Identification and Inference on the Correlation using Data from Two Independent Samples

David Pacini*

This Preliminary and Incomplete Version:

May 20, 2010

Abstract

We examine what can be learned about the correlation between Y and X when data are available from two independent random samples; the first sample gives information on variables (Y, Z), while the second sample gives information on (X, Z). The variable Z has the same distribution in both samples, but the samples have no common observational units. A difficulty arises because neither sample has joint information on the variables (Y, X). This situation applies, for instance, to the ecological correlation problem or in the measurement of impact heterogeneity in program evaluation. Our first contribution is to sharply characterize the set of all possible values of the correlation of interest that are compatible with hypothetical knowledge of the distribution of (Y, Z) and of (X, Z) (the identification result). Unlike the existing literature, our characterization does not rely on assumptions, other than regularity conditions, on the joint distribution of (Y, X, Z). The second contribution is to propose a series-based estimator for the later set, which turns to be consistent and asymptotically normal and thereby permitting relatively easy inference in applications. We evaluate the small sample properties of the proposed estimator by means of Monte-Carlo experiments.

*Phd. student, Toulouse School of Economics| **Corresponding Address:** Manufacture des Tabacs, 21 Allée de Brienne, Bureau MF003, (31000) Toulouse, France| **E-mail:** davidhpacini@yahoo.com| **First Version:** January 22, 2008| Please do not cite or circulate| **Comments are welcomed.**

1 Introduction

We examine what can be learned about the correlation between the variables Y and X when data are available from two independent random samples with a common variable Z. The first sample gives information on variables (Y, Z) but not X, while the second one gives information on (X, Z) and not Y. The variable Z has the same distribution in both samples, but the samples have no common observational units. A difficulty arises because joint realizations of the variables (Y, X) involved in the definition of the correlation of interest are not observed. Existing literature on the combination of independent samples with common variables (see Ridder and Moffit, 2007 for a survey) deals with this difficulty by assuming either that Y is independent of X conditionally on Z, or that the distribution of (Y, X, Z) is multivariate normal. Here we ask what can be ascertained about the correlation of interest without such assumptions.

As an example of application of the above framework, consider the following problem arising in media planning studies. Suppose that we are interested in learning about the correlation between consumers' purchase behavior of cookies, measured by Y, and consumers' exposure to advertisements on media, measured by X. For reasons of costs and focus, data on purchase behavior and media viewing are typically available from different random samples (c.f., The Nielsen Company, 2007). A first sample, say $\{Y_i, Z_i\}_{i=1}^{n_1}$, contains information about purchase behavior Y_i and socioeconomic characteristics Z_i for a group of consumers labeled $1, .., i, .., n_1$. A second sample, say $\{X_j, Z_j\}_{j=n_1+1}^{n}$, contains information about exposure to advertisements X_j and the same socioeconomic characteristics Z_j but for a different group of consumers $n_1 + 1, .., j, .., n$. We face a difficulty because the correlation of interest is a functional of the joint distribution of consumers' purchase behavior Y and consumers' exposure to advertisements X but the available data only reveal the distributions of (Y, Z) and of (X, Z). The issue is to determine to which extend the socioeconomic characteristics Z measured in both samples can help us to cope with the later difficulty without assuming either that Y is independent of X conditionally on Z, or that the distribution of (Y, X, Z) is multivariate normal.

We first ask which are the values of the correlation between Y and X compatible with hypothetical knowledge of the distributions of (Y, Z) and of (X, Z) when no assumptions, other than regularity conditions (e.g., existence of second moments), are placed on the distribution of (Y, X, Z). Following the terminology by Manski (2005), we call such set of feasible values the identified set for the correlation between Y and X and we denote it by Θ_I. We show that Θ_I is a closed interval of the real line, so it can be characterized by its two extreme points, say θ_l and θ_u. These extreme points represent the minimum and maximum value of the correlation between Y and X compatible with hypothetical knowledge of the distributions of (Y, Z) and of (X, Z). We derive analytical expressions for θ_l and θ_u in terms of the quantile function of X given Z and the distribution of Y conditional on Z by exploiting the conditional Frechet bounds on the joint distribution of (Y, X) proposed by Ridder and Moffit (2007) and the fact that the correlation between Y and X is a superadditive functional of later joint distribution. The resulting characterization of the identified set contains only the values of correlation between Y and X that are compatible with the maintained assumptions and the available data free of sample variation and *no others*. That is, our characterization is *sharp*. By contrast, the sharpness result is not guaranteed in the related literature (c.f., Rassler, 2002;Djebbari and Smith, 2008). Establishing whether a conjectured characterization of the identified set is sharp is a relevant question in identification analysis because *outer* characterizations -i.e., those ones including values of the parameter of interest incompatible with the maintained assumptions and the available data free of sample variation- may weaken our ability to perform tests and to make useful predictions.

We then ask how the identified set Θ_I previously characterized can be actually recovered from two samples on (Y, Z) and (X, Z) of finite size. To answer this question, we begin by proposing an estimator for the identified set (actually for its extreme points θ_l and θ_u), which builds on our previous characterization of this set. Our proposal involves to estimate by series the quantile function of X given Z and the distribution of Y conditional on Z in a first step, and then plug in such estimates in the analytical expressions of θ_l and θ_u. Although our method is fully non-parametric, our estimates of the extreme points of the identified set converge to their true counterparts at the usual parametric rate. We continue by approximating the distribution of the estimators using asymptotic theory. We show, under a set of sufficient conditions, that the estimators we propose for θ_l and θ_u are jointly asymptotically normal. This approximation could be useful in constructing asymptotic confidence intervals for the true value of the parameter of interest of the type proposed by Imbens and Manski (2004), and later refined by Stoye (2009).

We also show that our identification and inference results apply not only to the correlation between Y and X but also to any scalar parameter of interest θ defined by the equation $\theta = \mathbb{E}[c(Y, X; \mu)]$ where $\mathbb{E}$ denotes the expectation operator and $y, x \mapsto c(y, x; \mu)$ is superadditive function known up to the nuisance parameter μ depending only upon the distribution of (Y, Z) and of (X, Z). Though admittedly specific, the class defined by the later properties encompass many parameters of practical interest. In the text, we show that the problem of measuring the variance of the treatment effect in program evaluation (see Djebbari and Smith, 2008) or the covariance between individuals' voting behavior and their socioeconomic characteristics in a secret ballot (see Robinson, 1950) can be cast in our framework.

Related Literature. The identification and inference problems we deal with arise under different guises in different literatures. In the literature on the combination of independent samples with common variables, some authors (c.f., Kadane, 2001; Moriarity and Scheuren, 2001; Rassler, 2002) use the fact that the covariance matrix of (Y, X, Z) must be positive semidefinite to derived bounds on the feasible values for the correlation between Y and X. These bounds, nevertheless, are not guaranteed to be sharp when (Y, X, Z) does not follow a trivariate normal distribution. Sims (1972) notice that when Y is assumed to be independent of X conditionally on Z it turns that the correlation between Y and X is equal to that between $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$, so the correlation of interest is point-identified with two samples of the type described above. Many authors (see Ridder and Moffit, 2007 for a survey) exploit this later assumption as a vehicle to learn about the correlation between Y and X. By contrast, here we do not impose such a conditional independence assumption. A prominent application in the literature on the combination of independent samples with common variables is the so-called ecological correlation problem arising in political science (e.g., Robinson, 1950). In such applications, data provide with *estimates* of the distributions of (Y, Z) and of (X, Z), but nor necessarily with replications of such random variables, and knowledge is sought about the correlation between Y and X. Robinson (1950) criticizes the tacit interpretation of the correlation between $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$, the so-called ecological correlation, as the correlation between Y and X, and points out the fact that there are many values of the correlation between Y and X compatible with hypothetical knowledge of the distributions of (Y, Z) and (X, Z). Nevertheless, he neither characterizes such values nor proposes inference procedures. Our results thus extend the literature on the combination of independent samples with common variables to provide a sharp characterizations of the identified set of the correlation between Y and X without imposing assumption on the joint distribution of (Y, X, Z), and to propose inference procedures for such set.

