ or $\mathcal{R}_i = \mathbb{R}^+$ for some $i \in \{1, ..., I\}$. Then $g(U, \Sigma)$ is a mixture of chi-square distributions, and the weight of the chi-square of degree 0 is smaller than 1/2 (see, e.g., Kodde & Palm, 1986). Therefore, its quantile function is continuous on the interval $(1/2, 1)$. Combined with the convergence in distribution of $g(\widehat{U}, \widehat{\Sigma})$, this implies (see, e.g., van der Vaart, 2000, Theorem 21.2) that for any $\alpha \leq 1/2$, $\widehat{c}_\alpha \to c_\alpha$, the quantile of order $1 - \alpha$ of $g(U, \Sigma)$. Because the convergence of F_n, the cdf of W_n, towards F, the cdf of $g(U, \Sigma)$, is uniform (van der Vaart, 2000, Lemma 2.11), we have $F_n(\widehat{c}_\alpha) \to F(c_\alpha) = 1 - \alpha$. Thus, the test has the asymptotic level α.

Now, if $J = I$ and $\mathcal{R}_i = \mathbb{R}$ for all i, $g(U, \Sigma) = 0$ and the previous reasoning does not apply. On the other hand, remarking that $W_n = 0$ when $\widehat{\mathcal{H}(h_0)} = \mathcal{H}(h_0)$, we have

$$\Pr(W_n > \widehat{c}_\alpha) \leq \Pr(W_n > 0) \leq \Pr\left(\widehat{\mathcal{H}(h_0)} \neq \mathcal{H}(h_0)\right) \to 0.$$

Thus, the test has asymptotic level 0 in this case.

Finally, under the alternative, $h(p) \notin \mathcal{H}_0$. Then, by the continuous mapping theorem,

$$\min_{h \in \mathcal{H}(h_0)} \left(h - \widehat{h}\right)' \widehat{\Sigma}^{-} \left(h - \widehat{h}\right) \xrightarrow{P} \min_{h \in \mathcal{H}(h_0)} (h - h(p))' \Sigma^{-} (h - h(p)) > 0.$$

This implies that $W_n \xrightarrow{P} +\infty$, proving that the test is consistent $\square$

Proof of Theorem 3.3.3

The proof proceeds in two steps. In the first step we follow the approach of van der Vaart (2000, Chapter 25), who derive a necessary condition for the existence of a regular root-n estimator in semiparametric models. Intuitively, we exploit the fact that any score for the distribution of (D, DY, Z) is a projection of a score for the distribution of (D, Y, Z). This allows us to obtain a necessary condition for the existence of an influence function. In the second step, we characterize the set of such influence functions and derive the semiparametric efficiency bound of β_0.

Let us first introduce some notations. For the random variables U and V, we define $L^{20}(U)$ and $L^{20}(U|V)$ as the following sets of functions:

$$L^{20}(U) = L^2(U) \cap \{f | E(f(U)) = 0\},$$

$$L^{20}(U|V) = L^2(U,V) \cap \left\{f | E(f(U,V)^2|V) < \infty, E(f(U,V)|V) = 0 \ \ V\text{-almost surely}\right\}.$$

For any closed linear space $\mathcal{E} \subset L^2(U)$, we also let $\mathcal{P}_{\mathcal{E}}$ denote the orthogonal projection on $\mathcal{E}$.

Step 1. Assumption 15 is a necessary condition for existence of a regular root-n estimator.

Let $\mathcal{T}$ (respectively $\mathcal{S}$) denote the set of score function, for any subparametric model of the distribution of (D, Y, Z) (respectively of (D, DY, Z)). By Assumption 11, $\mathcal{T} = L^{20}(Y) + L^{20}(Z|Y) + L^{20}(D|Y) \subset L^{20}(D, Y, Z)$. Besides, because (D, DY, Z) is a function of (D, Y, Z), it follows from van der Vaart (2000, Section 25.5) that $\mathcal{S} = \{E(t|D, DY, Z) : t \in \mathcal{T}\}$. Hence, $\mathcal{T}$ and $\mathcal{S}$ are linear and closed here.

We define the score operator A by

$$\begin{array}{rccl} A: & \mathcal{T} & \to & L^{20}(D, DY, Z) \\ & h & \mapsto & [(d, u, z) \mapsto E(h(D, Y, Z)|D = d, DY = u, Z = z)]. \end{array}$$

Note that by definition, $\mathcal{R}(A) = \mathcal{S}$. The usual adjoint of A is the identity here. But, following van der Vaart (2000), we define the adjoint score operator A^* as the adjoint of A followed by the orthogonal projection onto $\mathcal{T}$:

$$\begin{array}{rccc} A^*: & L^{20}(D, DY, Z) & \to & \mathcal{T} \\ & \psi & \mapsto & \mathcal{P}_{\mathcal{T}}\psi. \end{array}$$

Because $L^{20}(Y), L^{20}(Z|Y)$ and $L^{20}(D|Y)$ are orthogonal for the usual inner product of $L^2(D, Y, Z)$, $\mathcal{P}_{\mathcal{T}}\psi = \mathcal{P}_{L^{20}(Y)}\psi + \mathcal{P}_{L^{20}(Z|Y)}\psi + \mathcal{P}_{L^{20}(D|Y)}\psi$.

Now let us consider a regular parametric submodel indexed by θ whose density with respect to an appropriate mesure is

$$f_Y(y, \theta) f_{Z|Y}(z|y, \theta)(P(y, \theta)d + (1 - P(y, \theta)(1 - d)).$$

Let also θ_0 denote the parameter corresponding to the true model. Defining $\mu(\theta) = E(g(Y, Z)|\theta)$, we have

$$\mu(\theta) = \int g(y, z) f_Y(y, \theta) f_{Z|Y}(z|y, \theta) dy dz.$$

The score of the submodel in θ_0 is, with obvious notations, $s_Y(y) + s_{Z|Y}(z|y) + \frac{P'(y)(d-P(y))}{P(y)(1-P(y))}$. Then

$$\frac{\partial \mu}{\partial \theta}(\theta_0) = E\left(g(Y, Z)(s_Y(Y) + s_{Z|Y}(Z|Y))\right).$$

It follows that the set of influence function is $\{g(Y, Z)\} + L^{20}(Y, Z)^{\perp}$. We can check that the second term is actually the set of constants. Thus, the efficient influence function corresponding to the complete model where (D, Y, Z) is observed, defined as the unique influence function that belongs to $\mathcal{T}$, is $g(Y, Z) - \beta_0$.

Theorem 25.32 of van der Vaart (2000) shows that if a regular root-n estimator exists, then $g(Y, Z) - \beta_0 \in \mathcal{R}(A^*)$. Let us now prove that this condition is equivalent to Assumption 15.

Let $\psi \in L^{20}(D, DY, Z)$ be such that $A^*(\psi) = g(Y, Z) - \beta_0$. Because A^* is a projector and satisfies $A^* = \mathcal{P}_{L^{20}(Y)} + \mathcal{P}_{L^{20}(Z|Y)} + \mathcal{P}_{L^{20}(D|Y)}$, we have

(a) $\mathcal{P}_{L^{20}(Y)}(\psi) = \mathcal{P}_{L^{20}(Y)}(g - \beta_0)$ or equivalently $E(\psi|Y) = \beta(Y) - \beta_0$,

(b) $\mathcal{P}_{L^{20}(D|Y)}(\psi) = \mathcal{P}_{L^{20}(D|Y)}(g - \beta_0)$ or equivalently $E(\psi|D, Y) - E(\psi|Y) = 0$

Let $m(Y,Z) = \psi(1,Y,Z)$ and $l(Z) = \psi(0,0,Z)$, we have :

$$E(\psi|D,Y) = \beta(Y) - \beta_0 \Rightarrow DE\left[m(Y,Z)|Y,D\right] + (1-D)E\left[l(Z)|Y,D\right] = \beta(Y) - \beta_0 \quad (3.6.7)$$

Hence, if $g(Y,Z) - \beta_0 \in \mathcal{R}(A^*)$, there exists $l \in L^2(Z|D=0)$ such that $E\left[l(Z)|Y,D=0\right] = \beta(Y) - \beta_0$ $P^{Y|D=0}-$ almost surely and $m = 1/P\left(g - \beta_0 - (1-P)l\right) \in L^2(Y,Z|D=1)$. Equivalently, because $Pm^2 + (1-P)l^2 = (g-\beta_0)^2 + \frac{1-P}{P}(g-\beta_0-l)^2$, there exists $l \in L^2(Z|D=0)$ such that $E\left[l(Z)|Y,D=1\right] = \beta(Y) - \beta_0$ $P^{Y|D=0}-$ almost surely and $E\left(\frac{1-P}{P}(g-\beta_0-l)(g-\beta_0-l)'\right) < \infty$ and $g \in L^2(Y,Z)$. Therefore, there exists $q \in L^2(Z|D=0)$ such that

$$T^*(q) = \beta(.) \text{ and } E\left(\frac{1-P}{P}(g-q)(g-q)'\right) < \infty.$$

Step 2. Characterization of the semiparametric efficiency bound.

First, recall that the semiparametric efficiency bound V^* satisfies

$$V^* = \min_{\psi \in \mathcal{I}} E\left[\psi\psi'\right], \quad (3.6.8)$$

where the minimum is understood in the partial order of symmetric nonnegative matrices and $\mathcal{I}$ is the set of influence functions, that is to say, the set of ψ satisfying, for all $s = A(\tau)$ $(\tau \in \mathcal{T})$, $E[\psi s] = E[(\beta(.) - \theta_0)\tau]$. Let us first show that

$$\mathcal{I} = \psi_0 + \mathcal{S}^\perp, \quad (3.6.9)$$

where $\psi_0 \in A^{*-1}\left(\{g(.,.) - \beta_0\}\right)$. Such an element exists under Assumption 15. First, for any $u \in \mathcal{S}^\perp$,

$$E[(\psi_0 + u)s] = E[\psi_0 s] = E[\psi_0 A(\tau)] = E[A^*(\psi_0)\tau] = E[(\beta(.) - \theta_0)\tau].$$

As a result, $\psi_0 + \mathcal{S}^\perp \subset \mathcal{I}$. Now, let $\psi \in \mathcal{I}$. By definition, $E[(\psi - \psi_0)s] = 0$. Thus, $\psi \in \psi_0 + \mathcal{S}^\perp$ and (3.6.9) holds. Note that we can also write any $\psi \in \mathcal{I}$ as $\mathcal{P}_\mathcal{S}(\psi_0) + u$, with $u \in \mathcal{S}^\perp$. By orthogonality,

$$E[\psi\psi'] - E[\mathcal{P}_\mathcal{S}(\psi_0)\mathcal{P}_\mathcal{S}(\psi_0)']$$

is nonnegative. Thus,

$$V^* = E[\mathcal{P}_\mathcal{S}(\psi_0)\mathcal{P}_\mathcal{S}(\psi_0)']. \quad (3.6.10)$$

Now, remark that $\mathcal{P}_{\mathcal{S}^\perp}(\psi_0) \in \mathcal{R}(A)^\perp = \text{Ker}\,(A^*)$. Hence, because $\psi_0 = \mathcal{P}_{\mathcal{S}}(\psi_0) + \mathcal{P}_{\mathcal{S}^\perp}(\psi_0)$, we have $\mathcal{P}_{\mathcal{S}}(\psi_0) \in A^{*-1}\left(\{g(.,.) - \beta_0\}\right)$. Combined with (3.6.10), this implies that

$$V^* \geq \min_{\psi \in A^{*-1}(\{g(.,.)-\beta_0\})} E\left[\psi\psi'\right] \geq V^*,$$

where the inequalities correspond to the partial order of symmetric matrices and the second inequality follows by (3.6.8) and the fact that $A^{*-1}\left(\{g(.,.) - \beta_0\}\right) \subset \mathcal{I}$.

Finally, note that ψ was associated to $l(Z)$, so that taking the minimum in ψ is equivalent to taking the minimum in $l(.) \in T^{*-1}\left(\{g(.,.) - \beta_0\}\right)$, or in $q(.) \in T^{*-1}\left(\{g(.,.)\}\right)$ (where $q = l + \beta_0$). Hence, because $V^* = E\left(\psi^*\psi^{*'}\right)$, we have

$$V^* = V(g(Y,Z)) + \min_{q \in T^{*-1}(\{\beta(.)\})} E\left(\frac{1 - P(Y)}{P(Y)}(q(Z) - g(Y,Z))(q(Z) - g(Y,Z))'\right) \quad \square$$

Chapter 4

On Pairwise Estimator of Honoré and Kyriazidou

4.1 Introduction

Disentangling state dependence and unobserved heterogeneity is an important problem in econometrics. The problem is particulary tedious when the model is not linear due to the endogeneity of the initial conditions and the problem of incidental parameters. In the case of a discrete variable, Honoré and Kyriazidou Honoré & Kyriazidou, 2000 (HK hereafter) have proposed an estimator that is consistent whatever the distribution of the individual fixed effects. Despite this attractive property and even if this paper is an important reference in the literature, this estimator is rarely used. Applied econometricians often prefer to use the estimator proposed by Wooldridge Wooldridge, 2007, although it is more restrictive, probably because this estimator is easily implementable with the standard procedures of econometric softwares. Indeed, the estimator of HK seems to require a specific programmation, in particular for inference since it is non root-N consistent. The aim of this note is to show that estimation and inference can actually be done with a simple weighted logit regression. For that, we only use an easy reshaping of the data. We also show that in the case of more than four periods, there exists two natural estimators of the asymptotic variance. Monte Carlo simulations on finite sample provide evidence that one of the two estimators of variance clearly outperforms the other, even with substantial sample sizes.

4.2 Theoretical results

Following HK, let us consider the fixed effect multinomial model with $M \geq 2$ alternatives depending of one lag of state dependence and k exogeneous regressors. We assume that variables are observed for a sample of individuals during $T+1$ periods ($t = 0..T$):

$$P(y_{it} = m | x_i, \alpha_i, y_{it-1} = j) = \frac{\exp(x_{mit}\beta_m + \alpha_{mi} + \gamma_{jm})}{\sum_{h=1}^{M} \exp(x_{hit}\beta_h + \alpha_{hi} + \gamma_{jh})}. \tag{4.2.1}$$

HK suggest the estimator

$$\left(\widehat{\beta}_n, \widehat{\gamma}_n\right) = \widehat{\theta}_n = \\ \arg\max_{\theta} \sum_{\substack{i=1..n \\ 1 \leq t < s \leq T-1 \\ m \neq l}} \mathbb{1}\left\{\{y_{it}, y_{is}\} = \{m, l\}\right\} K\left(\frac{x_{it+1} - x_{is+1}}{\sigma_n}\right) \ln\left(\frac{\exp(\mathbb{1}\left\{y_{it} = m\right\} Z_{itsml}\theta)}{1 + \exp(Z_{itsml}\theta)}\right) \tag{4.2.2}$$

with $\theta = (b', g')'$ a vector of size[1] $kM + (M-1)^2$ and Z_{itsml} is the vector of covariates such that:

$$\begin{aligned} Z_{itsml}\theta = (x_{mit} - x_{mis})b_m + (x_{lis} - x_{lit})b_l + g_{y_{it-1},m} + g_{l,y_{is+1}} - g_{y_{it-1},l} - g_{m,y_{is+1}} \\ + \mathbb{1}\left\{s - t = 1\right\}(g_{m,l} - g_{l,m}) + \mathbb{1}\left\{s - t > 1\right\}\left(g_{m,y_{it+1}} + g_{y_{is-1},l} - g_{l,y_{it+1}} - g_{y_{is-1},m}\right) \end{aligned} \tag{4.2.3}$$

HK show the asymptotic normality and compute the asymptotic variance of $\left(\widehat{\beta}, \widehat{\gamma}\right)$ when $M = 2$ and $T = 3$. We generalize their results to any M and T here. To simplify the forthcoming formulas, let $\Delta^{t,s}x_i = x_{it+1} - x_{is+1}$ and $f_{t,s}$ the density of $\Delta^{t,s}x_i$.