The identification and inference problems we deal with appear also in the econometric literature on the evaluation of social programs (see Heckman and Vytlacil, 2007 for a survey), so this paper is also related to some work there. Concurrent work by Fan and Zhu (2010) study identification on superadditive functionals of the joint distribution of (Y, X) when the distributions of (Y, Z) and of (X, Z) are given in the context of a potential outcome model. Our paper overlap with theirs in the class of parameters of interest and in the distributions considered as given. By contrast, we differ in the way to write the identified set and the proposed estimators. Fan and Zhu (2010) depict the extreme points of the identified set as statistical functionals of the quantile functions of Y given Z and of X given Z and propose a plug-in kernel-based estimator for identified. We depict that the extreme points of the identified set as moment conditions containing unknown nuisance functions which are the composition of quantile function of X given Z and the distribution function of Y given Z, and we propose a series-based estimator. Our characterization enhance the connection between our estimator and the class of semiparametric estimators studied, among others, by Newey (1994), and this facilitates the derivation of the asymptotic properties.

Our work also shows some overlap with the literature on identification of regression functions with given marginals. In this literature, Vitali (1976) characterizes the identified set for the regression function $x \mapsto \mathbb{E}(Y|x)$ when only the distributions of Y and of X are given. Cross and Manski (2002) depict the identified set for the regression function $x, z \mapsto \mathbb{E}(Y|x, z)$ when the distribution of (Y, Z) and of (X, Z) are given and Y has discrete support. Molinari and Peski (2006) extend such result to the continuous support case. Our work overlap with those of the later authors in terms of the distributions which are considered as given, however, we differ in the parameter of interest. While they focus on a regression function we focus on superadditive scalar parameters.

At a more general level, this paper belongs to the literature on set-identification and inference of scalar parameters with incomplete data. In this literature, the identified set for the parameters of interest is an interval whose lower and upper bounds can be estimated by the sample analog principle. Seminal papers in this literature includes Horowitz and Manski (1995, 1998, 2006). Imbens and Manski (2004) develop general testing procedures and confidence intervals, later refined by Stoye (2009), for this type of parameters. Although the problems we are concerned with here are different from those considered by aforementioned authors, our techniques are similar to theirs and we use several of their results.

Organization of the Paper. The outline of the paper is as follows. In the next section we set a general framework, we define the class of parameters to be studied and we describe the identification and inference problems. Section 3 is devoted to solve the identification problem. We first suggest a number of different ways to write the identified set. Then, we discuss applications of these characterizations. Section 4 is devoted to solve the statistical inference problem. We first propose an estimator for the identified set. Then, we approximate the distribution of the estimator using asymptotic theory and we discuss how to construct confidence intervals. We close section 4 with Monte-Carlo experiments aimed to explore the finite sample properties of the estimator and its implementation. Section 5 concludes. The appendix collects proofs and tables.

2 Notation, Assumptions, and the Identification and Inference Problems

In this section we introduce the parameters of interest, the available data and the problems we approach, namely the identification and inference of a superadditive scalar integral parameter with given overlapping marginal distributions. We employ two leading examples to illustrate our theoretical setup. The first example is a two-sample combination problem from political science. The second example relates to the measurement of the variance of the impact effect in the econometric evaluation of social programs.

Parameters of Interest. We begin by introducing the two assumptions defining the class of parameters of interest. We show below that the correlation between Y and X is particular member of this class. Let $(Y, X, Z) : \Omega \rightarrow \mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ be a real random vector defined on the probability space $(\Omega, \mathfrak{F}, P)$, where Ω is a non-atomic sample space, $\mathfrak{F}$ is a σ-algebra and P is an unknown probability measure, so-called *the population*; P must lie in the family $\mathcal{P}$ characterized by ex-ante constraints, so-called *the model.* Here the random variables Y and X are scalar while Z may be vector-valued, that is $\mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ for $d_Z \geq 1$. For the random variable $Y : \Omega \rightarrow \mathcal{Y}$, we define its support, $Supp(Y)$, as the smallest closed set included in $\mathcal{Y}$ such as its complement has measure zero. Similar definitions are adopted for the other random variables. The first assumption defining the class of parameters of interest is:

Assumption F (Integral Statistical Functional) The unknown parameter of interest, say θ, arising from the population P is defined as an integral functional:

[F] $$\theta := \int_{\mathcal{Y}\times\mathcal{X}} c(y, x; \mu) dF_{Y,X}(y, x)$$

or equivalently as $\theta := \mathbb{E}[c(Y, X; \mu)]$ where $y, x \mapsto c(y, x)$ is a known real-valued function from the Cartesian product $\mathcal{Y} \times \mathcal{X}$ into a subset of the real numbers, $y, x \mapsto F_{Y,X}(y, x)$ is the joint distribution function of (Y, X) induced by P, and μ is a vector of real-valued nuisance parameters depending only on the distributions of (Y, Z) and of (X, Z). Let $\mathcal{F}_{Y,X}$ denote the class of all possible distributions $y, x \mapsto F_{Y,X}(y, x)$ with support on $\mathcal{Y} \times \mathcal{X}$

induced by all possible P in $\mathcal{P}$. The range of $\mathbb{E}[c(Y,X;\mu)] : \mathcal{F}_{Y,X} \rightarrow \mathbb{R}$ defines the parameter space, Θ, which is assumed to be compact.

To illustrate assumption [F], we propose the following two examples. The first one is inspired by Robinson (1950):

Pilot Example 1 (Ecological Correlation). Suppose that one is interested in the correlation between voting behavior Y and the educational level X of individuals in a presidential election with secret ballot. Let $\mathcal{Y} := \{1, .., y, ..\}$ be the list of candidates, let $\mathcal{X} := \{1, .., y_2, ..\}$ be the possible levels of education for a given voter, and let $\mathcal{Z} := \{1, .., z, ..\}$ be the electoral precincts. For a given voter, the random variables Y, X and *Z map, respectively, states of the nature Ω into choices of candidate $\mathcal{Y}$, education level $\mathcal{X}$ and electoral precinct $\mathcal{Z}$. The correlation between voting behavior and educational level is*:

$$\rho := \int_{\mathcal{Y}\times\mathcal{X}} \frac{[y \times x - \mathbb{E}(Y) \times \mathbb{E}(X)]}{\sqrt{\mathbb{V}(Y)\mathbb{V}(X)}} dF_{Y,X}(y,x)$$

Set $\mu = (\mathbb{V}(Y), \mathbb{V}(X), \mathbb{E}(Y), \mathbb{E}(X))$ and $c(y,x;\mu) = [y \times x - \mathbb{E}(Y)\mathbb{E}(X)]/\sqrt{\mathbb{V}(Y)\mathbb{V}(X)}$. Thus, the correlation between educational level on voting behavior is an integral functional of the bivariate distribution $y, x \mapsto F_{Y,X}(y,x)$.$\square$

The next example is inspired by Heckman, Smith and Clements (1997)

Pilot Example 2 (Variance of the Treatment Effect). Consider a collection of households subject to a binary policy intervention. For a given household, let Y_0 and *Y_1 be random variables mapping states of the nature Ω into per capita consumption under intervention and no intervention, respectively. The vector Z may represent background variables such as number of members in the household or years of formal education of the head. Suppose that one is interested in the variance of the difference $Y_1 - Y_0$, that is in the variance of the treatment effect, which is defined by:*

$$\sigma^2_{TE} := \int_{\mathcal{Y}_1\times\mathcal{Y}_0} \{[y_1 - \mathbb{E}(Y_1)] - [y_0 - \mathbb{E}(Y_0)]\}^2 dF_{Y_1,Y_0}(y_1,y_0)$$

Set $\mu = (\mathbb{V}(Y_0), \mathbb{V}(Y_1), \mathbb{E}(Y_0), \mathbb{E}(Y_1))$ and $c(y_1, y_0) := (\mathbb{V}(Y_1) + \mathbb{V}(Y_0) - 2([y_1 y_0 - \mathbb{E}(Y_1)\mathbb{E}(Y_0)]))$.

Relabel Y_1 as Y, and Y_0 as X. Thus, the variance of the impact effect is an integral functional of the bivariate distribution $y, x \mapsto F_{Y,X}(y,x)$.$\square$

Since the parameter space Θ is compact, the integral in [F] exists. In many applications the object of interest is a finite vector of parameters rather than a scalar. In such a case, and following Horowitz and Manski (2006, pp. 447), we can consider θ as one of the components of the vector of interest. In order to rule out the over-identified case, we assume that $\mathbb{E}[c(Y,X;\mu)]$ is the only functional delivered by the model such that $\theta := \mathbb{E}[c(Y,X;\mu)]$.

Here is the second assumption defining the class of parameters of interest,

Assumption S (Strictly Superadditive). The function $y, x \mapsto c(y,x)$ satisfies:

[S.1] $c(y',x') + c(y,x) > c(y',x) + c(y,x')$ for every $y' > y$, $x' > x$; where $y', y \in Supp(Y)$ and $x', x \in Supp(X)$

[S.2] $y, x \mapsto c(y,x)$ is right-continuous.

The strictly superadditive assumption [S] is the key to obtain the results that follow. This assumption is convenient because it implies that $\theta := \mathbb{E}[c(Y,X;\mu)]$ is monotone in the joint distribution of (Y,X). Assumptions [F] and [S] define the class of population parameters we focus on. Examples of parameters belonging to this class include the Pearson's correlation coefficient introduced in the pilot example 1 and the Spearman's and Kendall's correlation coefficients (see Tchen, 1980). Our results shall apply as well to strictly subadditive integral functionals (i.e., parameters satisfying [F]-[S] when the inequality in [S.1] is reversed). An example of a strictly subadditive functional is the variance of the treatment effect introduced in the pilot example 2. The class [F]-[S] is similar to the class of D-parameters introduced by Horowitz and Manski (1995) and Manski (1997, pp. 1313) in the sense that both classes respect the stochastic dominance order. The difference between D-parameters and our class is that the former are functionals of univariate distributions while here the parameters are integral functionals of bivariate distributions. The variance of the impact effect, for instance, is not a D-parameter (Manski, 1997, pp 1313) but it belongs to the class studied here. On the other hand, the τ-quantile of the difference

$Y - X$, which is a parameter of interest in the literature on treatment effects, does not belong to the class studied here but it is a D-parameter (see Fan and Park, 2010).

The Available Data. We suppose that data are available from two samples, say $\{Y_i, Z_i\}_{i=1}^{n_1}$ and $\{X_j, Z_j\}_{j=n_1+1}^{n}$. We assume further that,

Assumption AD (Available Data). The samples $\{Y_i, Z_i\}_{i=1}^{n_1}$ and $\{X_j, Z_j\}_{j=n_1+1}^{n}$ are independent and identically distributed (iid) replications of the variables (Y, Z) and (X, Z) with distributions:

[AD] $\quad G_{Y,Z}(y, z) := P(Y \leq y, Z \leq z) \quad , \quad G_{X,Z}(y, z) := P(X \leq x, Z \leq z)$

We work with independent samples for simplicity although our results could be adapted to some non-independent cases. We refer $\{G_{Y,Z}, G_{X,Z}\}$ as the available data free of sample variation because they represent the data for a sample of infinite size. The absence of arguments for the functions $y, z \mapsto G_{Y,Z}(y, z)$ and $y, z \mapsto G_{X,Z}(y, z)$ denotes the entire function rather than its value at a point. Whenever convenient, we shall refer to the conditional distributions $\{G_{Y|Z}, G_{X|Z}\}$ also as the available data free of sample variation. The examples illustrate assumption [AD]:

Pilot Example 1 (Ecological Correlation cont'd). Because presidential elections employ the secret ballot it is impossible to jointly observe the voting behavior Y_i and the educational level X_i of an individual i. Election returns, however, allow us to estimate the distribution of the voting behavior by electoral precinct, $G_{Y|Z}$. Moreover, from census data we can estimate the distribution of educational level by electoral precinct, $G_{X|Z}$. Hence the available data free of sample variation consist of $\{G_{Y|Z}, G_{X|Z}\}$.□

Pilot Example 2 (Variance of the Treatment Effect cont'd). In a program evaluation with a binary intervention, each household i experiences only one of the two potential outcomes, so it is impossible to observe joint realizations of (Y_{i1}, Y_{i0}). Since interventions are assigned randomly to households, we can interpret the available data as two independent samples $\{Y_{i1}, Z_i\}_{i=1}^{n_1}$ and $\{Y_{j0}, Z_j\}_{j=n_1+1}^{n}$ from $\{G_{Y_1|Z}, G_{Y_0|Z}\}$ where Z are background variables not affected by the intervention and i, j are different households.□

The Identification and Inference Problems. At this point we can ask what can be learned about the parameter of interest θ from the available data. Following common practice in econometrics, we find useful to answer this question by treating two separate problems. The first problem consists in characterizing all the values of θ compatible with hypothetical knowledge of the distribution of (Y, Z) and of (X, Z). This is an instance of the so-called identification problem. The second one consists in determining how the feasible values of θ previously characterized can be actually recovered from two samples, say $\{Y_i, Z_i\}_{i=1}^{n_1}$ and $\{X_j, Z_j\}_{j=n_1+1}^{n}$, of finite size. This is an instance of the so-called statistical inference problem and its treatment logically precedes the treatment of the identification problem.

Before approaching these two problems, we introduce the concept of identified set, Θ_I. Such set contains all the values of the parameter of interest that are compatible with the model and hypothetical knowledge of the available data free of sample variation. For identification purposes thus, we shall regard $G_{Y|Z}$ and $G_{X|Z}$, and consequently μ, as given. Formally, we define the identified set by:

$$\Theta_I := \left\{ \begin{array}{l} \theta \in \Theta : \theta = \int_{\mathcal{Y}} \int_{\mathcal{X}} c(y, y; \mu) f_{Y,X}(y, x) dy dx \\ f_{Y,X}(y, x) = \int_{\mathcal{Z}} f_{Y,X|Z}(y, x|z) g_Z(z) dz \\ g_{Y|Z}(y|z) = \int_{\mathcal{X}} f_{Y,X|Z}(y, x|z) dy \quad \forall x \in \mathcal{X}, z \in \mathcal{Z} \\ g_{X|Z}(x|z) = \int_{\mathcal{Y}} f_{Y,X|Z}(y, x|z) dy \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z} \end{array} \right\} \tag{1}$$

where g_Z denotes the density of Z. Heuristically, θ belongs to Θ_I if and only if there exists a sequence of conditional densities $\left\{y, x \mapsto f_{Y,X|Z}(y, x|z)\right\}_{z \in \mathcal{Z}}$ matching the density of (Y, X) induced by the model, that is $f_{Y,X}$, with the available data free of sample variation represented by $g_{Y|Z}$ and $g_{X|Z}$. The conditional densities $\left\{y, x \mapsto f_{Y,X|Z}(y, x|z)\right\}_{z \in \mathcal{Z}}$ are unknown because we never observe joint realizations of (Y, Z). Challenges for identification arise because $\left\{y, x \mapsto f_{Y,X|Z}(y, x|z)\right\}_{z \in \mathcal{Z}}$ are not uniquely determined by hypothetical knowledge of $G_{Y|Z}$ and $G_{X|Z}$. To gain some intuition about we resort to our examples:

Pilot Example 1 (Ecological Correlation cont'd). The identified set for ρ contains all the *possible values for the correlation between voting behavior and educational level that are compatible with knowledge of the distributions $G_{Y|Z}, G_{X|Z}$ obtained from elections returns and the census. Characterizing the identified set in this case is an instance of the so-called ecological correlation problem (Robinson, 1950).*□

Pilot Example 2 (Variance of the Treatment Effect cont'd). The identified set for σ^2_{TE} contains all the possible values for the variance of the treatment effect that are compatible with hypothetical knowledge of the distributions $G_{Y_1|Z}, G_{Y_0|Z}$. Characterizing the identified set in this case is an instance of the so-called program evaluation problem (Heckman and Vytlacil, 2007).□

The above definition of the identified set is not an operational one, in the sense that it does not allow the computation of Θ_I based on hypothetical knowledge of $G_{Y|Z}$ and $G_{X|Z}$. The identification problem here consists in finding an equivalent operational characterization of Θ_I.