Theorem 4.2.1 (Asymptotics of the multinomial logit case) *Under the assumption that $x_{it} - x_{it'}$ has a positive density $f_{t,t'}$ in a neighborhood of 0 for all $t \neq t'$, σ_n is a positive sequence such that $\sqrt{n}\sigma_n^{2+k/2} \to \sigma$ and $K(.)$ is a smooth symmetric kernel, and assumptions (A1)-(A11) detailed in appendix:*

$$\sqrt{n}\sigma_n^{k/2}\left(\widehat{\theta}_n - \theta_0\right) \to \mathcal{N}\left(B, J^{-1}VJ^{-1}\right)$$

With

$h_{itsml}(\theta) = \mathbb{1}\{\{y_{it}, y_{is}\} = \{m, l\}\} \ln\left(\frac{\exp(\mathbb{1}\{y_{it}=m\}Z_{i,t,s,m,l}\theta)}{1+\exp(Z_{i,t,s,m,l}\theta)}\right)$,

$h^{(1)}_{itsml}(\theta) = \frac{\partial^1 h_{itsml}(\theta)}{(\partial\theta)}$,

[1] Without normalisation on the distribution of $(\alpha_m)_{1 \leq m \leq M}$, only $(M-1)^2$ component of γ are identified. Without loss of generality, one can assume that $\gamma_{1m} = \gamma_{m1} = 0$ for $1 \leq m \leq M$.

$h^{(2)}_{itsml}(\theta) = \frac{\partial^2 h_{itsml}(\theta)}{(\partial\theta\partial\theta')}$
and
$B = O\left(\sqrt{n}\sigma_n^{2+k/2}\right)$,
$J = -\sum_{\substack{1\leq t<s\leq T-1 \\ m\neq l}} f_{t,s}(0)E\left(h^{(2)}_{itsml}|x_{it+1} = x_{is+1}\right)$,
$V = \sum_{\substack{1\leq t<s\leq T-1 \\ m\neq l}} f_{t,s}(0)E\left(h^{(1)}_{itsml}h^{(1)'}_{itsml}|x_{it+1} = x_{is+1}\right)\int K^2(u)du.$
J can be consistently estimated by
$\widehat{J}_n = -\frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\sum_{\substack{1\leq t<s\leq T-1 \\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)h^{(2)}_{tsmli}(\widehat{\theta}_n).$
V can be consistently estimated by
$\widehat{V}_n = \frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\left[\sum_{\substack{1\leq t<s\leq T-1 \\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)h^{(1)}_{itsml}(\widehat{\theta}_n)\right]\left[\sum_{\substack{1\leq t<s\leq T-1 \\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)h^{(1)}_{itsml}(\widehat{\theta}_n)'\right]$
or by
$\widetilde{V}_n = \frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\sum_{\substack{1\leq t<s\leq T-1 \\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)^2 h^{(1)}_{itsml}(\widehat{\theta}_n)h^{(1)}_{itsml}(\widehat{\theta}_n)'.$

This theorem calls four remarks, which are at the basis of our simple estimation procedure. First, following HK, by considering pairwise estimation, we are led back to maximize the weighted likelihood of a binary logit, the weights being $K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)$.
Second, the "sandwich" structure of asymptotic covariance matrix looks like the structure of a robust covariance matrix.
Third, $\widehat{V}_n$ seems to be the "natural" candidate to estimate V, since $\widehat{V}_n$ appears in the Taylor expansion of the first order condition. $\widehat{V}_n$ depends on some intra-individual correlations of $\Delta^{t,s}x_i$ and $h^{(1)}_{itsml}$, when s and t vary. However, after expansion of the products appearing in $\widehat{V}_n$, the terms depending of two distincts pairs $((s,t) \neq (s',t'))$ vanish when $\sigma_n \to 0$. So, $\widetilde{V}_n$ is an alternative estimate of V, which does not depend on the intra-individual correlation of $\Delta^{t,s}x_i$ and $h^{(1)}_{itsml}$. In Section 4.4, the accuracy of $\widehat{V}_n$ and $\widetilde{V}_n$ will be compared using Monte Carlo simulations.
Fourth, even though the optimal rate of consistence is $n^{2/(4+k)} < n^{1/2}$ (for $\sigma_n \sim n^{-1/(4+k)}$), $\widehat{V}\left(\widehat{\theta}_n - \theta_0\right) \sim \frac{1}{n\sigma_n^k}\widehat{J}_n^{-1}\widehat{V}_n\widehat{J}_n^{-1} = \frac{1}{n}\left(\sigma_n^k\widehat{J}_n\right)^{-1}\left(\sigma_n^k\widehat{V}_n\right)\left(\sigma_n^k\widehat{J}_n^{-1}\right)$ only depends on k and σ_n through $K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)$. So the estimated standard deviations, t-statistics, confidence intervals and p-values also depend on k and σ_n only through $K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)$. Making inference using the wrong consistence rate $\sqrt{n}$ and the wrong estimators $\sigma_n^k\widehat{J}_n$ and $\sigma_n^k\widehat{V}_n$ (respectively $\sigma_n^k\widetilde{V}_n$), will give correct results.

4.3 Simple computation and inference

To understand our implementation strategy, let us consider a weighted binary logit in the presence of clustering. Let $i = 1..n$ denote the clusters, j a unit within the cluster, w_j a weight, y_j the outcome and X_j a set of exogeneous regressors. The weighted logit estimator is given by:

$$\widehat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j \in i} w_j \ln \left(\frac{\exp(y_j X_j \theta)}{1 + \exp(X_j \theta)} \right) \tag{4.3.1}$$

$\widehat{\theta}_n$ is a consistent estimator of $\theta_0 = \arg\max_\theta E\left[w \ln\left(\frac{\exp(yX\theta)}{1+\exp(X\theta)}\right)\right]$, and following Binder (1983) :

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \to \mathcal{N}\left(0, Q^{-1}GQ^{-1}\right)$$

With $h_j(\theta) = \ln\left(\frac{\exp(yX\theta)}{1+\exp(X\theta)}\right)$, $h_j^{(1)}(\theta) = \frac{\partial h_j(\theta)}{\partial \theta}$, $h_j^{(2)}(\theta) = \frac{\partial^2 h_j(\theta)}{\partial \theta \partial \theta'}$,
$G = E\left(\sum_{j\in i} w_j h_j^{(1)}(\theta) \sum_{j'\in i} w_{j'} h_{j'}^{(1)}(\theta)'\right)$ and $Q = E\left(\sum_{j\in i} w_j h_j^{(2)}(\theta)\right)$, that can be consistently estimated by their empirical counterpart $\widehat{G}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j\in i} w_j h_j^{(1)}(\theta) \sum_{j'\in i} w_{j'} h_j'^{(1)}(\theta)'\right]'$ and $\widehat{Q}_n = \sum_{i=1}^{n} w_j h_j^{(2)}$.

Such a "sandwich" formula of variance is frequent in econometrics, and is often referred to as the robust estimation of standard deviation. Most of current statistical or econometrics softwares offer procedures that quickly compute such estimators and provide t-statistics, confidence intervals and p-value.

A remarkable analogy exists between Programs 4.2.2 and 4.3.1. The programs will be the same if units j in Program 4.3.1 correspond to the 5-uplets (i, s, t, m, l) such that $1 \leq t < s \leq T-1$ and $m \neq l$ and $\mathbb{1}\{y_{it} = m\}\mathbb{1}\{y_{is} = l\} + \mathbb{1}\{y_{it} = l\}\mathbb{1}\{y_{is} = m\} = 1$ in the initial Program 4.2.2.

As mentioned above, V can be estimated by two asymptotically equivalent estimators $\widehat{V}_n$ or $\widetilde{V}_n$. Following the analogy, the asymptotic variance of the weighted logit estimate can be estimated under the assumption of clustering (identified by i) or not.

The forms of the asymptotic covariance matrix are the same except for the presence of a $\frac{1}{\sigma_n^k}$ factor in the three terms of the "sandwich" formula. But as mentioned previously, this factor is compensated by the slower rate of consistence of the HK estimator. And so, estimation of the variance of the HK estimator can be correctly performed using the naive root-N rate of consistence and the naive estimators $\widehat{Q}_n$ and $\widehat{G}_n$ for J and V.

As a result, we propose the following simple procedure for the estimation and inference of the model:

1. Create a new data set, where each line corresponds to a 5-uplet (i, t, s, m, l) such that $1 \leq t < s \leq T-1$ and[2] $m < l$ and $\mathbb{1}\{y_{it} = m\}\mathbb{1}\{y_{is} = l\} + \mathbb{1}\{y_{it} = l\}\mathbb{1}\{y_{is} = m\} = 1$ in the original data set. Each line contains the binary variable $Y = \mathbb{1}_{\{y_{it}=m\}}$ and the variables Z_{itsml} define as in Equation 4.2.3 (see also Table 4.1).

2. Choose a kernel K and a bandwidth σ_n and compute the weights[3] $W = K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)$. Example of original and new dataset is presented in Table 4.2, for the case where $M = 3$, $T = 5$ and a gaussian kernel.

3. Use a weighted binary logit procedure (the procedure must use a robust estimator of the covariance matrix) to regress Y on Z using weights W. The absence of intercept in the regressors can be specified.

   - Specify a clustering (identified by i), to estimate V by $\widehat{V}_n$.
   - Do not specify a clustering, to estimate V by $\widetilde{V}_n$.

4. Then the estimators of the parameters, and of standard deviations, as well as t-statistics and p-values are asymptotically valid.

4.4 Monte Carlo simulations

In this section, we use the previous method to estimate a logit model with state dependence on several data generating processes.
We draw the data according to (4.2.1), with $M = 3$, one exogeneous regressor ($k = 1$), $T = 4$ or $T = 8$ and $n = 250, 1000$ or 4000. We let $x_{it} \sim \mathcal{N}(0, 1)$, $\alpha_{i1} = 0$, $\alpha_{i2} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ and $\alpha_{i3} = -sgn(\sum_{t=1}^{T} x_{it}^3)\left(\frac{1}{T}\sum_{t=1}^{T} x_{it}^3\right)^{1/3}$. State 1 is chosen as the reference, so that $\beta_1 = \gamma_{1m} = \gamma_{m1} = 0$. We set: $\beta_2 = 0.7$, $\beta_3 = 0.7$, $\gamma_{22} = 0.6$, $\gamma_{23} = 0.3$, $\gamma_{32} = 0.2$, $\gamma_{33} = 0.2$. The initial conditions are drawn following the multinomial logit model (without state dependence).

$$P(y_{i0} = m|x_i, \alpha_i) = \frac{\exp(x_{mi0}\beta_m + \alpha_{mi})}{\sum_{h=1}^{3}\exp(x_{hi0}\beta_h + \alpha_{hi})}$$

[2] If pairs such that $m > l$ are also included, some observations in the new dataset are redondant and the variance of estimators is underestimate. In the definition of the HK estimator, the index $m \neq l$ of the sum describes the non ordered pairs (cf. the binary case).

[3] In theory, estimates are invariant to a multiplicative change of the weights. However, in practice large weights can cause computational problems. For this reason we recommend to use $K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)$ as weight instead of $\frac{1}{\sigma_n^k}K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)$.

Table 4.1: Construction of regressor in the new dataset used to estimate. State 1 choose as reference, $2 \leq m < l \leq M$, $k, k' \notin \{1, m, l\}$

$$
\begin{array}{rcl}
Z_{itsml}(\beta_m) & = & x_{imt} - x_{ims} \\
Z_{itsml}(\beta_l) & = & x_{ils} - x_{ilt} \\
Z_{itsml}(\beta_k) & = & 0 \\
Z_{itsml}(\gamma_{mm}) & = & \mathbb{1}_{\{y_{it-1}=m\}} - \mathbb{1}_{\{y_{is+1}=m\}} + \mathbb{1}_{\{s-t>1\}} \left(\mathbb{1}_{\{y_{it+1}=m\}} - \mathbb{1}_{\{y_{is-1}=m\}}\right) \\
Z_{itsml}(\gamma_{ll}) & = & \mathbb{1}_{\{y_{is+1}=l\}} - \mathbb{1}_{\{y_{it-1}=l\}} + \mathbb{1}_{\{s-t>1\}} \left(\mathbb{1}_{\{y_{is-1}=l\}} - \mathbb{1}_{\{y_{it+1}=l\}}\right) \\
Z_{itsml}(\gamma_{lm}) & = & \mathbb{1}_{\{y_{it-1}=l\}} + \mathbb{1}_{\{y_{is+1}=m\}} - \mathbb{1}_{\{s-t=1\}} - \mathbb{1}_{\{s-t>1\}} \left(\mathbb{1}_{\{y_{it+1}=m\}} - \mathbb{1}_{\{y_{is-1}=l\}}\right) \\
Z_{itsml}(\gamma_{ml}) & = & -\mathbb{1}_{\{y_{it-1}=m\}} - \mathbb{1}_{\{y_{is+1}=l\}} + \mathbb{1}_{\{s-t=1\}} + \mathbb{1}_{\{s-t>1\}} \left(\mathbb{1}_{\{y_{it+1}=l\}} - \mathbb{1}_{\{y_{is-1}=m\}}\right) \\
Z_{itsml}(\gamma_{km}) & = & \mathbb{1}_{\{y_{it-1}=k\}} - \mathbb{1}_{\{s-t=1\}} \mathbb{1}_{\{y_{is-1}=k\}} \\
Z_{itsml}(\gamma_{mk}) & = & -\mathbb{1}_{\{y_{is+1}=k\}} + \mathbb{1}_{\{s-t=1\}} \mathbb{1}_{\{y_{it+1}=k\}} \\
Z_{itsml}(\gamma_{kl}) & = & -\mathbb{1}_{\{y_{it-1}=k\}} + \mathbb{1}_{\{s-t=1\}} \mathbb{1}_{\{y_{is-1}=k\}} \\
Z_{itsml}(\gamma_{lk}) & = & \mathbb{1}_{\{y_{is+1}=k\}} - \mathbb{1}_{\{s-t=1\}} \mathbb{1}_{\{y_{it+1}=k\}} \\
Z_{itsml}(\gamma_{kk'}) & = & 0
\end{array}
$$

We use the optimal rate $\sigma_n = cn^{-1/(k+4)} = cn^{-1/5}$, a gaussian kernel, and choose three values (0.1, 1 and 10) for the parameter c. For $T = 3$ there is at most one pair of periods for each individual that is used in the estimation, so $\widehat{V}_n = \widetilde{V}_n$. For $T = 7$, even if $\widehat{V}_n$ and $\widetilde{V}_n$ are asymptotically equivalent, they are different in small samples.

We reshape the data as explained above and use the SURVEYLOGISTIC procedure in SAS[4], which gives an estimation of the parameters as well as the associated inference. The number of individuals that are used in the estimation is close to 54% of n. With our method the results are obtained almost instantaneous.

The main results of the simulation are the following (see Table 4.3). First, the bias on small sample depends only weakly on the constant c (except for small c and sample size $n = 250$) and is negligible compared to the standard deviation.

Second, the level of the tests on small samples are close to the asymptotic level for the estimation using clustering. However, when we use the estimate of $\widetilde{V}_n$, the actual level of tests are widely above their nominal level. Additional simulations (not reproduced here) show that the disturbance increases with T, or, equivalently, with ratio of the number of pairs of periods used to the number of individual observations used.

Third, the actual level of the tests using $\widetilde{V}_n$ decreases slightly with c and dramatically slowly as n increases. In Table 4.3, when c is divided by 100, the actual level only slowly decreases to the nominal level. To obtain a similar reduction with an increase of n while keeping c constant, the sample size must be multiplied by $100^5 = 10^{10}$. This explains the poor behavior of confidence interval using $\widetilde{V}_n$ even on samples of substantial size. So, for

[4]Same results can be obtained using instruction LOGIT with options PWEIGHTS and CLUSTER in Stata or using functions SVYDESIGN and SVYGLM of package SURVEY in R.