Regularity Conditions. We close this section with some regularity conditions. Even if these regularity assumptions have little bearing on applied work, they are essential to prove the results of this paper. They will be assumed valid in the rest of the text.

Assumption R - (Regularity): We assume that the following regularity conditions hold:
[R.1] The elements of the vector of functions $[Y, X, Z, c(Y, X; \mu)]$ from Ω to $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \times \mathbb{R}$ belongs to the space $L^2(\Omega, \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \times \mathbb{R})$ of square integrable functions.
[R.2] The distribution of Z, G_Z, has a strictly positive density g_Z with respect to P, that is $g_{\mathbf{z}} > 0$ for $G_Z = \int g_z dP$, $P \in \mathcal{P}$.

Assumption [R.1] ensures measurability of the objects we define on $(\Omega, \mathfrak{F}, P)$, while [R.2] ensures that the functions $G_{Y|Z}$ and $G_{X|Z}$ are well-defined.

3 Identification Results

In this section we ask which are the values of the unknown parameter of interest θ compatible with hypothetical knowledge of the available data free of sample variation (i.e., compatible with the distribution of (Y, Z) and of (X, Z)). We show that there are three equivalent ways to answer this question, which take the form of equivalent sharp characterizations of the identified set Θ_I. These characterizations are equivalent solutions to the identification problem. We use these characterizations to derive two types of related results. On one hand, we show that the parameter of interest is not uniquely defined by the available data free of sample variation, that is, the identified set contains more than one element. We discuss the implications of this result for several applications including sample combination in media planning studies, the ecological correlation problem in political science and the measurement of the variance of impact effect in the econometric evaluation of social programs. On the other hand, we use one of the characterizations to solve the inference problem but in the next section.

3.1 Geometric Properties of the Identified Set

We start by showing that the identified set Θ_I is a closed segment of the real line. This is useful to derive the main result of the paper, namely the equivalent characterizations of Θ_I. For notational convenience we focus on the case where Z is a scalar. Allowing for Z being a vector would complicate the exposition, without adding much insight. Since the distribution of (Y, Z), $G_{Y,Z}$, and of (X, Z), $G_{X,Z}$, are given we treat the nuisance parameters μ in the equation $\theta := \mathbb{E}[c(Y, X; \mu)]$ also as given, so we suppress it. We do this without loss of generality because μ only depends on $G_{Y,Z}$ and on $G_{X,Z}$.

Let $\mathcal{F}_{Y,X,Z}$ be the class of trivariate distribution functions with support on $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}$ for which the marginals $G_{Y,Z}$ and $G_{X,Z}$ are given. The range of the map $F_{Y,X,Z} \mapsto \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y, x) dF_{Y,X,Z}(y, x, z)$ is equal to Θ_I. Since this map is linear and $\mathcal{F}_{Y,X,Z}$ is non-empty and convex because of the regularity conditions, it follows that Θ_I is non-empty and convex. Moreover, Θ_I is bounded because it is a subset of the compact set parameter space Θ. That is, Θ_I is a segment of the real line. The following proposition sums up these geometric properties.

Proposition 1 *(Geometric Properties of the Identified Set) Let assumptions [F], [S], [AD] and [R] hold. Define the identified set Θ_I as in (1). Then, Θ_I is a non-empty, bounded, closed, convex subset -i.e., a segment- of the real line.*

Proof. See Appendix A ■

As a consequence of Proposition 1, Θ_I is included in the segment $[\theta_l, \theta_u]$, where θ_l and θ_u denote the infimum and supremum of the map $\mathbb{E}[c(Y, X; \mu)] : \mathcal{F}_{Y,X,Z} \rightarrow \Theta$ over the space of functions $\mathcal{F}_{Y,X,Z}$. In order to characterize Θ_I we need to find analytical expressions for θ_l and θ_u in terms of the available data free of sample variations, that is in terms of $G_{Y|Z}$ and $G_{X|Z}$-. We perform such task next.

Notice that we can determine θ_l and θ_u by solving the programming problems:

$$\theta_l := \min_{F_{Y,X,Z}} \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y, x) dF_{Y,X,Z}(y, x, z)$$
$$\text{s.t. } G_{Y,Z}(y, z) = \lim_{x\rightarrow\infty} F_{Y,X,Z}(y, x, z) \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}$$
$$G_{X,Z}(x, z) = \lim_{y\rightarrow\infty} F_{Y,X,Z}(y, x, z) \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

for the lower bound, θ_l, and the corresponding maximization problem for the upper bound, θ_u. These programming problems have a linear objective function with linear constraints and they are the object of study in the literature on mass transportation. Solving this type of problem is a potentially delicate issue, that have attracted considerable attention in the literature. However, since here $y, x \mapsto c(y, x)$ in the objective function is strictly superadditive our problem is much less complicated and has a well known unique closed form solution (see Ruschendorf, 1991; Rachev and Ruschendorf, 1998). In particular, we have:

$$\begin{aligned} G^l_{Y,X|Z}(y, x, z) &= \max\{0, G_{Y|Z}(y|z) + G_{X|Z}(x|z) - 1\} \\ G^u_{Y,X|Z}(y, x, z) &= \min\{G_{Y|Z}(y|z), G_{X|Z}(x|z)\} \end{aligned}$$

The functions $G^l_{Y,X1Z}$ and $G^u_{Y,X|Z}$ are the so-called Frechet distributions and they have been used to bound the distribution $F_{Y,X|Z}$ by Ridder and Moffit (2007). We next use the Frechet distributions to characterize Θ_I.

3.2 Equivalent Characterizations of the Identified Set

The main result of this paper is the following theorem which suggest equivalent ways to write the identified set Θ_I.

Theorem 1 *(Equivalent Characterizations of the Identified Set) Let assumptions [F], [S], [AD] and [R] hold. Define the identified set Θ_I as in (1). Let $Q_{X|Z}$ denote the quantile function of X given Z, and $G_{Y|Z}$ denote the distribution of Y given Z. Define the random variables $\xi^l(Y,Z) := Q_{X|Z}[1 - G_{Y|Z}(Y|Z)|Z]$ and $\xi^u(Y,Z) = Q_{X|Z}[G_{Y|Z}(Y|Z)|Z]$. Then, the following characterizations of Θ_I are equivalent:*

(i) $\Theta_I = [\theta_l, \theta_u]$, *where θ_l and θ_u are defined by:*

$$\begin{aligned}
\theta_l &= \int_{\mathcal{Z}} \int_0^1 c(Q_{Y|Z}(\tau|z), Q_{X|Z}(1-\tau|z)) d\tau dG_Z(z) \\
\theta_u &= \int_{\mathcal{Z}} \int_0^1 c(Q_{Y|Z}(\tau|z), Q_{X|Z}(\tau|z)) d\tau dG_Z(z)
\end{aligned}$$

(ii) $\Theta_I = \mathbb{A}[\mathbb{A}(R|Z)]$, *where $\mathbb{A}$ denotes the Aumann expectation and R the random set*

$$R(z) := \{c(Y,\xi) : \xi \in [\xi^l(Y,z), \xi^u(Y,z)]\} \tag{2}$$

(iii) $\Theta_I = \left\{\theta \in \Theta : \mathbb{E}[c(Y,\xi^l(Y,Z))] \le \theta \le \mathbb{E}[c(Y,\xi^u(Y,Z))]\right\}$.

Proof. See Appendix A. ■

Each of the characterizations of Θ_I in Theorem 1 can be interpreted as conceptually distinct representations of the same object, so that their equivalence is a result in itself. All of these characterizations are sharp because they contain the feasible values of the parameter of interest and no others. Our first characterization of the identified set $\Theta_I = [\theta_l, \theta_u]$ is obtained from plug the conditional Frechet distributions in the definition of θ. This characterization is new in the form stated, but literature on treatment effects contains numerous specializations that anticipate the intuition (e.g., Heckman, Smith and Clements, 1997).[1] The characterization given in expression Theorem 1(ii) depicts Θ_I as the Aumann expectation of the random set R.[2] The random set R is made up of the random variables $c(Y, \xi)$ that are compatible with the data and the maintained assumptions. The extreme elements of this random set are given by the map $y, x \mapsto c(y, x)$ evaluated at a pair of random vectors having Frechet distributions (1)-(2). Heuristically, Θ_I can be though as the "expectation" of a bundle of random variables -i.e., the random set R-. The expression "expectation" used above is in quotation marks because when working with random sets, a particular expectation operator needs to be used, the Aumann expectation. The key insight leading to this representation is the observation that Θ_I is a convex set. The characterization of the identified set given in expression Theorem 1(iii) depicts Θ_I in terms of moment inequalities. This characterization follows from Theorem 1(ii) by means of Hormander's embedding (Molchanov, 2005 pp.157). In the next section, we use the characterization in Theorem 1(iii) to propose inference procedures for θ.