Table 4.2: Rearrangements with 6 time periods ($T = 5$), $M = 3$ states, $k = 1$ regressor

(a) Original Dataset

i	T	y	x
Bo	0	1	0.21
Bo	1	2	1.44
Bo	2	2	0.94
Bo	3	1	−1.25
Bo	4	3	0.13
Bo	5	2	−0.15
Ekaterini	0	2	−0.83
Ekaterini	1	1	0.10
Ekaterini	2	2	0.25
Ekaterini	3	1	0.69
Ekaterini	4	1	0.71
Ekaterini	5	3	0.21

(b) Dataset used to estimate

i	t	s	m	l	$Z(\beta_2)$	$Z(\beta_3)$	$Z(\gamma_{22})$	$Z(\gamma_{23})$	$Z(\gamma_{32})$	$Z(\gamma_{33})$	y	W
Bo	1	3	1	2	−2.69	0.00	0	1	0	0	0	0.44628
Bo	1	4	2	3	1.31	−1.31	0	0	0	0	1	0.3691
Bo	2	3	1	2	−2.20	0.00	−1	1	0	0	0	0.2827
Bo	2	4	2	3	0.81	−0.81	0	−1	1	0	1	0.36301
Bo	3	4	1	3	0.00	1.39	0	−1	1	0	1	0.54863
Ekaterini	1	2	1	2	0.15	0.00	−1	0	0	0	1	0.5268
Ekaterini	2	3	1	2	0.44	0.00	0	0	0	0	0	0.5641
Ekaterini	2	4	1	2	0.46	0.00	0	1	0	0	0	0.52065

$$W = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_{t+1}-x_{s+1})^2}{2\sigma_n^2}} \quad \text{with} \quad \sigma_n = 1000^{-\frac{1}{k+4}} \simeq 0.25$$

$T > 3$ using clustering for inference clearly outperforms the inference without clustering.

4.5 Proof

Theorem 4.5.1 (detailed assumptions and extended proof)

(A1) $\{(y_{it})_{t=1..T}, (x_{mit})_{m=1..M,t=1..T}\}_{i=1}^{n}$ *is a random sample of n observations from a distribution satisfying Equation (1) in our main paper*
(A2) $\theta_0 \in int(\Theta)$*, with* Θ *a compact subset of* $\mathbb{R}^{(k+1)(M-1)}$
(A3) $\forall 2 \leq t < s \leq T-1$, $\Delta^{t,s}x_i$ *is absolutely continuously distributed with a density* $f_{t,s}$*, bounded above on its support, and strictly positive at zero, twice differentiable on its support, with bounded derivatives.* $\forall 2 \leq t < s \leq T-1$, $\forall 2 \leq t' < s' \leq T-1$ *such that* $(t,s) \neq (t',s')$, $(\Delta^{t,s}x_i, \Delta^{t',s'}x_i)$ *is absolutely continuously distributed with a density* $f_{t,s,t',s'}$*, bounded above on its support, and strictly positive at zero, twice differentiable on its support, with bounded derivatives.*
(A4) $\forall 2 \leq t < s \leq T-1$, $E\left[||x_{it} - x_{is}||^6 | \Delta^{t,s}x_i\right]$ *is bounded on its support.*
(A5) Let $h_{itsml}(\theta) = \mathbb{1}\{\{y_{it}, y_{is}\} = \{m,l\}\} \ln\left(\frac{\exp(\mathbb{1}\{y_{it}=m\}Z_{i,t,s,m,l}\theta)}{1+\exp(Z_{i,t,s,m,l}\theta)}\right)$. $\forall 2 \leq t < s \leq T-1$, $\forall m,l \in [1..M]$, $E\left[h_{itsml}(\theta)|\Delta^{t,s}x_i\right]$ *is continuous in a neighborhood of zero for all* $\theta \in \Theta$.
(A6) $\forall 2 \leq t < s \leq T-1$, $\forall m,l \in [1..M]$, $E\left[(x_{it}-x_{is})'(x_{it}-x_{is})|\Delta^{t,s}x_i\right]$ *has full rank k in a neighborhood of zero.*
(A7) $K : \mathbb{R}^k \to \mathbb{R}$ *is a bounded and symmetric kernel such that* $\int K(u)du = 1$.
(A8) $\sqrt{n}\sigma_n^{2+k/2} \to \sigma \in \mathbb{R}^+$.
(A9) Let $h^{(1)}_{itsml}(\theta) = \frac{\partial h_{itsml}(\theta)}{(\partial\theta)}$. $\forall 2 \leq t < s \leq T-1$, $\forall m,l \in [1..M]$, $E\left[h^{(1)}_{itsml}(\theta)|\Delta^{t,s}x_i\right]$ *is continuous in a neighborhood of zero for all* $\theta \in \Theta$.
(A10) Let $h^{(2)}_{itsml}(\theta) = \frac{\partial^2 h_{itsml}(\theta)}{(\partial\theta\partial\theta')}$. $\forall 2 \leq t < s \leq T-1$, $\forall m,l \in [1..M]$, $E\left[h^{(2)}_{itsml}(\theta)|\Delta^{t,s}x_i\right]$ *is continuous in a neighborhood of zero for all* $\theta \in \Theta$.
(A11) $E\left(h^{(1)}(\theta_0)h^{(1)}(\theta_0)'|\Delta^{t,s}x_i\right)$ *is continuous in a neighborhood of zero.*
If (A1)-(A11) hold, then

$$\sqrt{n}\sigma_n^{k/2}\left(\widehat{\theta}_n - \theta_0\right) \to \mathcal{N}\left(B, J^{-1}VJ^{-1}\right)$$

With $B = O\left(\sqrt{n}\sigma_n^{2+k/2}\right)$, J *and* V *being consistently estimated by*
$\widehat{J}_n = -\frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h^{(2)}_{itsml}(\widehat{\theta}_n)$ *and*

Table 4.3: Estimation and inference on finite sample

			4 Periods ($T = 3$)			8 Periods ($T = 7$)			
Parameter	c	n	Bias	Std.Dev.	level of test	Bias	Std.Dev.	level of test using $\widehat{V}_n$	level of test using $\widetilde{V}_n$
	0.1	250	0.0987	0.5377	0.077	0.0065	0.1170	0.064	0.137
	0.1	1000	0.0170	0.2106	0.074	0.0042	0.0570	0.062	0.134
	0.1	4000	0.0048	0.0948	0.045	0.0005	0.0273	0.056	0.119
	1	250	0.0437	0.2733	0.061	0.0066	0.0907	0.049	0.216
$\beta_1 = 0,7$	1	1000	0.0175	0.1261	0.052	0.0053	0.0443	0.051	0.212
	1	4000	0.0107	0.0572	0.040	0.0025	0.0215	0.048	0.205
	10	250	0.0407	0.2283	0.056	0.0079	0.0843	0.049	0.242
	10	1000	0.0213	0.1082	0.056	0.0069	0.0416	0.046	0.256
	10	4000	0.0169	0.0499	0.057	0.0046	0.0205	0.053	0.247
	0.1	250	0.1958	0.5868	0.085	0.0104	0.1214	0.056	0.124
	0.1	1000	0.0349	0.2110	0.072	0.0037	0.0614	0.058	0.143
	0.1	4000	0.0097	0.0996	0.051	0.0018	0.0286	0.045	0.116
	1	250	0.0790	0.2851	0.075	0.0067	0.0945	0.054	0.203
$\beta_2 = 0,7$	1	1000	0.0200	0.1240	0.053	0.0037	0.0482	0.067	0.224
	1	4000	0.0092	0.0619	0.057	0.0028	0.0224	0.045	0.186
	10	250	0.0666	0.2399	0.065	0.0073	0.0883	0.056	0.226
	10	1000	0.0225	0.1083	0.057	0.0050	0.0448	0.055	0.259
	10	4000	0.0143	0.0541	0.077	0.0044	0.0210	0.051	0.222
	0.1	250	0.1621	1.1644	0.100	-0.0098	0.2554	0.059	0.190
	0.1	1000	0.0048	0.4208	0.059	-0.0044	0.1257	0.062	0.192
	0.1	4000	-0.0066	0.1951	0.051	-0.0004	0.0604	0.047	0.189
	1	250	0.0246	0.5956	0.068	-0.0144	0.2090	0.048	0.290
$\gamma_{11} = 0,6$	1	1000	-0.0132	0.2602	0.046	-0.0065	0.1042	0.051	0.297
	1	4000	-0.0073	0.1282	0.043	-0.0029	0.0502	0.046	0.285
	10	250	0.0020	0.5121	0.055	-0.0165	0.1991	0.051	0.325
	10	1000	-0.0182	0.2383	0.050	-0.0089	0.0989	0.053	0.318
	10	4000	-0.0098	0.1163	0.042	-0.0058	0.0487	0.043	0.350
	0.1	250	0.0983	1.0842	0.078	-0.0040	0.2577	0.044	0.186
	0.1	1000	-0.0078	0.3991	0.055	-0.0041	0.1254	0.044	0.185
	0.1	4000	-0.0035	0.1866	0.051	-0.0011	0.0650	0.048	0.208
	1	250	0.0099	0.5719	0.054	-0.0118	0.2192	0.049	0.299
$\gamma_{12} = 0,3$	1	1000	-0.0241	0.2626	0.058	-0.0085	0.1072	0.041	0.296
	1	4000	-0.0089	0.1263	0.054	-0.0047	0.0550	0.051	0.315
	10	250	0.0002	0.4969	0.055	-0.0183	0.2094	0.050	0.335
	10	1000	-0.0257	0.2405	0.071	-0.0137	0.1032	0.047	0.346
	10	4000	-0.0122	0.1146	0.044	-0.0095	0.0523	0.047	0.352
	0.1	250	0.0418	1.0408	0.096	-0.0065	0.2516	0.043	0.181
	0.1	1000	-0.0020	0.4054	0.069	-0.0052	0.1235	0.039	0.190
	0.1	4000	-0.0041	0.1795	0.048	-0.0020	0.0625	0.055	0.196
	1	250	0.0144	0.5582	0.048	-0.0104	0.2051	0.036	0.299
$\gamma_{21} = 0,2$	1	1000	-0.0074	0.2610	0.054	-0.0056	0.1048	0.039	0.312
	1	4000	-0.0062	0.1201	0.045	-0.0027	0.0528	0.049	0.298
	10	250	0.0000	0.4893	0.045	-0.0135	0.1964	0.037	0.335
	10	1000	-0.0117	0.2352	0.054	-0.0068	0.1007	0.042	0.342
	10	4000	-0.0082	0.1103	0.049	-0.0041	0.0509	0.050	0.328
	0.1	250	0.1081	1.1154	0.085	0.0085	0.2642	0.048	0.191
	0.1	1000	0.0373	0.4117	0.051	0.0003	0.1274	0.039	0.184
	0.1	4000	0.0324	0.1879	0.051	0.0032	0.0628	0.050	0.170
	1	250	0.0486	0.5757	0.055	0.0031	0.2216	0.050	0.311
$\gamma_{22} = 0,2$	1	1000	0.0230	0.2679	0.061	-0.0007	0.1066	0.045	0.304
	1	4000	0.0194	0.1255	0.063	0.0008	0.0520	0.038	0.298
	10	250	0.0377	0.5022	0.048	-0.0039	0.2119	0.053	0.337
	10	1000	0.0159	0.2429	0.049	-0.0048	0.1019	0.049	0.331
	10	4000	0.0140	0.1145	0.049	-0.0026	0.0498	0.043	0.311

Note : Computation obtained with 1000 simulations.
The level of test reported are the estimation of the actual tests for a nominal level of 5%.

$$\widehat{V}_n = \frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\left[\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h^{(1)}_{itsml}(\widehat{\theta}_n)\right]\left[\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h^{(1)}_{itsml}(\widehat{\theta}_n)'\right]$$

or

$$\widetilde{V}_n = \frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\left[\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)^2 h^{(1)}_{itsml}(\widehat{\theta}_n)h^{(1)}_{itsml}(\widehat{\theta}_n)'\right]$$

The proof is very close to the proofs of theorems 1 to 3, of Honoré and Kyriazidou Honoré & Kyriazidou, 2000 :
We will use two useful results :

- (Kernel estimators) If Z is a random variable iid across individuals such that $E\left(Z|\Delta^{t,s}x_i = x\right)$ exists and is twice differentiable in a neighborhood of $x = 0$, and such that $E\left(Z|\Delta^{t,s}x_i = x; \Delta^{t',s'}x_i = x'\right)$ exists and is twice differentiable in a neighborhood of $(x, x') = (0, 0)$ for $(t, s) \neq (t', s')$. Then

$$\forall \nu \geq 0, \quad E\left(\frac{1}{\sigma_n^k}K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)^{\nu} Z_i\right) =$$

$$f_{t,s}(0)E\left(Z|x_{it+1} = x_{is+1}\right)\int K(u)^{\nu}du + O\left(\sigma_n^2\right)$$

$$E\left(\frac{1}{\sigma_n^{2k}}K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)K\left(\frac{\Delta^{t',s'}x_i}{\sigma_n}\right) Z_i\right) =$$

$$f_{t+1,s+1,t'+1,s'+1}(0,0)E\left(Z_i|x_{it+1} = x_{is+1}; x_{it'+1} = x_{is'+1}\right) + O\left(\sigma_n^2\right)$$

- (Corollary 2.2, Newey (1991)) If $\mu_n(\theta)$ is a sequence of random differentiable function such that for all $\theta \in \Theta$, $\mu_n(\theta) \underset{P}{\to} \mu(\theta)$, and if the derivative of $\mu_n(\theta)$ are dominated by a random variable U_n such that $U_n = O_p(1)$ and $E(U_n) < \infty$, the convergence in probability is uniform on the compact Θ. For $j = 0, 1, 2$, note that the sequence of random function $\frac{1}{n\sigma_n^k}\sum_{i=1}^{n}\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h^{(j)}_{itsml}(\theta)$ verify the condition of domination.

Let $m_i(\sigma_n, \theta) = \sum_{2\leq t<s\leq T-1, m\neq l} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h_{itsml}(\theta)$ and $M_n(\theta) = \frac{1}{n\sigma_n^k}\sum_{i=1}^{n} m_i(\sigma_n, \theta)$.
We have : $E\left[M_n(\theta)\right] \to \sum_{2\leq t<s\leq T-1, m\neq l} f_{t,s}(0)E\left[h_{itsml}(\theta)|x_{i,t+1} = x_{i,s+1}\right] = M(\theta)$
To prove consistency we use Theorem 5.7 of Van Der Vaart (1998), the first assumption we need to verify is a stochastic uniform convergence $\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \xrightarrow{P} 0$.
Because

$$\begin{aligned} V\left(M_n\left(\theta\right)\right) &= \tfrac{1}{n\sigma_n^{2k}} V\left(m_i(\sigma_n,\theta)\right) \\ &\leq \tfrac{1}{n}\sum_{\substack{2\leq s<t\leq T-1 \\ 2\leq s'<t'\leq T-1 \\ (s,t)\neq(s',t')}} \sum_{\substack{m\neq l \\ m'\neq l'}} \tfrac{1}{\sigma_n^{2k}} E\left(K\left(\tfrac{\Delta^{t,s}x_i}{\sigma_n}\right) K\left(\tfrac{\Delta^{t',s'}x_i}{\sigma_n}\right) h_{itsml}(\theta) h_{it's'm'l'}(\theta)\right) \\ &\leq \tfrac{1}{n}\left(O(\tfrac{1}{\sigma_n^k}) + O(1)\right) = O\left(\tfrac{1}{n\sigma_n^k}\right) = o(1) \end{aligned}$$

$M_n\left(\theta\right) \xrightarrow{L_2} M\left(\theta\right)$ for all $\theta \in \Theta$, so $M_n\left(\theta\right) \xrightarrow{P} M\left(\theta\right)$ for all $\theta \in \Theta$. Uniformity of the convergence on Θ is ensured by domination.

The second assumption of Theorem 5.7 of Van Der Vaart (1998) is that θ_0 is a well-separated point of maximum of M : $\sup_{\theta:d(\theta,\theta_0)\geq\varepsilon} M\left(\theta\right) < M\left(\theta_0\right)$

$$M\left(\theta\right) = \sum_{2\leq t<s\leq T-1, m\neq l} f_{t,s}(0) P\left(\left\{y_{it}, y_{is}\right\} = \left\{l,m\right\} | x_{t+1} = x_{s+1}\right) g_{tsml}\left(\theta\right)$$

with $g_{tsml}\left(\theta\right) = E\left(ln\left(\frac{\exp(y_{mit}Z\theta)}{1+\exp(Z\theta)}\right) | x_{t+1} = x_{s+1}; \left\{y_{it}, y_{is}\right\} = \left\{l,m\right\}\right)$.

If $s = t+1$, $\theta \mapsto E\left(ln\left(\frac{\exp(y_{mit}Z\theta)}{1+\exp(Z\theta)}\right) | x_{t+1} = x_{s+1}; \left\{y_{it}, y_{is}\right\} = \left\{l,m\right\}; \left(y_{i\tau}\right)_{\tau\neq t,s}\right)$ is well-separated for the component $\beta_m, \beta_l, \gamma_{y_{i,t-1}m}, \gamma_{y_{i,t-1}l}, \gamma_{my_{i,s+1}}, \gamma_{ly_{i,s+1}}, \gamma_{lm}, \gamma_{ml}$ and does not depend on other component of θ. It is well-separated for the component β_m, β_l, $\gamma_{y_{i,t-1}m}$, $\gamma_{y_{i,t-1}l}$, $\gamma_{my_{i,s+1}}$, $\gamma_{ly_{i,s+1}}$, $\gamma_{my_{i,t-1}}$, $\gamma_{ly_{i,t-1}}$, $\gamma_{y_{i,s+1}m}$, $\gamma_{y_{i,s+1}l}$ and does not depend on other component of θ if $s > t+1$. Then, $g_{tsml}(\theta)$ is well-separated for the component β_m, β_l and $\left(\gamma_{qm}, \gamma_{ql}, \gamma_{mq}, \gamma_{lq}\right)_{q\in[1,M]}$ and does not depend on the other components. Because $f_{t,s}(0)P\left(\left\{y_{it}, y_{is}\right\} = \left\{l,m\right\} | x_{t+1} = x_{s+1}\right)$ are positive quantities for every 4-uplet s,t,l,m, $M\left(\theta\right)$ is well-separated for θ.

Theorem 5.7 of Van Der Vaart (1998) implies the consistency of the estimate.

Now let's focus on the asymptotic normality. For that, we use Taylor expansion of the first order condition :

$$\begin{aligned} 0 &= \tfrac{1}{\sqrt{n\sigma_n^k}}\sum_{i=1}^n \left\{\tfrac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta} - E\left(\tfrac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}\right)\right\} \\ &+ \tfrac{1}{\sqrt{n\sigma_n^k}}\sum_{i=1}^n E\left(\tfrac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}\right) \\ &+ \tfrac{1}{\sqrt{n\sigma_n^k}}\sum_{i=1}^n \tfrac{\partial^2 m_i(\sigma_n,\theta_n^\star)}{\partial\theta\partial\theta'}(\widehat{\theta}_n - \theta_0) \end{aligned}$$

Let $\xi_{in} = \frac{1}{\sqrt{n\sigma_n^k}}\frac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}$. We use the Lindeberg-Feller central limit theorem (see Van Der Vaart (1998), proposition 2.27), to show that $\sum_{i=1}^n \left(\xi_{in} - E\left(\xi_{in}\right)\right)$ converge in distribution to $\mathcal{N}\left(0,V\right)$.

$$\begin{aligned}
\sum_{i=1}^{n} Cov(\xi_{in}) &= \tfrac{1}{\sigma_n^k} E\left(\tfrac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}\tfrac{\partial m_i'(\sigma_n,\theta_0)}{\partial\theta'}\right) - \tfrac{1}{\sigma_n^k} E\left(\tfrac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}\right) E\left(\tfrac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}\right)' \\
&= \sum_{\substack{2\leq t<s\leq T-1 \\ m\neq l}} f_{t,s}(0) E\left(h^{(1)}_{itsml} h^{(1)'}_{itsml} | x_{it+1} = x_{is+1}\right) \int K^2(u)du \\
&\quad + \sigma_n^k \sum_{\substack{2\leq s<t\leq T-1 \\ 2\leq s'<t'\leq T-1 \\ (s,t)\neq(s',t')}} \sum_{\substack{m\neq l \\ m'\neq l'}} f_{t,s,t',s'}(0,0) E\left(h^{(1)}_{itsml} h^{(1)'}_{it's'm'l'} | \Delta^{t,s} x_i = \Delta^{t',s'} x_i = 0\right) \\
&\quad + O(\sigma_n^2) + O(\sigma_n^{k+2}) + O(\sigma_n^{k+4}) \\
&= \sum_{\substack{2\leq t<s\leq T-1 \\ m\neq l}} f_{t,s}(0) E\left(h^{(1)}_{itsml} h^{(1)'}_{itsml} | x_{it+1} = x_{is+1}\right) \int K^2(u)du + O(\sigma_n^{\min(2,k)})
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} E(||\xi_{in}||^2 \mathbb{1}_{\{||\xi_{in}||>\varepsilon\}}) &\leq n\varepsilon^{-\delta} E\left[||\xi_{in}||^{2+\delta}\right] && \text{(Markov)} \\
&\leq \tfrac{n^{-\delta/2}\varepsilon^{-\delta}}{(\sigma_n^k)^{(2+\delta)/2}} \left(\tfrac{(T-2)(T-3)}{2}\right)^{1+\delta} \\
&\quad E\left[\sum_{\substack{2\leq t<s\leq T-1 \\ m\neq l}} K\left(\frac{\Delta^{t,s} x_i}{\sigma_n}\right)^{2+\delta} ||h^{(1)}_{itsml}(\theta)||^{2+\delta}\right] && \text{(Hölder)} \\
&\leq O\left(\tfrac{1}{\sqrt{n\sigma_n^k}^{\delta}}\right) = o(1)
\end{aligned}$$

It follows from the Lindeberg-Feller theorem (see, for instance Van Der Vaart (1998), chapter 2), that $\sum_{i=1}^{n} \xi_{in} - E(\xi_{in}) \to \mathcal{N}(0, V)$ with
$V = \sum_{\substack{2\leq t<s\leq T-1 \\ m\neq l}} f_{t,s}(0) E\left(h^{(1)}_{itsml} h^{(1)'}_{itsml} | x_{it} = x_{is}\right) \int K^2(u)du.$

Let $b_n = \frac{1}{n\sigma_n^k} \sum_{i=1}^{n} E\left(\frac{\partial m_i(\sigma_n,\theta_0)}{\partial\theta}\right)$, we have : $b_n = O(\sigma_n^2)$

And then $\sqrt{n\sigma_n^k} b_n = o_p(1)$.

Let $J_n(\theta) = \frac{1}{n\sigma_n^k} \sum_{i=1}^{n} \frac{\partial^2 m_i(\sigma_n,\theta)}{\partial\theta\partial\theta'}$. For all $\theta \in \Theta$, we have :

$$\begin{aligned}
E(J_n(\theta)) &= \frac{1}{\sigma_n^k} E\left(\sum_{\substack{2\leq t<s\leq T-1 \\ m\neq l}} K\left(\frac{\Delta^{t,s} x_i}{\sigma_n}\right) h^{(2)}_{itsml}(\theta)\right) \\
&= \sum_{\substack{2\leq t<s\leq T-1 \\ m\neq l}} f_{t,s}(0) E\left[h^{(2)}_{itsml}(\theta) | x_{it+1} = x_{is+1}\right] + o_p(1) \\
&= J(\theta) + o_p(1)
\end{aligned}$$

The variance of the jj''th component of $J_n(\theta)$ decrease to 0.

$$\begin{aligned} Var\left(J_n^{[j,j']}(\theta)\right) &\leq \frac{1}{n}E\left[\left(\frac{1}{\sigma_n^k}\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h_{itsml}^{(2)[j,j']}(\theta)\right)^2\right] \\ &\leq \frac{1}{n\sigma_n^k}\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} f_{t,s}(0)E\left(h_{itsml}^{(2)[j,j']}(\theta)^2 | x_{it+1}=x_{is+1}\right)\int K^2(u)du \\ &\quad +\frac{1}{n}\sum_{\substack{2\leq s<t\leq T-1\\ 2\leq s'<t'\leq T-1\\ (s,t)\neq(s',t')}}\sum_{\substack{m\neq l\\ m'\neq l'}} f_{t,s,t',s'}(0,0)E\left(h_{itsml}^{(2)[j,j']}h_{is't'm'l'}^{(2)[j,j']}|\Delta^{t,s}x_i=\Delta^{t',s'}x_i=0\right) \\ &\quad +O(\tfrac{\sigma_n^2}{n\sigma_n^k})+O(\tfrac{\sigma_n^2}{n}) \\ &= O\left(\frac{1}{n\sigma_n^k}\right)=o(1) \end{aligned}$$

So for all $\theta\in\Theta$ $J_n(\theta)=J(\theta)+o_p(1)$. Using the second part of the preliminary remark, the convergence is uniform for $\theta\in\Theta$. We deduce that $\widehat{J}_n=J_n\left(\widehat{\theta}_n\right)=J(\theta_0)+o_p(1)=J+o_p(1)$
Similarly, if we note

$$V_n(\theta)=\frac{1}{n\sigma_n^k}\sum_{i=1}^n\left[\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h_{itsml}^{(1)}(\theta)\right]\left[\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right) h_{itsml}^{(1)}(\theta)'\right]$$

or

$$V_n(\theta)=\frac{1}{n\sigma_n^k}\sum_{i=1}^n\left[\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} K\left(\frac{\Delta^{t,s}x_i}{\sigma_n}\right)^2 h_{itsml}^{(1)}(\theta)h_{itsml}^{(1)}(\theta)'\right]$$

Then $V_n(\theta)$ converges uniformly on Θ to
$V(\theta)=\sum_{\substack{2\leq t<s\leq T-1\\ m\neq l}} f_{t,s}(0)E\left(h_{itsml}^{(1)}(\theta)h_{itsml}^{(1)'}(\theta)|x_{it}=x_{is}\right)\int K^2(u)du$
And we conclude that $V_n\left(\widehat{\theta}_n\right)=V(\theta_0)+o_p(1)=V+o_p(1)$.

Chapter 5

Evaluation of the "Ambition Success Network" (Réseaux Ambitions Réussite)

Education pursues two goals: offering equal opportunities to all, and promoting success for each student. In France, however, as in other countries, large achievement inequality has persisted since the 60's, mainly due to heterogenous parental backgrounds.(see for instance Coleman et al. (1966) for the USA and Plowden & the Central Advisory Council for Education (1967) for the UK). As a result many countries have set up compensatory education programs to foster equality between pupils (van der Klaauw (2008), Heckman & al. (2010)).

In France the main compensatory policy - the Politique d'Education Prioritaire, or Preferential Educational Policy - has targeted schools rather than students since its introduction in 1982.[1] This school-based program, *Educational Priority Zones* (ZEP), was overhauled twice and eventually replaced in 2006 by the *Réseaux Ambition Réussite* (RAR) policy, or *Ambition Success Network* program, the focus of this paper.

The ZEP policy came under criticism for spreading itself too thinly: too many schools were given too few funds. As a result, many evaluations of the ZEP policy conclude that this costly program had no positive effect on student achievement (see Benabou et al. (2009), and Meuret (1994)) but a negative signaling effect instead. Hence the RAR program was intended to concentrate its funds on fewer schools: 249 junior high schools in 2006-2007

[1]Though pupil- or class-oriented policies are more widespread in developed countries, a few noticeable educational policies target schools. In the United States, Title I of the Elementary and Secondary Education Act funded schools and school districts with a high percentage of disadvantaged students in 1965. In the United Kingdom the *Education Priority Areas* (EPA) were launched in 1967, disappeared at the end of the 70's but following the victory of New Labour in the 90's were reborn in two new programs: *Education Action Zones* and *Excellence in the Cities*. They both targeted schools (see Machin & Vignoles (2005)).

against 900 for the ZEP policy.[2]

In this paper we provide a first evaluation of the recently implemented RAR program.

Any evaluation of school-targeting policies must deal with several methodological issues. First, treated schools are not randomly selected: they differ greatly from non-selected schools. We overcome this issue by using a regression discontinuity design that allows us to find valid counterfactual schools for both groups of treated schools. Second, it is not without interest to describe some aspects of the actual policy. Each school was supposed to get supplementary teachers. However, the real and announced policies may differ: we therefore need to study the number of teachers in treated and untreated schools. Third, the RAR program may have had negative signaling effects on both teachers and students. Labeling a junior high school as a RAR school may induce poor-quality teacher assignment and the departure of the best students. Therefore we study the program effect on both pupil-school and teacher-school matchings. Finally, we also compare final exam results between treated and non-treated schools. This mixes two effects: the change in teaching team efficiency, and the change in sorting on unobserved characteristics of pupils across schools.

We find negligible or adverse effects of the policy on the two groups of treated schools for which we are able to find credible counterfactuals: the first is centered on the disadvantaged student proportion threshold, the second on the repeating student rate threshold. For the first group of treated schools, the policy has had no effect on the teacher per pupil ratio, nor on teacher and pupil characteristics, but it has had a negative effect on pupil achievement. For the second group, we find an increased number of teachers in 6th and 7th grades (but not for other grades). This effect comes from an increase in the proportion of older teachers, and teachers with an unusual degree. Pupil structure has also been modified in these schools: we find an increase in the proportion of children from blue-collar backgrounds, counteracted by a decrease in the proportion of children with self-employed parents. Finally, for both groups of treated schools, final junior high-school scores have worsened in RAR assigned schools.

The rest of this paper is organized as follows. Section 2 describes the RAR program. Section 3 develops the identification and estimation strategies. Section 4 presents the results and Section 5 concludes.

[2] A network was made up of a junior high-school (from 6th to 9th grade) and a few primary and/or nursery schools (from 1st to 5th grade, or before 1st grade).

5.1 RAR program: design and background

5.1.1 A brief history

In 1982 a new program - *Educational Priority Zones* (ZEP) - was conducted in France. This policy's main objective was to increase efforts in unsuccessful zones to reduce inequalities. These new education priority zones aimed to set up an educational project that would provide support to underachieving students. In 1985, its main focus was redirected to address deficiencies in 'core learning' such as reading and French. Afterwards successive reforms were introduced, each one expanding the number of schools concerned, but without increasing further inputs in already treated schools.

Benabou et al. (2004) draw three main conclusions concerning these Education Priority Zones. First they argue that the subsidies were divided between too many schools and mainly given to teachers via supplementary wages without any overtime teaching. Hence the actual per capital allocation of funds to pupils was scarce. Secondly, the authors find that the treated junior high schools experienced a decrease in their total number of students and an increase in the proportion of socially disadvantaged pupils. Teachers also migrated from these schools. Their turnover increased after the assignment of these schools to priority zones. Finally, Benabou et al. (2004) found no significant effect on different measures of student achievement, or on high-school graduation. These disappointing results suggest a restructuring of the ZEP program that better targets efforts and funds.

5.1.2 The Ambition Success Network policy

As a consequence, the reform introduced in 2006 pursued the goal of better targeting funds and efforts. The education priority map was reshaped, and the resources were given to a smaller number of schools (249 junior high schools in 2006-2007 against around 900 with the previous policy). New zones were defined and named "Réseaux Ambition Réussite" (Ambition Success Networks). These networks are made up of one junior high-school (6th to 9th grade) and some primary and/or nursery schools (1st to 5th grade, or before 1st grade). These entities share a common project under the guidance of a committee composed of the heads of the junior high-schools and some representatives of the primary or nursery schools. At the beginning of the 2006-2007 school year, 249 networks were created, consisting of 249 junior high-schools and 1,715 elementary schools. It represented 126,000 pupils in junior high schools: one junior high school student in twenty was enrolled

in a RAR targeted school.
The Ministry of Education recommended to its regional heads ("recteurs") that each network get three to four supplementary teachers, some teaching assistants and at least one full-time nurse. On a national level, this amounted to 1,000 supplementary teachers, and 3,000 teaching assistants. However considerable discretion was left to regional heads. Although the list of "RAR" junior high-schools was decided in a concerted way, the list of primary and nursery schools, and the resources devoted to each network were chosen by the regional heads. Moreover the type of educational services provided was left to the discretion of the junior high schools and regional heads, with no requirement for accountability.
The number of RAR networks has evolved since their introduction. From 249 in 2006-2007, there were 254 public junior high-schools at the beginning of the 2008-2009 school year and 118,000 junior students. There are large regional discrepancies: the proportion of "RAR" junior high-schools ranges from 0.4% for the regional area of Grenoble to 13% for the regional area of Aix-Marseille. There are also a few private "RAR" junior high schools (11 in 2008-2009). Networks or zones that were not targeted by the new policy in 2006 constitute a different category of networks called "networks of school success" ("Réseaux de Réussite Scolaire", RRS). In 2007-2008, of the 253 RAR junior high schools, 238 had previously been "ZEP". The previous denomination "ZEP" (theoretically) disappeared in September 2008.