Before presenting the inference procedures, we discuss three issues related to Theorem 1, namely potential applications of this result, a remark on set identification, and the role played by the common variable Z.

Applications. The following applications illustrate potential uses of Theorem 1 in concrete situations. The case from media planning introduced at the beginning of the paper is an example of the type of problem treated in the literature on data fusion (e.g., Rassler, 2002). The current approach to solve this problem (e.g., Gilula, McCulloch and Rossi, 2006), is to brought to bear additional assumptions, in particular independence of Y and X conditional on Z, to point-identify the correlation of interest. By contrast, here we do not impose such type of assumptions to derive results. Therefore, Theorem 1 enable the researcher to see what is lost when the conditional independence assumption is jettisoned. It turns out that what is lost is point-identification.

[1] Concurrent work by Fan and Zhu (2010) presents this characterization.

[2] See Molchanov (2005) for a general introduction to the theory of random sets. We refer to Beresteanu and Molinari (2008) for a concise and gentle introduction to this theory with applications in econometrics.

Other potential application is the ecological correlation problem presented in our pilot example 1. Stimulated by Ogburn (1919), political scientists have sought to learn about the correlation between voting behavior Y and socioeconomic characteristics, such as educational level X, in elections with secret ballot using administrative records by voting district Z and census data on the educational level of individuals in each district. One of the current approaches to solve this problem (e.g., Gentzkow, 2006) is to aggregate the voting behavior and socioeconomic characteristic by district into shares, calculate the correlation between the aggregate shares and make assumptions ensuring the equivalence between the correlation at individual and aggregate level. By contrast, we do not impose such type of assumptions. Our results therefore can be used to carry out a conservative analysis on the correlation between voting behavior and socioeconomic characteristics in secret ballots, in the spirit of Manski (2003), to analyze the sensitivity of inferences to failure of the assumptions currently adopted. This analysis should help to make plain the limitations of the available data while highlighting the identification power of the ancillary assumptions.

The identification and estimation of the variance of the treatment effect presented in the our pilot example 2 is another potential application we review. Djebbari and Smith (2008) seek to identify and estimate such variance. They acknowledge that their characterization for the identified set is not sharp. By contrast, the characterization in Theorem 1 is sharp. Interesting enough, Theorem 1 enables us to phrase Djebbari and Smith (2008)´s identification problem in terms of moment inequalities. In a related type of application, several researchers have sought to restrict the across regime correlation in switching regression models.[3] Our results apply with similar concerns to this literature.

The last application we review is the linear regression model. This application illustrates the limitations of Theorem 1 to handle the case where the parameter of interest is overidentified. Knowledge in this example is sought about the conditional mean of Y on X, which is modeled using the linear function $\alpha + X\beta$, where α and β are real-valued parameters. The parameter of interest is defined by $\beta := \mathbb{E}(X\mathbf{W})^{-1}\mathbb{E}(Y\mathbf{W})$ where $\mathbf{W}$

[3] See Fan and Wu (2010) and references therein.

is a vector of positive transformations of X, for example, $\mathbf{W} = \{I(X \in \mathcal{X}_s), s = 1, .., S\}$ for a suitable collection of sets $\mathcal{X}_s$. Notice that β is overidentified. Let $\underline{y}$ and $\bar{y}$ denote, respectively, the infimum and supremum of $supp(Y)$. With loss of generality, fix $S = 2$, $\mathcal{X}_1 = [\underline{y}, \bar{y}/2]$ and $\mathcal{X}_2 = (\bar{y}/2, \bar{y}]$, so $\mathbf{W} = (W_1, W_2)$. Since $\mathbb{E}(YW_1)$ is superadditive so is the parameters of interest β. Then Theorem 1 (iii) implies the inequalities

$$\mathbb{E}(\xi^u \times W_1)\mathbb{E}(YW_1) \leq \beta \leq \mathbb{E}(\xi^l \times W_1)^{-1}\mathbb{E}(YW_1)$$

Since ξ^l, ξ^u and W_1 are revealed by the data free of sample variation, the later inequalities can be used to estimate bounds on β. We can not, however, claim that such bounds are sharp because we have not used the all the available inequalities on β provided by our assumptions.

Set-Identification. In the recent econometric literature (c.f., Tamer, 2009), a parameter of interest is said to be *set-identified* or *partial-identified* when the identified set is not a singleton but it is a proper subset of the parameter space. According to Theorem 1, the identified set in our case is not a singleton. However, we can not claim that Θ_I is a proper subset of Θ and thus two independent samples on (Y, Z) and (X, Z) may not contain any information about some superadditive parameters defined in terms of $F_{Y,X}$. For this reason, we refrain from saying that superadditive parameters are set-identified when the available data are two-samples of the type considered here. An example helps to clarify this issue. The Kendall's tau between Y and X is a superadditive parameter measuring the "concordance" between the random variables (see Tchen, 1980). The parameter space for Kendall's tau is $\Theta = [-1, 1]$. Using Theorem 1(i) it is possible to show that, when the available data are two independent sample on (Y, Z) and (X, Z), the identified set for this parameter is $\Theta_I = [-1, 1]$. That is, the Kendall tau between Y and X is neither point nor set-identified with our available data.

The Role of the Common Variable Z. The presence of a common variable Z measured in both samples is the main feature that distinguishes our setting from the one studied by Cambanis, Simons and Stout (1976). The results by these authors can be applied to characterize the identified set for a superadditive parameter when only the marginals distributions of Y and X, G_Y and G_X, are given. Let Θ_C denote such a set. The next corollary to theorem 1 shows that the presence of Z may shrink the identified set derived using the results by Cambanis et. al. (1976).

Corollary 1 *Let Θ_C be the identified set for the parameter of interest without considering common variables measured in both samples:*

$$\Theta_C = \{\theta \in \Theta : \int_0^1 c(Q_Y(\tau), Q_X(1-\tau))d\tau \leq \theta \leq \int_0^1 c(Q_Y(\tau), Q_X(\tau))d\tau\}$$

Under the assumptions of Theorem 1, the identified set considering common variables Θ_I is a subset of Θ_C -i.e., $\Theta_I \subseteq \Theta_C$-.

Proof. See Appendix A. ■

The following example, adapted from Ridder and Moffit (2007), illustrates Corollary 1. Suppose that the joint distribution of (Y, X, Z) is trivariate normal and the superadditive parameter of interest θ is the correlation between Y and X. Let r_{YZ} and r_{XZ} denote, respectively, the correlation between Y and Z and between X and Z. Since G_Y and G_X belong to the same location-scale family, we have $\Theta_C = [-1, 1]$. Using our theorem it is possible to show (see also Rachev and Ruschendorf, 1998) that $\Theta_I = [r_{YZ}r_{XZ} - \sqrt{(1-r_{YZ}^2)(1-r_{XZ}^2)}, r_{YZ}r_{XZ} + \sqrt{(1-r_{YZ}^2)(1-r_{XZ}^2)}]$. If either $r_{XZ} \neq 0$ or $r_{YZ} \neq 0$, then Θ_I is a proper subset of Θ_C. That is, the correlation between Y_1 and Y_2 is set-identified when, for instance, Y and Z are correlated. In particular, if either $|r_{YZ}| = 1$ or $|r_{XZ}| = 1$, then $\Theta_I = \{r_{YZ} \times r_{XZ}\}$ is a singleton. That is, the correlation between Y and X is point identified if Y and Z are perfectly correlated. The message to take away from this example is that the presence of common variable Z may help in some cases to restrict the correlation between Y and X.

4 Inference Procedures

In this section we ask how the identified set Θ_I characterized in the previous section can be actually recovered from two samples on (Y, Z) and (X, Z) of finite size. To answer this question, we start by proposing an estimator for the identified set (actually for its extreme points θ_l and θ_u), which builds on Theorem 1. We continue by approximate the distribution of the proposed estimator using asymptotic theory. This approximation could be useful in constructing asymptotic confidence intervals for the true value of the parameter of interest of the type proposed by Imbens and Manski (2004), and later refined by Stoye (2009). Finally, we evaluate the small sample properties of the proposed estimator by means of Monte-Carlo experiments.

4.1 A Multi-step Estimator of the Identified Set based on Series

Here we propose an estimator for the extreme points θ_l and θ_u of the identified set $\Theta_I = [\theta_l, \theta_u]$. According to Theorem 1 such extreme points are defined by the moment conditions

$$\theta_l := \mathbb{E}[c(Y, \xi^l(Y, Z); \mu_o)] \quad , \quad \theta_u := \mathbb{E}[c(Y, \xi^u(Y, Z); \mu_o)] \tag{MC}$$

where $\xi^l(Y, Z) := Q_{X|Z}(1 - G_{Y|Z}(Y|Z)|Z)$ and $\xi^u(Y, Z) := Q_{X|Z}(G_{Y|Z}(Y|Z)|Z)$ are compositions of the quantile function of Y given Z, $\tau \mapsto Q_{X|Z}(\tau|z)$, and the distribution function of Y given Z, $y \mapsto G_{Y|Z}(y|z)$, and μ_o is the true value of the vector of parameters μ. If we knew $Q_{X|Z}$ and $G_{Y|Z}$, we could determine the extreme points of the identified set. When dealing with the inference problem, however, this is not the case and thus (μ, ξ^l, ξ^u_o) are unknown and they must be treated as nuisance parameters to be estimated from the available data. Therefore, our inference problem can be interpreted as one in which the objects of interest, θ_{lo} and θ_{uo}, are defined by the unconditional moment restrictions (MC) containing unknown finite-dimensional nuisance parameters μ, and unknown infinite-dimensional nuisance functions (ξ^l, ξ^u). Similar inference problems have attracted considerable interest in recent literature on semiparametric models (see

Andrews, 1994; Newey, 1994 and Chen, Linton and van Keilegom, 2003). Following this literature, we introduce a multistep estimator for the extreme points θ_l and θ_u of the identified set, where in a first step estimators of the nuisance parameters (μ, ξ^l, ξ^u) are formed. In a second step, an estimator of $[\theta_l, \theta_u]$ based on the sample analog of (MC) are computed but using the predicted values of (μ, ξ^l, ξ^u).

Nonparametric Estimation, **Support Conditions and Set-Identification.** To implement the estimator of $[\theta_l, \theta_u]$ we propose, we need to be specific about the estimators of the nuisance parameters (μ, ξ^l, ξ^u). In order to focus in a situation where estimation is more challenging, we start by making assumptions which rule out cases where existing results can be applied to solve the estimation problem we face. These assumptions concern the non-parametric nature of ξ^l and ξ^u, support conditions on (Y, X, Z), and set-identification.

We start by ruling out the case where $G_{Y|Z}$ and $Q_{X|Z}$, and consequently the nuisance functions (ξ^l, ξ^u), can be estimated using parametric methods. In some applications (e.g., Fan and Wu, 2010), additional assumptions on the population P are imposed, such as normality, so that the unknown functions $G_{Y|Z}$ and $Q_{X|Z}$ are restricted to live in a given parametric family. In such a case existing parametric methods can be applied to make inference about the extreme points of the identified set. By contrast, here we do not impose any parametric restriction on P so the later parametric inference methods are not an option for our first step estimator.

We now impose conditions on the support of the random variables so as smoothing methods for the first step estimator need to be used. When the common variable Z is discrete (as it is usually the case in the ecological correlation problem), inference on the extreme points of the identified set can be performed using existing techniques. First, the two samples should be split into subgroups according to the values of Z. Second, the functions $y \mapsto G_{Y|Z}(y|z)$ and $\tau \mapsto Q_{X|Z}(\tau|z)$ should be estimated, respectively, using the empirical distribution function and its generalized inverse for each subgroup so as to construct estimators for the nuisance parameters μ, ξ^l and ξ^u. Third, a method-of-moment estimator for the identified set should be constructed using the predicted values of (μ, ξ^l, ξ^u) from the previous step. Asymptotic properties of the resulting estimators of θ_l and θ_u can be derived by extending the results in Athey and Imbens (2006, Theorem 5.1) to the case at hand. By contrast, we focus throughout the case where (Y, X, Z) have compact supports:

Assumption - Estimation: Support Conditions.

[E.1] *Supp*(Y), *Supp*(X) *and Supp*(Z) *are compact sets.*

We end by ruling out the case where the parameter of interest is point-identified. When θ is point-identified there is no need to implement our two-step estimator. To avoid such a situation we assume that,

Assumption - Estimation: - Set Identification

[E.2] *The conditional variances* $z \mapsto \mathbb{V}(Y|z)$ *and* $z \mapsto \mathbb{V}(X|z)$ *are bounded away from zero.*

[E.3] θ_l *and* θ_u *are in the interior of the parameter space* Θ

Assumption [E.1] rules out the case where Y or X are perfectible predictive from Z. The parameter of interest then can not be point-identified. Assumption [E.2], on the other hand requires that the available data have some information about the parameter of interest. Assumption [ESI] therefore ensures that the parameter of interest is set-identified. Assumptions [E.1]-[E.3] together justify the introduction of the estimator we propose below.

The Estimator. In this subsection we present a two-step estimator for the identified set under assumptions [E.1]-[E.3]. Recall that our problem is to estimate the extreme points of the identified set, θ_l and θ_u, using the moment conditions [MC] and we would use a method-of-moment estimator if the nuisance parameters μ, ξ^l and ξ^u were known. Because these parameters are not known, we plug preliminary nonparametric estimates of them into moment conditions defining the extreme points of the identified set. Estimators of θ_l and θ_u are then formed by replacing the later moments with their sample analogs.

We start the presentation by assuming that a first stage procedure has obtained an estimate of the vector of nuisance parameters μ, say $\hat{\mu}$. In the next subsection, we illustrate this assumption using our pilot examples. We continue by providing details on the nonparametric estimates of the nuisance functions ξ^l and ξ^u. Recall that ξ^l and ξ^u are the composition of the conditional quantile function $\tau, z \mapsto Q_{X|Z}(\tau|z)$ and conditional distribution function $y, z \mapsto G_{Y|Z}(y|z)$. To estimate nonparametrically ξ^l and ξ^u we compose series estimators of the later two functions. We make this choice because we find this series estimator relatively simple to implement with respect to alternative procedures, such as a plug-in method based on kernels. Indeed, once $y, z \mapsto G_{Y|Z}(y|z)$ and $\tau, z \mapsto Q_{X|Z}(\tau|z)$ have been replaced by their series estimators our estimation problem effectively becomes a parametric one; hence commonly used software can be used to estimate θ_l and θ_u.