5.1.3 How were the RAR schools selected?

Internal notes of the Statistical Service of the Education Ministry provide information about the selection process of secondary schools. RAR networks were chosen on the basis of three main criteria evaluated for the 2004-2005 school year. First, the proportion of socially disadvantaged[3] students in 6th grade had to be equal to or above 67%. Second, either of the two following criteria had to be met: the proportion of students who have repeated two grades or more when they entered 6th grade had to be at or above 10%; or the school average score at the entrance evaluation of 6th grade had to be at or below 47%. Additional criteria that were used to define the final list of RAR junior high schools included the local unemployment rate and the proportion of people benefiting from social assistance. Instead of the 164 initially chosen with the three criteria previously mentioned, 249 were finally selected after an agreement between the French Ministry of Education and

[3]A child was classified as 'disadvantaged' when their referring parent was either a blue collar worker, or retired from a blue collar or white collar occupation, or out of work. This criterion was calculated from the occupation covariate coded on two digits in the data files.

its regional heads.

5.1.4 Expected results

If we assume that 1,000 supplementary teachers were uniformly assigned across new RAR junior high schools, then we would expect an 8% increase in the total number of teachers in RAR schools versus non-RAR schools. As mentioned before however, regional head masters have considerable discretion. For instance, even if the Ministry of Education gives them additional teachers that must be specifically assigned within the RAR junior-high-schools, the regional head officials can assign them as they see fit. In particular, teachers already working in a RAR school may be transferred to a non-RAR school to ensure the stability of the total number of teachers both in RAR and non-RAR schools. This eviction phenomenon is credible if regional head officials are reluctant to treat similar schools differently: e.g. two schools just below and above the selection thresholds were alike before program assignment.

Some parents may have interpreted the RAR assignment as a negative signal, as a consequence the number of pupils in treated schools may have decreased. Benabou et al. (2004) and Benabou et al. (2009) have discussed such a decrease for ZEP-schools during the 90's. In that case, the teacher-pupil ratio may increase more than expected.

If the teacher-pupil ratio increases enough, the average class size may decrease. The effect of class-size on pupil achievement, using the seminal method of Angrist & Lavy (1999), is often reported as negative (see Angrist & Lavy (1999), Leuwen et al. (2008), and for the French case, Gary-Bobo & Mahjoub (2006) and Piketty & Valdenaire (2006)). Another group of estimates comes from the Tennesse STAR class size experiment. Word et al., (1990), Finn & Achilles (1990), Krueger (1999), and Krueger & Whitmore (2001) all found that smaller class sizes have a significant and lasting impact on academic achievement and educational attainment.

Another issue relates to the teacher quality in RAR schools. First, supplementary teachers can have some specific characteristics. Previous studies on ZEP policies show that the new teachers in treated schools are often younger and have less experience. It can be argued that the assignment of schools to the RAR-program acts as a negative signal for the teachers as well, who would then prefer to be migrate to non-treated schools. However, teacher mobility and tenure are closely related: the less experience they have, the lower their bargaining power to move. Moreover in the previous ZEP program, teachers willing to work within ZEP schools were promoted quicker to encourage them to teach in such

schools. However bigger promotions ease the transfer to another school if requested. And it was noticed indeed that the turnover of the teaching teams was higher in ZEP schools, which can be harmful to school management. Ly (2010) observed that in 1999, the ZEP assignment change induced the mobility of oldest teachers from ZEP to non-ZEP schools. Finally the policy can modify the sorting of pupils across schools. Benabou et al. (2004) and Benabou et al. (2009) found that school discrepancy between treated and untreated schools increases after assignment: they measure school discrepancy through the proportion of pupils not enrolled in the school cafeteria. Other studies show that parents take into account school choice in their decision to move or relocate (Fack & Grenet (2010) for the case of Paris). More generally, sociologists have reported the existence of parental strategies concerning the schooling of their children (François & Poupeau (2004)) and economists have tried to quantify the valuation of schools (Black (1999), Fack & Grenet (2010), Gibbons & Machin (2003)). All these parental strategies are based on available information about schools and assignment to the RAR program can affect parental choice. Even if one can not *a priori* exclude the possibility that parents give a higher valuation to RAR-schools because such schools get higher resources, empirical results in the literature suggest that school-based discrimination increases segregation. Should there be such an increase in segregation the effect on average test scores in the final exam is muddled: the policy may increase the efficiency of treated schools, but such an increase in efficiency may be compensated for by a sorting effect. Our data only allow us to compare average results across schools. This is a clear limit of this paper and the reader must bear in mind that the estimates on the final score exam mix the two effects.

5.2 Data and some descriptive statistics

To analyze the effectiveness of the RAR program, we use school-level data collated from various administrative sources of the Ministry of Education:

- The first dataset is an exhaustive pupil-level cross-sectional dataset (*Scolarité*) for every student in junior or secondary high school. This data provides cross-sectional information about age, nationality, residence location, main parent's occupation, class, languages and other options, school lunch status, and the same variables for the previous year.

- The second dataset is an exhaustive teacher-level panel dataset (*Relais*) for every teacher in junior or secondary high school. This provides information on the total

number of hours taught by each teacher, for each subject they teach, in each school, along with their age and degree.

- We supplement the pupil-level data with a third, exhaustive, pupil-level dataset that contains their national exam scores, (*Brevet des Colleges*), taken at the end of the 9th grade. We are unable to fully merge the two pupil-level datasets due to the absence of a unique student identification number that could combine them both. We are however able to provide a distribution of test scores for every school and for every year by combining these two datasets.

- The fourth dataset is an exhaustive panel dataset of French junior and secondary high schools that includes information on their Educational Priority policy status and their RAR eligibility criteria.

We construct a school-level dataset for school years 2003/2004 up to 2008/2009. We restrict our attention to public junior high schools in mainland France. This is because private schools are almost never assigned to treatment, and because in French overseas departments and territories, junior-high schools are almost always assigned to treatment. Moreover private schools, which represent around 20% of pupils, differ substantially from public schools on various dimensions. This is also true for overseas schools. We thus obtained an exhaustive panel of around 5000 public junior high schools in metropolitan France. Among these schools, 206 were affected to the RAR program.
We exclude educationally disadvantaged students who have severe and long-running problems with core learning as they belong to special classes called "Segpa" (*section d'enseignement général et professionnel adapté*). Special funds are dedicated to them, but they do not interfere with the resources allocated to the educational priority programs. Similarly, when we evaluate the average number of pupils per class, or the social structure of pupils moving up into 6th grade, we exclude these students.[4]
Finally, for final exam test scores, we focus our attention on formal written literacy and maths tests. Every year from 2003 to 2009, final exams took place in June for pupils in 9th grade. For each school, we calculate average and quantiles of test score distribution. For these computations, we exclude disabled pupils[5] and pupils following vocational curricula.

[4]Schools have indeed specific resources for SEGPA classes and we're interested in the extra resources allocated to schools by the RAR program that are not allocated to SEGPA students. SEGPA pupils are within junior high schools but in different teaching structures, in separate classes, with different teachers. Moreover the proportion of SEGPA pupils has the same distribution below and above both considered thresholds. Hence excluding SEGPA students from teaching, score, and resource indicators does not lead to biased estimates.

[5]The way they take the national exam is different from the non-disabled.

Table 5.1 presents pre-treatment descriptive statistics for the whole sample of public junior high schools in mainland France. RAR public junior high schools have a higher proportion of pupils entering 6th grade after repeating a grade. A junior entering 6th grade is usually 11 years old. Four pupils out of 10 in RAR schools had repeated at least one grade when entering their 6th grade in September 2005, only two out of 10 were in this case in non-RAR schools. This discrepancy is roughly the same after 2006, even though the proportion of students having repeated a grade has fallen since 2003 due to a no repeating grade policy in the French educational system. In the 6th and 9th grades, socially disadvantaged pupils are overrepresented in RAR schools. Only 2.1% of juniors entering the 6th grade in RAR schools belong to a family whose head[6] is an executive. This proportion is to 15.9% for non-RAR schools. Unlike 'executive-children', 'unemployed-children' and 'blue-collar-children' are over-represented in RAR schools. In grade 6 in 2006/2007, over 27% of children in RAR schools have an unemployed parent compared to non-RAR schools where this figure is approximatively 9%.

Table 5.1 also presents proxies of average expenditure per pupil, namely average class size and average weekly hours of teaching per pupil. One year before treatment, average class size was around 21 in RAR schools and 24 in non-RAR schools. Similarly, average per-pupil weekly teaching hours was about 12% higher in RAR schools (1.41 versus 1.24 in non-RAR schools). In non reported results we find that this figure holds for all grades: many RAR schools previously benefited from the ZEP program, hence from a higher teacher-pupil ratio.

Teachers are less qualified and are younger in RAR junior high schools. The proportion of physical education teachers was larger in RAR schools.

Table 5.1 provides a comparison between RAR and non-RAR schools in maths and literacy scores as obtained in the final national exam. Each student at the end of his/her 9th grade takes written tests in maths and French. Scores range between 0 and 40. For each junior high school, we computed average test scores in maths and French. The comparison of the average scores highlights the huge discrepancy between RAR and non-RAR schools prior to treatment. For the exam taken in 2006, the average test score in French for RAR schools was 14.32 against 19.01 for non-RAR schools. The same holds in maths. Comparison of the quantiles (not reported here) shows that this result holds for the entire distribution of test scores.

The descriptive statistics show that a naive comparison between the treated and non-treated schools is unsatisfactory to test efficiency of the RAR program. Conditioning

[6]In the data, only the occupation of one of the two parents is available. This parent, who is usually the father, is called the family head.

on some important observable covariates may mitigate bias but even in this case, the interpretation is likely to be misleading: assignment to treatment may have been driven by some unobserved variables. To overcome this issue, one difference-in-difference strategy may be adopted using the panel dimension, but this would be valid only if temporal trends are the same between treated and non-treated schools. Simple panel-data regressions reported in Table 5.2 show that temporal trends between 2003 and 2005 differ significantly (at the 5% level) for many variables.[7] At baseline, we find that the gap in the proportion of 13-year-olds enrolling into grade 6 falls. This however is not true for children whose parents are executives or blue-collar workers. The gap in average class size between RAR and non-RAR schools also increases, and structures of teacher qualifications and age also evolve differently in the treated and non-treated schools before the beginning of the policy. These different evolutions rule out the use of a difference-in-difference estimation strategy.

5.3 Identification and estimation strategy

5.3.1 A fuzzy regression discontinuity design

As explained in the previous section, the assignment to RAR treatment is based on thresholds of some predetermined variables. These key features of the RAR-policy allow us to use a fuzzy regression discontinuity design to estimate the causal effects of the RAR program on different outcomes such as means, but also signaling and sorting. Such a discontinuity is clearly supported by Figure 5.1.

This discontinuity in the probability of being treated implies that local average treatment effects ($LATE$) are nonparametrically identified (Hahn et al. (2001)). The basic idea is to compare the outcomes of junior high schools just above and just below the thresholds used to assign treatment. Multiple assignment criteria allow us to identify several parameters. We can identify and estimate nonparametrically the LATE for junior high schools having around 67% of socially disadvantaged pupils and having more than 10% of pupils having repeated a grade twice before Grade 6. We can also identify the LATE for junior high schools having around 10% of repeating pupils and with more than 67% of socially disadvantaged pupils.

To describe the identification approach used in this paper, let T_i be an indicator equal

[7] To be more precise, trends of treated and untreated schools differ significantly for at least one modality of every categorical variable considered: parental profession, teacher characteristics and age of pupils at beginning of 6th grade.

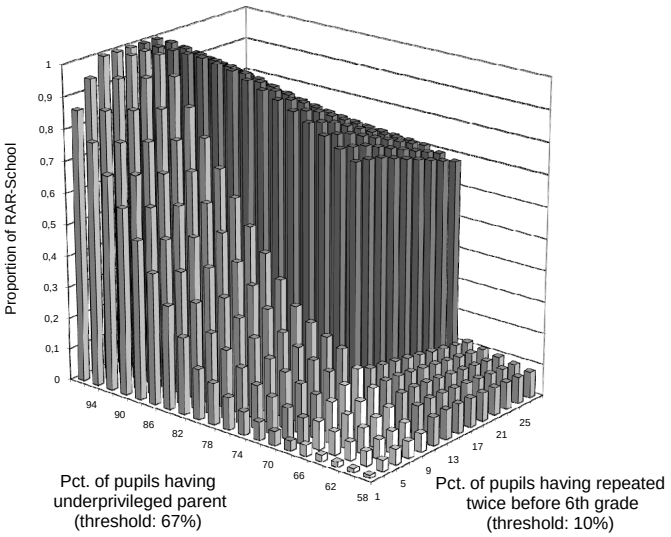
Table 5.1: School Characteristics in September 2005, for treated and non-treated schools

School Characteristics	non-RAR	RAR	t-stat
ZEP before 2006	0.13	0.99	36.36
Number of pupils	461.46	428.07	-2.63
Pupils entering sixth grade			
% of male	50.73	50.87	0.35
% 10 years old or younger	2.95	1.48	-9.53
% 11 years old	76.85	58.30	-30.62
% 12 years old	19.15	36.72	29.71
% 13 years old or older	1.05	3.50	19.93
Farmer	2.95	0.17	-7.62
Self-employed	8.66	3.78	-13.57
Executive	15.88	2.08	-15.46
Intermediate	14.68	5.44	-20.49
Employee	15.77	11.21	-9.01
Blue collar	29.57	41.26	12.05
Retired	1.22	3.63	19.68
Unemployed	8.60	27.31	36.89
Pupil supervision			
Hours of teaching per pupil	1.24	1.41	17.75
Class size	24.00	20.96	-20.16
Teaching Staff (% of hours dispensed by)			
Highest teaching degree	0.04	0.04	-2.74
Qualified teacher	0.72	0.72	-0.56
PE teacher	0.10	0.11	2.21
Other teacher	0.13	0.14	1.17
Teacher under age 30	0.15	0.27	14.73
Teacher between 30 and 40	0.31	0.37	7.27
Teacher between 40 and 55	0.35	0.25	-11.87
Teacher over 55	0.44	0.27	-16.59
Average results at the Brevet exam (Grade 9, June 2006)			
French Score	19.01	14.32	-24.47
Maths Score	18.32	11.85	-18.93
Number of schools	4,795	205	

Table 5.2: Temporal trend of school characteristics before September 2005, for treated and non-treated schools

School Characteristics	non-RAR	RAR	t-stat
Number of pupils	-10.00	-11.80	-1.56
Pupils entering sixth grade			
% of male	-0.06	0.16	0.80
% 10 years old or younger	0.16	0.08	-0.84
% 11 years old	0.41	0.74	1.42
% 12 years old	-0.48	-0.39	0.45
% 13 years old or older	-0.09	-0.44	-5.33
Farmer	-0.07	-0.04	0.30
Self-employed	0.24	0.15	-0.55
Executive	0.31	-0.17	-2.55
Intermediate	-0.14	-0.06	0.42
Employee	-0.02	0.33	1.59
Blue collar	-0.39	-1.00	-2.16
Retired	0.04	0.00	-0.59
Unemployed	0.06	0.91	4.71
Pupil supervision			
Hours of teaching per pupil	0.00	0.00	0.10
Class size	0.02	-0.09	-2.30
Teaching Staff (% of hours dispensed by)			
Highest teaching degree	0.00	0.00	-2.23
Qualified teacher	0.01	0.02	2.78
PE teacher	0.00	0.00	-0.08
Other teacher	-0.01	-0.02	-1.78
Teacher under age 30	-0.01	-0.01	-0.60
Teacher between 30 and 40	0.01	0.01	2.31
Teacher between 40 and 55	-0.02	-0.01	2.26
Teacher over 55	0.00	0.00	-1.17
Average results at the Brevet exam (Grade 9, before June 2005)			
French Score	-0.45	-0.59	-1.85
Maths Score	-0.46	-0.64	-1.53
Number of schools	4,795	205	

Figure 5.1: RAR recipiency rate in 2006-2007



to 1 if school i is treated. $Y_i(0)$ denotes the potential outcome if the junior high school i was non-treated ($T_i = 0$), and $Y_i(1)$ if it was treated ($T_i = 1$). Then the actual and observed outcome is $Y_i = Y_i(0) + T_i(Y_i(1) - Y_i(0))$. Let Z_i^F be the percentage of pupils coming from a disadvantaged family and Z_i^L be the percentage of pupils in 6th grade having repeated a grade at least twice. Discontinuities of the conditional regression function of the treatment T_i on the running variables (Z_i^F, Z_i^L) imply the existence of some complying schools for each running variable. This means that some schools are not treated when the values of the running variables are under thresholds, but are treated otherwise. For these complying schools, if Z_i^F and Z_i^L cannot be manipulated, the rules of assignation to the treatment generate an "as-good-as random assignment". We denote by the dummy C_L (respectively C_F) the dummies of being a complying school for the threshold $Z^L = 10\%$ (respectively for the threshold $Z^F = 67\%$). Assuming continuity of conditional regression functions of potential outcomes on running variables, and monotonicity of treatment in a neighborhood of discontinuities (see Imbens & Lemieux (2008)), we can identify the

following local average treatment effects:

$$LATE_{F67} = \mathbb{E}\big(Y(1) - Y(0) | Z^F = 67\%, Z^L \geq 10\%, C_F = 1\big)$$

$$LATE_{L10} = \mathbb{E}\big(Y(1) - Y(0) | Z^L = 10\%, 80\% \geq Z^F \geq 67\%, C_L = 1\big)$$

We impose the restriction $80\% \geq Z^F$ on the second parameter because discontinuity vanishes when $Z_L = 10\%$ and Z_F close to 100% (see Figure 5.1).