In order to describe the series estimator we need to introduce some notation. For a fixed y, let $\mathcal{G}$ be the space of all bounded continuous functions $z \mapsto G_{Y|Z}(y, z)$. Similarly, let $\mathcal{Q}$ be the space of all bounded continuous functions $z \mapsto Q_{X|Z}(\tau|z)$. Given the function spaces $\mathcal{G}$, our first task is to construct a sequence of approximating spaces $\mathcal{G}_1, \mathcal{G}_2, .., \mathcal{G}_{n_1}$ for $\mathcal{G}$ -i.e. a sieve-. There are many sieves that could be used -such as Fourier series, power series, spline, wavelets, etc.- whose approximation properties are already known. The choice of a particular one depends on how well the sieve approximates $\mathcal{G}$ and how easily we can compute this approximation (Chen, 2007 pp. 5579). Following these two criteria, we find suitable for our purposes to choose the sieve defined by a univariate spline basis. Formally,

$$\mathcal{G}_{n_1} = \left\{ G(y,z) : \mathcal{Z} \mapsto [0,1] \ ; \ G(y,z) = \sum_{k=1}^{K_{n_1}+m_1+1} \alpha_k(y) \times p_k(z); \ \alpha_k \in \mathbb{R}, \ y \in \mathcal{Y} \right\}$$

where $p_k(z)$ is a piecewise polynomial of degree m_1 that is smoothly connected at its knots -i.e., a spline-.[4] The integer K_{n_1} represents the number of knots at which the spline is defined. K_{n_1} is is required to grow with n_1 so that $\mathcal{G}_{n_1}$ becomes dense in $\mathcal{G}$, and it can be view as a smoothing parameter. The vector $\boldsymbol{\alpha}(y) = (\alpha_1(y), ..., \alpha_{K_{n_1}+m_1+1}(y))'$ contains

[4] See Chen (2007) for futher details.

unknown parameters to be estimated.

Given the function space $\mathcal{G}$ and its associated sieve $\mathcal{G}_{n_1}$, we can construct the estimator for $y, z \mapsto G_{Y|Z}(y|z)$. In order to do so, define the variable $W_i(y) = \mathbb{I}(Y_i \leq y)$ where $\mathbb{I}(Y_i \leq y)$ is an indicator function that equals one if $Y_i \leq y$ and zero otherwise. The sieve estimator for $y, z \mapsto G_{Y|Z}(y|z)$ we propose is

$$\hat{G}_{Y|Z}(y, z) = \sum_{k=1}^{K_{n_1}+m_1+1} \hat{\alpha}_k(y) \times p_k(z) \tag{3}$$

where $\hat{\alpha}_k(y)$ solves the minimization problem

$$\hat{\alpha}_k(y) = \arg \min_{\alpha_1,..,\alpha_{K_{n_1}+m_1+1}} \sum_{i=1}^{n_1} \left(w_i(y) - \sum_{k=1}^{K_{n_1}+m_1+1} \alpha_k \times p_k(Z_i) \right)^2 \tag{4}$$

The estimator $\hat{G}_{Y|Z}(y, z)$ can be interpreted as a least square estimator of the conditional expectation $y, z \mapsto \mathbb{E}(\mathbb{I}(Y \leq y)|Z = z)$ over the sequence of approximating spaces $\mathcal{G}_{n_1}$, or, alternatively, as a weighted version of the usual empirical distribution function.[5] Asymptotic properties of this estimator has been studied, among others, by Song (2008).

We now turn to the estimation of the conditional quantile function $\tau \mapsto Q_{X|Z}(\tau|z)$. To describe the estimator of this function, let $\rho_\tau(t) = |t| - (2\tau - 1)t$ be the czech function as in Koenker (2005). Using a procedure similar to the one employed to estimate $y, z \mapsto G_{Y|Z}(y|z)$, we approximate $\mathcal{Q}$ by the sieve

$$\mathcal{Q}_{n_2} = \left\{ Q(\tau, z) : \mathcal{Z} \mapsto \mathcal{X} \ ; \ Q(\tau, z) = \sum_{l=1}^{L_{n_2}+m_2+1} \beta_l(\tau) \times p_l(z); \ \beta_l \in \mathbb{R}, \ \tau \in [0, 1] \right\}$$

where L_{n_2} is the smoothing parameter and m_2 is the order of the spline. We propose the following estimator for the the function $\tau, z \mapsto Q_{X|Z}(\tau|z)$,

$$\hat{Q}_{X|Z}(\tau, z) = \sum_{l=1}^{L_{n_2}+m_2+1} \hat{\beta}_l(\tau) \times p_l(z) \tag{5}$$

[5] One can estimate $H_{1|Z}$ by any nonparametric method such as kernel, nearest neighbor or local linear regression.

where $\hat{\beta}_l(\tau)$ solves the minimization problem

$$\hat{\beta}(\tau) := \arg \min_{\beta_1,..,\beta_{L_{n_2}+m_2+1}} \sum_{j=1}^{n_2} \rho_\tau \left(X_j - \sum_{l=1}^{L_{n_2}+m_2+1} \beta_l \times p_l(Z_j) \right) \tag{6}$$

Asymptotic properties of this estimator have been studied, among others, by Horowitz and Lee (2005). To get the series estimator for the nuisance functions ξ^l and ξ^u we compose (6) and (7),

$$\hat{\xi}^l(y,z) = \sum_{l=1}^{L_{n_2}+m_2+1} \hat{\beta}_l \left(1 - \sum_{k=1}^{K_{n_1}+m_1+1} \hat{\alpha}_k(y) \times p_k(z) \right) \times p_l(z) \tag{7}$$

$$\hat{\xi}^u(y,z) = \sum_{l=1}^{L_{n_2}+m_2+1} \hat{\beta}_l \left(\sum_{k=1}^{K_{n_1}+m_1+1} \hat{\alpha}_k(y) \times p_k(z) \right) \times p_l(z) \tag{8}$$

Since $y, z \mapsto \hat{G}_{Y|Z}(y,z)$ and $\tau, z \mapsto \hat{Q}_{X|Z}(\tau, z)$ are estimators arising from minimizing concave criterion functions over finite linear sieve spaces, we say that $\hat{\xi}^l(y,z)$ and $\hat{\xi}^u(y,z)$ are *series estimators.*

We propose to estimate θ_l and θ_u by the empirical counterpart of the moments (MC)

$$\hat{\theta}_l = n_1^{-1} \sum_{i=1}^{n_1} c[Y_i, \hat{\xi}^l(Y_i, Z_i); \hat{\mu}] \quad ; \quad \hat{\theta}_u = n_1^{-1} \sum_{i=1}^{n_1} c[Y_i, \hat{\xi}^u(Y_i, Z_i); \hat{\mu}]$$

where the unknown nuisance parameters (μ, ξ^l, ξ^u) have been replaced by their estimated values $(\hat{\mu}, \hat{\xi}^l, \hat{\xi}^u)$. The estimator for the identified set is $\hat{\Theta}_I = [\hat{\theta}_l, \hat{\theta}_u]$. This estimator has an appealing interpretation. Observe that $\hat{\xi}^u(Y_i, Z_i)$ computes a unit i location on the conditional distribution of Y given Z and reassigns it the corresponding quantile of the conditional distribution of X given Z. When $\hat{G}_{Y|Z}(y_i, z_i) > .5$, the quantity $\hat{\xi}^u(Y_i, Z_i)$ is an estimate of the least upper bound for the value that the unobserved variable X_i can take. Similarly, $\hat{\xi}^l(Y_i, Z_i)$ is an estimate of the greatest lower bound for X_i. That is, we can think of X_i as an unknown real random variable bracketed below by $\hat{\xi}_i^l$ and above by $\hat{\xi}_i^u$, both of which are observed real random variables. Therefore, $\hat{\theta}_l$ and $\hat{\theta}_u$ are the sample analog estimator of the parameter of interest when we impute to the unobserved variable

X_i, respectively, its minimum and supremum values compatible with the available data and the maintained assumptions.

We end this subsection by mentioning computational aspects of our two steps estimator. Minimization problems (4) and (6) are, respectively, unconstrained quadratic and linear programming problems. The quadratic minimization problem (6) has an analytical solution, so estimation of $\boldsymbol{\alpha}(y)$ can be performed without numerical optimization routines. Indeed, once the spline basis has been constructed estimation of the vector $\boldsymbol{\alpha}(y)$ can be carried out by ordinary least squares. The linear minimization problem (6) has not an analytical solution, but can be solved by using computation methods developed for linear quantile regression methods.[6] Importantly enough, our two-step estimator does not require numerical inversion of estimated conditional distribution functions as it is the case for some kernel-based estimator. From a computational point of view, this is an advantage because such type of numerical inversion is known to be a delicate issue (see Yu, Lu and Stander, 2003 pp. 341). The estimators $(\hat{\theta}_l, \hat{\theta}_u)$ depend on the number of approximating functions K_{n_1} and L_{n_2}. Derivation of optimal of optimal values for these tuning parameters is beyond the scope of this paper.