5.3.2 Can the threshold be manipulated?

To get consistent estimates, the conditional regression functions of potential outcomes on the running variable (see Imbens & Lemieux (2008)) have to be continuous. This assumption may be violated if some agents, for instance headmasters or regional heads, are able to manipulate the running variables Z^L or Z^F. This is highly unlikely for the following reasons.

First, the statistical service of the Education Ministry collects family information about each pupil in junior high schools. The running variables Z^L and Z^F are calculated by aggregating this information at the school level and are not available within junior high schools.

Second, the statistical service has had considerable autonomy in choosing the eligibility criteria and the threshold values. It is unlikely that headmasters had access to or were aware of the various selection criteria chosen by the ministry. Moreover, the Education Ministry had ordered the release of the RAR list in 2006 based on information collected in September 2004. Therefore, in 2004, headmasters, collecting individual information about pupils, could not have anticipated the introduction of a policy two years beforehand, especially as the RAR program was the first one to be threshold-based. If by any chance they had heard of a future policy, the eligibility thresholds were not known in September 2004. As a consequence, we can be confident that manipulating parents' occupation, or entrance results to be assigned to RAR treatment was not feasible.

Finally, following the approaches of Saez (2010) and McCrary (2008), we studied the densities of Z^L and Z^F near the thresholds. Indeed, a monotonic manipulation of the running variable density-for instance if some schools just under the threshold manipulate the data and report being just over the threshold- must display a local minimum and a local maximum on either side of the threshold. Without any manipulation, there is no reason to

see any local extremum around the threshold.[8] Figures 5.2 and 5.3 provide strong evidence that potential manipulations of running variables are not troublesome in our case.

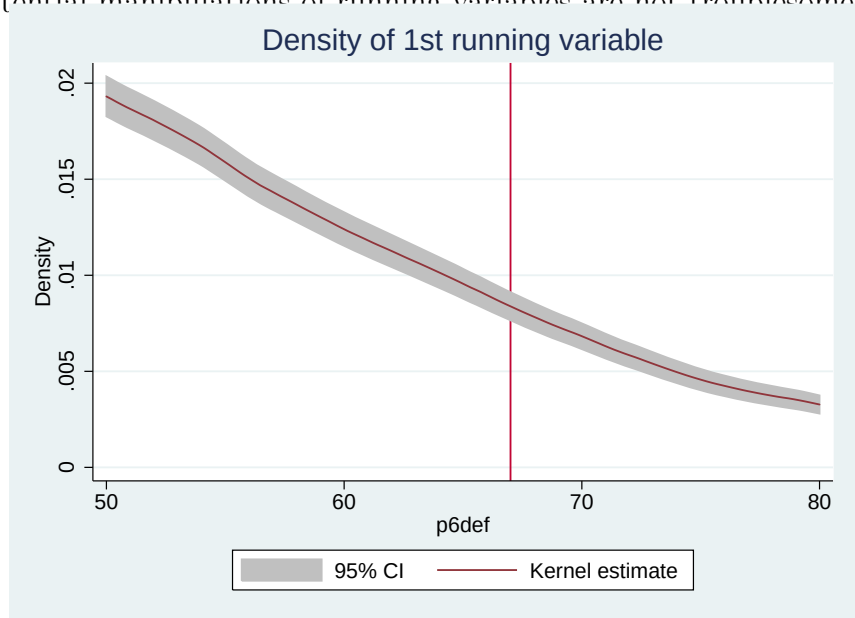


Figure 5.2: Density of the running variable Z^F around the threshold (67%)

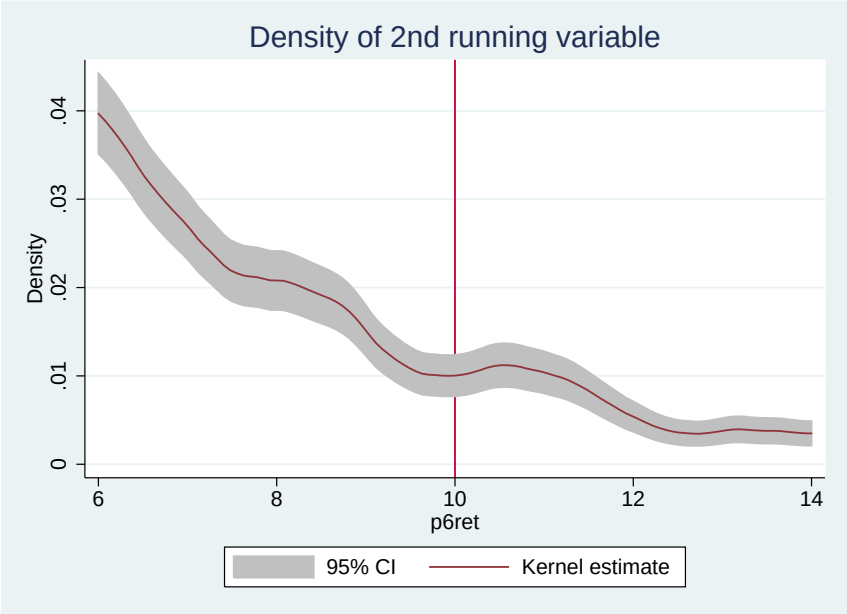


Figure 5.3: Density of the running variable Z^L around the threshold (10%)

[8]The absence of local extremum can also occur if upward manipulations compensate for downward manipulations. However in our case, such a coincidence seems very unlikely.

5.3.3 Advantages and drawbacks of the identification strategy

Some former evaluations of the ZEP program relied on the assumption that the treatment was exogenously assigned given a set of observable covariates. Some articles, such as Benabou et al. (2009), stand out because the authors use a difference-in-difference approach. Unless the treated and non-treated schools's different outcomes share common trends, this strategy is not valid. This assumption is indeed questionable: treated schools can be located within deprived areas. In France, the increasing geographical segregation during last the two decades has been demonstrated by many researchers (see, for instance, Maurin (2004)). We test the assumption of common trends before the RAR policy and, as mentioned in the previous section, Table 5.2 clearly rejects the assumption of common trends between treated and non-treated schools. Hence, as noticed by Gurgand in Benabou et al. (2004), though the difference-in-difference estimation may reduce bias, it cannot be measured.

Unlike the difference-in-difference approach, the regression-discontinuity method does not rely on such an assumption of common trend. Nonetheless the regression discontinuity estimates cannot be extrapolated to the whole population of treated schools, and they may be not precise enough. This is a concern when the effects we want to measure have a low magnitude, or when few observations are located near the threshold. To compensate for the lack of precision, we can use observations further away from the threshold, but at the cost of an increased bias.[9] Our strength is the panel data at hand. We thus add school fixed effects and time dummies, and they turn out to explain more than 75% of the variance of the outcomes we studied. As a result, we are able to statistically measure small effects. However, we need to check whether complying schools around the thresholds display common trends in the absence of treatment to ensure our approach is valid. We tested this assumption using observations before the RAR program was set up, from 2004 to 2006: no trend difference protrudes around both thresholds before September 2006.[10]
We estimate the LATE by a two stage panel least squares (TSLS) regression after selecting our data around both thresholds. This is equivalent to using a uniform kernel for the local linear regression, as suggested by Hahn et al. (2001). For observations such that $Z^L \in [10-h, 10+h]$ and $80\% \geq Z^F \geq 67\%$ and with $c = 10\%$, and for observations such that $Z^F \in [67-h, 67+h]$ and $Z^L \geq 10\%$ and with $c = 67\%$, we compute the following

[9] Adding covariates in the regressions decreases the variance of the estimates by the proportion of the explained variance of the outcomes by covariates.
[10] Common trend tests around both thresholds are available upon request.

TSLS estimates:

$$Y_{it} = \alpha_i + \beta_t + \gamma T_{it} 1\{t \geq 2006\} + \delta' V_{it} + \varepsilon_{it} \quad (5.3.1)$$

where $T_i.1\{t \geq 2006\}$ is the endogeneous covariate, $V_{it} = \begin{pmatrix} 1\{Z_i < c\}1\{t \geq 2006\}(Z_i - c) \\ 1\{Z_i \geq c\}1\{t \geq 2006\}(Z_i - c) \end{pmatrix}$ are the exogenous covariates and $1\{Z_i \geq c\}1\{t \geq 2006\}$ is the instrument.

The regressors V_{it} are introduced to avoid asymptotic bias in the estimates (Hahn et al. (2001), Imbens & Lemieux (2008)). Standard tests remain asymptotically valid when the regressors V_{it} are added to regressions.

It is important to notice that we have not induced any estimation bias by selecting our sub samples having checked that the repeating rate distribution was continuous around 67% disadvantaged threshold, (resp. the disadvantaged student proportion around 10% repeating threshold). The estimates we obtained are only local average treatment effects and we cannot answer for the efficiency of the global policy since we would have to impose unrealistic parametric assumption to identify the ATE on the whole public school set.

5.3.4 The outcomes

We provide an assessment of RAR treatment along different dimensions:

1. First, we consider outcomes that proxy for expenditures, to estimate the intensity of positive discrimination: average class size and weekly per-pupil teaching hours at the school level and for different grades (from Grade 6 to Grade 9).

2. A second set of outcomes considered are observable characteristics of pupils at the beginning of Grade 6. Hence, we examine pupil sorting across treated and non-treated schools. These are: family head occupation[11], dummy for attending cafeteria at lunch and the total number of pupils.

3. A third set of outcomes examines changes to the teaching structure. We focus on teacher-related outcomes such as the percentage of teaching hours per qualification, and the teacher age structure. We distinguish between four types of qualification:

[11] We distinguish 9 main occupations: farmers, self-employed occupations (artisans, shopkeepers, company managers), executives, intermediate occupations, administrative, sales or service occupation, blue collar, retired parents, out of the labor force

physical education (PE) teachers, post-graduate teachers ("agrégation", the top competitive examination), junior high school teachers ("certified" teachers recruited by a more open competitive examination) and finally, other teachers that have not been recruited by the usual competitive examinations (a priori less qualified). We also distinguish teachers by age, as it proxies tenure: less than 30, between 30 and 39, between 40 and 54, 55 and over.

4. The last set of outcomes is related to the academic achievement of pupils: we study the means and quantiles (Q10-Q25-Q50-Q75-Q90) of French and maths score distributions, at the final exam in Grade 9.

Studying these outcomes will provide us with an answer about any potential negative signaling effects of RAR treatment.
For all these outcomes, we test sensitivity to bandwidth choice h. We select the observations around both thresholds. For the percentage of disadvantaged pupils entering 6th grade, results are reported for $h = 4$, 6 and 8. For percentage of grade repeaters, the results are reported for $h = 2$, 3 and 4. Our results are robust to the choice of some alternative bandwidth values.

5.4 Results

5.4.1 Per pupil expenditure

We begin by reporting per pupil expenditure across RAR and non-RAR schools. If 1,000 supplementary teachers had been uniformly assigned across new RAR junior high schools, we would expect an 8% increase in the total number of teachers in RAR schools versus non-RAR schools. Tables 5.3a and 5.3b test this claim and present two types of indicators for different grades: the number of pupils per class and the number of teaching hours per week divided by the number of pupils (denoted resp. P/C and H/P, where P stands for *pupil*, H for *teaching hour*, and C for *class*).
We actually see that the 1,000 supplementary teachers had not been uniformly distributed across RAR junior high schools: the recommendation that each RAR junior high school receive 4 supplementary teachers was not strictly followed by each regional head. It can be inferred that teacher assignment was conducted on an ad-hoc basis, driven partly by regional heads and teachers' preferences. Furthermore regional heads and headmasters of junior high schools have discretion on how to allocate additional teachers to different

grades, which is why it is important to study the effect of RAR assignment on school resources, since we do not clearly know whether the policy has been followed by regional heads and school headmasters.

Average class size and per-pupil teaching hour estimations do not match political commitments. Specifically, results differ only very slightly across the two discontinuities, implying that treatment has only a weak effect on resources if any. For schools where $Z^F = 67\%$ and $Z^L \geq 10\%$ in 2004, we notice a small effect of the policy on class size reduction but the magnitude of this effect drastically decreases when estimation is made using a larger bandwidth (see Table 5.3a). The robustness of this measured effect is therefore questionable. A more robust and significant result concerns per-pupil teaching hours for Grades 6 and 7. The magnitude of this effect is more stable across bandwidth changes (see Table 5.3a). For schools with $Z^L = 10\%$ and $67\% \leq Z^F \leq 80\%$ in 2004, we find no significant effect on per-pupil teaching hours or on class size (see Table 5.3b). Per-pupil teaching hours decrease less than expected for all grades.[12] This may be explained by the preference of headmasters for concentrating compensatory policy on the most disadvantaged schools, that are schools far from both thresholds.

To sum up, we can conclude that the treatment has only a weak effect on per-pupil teaching hours and class size in borderline schools.

What therefore has happened for schools away from the thresholds? Though we do not have a reliable identification strategy to evaluate the RAR effects on these schools, we can nonetheless provide some descriptive statistics. Graphs 5.4 and 5.5 support that extra resources may have been allocated to severely disadvantaged schools. Per-pupil teaching hours increased after RAR assignment in September 2006. The average class size was also affected though moderately. However, the literature suggests that such a decrease in class size would have only small effects on achievement : for junior high schools Piketty & Valdenaire (2006) found that dividing the class size by 2 increases the scores of pupils in the National Exam in Grade 9 by only 10% of a standard deviation (see also Gary-Bobo & Mahjoub (2006)).

Though the program may have had no effect on the resources of borderline schools, it may anyway have induced some teachers to relocate, or some parents to withdraw their children from the treated schools, and this would impact on school achievement.

[12]For the number of hours per pupil, one can expected an increase of 1.4*8%=0.11. The corresponding estimates only range between -0.09 and 0.09 (with strong variations depending on the grade). For the number of pupils per class, estimates range between -1.38 and 1.09 for an expected value close to -20*8%=-1.6.

Table 5.3: RAR effect on pupils per class (P/C) and per-pupil teaching hours (H/P) indicators

(a) Disadvantaged discontinuity: $Z^F = 67\%$ and $10\% \leq Z^L$

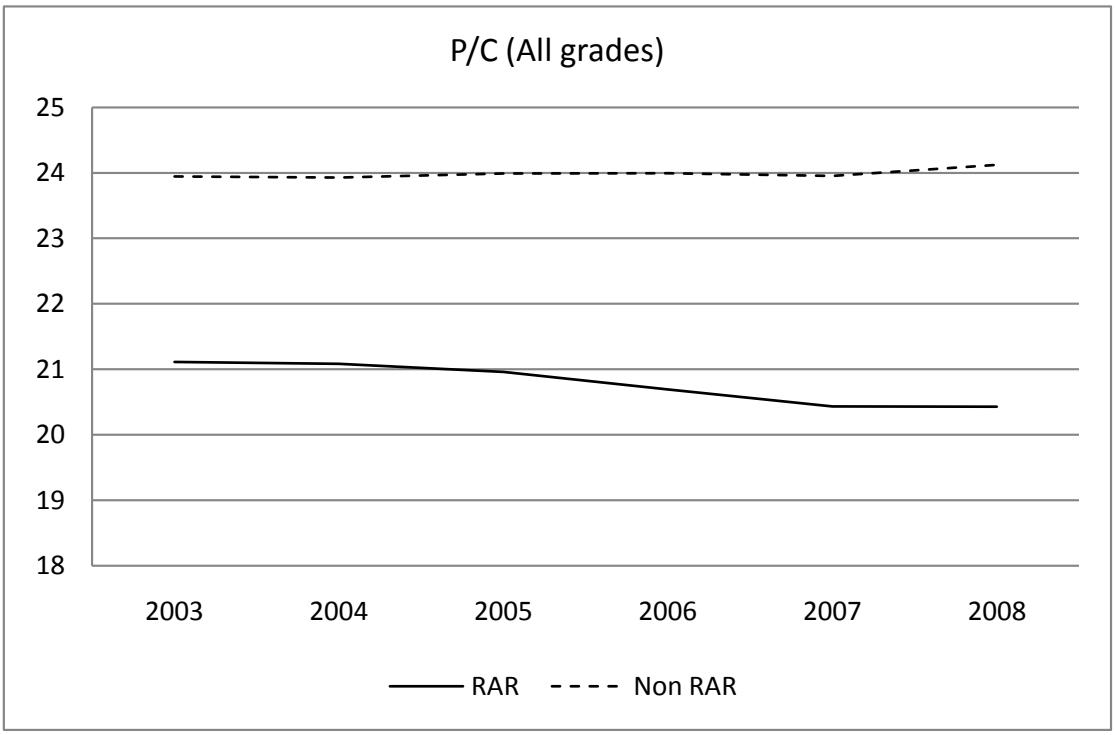
	h= 4		h= 6		h= 8	
P/C (all grades)	-1.38 *	[1.02]	-0.03	[0.85]	1.09	[0.83]
P/C (9th grade)	-2.81 *	[1.95]	-0.60	[1.70]	-0.89	[1.59]
P/C (8th grade)	0.09	[2.00]	0.96	[1.75]	3.95	[1.72]
P/C (7th grade)	-0.43	[1.89]	0.60	[1.59]	1.60	[1.56]
P/C (6th grade)	-0.82	[1.87]	0.11	[1.60]	0.60	[1.55]
H/P (all grades)	0.07	[0.07]	0.07	[0.06]	-0.02	[0.06]
H/P (9th grade)	-0.10	[0.13]	-0.11	[0.13]	-0.04	[0.12]
H/P (8th grade)	0.03	[0.12]	0.01	[0.11]	-0.16	[0.10]
H/P (7th grade)	0.13	[0.13]	0.14 *	[0.10]	0.06	[0.10]
H/P (6th grade)	0.16	[0.14]	0.21 **	[0.12]	0.05	[0.14]
N° Schools	29		52		76	
N° Obs.	174		312		456	

(b) Repeating discontinuity: $Z^L = 10\%$ and $67\% \leq Z^F \leq 80\%$

	h= 2		h= 3		h= 4	
P/C (all grades)	-0.68	[2.19]	-0.45	[3.98]	0.10	[1.06]
P/C (9th grade)	-1.55	[4.62]	-1.85	[8.15]	0.71	[2.25]
P/C (8th grade)	0.04	[4.82]	1.83	[8.27]	1.26	[2.10]
P/C (7th grade)	-0.44	[4.96]	5.15	[9.50]	1.06	[2.12]
P/C (6th grade)	-3.21	[4.39]	-5.81	[8.98]	-1.96	[2.00]
H/P (all grades)	0.01	[0.15]	0.09	[0.27]	-0.09	[0.09]
H/P (9th grade)	0.08	[0.34]	-0.14	[0.61]	-0.22	[0.17]
H/P (8th grade)	-0.07	[0.31]	-0.16	[0.54]	-0.17	[0.14]
H/P (7th grade)	-0.10	[0.31]	-0.33	[0.64]	-0.11	[0.15]
H/P (6th grade)	0.33	[0.39]	1.15	[1.12]	0.14	[0.18]
N° Schools	33		50		77	
N° Obs.	198		300		460	

Estimated treatment effect, standard error in bracket, unilateral test of equality between treated and non-treated schools. Level: $*$ 10%, $**$ 5%, $***$ 1%

Figure 5.4: Pupils per Class (All Grades)

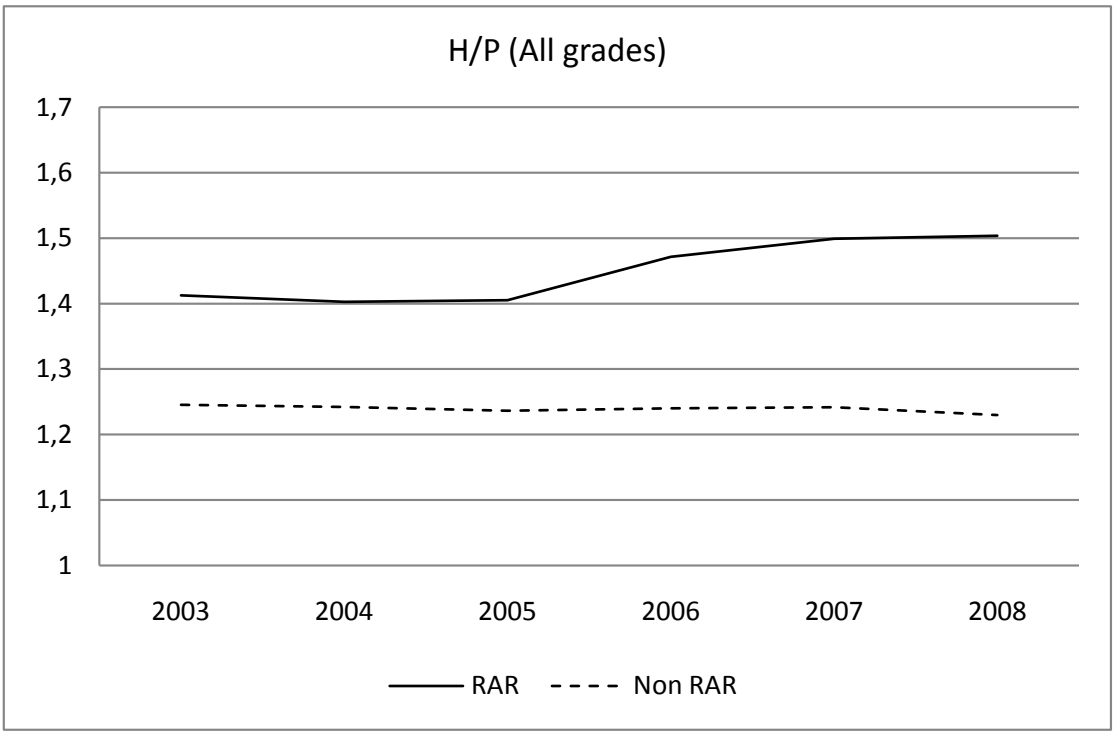


5.4.2 Pupils entering Grade 6

We now present results on student enrollment. Before September 2007, pupils were assigned to junior high schools on the basis of their residence. Parents could avoid this assignment by either choosing a private junior high school, or a specific language or option not available in the assigned school when their child entered Grade 6.[13] Previous studies suggest that parents are sensitive to junior high school districts when choosing their residence (Fack & Grenet (2010), Black (1999), Gibbons & Machin (2003), Maurin (2004)). Since September 2007, in the aftermath of the Presidential election, the number of exemptions to assignment has increased and it has been easier for parents to choose their children's junior high school. In this context, that of parental junior high school choice, the "RAR" label could have had

[13]In France, in Grade 6, pupils have to choose options that they did not follow before, usually a language. If a pupil wants to avoid his or her assigned junior high school, he or she can ask for learning Russian or Chinese.

Figure 5.5: Per-pupil Teaching Hours (All Grades)



mixed consequences. On the one hand, it could act as a positive signal for parents - additional school inputs or resources per child- or as a negative signal - indicating severe difficulties at the treated school. For instance, Benabou et al. (2004, 2009) found that the proportion of students who have lunch at school decreases when junior high schools receive the "ZEP" label (Education Priority Zones). This evolution may indicate an increase in segregation of ZEP and non-ZEP junior high schools: poorer kids do not go to the cafeteria for school lunch. The authors also found mixed effects on the total number of pupils in ZEP junior high schools: for junior high schools treated in 1989, the total number of junior pupils decreased relative to non-ZEP schools. However for those treated in 1990, they did not find any effect. For our part, we did not find any significant difference between RAR and non-RAR schools near the second discontinuity concerning the total number of pupils entering Grade 6 (see Table 5.4b), while for the first discontinuity, we found an increase in pupils entering Grade 6 (see Table 5.4a). Furthermore, we found no significant effect

of treatment on the proportion of pupils entering Grade 6 who have lunch at school (see Tables 5.4a and 5.4b).

We also use parental occupations at the beginning of Grade 6 to test whether social segregation increases in "RAR" junior high schools. Results differ along both discontinuities.

For schools around the 'disadvantaged' discontinuity, we found a small negative effect in the enrollment of children whose parents are self-employed, and to some extent, on enrollment of children whose parents are involved in intermediate occupations. This reduction goes along with an increase in the proportion of blue collar parents (see Table 5.4a). For schools near the 'repeating' discontinuity, we do not observe such an effect (see Table 5.4b).

We therefore conclude that treatment effects are heterogeneous, since we get different results on both discontinuities. For the first one, RAR policy has had an adverse effect, while for the second no significant effect is detected.

5.4.3 Teaching structure

In this subsection, we report results on another set of outcomes related to teaching structure. The outcomes we consider are teacher mobility, their seniority and their qualifications. For instance, some incentives, such as faster access to promotion, were implemented to encourage teachers to teach longer within RAR schools.

For the ZEP program, these issues were studied by Benabou et al. (2004, 2009) with a different identification strategy. The authors found that the incentives were insufficient to cap the high turnover of teachers in treated schools. Ly (2010) found that the reform of the ZEP program in 1999 had had an adverse effect on the age and the experience of teachers in treated schools. To study this issue, we compare the structure of age and qualifications of teachers in schools close to the discontinuities. Tables 5.5a and 5.5b report regression discontinuity estimates. These results are novel with respect to the Preferential Educational Policy litterature and differ from the results obtained by Benabou et al. (2004, 2009) and Ly (2010).

For schools where $Z^F = 67\%$, the proportion of teachers over 55 has increased significantly with the introduction of the RAR program. For these schools, the proportion of highly qualified teachers ("agrégation") has decreased while the proportion of teachers having a non-standard qualification in junior high schools has increased. This may be due to the fact that RAR policy has encouraged primary school teachers to work in junior high schools. The new teachers assigned to treated junior high schools may more often have been primary school teachers, and often the oldest, being the most experienced, are more

Table 5.4: RAR effect on Pupils entering Grade 6

(a) Disadvantaged discontinuity: $Z^F = 67\%$ and $10\% \leq Z^L$

	h= 4		h= 6		h= 8	
Parents' occupation (%):						
Farmer	1.47	[1.27]	1.43	[1.34]	0.31	[1.14]
Executive	-9.62 ***	[3.26]	-8.41 ***	[2.76]	-8.42 ***	[2.52]
Manager	0.62	[2.42]	0.03	[2.14]	-1.09	[1.91]
Intermediate	-6.78 *	[3.50]	-6.26 **	[3.09]	-3.52	[3.04]
Employee	3.60	[4.68]	4.24	[4.28]	2.69	[3.99]
Worker	11.96 **	[5.89]	10.67 *	[6.14]	12.16 **	[6.16]
Retired	1.23	[1.50]	0.61	[1.36]	0.87	[1.33]
Unemployed	0.30	[5.31]	-2.89	[4.58]	-5.95	[4.59]
Not known	-2.78	[4.93]	0.58	[4.23]	2.94	[4.03]
Lunch at school	1.17	[6.61]	2.45	[5.86]	-1.79	[5.73]
Tot. Entering	6.94	[12.51]	20.14 *	[11.92]	34.80 ***	[12.80]
N° Schools	29		52		76	
N° Obs.	174		312		456	

(b) Repeating discontinuity: $Z^L = 10\%$ and $67\% \leq Z^F \leq 80\%$

	h= 2		h= 3		h= 4	
Parents' occupation (%):						
Farmer	0.38	[4.22]	-2.90	[6.42]	-0.90	[1.56]
Executive	-6.72	[6.70]	-9.99	[13.48]	-4.31	[2.91]
Manager	5.41	[5.49]	3.39	[8.82]	3.64	[2.61]
Intermediate	2.33	[7.90]	-3.08	[14.40]	0.29	[3.77]
Employee	7.34	[11.13]	-23.10	[25.60]	-4.21	[4.86]
Worker	-3.40	[17.39]	68.61	[67.20]	10.32	[8.26]
Retired	-1.70	[3.98]	-2.00	[6.91]	0.77	[1.78]
Unemployed	2.53	[14.10]	8.21	[24.59]	3.76	[6.23]
Not known	-6.17	[10.78]	-39.14	[39.45]	-9.36 *	[4.86]
Lunch at school	-0.56	[10.80]	6.23	[16.71]	-2.35	[7.08]
Tot. Entering	-6.26	[25.59]	-27.84	[58.00]	13.11	[15.59]
N° Schools	33		50		77	
N° Obs.	198		300		460	

Estimated treatment effect, standard error in bracket, bilateral test of equality between treated and non-treated schools. Level: ∗ 10%, ∗∗ 5%, ∗∗∗ 1%

likely to have been selected. It could also be explained by lower mobility among older teachers who would like to leave their newly assigned RAR school, but cannot for family reasons. For schools where $Z^L = 10\%$, not a single treatment effect can be displayed on the teaching structure. Once again, such a difference between the two discontinuities highlights the heterogeneity of treatment effects.

5.4.4 Maths and Literacy scores

Finally, we present results on student achievement. Table 5.6 reports estimates of treatment effects on junior high school score distributions. We find that treatment has a negative effect on scores, when significant. This result is more pronounced for schools where $Z^F \approx 67\%$ and $Z^L > 10\%$. For these schools, the treatment effect differs for maths and French (Tables 5.6a). The negative effect of treatment on scores is mainly concentrated at the bottom of the French distribution, whereas, for maths, it is at the top of the distribution. For schools where $Z^L \approx 10\%$ and $80\% \geq Z^F \geq 67\%$, estimates are often insignificant (especially in French) but other than that, these effects are quite large and negative. For these schools, treatment effects are mainly visible in the middle of the Math score distribution.

How can we interpret these results on the change in pupil achievement as a result of treatment? The results may be due to a combination of two distinct effects: a potential increase in the sorting of pupils across schools based on ability and/or parental schooling choice, and the potential inefficiency of the educational policy within treated schools. Our results imply the existence of at least one of these two effects. It cannot be excluded that the policy has had a positive effect on pupil achievement for those that remain in treated schools, but this effect might be compensated for by an increase in a pupil selection effect across schools. Moreover, we note that we can only observe the results of pupils up until 2009. Therefore observed final scores are relevant to pupils who began junior high schools before the beginning of the RAR program. As a consequence this result can differ from the effect of policy on pupils having a full scholarship into treated schools.

A referee suggested that the test results should not be pooled to assess the program impact on score results. We have thus estimated the effect of RAR treatment with an unequal length of treatment:

$$Y_{it} = \alpha_i + \beta_t + \gamma_{2007} T_{it} \mathbb{1}_{\{t=2007\}} + \gamma_{2008} T_{it} \mathbb{1}_{\{t=2008\}} + \gamma_{2009} T_{it} \mathbb{1}_{\{t=2009\}}$$

Table 5.5: RAR effect on teachers' characteristics

(a) Disadvantaged discontinuity: $Z^F = 67\%$ and $10\% \leq Z^L$

	h= 4		h= 6		h= 8	
Tot. of hours dispensed	68.45	[56.68]	135.08 ***	[49.33]	167.77 ***	[53.47]
% of hours dispensed by :						
Highest teaching degree	-0.03	[0.02]	-0.03 *	[0.02]	-0.03 *	[0.02]
Qualified teacher	-0.02	[0.05]	-0.04	[0.04]	-0.04	[0.04]
PE teacher	0.01	[0.01]	0.00	[0.01]	0.00	[0.01]
Other teacher	0.04	[0.05]	0.07 *	[0.04]	0.07 *	[0.04]
Teachers under age 30	-0.07	[0.05]	0.01	[0.05]	0.01	[0.05]
Teachers between 30 and 39	-0.01	[0.05]	-0.03	[0.05]	-0.07	[0.05]
Teachers between 40 and 55	0.02	[0.05]	-0.02	[0.05]	0.04	[0.04]
Teachers over 55	0.12 **	[0.05]	0.09 **	[0.04]	0.11 **	[0.04]
N° Schools	29		52		76	
N° Obs.	174		312		456	

(b) Repeating discontinuity: $Z^L = 10\%$ and $67\% \leq Z^F \leq 80\%$

	h= 2		h= 3		h= 4	
Tot. of hours dispensed	-35.91	[106.86]	154.44	[283.97]	62.58	[63.67]
% of hours dispensed by :						
Highest teaching degree	0.01	[0.05]	0.01	[0.09]	-0.01	[0.02]
Qualified teacher	-0.09	[0.11]	-0.24	[0.25]	0.00	[0.05]
PE teacher	0.04	[0.03]	-0.03	[0.05]	0.00	[0.01]
Other teacher	0.04	[0.10]	0.26	[0.25]	0.02	[0.05]
Teachers under age 30	-0.04	[0.13]	-0.08	[0.23]	0.01	[0.06]
Teachers between 30 and 39	0.02	[0.14]	0.36	[0.40]	0.04	[0.06]
Teachers between 40 and 55	0.04	[0.13]	-0.26	[0.33]	-0.07	[0.06]
Teachers over 55	-0.12	[0.13]	-0.01	[0.19]	0.06	[0.05]
N° Schools	33		50		77	
N° Obs.	198		300		460	

Estimated treatment effect, standard error in bracket, bilateral test of equality between treated and non-treated schools. Level: ∗ 10%, ∗∗ 5%, ∗ ∗ ∗ 1%

where T_{it} equals 1 if the junior high school i enters the RAR program at date t. These variables $T_{it}\mathbb{1}_{\{t=j\}}$ are instrumented by $\mathbb{1}_{\{Z_{i\geq c}\}}\mathbb{1}_{\{t=j\}}$. We obtain similar results which are robust to a differentiating length treatment effect.

5.5 Conclusion

In this paper, we have evaluated the effect of the RAR educational policy, introduced in French public junior high schools in September 2006. To do so, we used two strong discontinuities in assignment to the treatment. For schools close to these two discontinuities, no substantial desired treatment effects stand out. Precisely, resources allocated to schools were disappointing around the thresholds. Second, we found that the policy worsens social segregation across schools when measured by parental occupation. This may induce the reduced achievement we observed in treated schools. The RAR assignment would therefore appear to have had a strong negative signaling effect, encouraging better students to move from their junior high school. An alternative, but discouraging, explanation would be that the policy has had a negative effect on achievement in treated schools. Our results on teachers indicate an increase in the proportion of older and less qualified teachers. This could be explained by a reassignment of older primary teachers to RAR schools, in order for them to obtain promotion more easily.

Table 5.6: RAR effect on school level distribution of scores at the National Exam in Grade 9

(a) Disadvantaged discontinuity: $Z^F = 67\%$ and $10\% \leq Z^L$

	h= 4		h= 6		h= 8	
French - Mean	0.74	[1.29]	-1.79	[1.25]	-2.03	[1.18]
French - Q10	-0.14	[1.55]	-1.75	[1.46]	-2.18 ***	[1.41]
French - Q25	0.14	[1.41]	-2.01	[1.34]	-2.48 ***	[1.28]
French - Median	-0.05	[1.47]	-2.32	[1.47]	-2.71 *	[1.39]
French - Q75	1.90	[1.66]	-1.11	[1.51]	-1.13	[1.42]
French - Q90	3.23 *	[1.86]	-0.66	[1.70]	-0.62 *	[1.58]
Maths - Mean	-1.13	[2.23]	-2.60	[1.94]	-4.75 *	[1.92]
Maths - Q10	1.20	[2.19]	-0.17	[1.88]	-1.02	[1.68]
Maths - Q25	0.70	[2.56]	-1.86	[2.20]	-4.49	[2.11]
Maths - Median	-1.69	[2.64]	-2.88	[2.27]	-5.40 **	[2.29]
Maths - Q75	-3.59	[2.70]	-4.68 *	[2.41]	-6.95	[2.46]
Maths - Q90	-3.09	[2.82]	-3.93	[2.52]	-6.47 **	[2.54]
N° Schools	29		52		76	
N° Obs.	174		312		456	

(b) Repeating discontinuity: $Z^L = 10\%$ and $67\% \leq Z^F \leq 80\%$

	h= 2		h= 3		h= 4	
French - Mean	-6.77	[4.80]	-0.37	[5.63]	-0.49	[1.52]
French - Q10	-0.97	[3.72]	4.67	[7.50]	0.06	[1.76]
French - Q25	-2.43	[3.95]	1.56	[6.54]	0.02	[1.71]
French - Median	-7.24	[5.38]	0.02	[6.38]	-0.84	[1.73]
French - Q75	-13.23 *	[7.75]	-4.72	[7.78]	-1.02	[1.84]
French - Q90	-11.01	[7.01]	0.18	[7.94]	1.13	[2.13]
Maths - Mean	-7.74	[6.23]	-5.59	[9.52]	-3.61	[2.42]
Maths - Q10	-5.76	[5.23]	-5.29	[8.81]	-1.91	[2.10]
Maths - Q25	-4.13	[5.43]	-8.99	[11.77]	-5.32 *	[2.81]
Maths - Median	-8.14	[6.92]	-13.21	[15.00]	-6.27 **	[3.13]
Maths - Q75	-11.59	[8.69]	-3.24	[11.20]	-2.63	[2.99]
Maths - Q90	-7.65	[8.09]	9.70	[14.51]	-0.04	[3.10]
N° Schools	33		50		77	
N° Obs.	198		300		460	

Estimated treatment effect, standard error in bracket, bilateral test of equality between treated and non-treated schools. Level: $*$ 10%, $**$ 5%, $***$ 1%

Bibliography

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.

Abowd, J. M., Crépon, B. & Kramarz, F. (1999), 'Moment estimation with attrition: An application to economic models', *Journal of the American Statistical Association* **96**, 1223–1231.

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.

Andrews, D. K. (1999), 'Estimation when a parameter is on a boundary', *Econometrica* **67**, 1341–1383.

Andrews, D. K. (2000), 'Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space', *Econometrica* **68**, 399–405.

Andrews, D. K. (2011), Examples of l2-complete and boundedly-complete distributions. Working Paper.

Andrews, D. K. & Soares, G. (2010), 'Inference for parameters defined by moment inequalities using generalized moment selection', *Econometrica* **78**, 119–157.

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.

Angrist, J. D. & Lavy, V. C. (1999), 'Using mainmonides' rule to estimate the effect of class size on scholastic achievement', *Quaterly Journal of Economics* **114**(2), 533–575.

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

Benabou, R., Kramarz, F. & Prost, C. (2004), 'Zones d'éducation prioritaire : quels moyens pour quels résultats ? une évaluation sur la période 1982-1992', *Economie et Statistique* (380), 3–30.

Benabou, R., Kramarz, F. & Prost, C. (2009), 'The french zones d'éducation prioritaire : Much ado about nothing?', *Economics of Education Review* **28**, 345–356.

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

Bhattacharya, D. (2008), 'Inference in panel data models under attrition caused by unobservables', *Journal of Econometrics* **144**, 430–446.

Biemer, P. P. (2001), 'Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing', *Journal of Official Statistics* **17**, 295–320.

Bierens, H. J. (1982), 'Consistent model specification tests', *Journal of Econometrics* **20**, 105–134.

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

Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review / Revue Internationale de Statistique* **51**(3), 279–292.

Black, S. (1999), 'Do better schools matter? parental valuation of elementary education', *Quaterly Journal of Economics* **114**(2), 577–599.

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.

Carter, M. (2001), *Foundations of Mathematical Economics*, MIT Press.

Chen, C. (2001), 'Parametric models for response-biased sampling', *Journal of the Royal Statistical Society, Series B* **63**, 775–789.

Chernozhukov, V., Fernandez-Val, I., Hahn, J. & Newey, W. (2012), 'Average and quantile effects in nonseparable panel models', *Econometrica* **Forthcoming**.

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. (2011), An instrumental variable model of multiple discrete choice. CEMMAP Working Paper CWP/12.

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.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McParland, J., Mood, A. M., Weinfeld, F. D. & York, R. L. (1966), Equality of educational opportunity, report to the president and the congress, Technical report, National Center for Educational Statistics.

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'Haultfœuille, X. (2011), 'On the completeness condition in nonparametric instrumental regression', *Econometric Theory, forthcoming* .

D'Haultfoeuille, X. & Rathelot, R. (2011), 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.

Evdokimov, K. (2011), Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells. Working paper.

Fack, G. & Grenet, J. (2010), 'When do better schools raise housing prices? evidence from paris public and private schools', *Journal of Public Economics* **94**(1-2), 59–77.

Finn, J. D. & Achilles, C. M. (1990), 'Answers and questions about class size: A statewide experiment', *American Educational Research Journal* **27**(3), 557–577.

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

François, J.-C. & Poupeau, F. (2004), 'L'évitement scolaire et les classes moyennes à paris', *Education et Sociétés* **14**(2), 51–66.

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.

Gary-Bobo, R. J. & Mahjoub, M.-B. (2006), 'Estimation of class-size effects, using "maimonides' rule": the case of french junior high schools', *Working paper CEPR* (DP5754).

Gibbons, S. & Machin, S. (2003), 'Valuing english primary schools', *Journal of Urban Economics* **53**(2), 197–219.

Gonzalez-Demichel, C. & Nauze-Fichet, E. (2003), 'Les contours de la population active : aux frontières de l'emploi, du chômage et de l'inactivité', *Economie et Statistique* **362**, 85–103.

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

Guillemot, D. (1996), 'La population active : une catégorie statistique difficile à cerner', *Economie et Statistique* **300**, 39–53.

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

Hahn, J., Todd, P. & Van Der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design', *Econometrica* **69**(1), 201–209.

Hall, P. & Horowitz, J. L. (2005), 'Nonparametric methods for inference in the presence of instrumental variables', *Annals of Statistics* **33**, 2904–2929.

Hall, R. & Mishkin, F. S. (1982), 'The sensitivity of consumption to transitory income: estimates from panel data on households', *Econometrica* **50**, 461–481.

Hausman, J. A. & Wise, D. A. (1979), 'Attrition bias in experimental and panel data: the Gary income maintenance experiment', *Econometrica* **47**, 455–473.

Heckman, J. (2001), 'Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture', *Journal of Political Economy* **109**, 673–748.

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

Heckman, J. J. & al. (2010), 'The rate of return to the highscope perry preschool program', *Journal of Public Economics* **94**, 114–128.

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**, 669–738.

Hirano, K., Imbens, G. W., Ridder, G. & Rubin, D. B. (2001), 'Combining panel data sets with attrition and refreshment samples', *Econometrica* **69**, 1645–1659.

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

Honoré, B. & Kyriazidou, E. (2000), 'Panel data discrete choice models with lagged dependent variables', *Econometrica* **68**(4), 839–874.

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

Hu, Y. & Shiu, J. (2013), Nonparametric identification using instrumental variables: sufficient conditions for completeness. Working Paper.

Imbens, G. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: a review', *The Review of Economics and Statistics* **86**, 4–29.

Imbens, G. & Lemieux, T. (2008), 'Regression discontinuity designs: A guide to practice', *Journal of Econometrics* **142**, 615–635.

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.

Kodde, D. A. & Palm, F. C. (1986), 'Wald criteria for jointly testing equality and inequality restrictions', *Econometrica* **54**, 1243–1248.

Krueger, A. B. (1999), 'Experimental estimates of education production functions', *The Quarterly Journal of Economics* **114**(2), 497–532.

Krueger, A. B. & Whitmore, D. M. (2001), 'The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star', *The Economic Journal* **111**, 1–28.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*, Cambridge University Press.

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.

Leuwen, E., Hessel, O. & Marte, R. (2008), 'Quasi-experimental estimates of the effect of class size on achievement in norway', *The Scandinavian Journal of Economics* **110**(4), 663–693.

Little, R. & Rubin, D. B. (1987), *Statistical analysis with Missing Data*, John Wiley & Sons, New York.

Ly, S. (2010), 'La répartition des ressources scolaires entre les collèges français. les effets paradoxaux de la relance des zones d'education prioritaire de 1999.', *Unpublished master report, directed by Eric Maurin* .

Machin, S. & Vignoles, A. (2005), Education policy in the uk, Technical report, CESifo DICE report 4.

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.

Mattner, L. (1992), 'Completeness of location families, translated moments, and uniqueness of charges', *Probability Theory and Related Fields* **92**, 137–149.

Maurin, E. (2004), *Le ghetto français, enquête sur le séparatisme social*, La république des idées, Le seuil.

McCrary, J. (2008), 'Manipulation of the running variable in the regression discontinuity design: A density test', *Journal of Econometrics* (142), 698–714.

Meuret, D. (1994), 'L'efficacité de la politique des zones d'éducation prioritaire dans les collèges', *Revue Française de Pédagogie* **109**, 41–64.

Newey, W. K. (1991), 'Uniform convergence in probability and stochastic equicontinuity', *Econometrica* **59**(4), 1161–1167.

Newey, W. & Powell, J. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**, 1565–1578.

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

Piketty, T. & Valdenaire, M. (2006), 'L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français : Estimations à partir du panel primaire 1997 et du panel secondaire 1995', *Les dossiers de la DEP, Ministère de l'Education Nationale* (173).

Plowden, B. & the Central Advisory Council for Education (1967), Children and their primary schools, Technical report, Central Advisory Council for Education (England).

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

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

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

Rosen, A. (2008), 'Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities', *Journal of Econometrics* **146**, 107–117.

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

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.

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

Saez, E. (2010), 'Do taxpayers bunch at kink points?', *American Economic Journal* **2**(3), 180–212.

Santos, A. (2011), 'Instrumental variables methods for recovering continuous linear functionals', *Journal of Econometrics* **161**, 129–146.

Sasaki, Y. (2012), Heterogeneity and selection in dynamic panel data. Working paper.

Severini, T. & Tripathi, G. (2006), 'Some identification issues in nonparametric linear models with endogenous regressors', *Econometric Theory* **22**, 258–278.

Severini, T. & Tripathi, G. (2011), 'Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors', *Journal of Econometrics, forthcoming* .

Stewart, G. W. (1969), 'On the continuity of the generalized inverse', *SIAM Journal on Applied Mathematics* **17**, pp. 33–45.

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

Tang, G., Little, R. J. A. & Raghunathan, T. E. (2003), 'Analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **90**, 747–764.

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 Klaauw, W. (2008), 'Breaking the link between poverty and low student achievment: An evaluation fo title i', *Journal of Econometrics* **142**, 731–756.

van der Vaart, A. (2000), *Aymptotic Statistics*, Cambridge University Press.

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.

Wooldridge, J. (2007), 'Inverse probability weighted estimation for general missing data problems', *Journal of Econometrics* **141**, 1281–1301.