4.2 Asymptotic Properties

We now turn to the large-sample properties of the estimators of the extreme points of the identified set, $\hat{\theta}_l$ and $\hat{\theta}_u$. We begin by stating a number of sufficient conditions under which $\hat{\theta}_l$ and $\hat{\theta}_u$ are consistent and asymptotically normal. Using these asymptotic properties and the fact that $\hat{\theta}_u$ is almost sure greater than $\hat{\theta}_l$, we continue by discussing how to construct confidence set of the type proposed by Imbens and Manski (2003) for the true value of the parameter of interest.

Assumptions. In order to derive the asymptotic properties of $\hat{\theta}_l$ and $\hat{\theta}_u$, we first note that these estimators are a particular example of the general class of semiparametric estimators studied, among others, by Andrews (1994), Newey (1994) and Chen, Linton and van

[6]Methods to estimate linear quantile regressions are widely available in standard statistical softwares. See the package "quantreg" in R or the function "qreg" in STATA.

Keilegom (2003). These authors study the asymptotic properties of estimator defined as the "zeros" of a system of equations which are both indexed by a finite-dimensional vector of parameter of interest and an infinite-dimensional nuisance parameter. In our case, the equations are the moments (MC), the vector of parameter of interest are the extreme points of the identified set $[\theta_l, \theta_u]$, and the nuisance parameters are (μ, ξ^l, ξ^u). The later connection is useful because it allows us to use the results from the literature on semiparametric estimation to derive the asymptotic properties of $\hat{\theta}_l$ and $\hat{\theta}_u$.

Some additional notation is required to state the asymptotic properties of the estimators. Vectors are denoted in bold face. All vectors are column vectors. We denote $\boldsymbol{\theta}_o = (\theta_{lo}, \theta_{uo})$ for true values of the extreme points of the identified set, μ_0 for the true value of the finite-dimensional nuisance parameters μ and $\boldsymbol{\xi}_o = (\xi_o^l, \xi_o^u)$ for true value of the infinite-dimensional nuisance parameters ξ^l and ξ^u. Let $\mathcal{M} \subseteq \mathbb{R}^{d_\mu}$ denote the parameter space for μ, where $d_\mu \leq \infty$ denotes the dimension of μ. Similarly, let Ξ denote the space of functions where $\boldsymbol{\xi}$ lives. Equip $\mathcal{M}$ and Ξ with norms $||\cdot||$ and $||\cdot||_\Xi$, respectively. Define the nonrandom vector-valued function $\mu, \boldsymbol{\xi} \mapsto \mathbb{E}\left[c(Y, \boldsymbol{\xi}(Y,Z), \mu)\right]$ from $\mathcal{M} \times \Xi$ into $\mathbb{R}^2$ by $\mathbb{E}\left[c(Y, \boldsymbol{\xi}(Y,Z), \mu)\right] := (c[Y, \xi^l(Y,Z); \mu], c[Y, \xi^u(Y,Z); \mu])$. Using this notation, the extreme points of the identified set $\boldsymbol{\theta}_o$ are defined by the equation $\mathbb{E}\left[c(Y, \boldsymbol{\xi}_o(Y,Z); \mu_o) - \boldsymbol{\theta}_o\right] = 0$.

The first of the assumptions we impose regards smoothness restrictions on the functions $G_{Y|Z}$, $G_{X|Z}$ and g_Z.

Assumption - Differentiability Conditions on $G_{Y|Z}$ and $G_{X|Z}$

[E.4] $G_{Y|Z}(y|z) : \mathcal{Y} \times \mathcal{Z} \rightarrow [0,1]$ and $G_{X|Z}(x|z) : \mathcal{X} \times \mathcal{Z} \rightarrow [0,1]$ are p-times uniformly continuously differentiable with respect to all of their arguments at $y \in int\{Supp(Y)\}$, $x \in int\{Supp(X)\}$, $z \in int\{Supp(Z)\}$ with $p \geq 2$ and has uniformly continuous p-order derivatives at the boundary of $Supp(Y)$, $Supp(X)$ and $Supp(Z)$.

[E.5] The density of Z, $z \mapsto g_Z(z)$, is twice continuously differentiable at $z \in int\{Supp(Z)\}$ and has continuous second-order derivatives at the boundary of $Supp(Z)$.

The continuity assumptions [E.4] and [E.5] are standard in the literature on nonparametric estimation of distribution functions (see Imbens and Newey, 2009). They have major implications in terms of convergence rates. Without them, the rate of convergence could be slower and the asymptotic distribution could be non-normal. We continue by imposing assumptions on how well the sieves approximate the target spaces.

Assumption - Series Approximation

[E.6] $\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'\boldsymbol{\alpha}(y) - G_{Y|Z}(y|z)\right| = O(K_{n_1}^{-1/2})$

[E.7] $\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'\boldsymbol{\beta}(\tau) - Q_{X|Z}(\tau|z)\right| = O(L_{n_1}^{-1/2})$

[E.8] The conditional variance $z \mapsto \mathbb{V}[\mathbb{I}(Y < y)^2|z]$ is bounded uniformly in y.

[E.6] and [E.7] requires that the approximation error to the target functions $G_{Y|Z}$ and $Q_{X|Z}$ should vanish at a polynomial rate of the number of functions in the corresponding sieve uniformly over both $\mathcal{Y}$ and $\mathcal{Z}$. This assumption controls the asymptotic bias of the estimators of $G_{Y|Z}$ and $Q_{X|Z}$.[7] The bounded conditional variance assumption [E.8] is standard in series estimation (c.f., Newey, 1997) and it is difficult to relax without affecting the convergence rates.

To obtain convergence rates for the series estimators it is necessary to restrict the rate at which the number of functions in the sieves increases when the sample sizes increases. The following condition fulfills this purpose.

Assumption - Smoothing Parameters. Let $\bar{C}$ denote a constant and let $n = n_1 + n_2$. As $n_1 \to \infty$, $n_2 \to \infty$ and $n_1/n = n_2/n \to \bar{C}$

[E.9] $K_{n_1} \to \infty$, $K_{n_1} = o(n_1^{-1/2})$, $n_1^{-3/4}K_{n_1} \to \infty$

[E.10] $L_{n_2} \to \infty$, $L_{n_2} = o(n_2^{-1/2})$, $n_2^{-3/4}L_{n_2} \to \infty$

Assumptions [E.4]-[E.10] are used in the following proposition to derive properties of the estimator of the nuisance parameter $\boldsymbol{\xi}$, $\hat{\boldsymbol{\xi}}$, which are useful for showing asymptotic normality.

Proposition 2 *(Consistent Estimation of $\boldsymbol{\xi}$) Let assumptions [AD], [E.4]-[E.10] holds. Define the estimators of $\boldsymbol{\xi} = (\xi^l, \xi^u)$, $\hat{\boldsymbol{\xi}} = (\hat{\xi}^l, \hat{\xi}^u)$, as in (7)-(8). Then,*

(i) $||\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}||_{\Xi} = o_P(n^{-1/4})$

[7] [E.7] is analogous to assumption 4 in Song (2008, pp. 1470), condition B(iv) of Lemma A.1 in Ai and Chen (2003), Assumption 3 in Newey (1997, pp. 150) and Assumption 3 in de Jong.(2002, pp. 2).

(ii) $\hat{\boldsymbol{\xi}}$ *belongs to* Ξ *with probability tending to one.*

Proof. See Appendix A. ■

We now impose conditions on the estimators of the nuisance parameters μ, $\hat{\mu}$, similar to those derived for $\hat{\xi}$:

Assumption - Consistent Estimation of μ. *There is* $\hat{\mu}$ *such that:*

[E.11] $||\hat{\mu} - \mu|| = o_P(n^{-1/4})$ and $\hat{\mu}$ *belongs to* $\mathcal{M}$ *with probability tending to one.*

[E.12] $n^{1/2}(\hat{\mu} - \mu_o) = n^{-1/2} \sum_{i=1}^{n_1} \phi(Y_i, Z_i) + o_P(1)$

To illustrate this later assumption, we lean on our example.

Pilot Examples 1 and 2. When the parameter of interest is the correlation between Y *and* X *or the variance of the difference* $Y - X$ *the nuisance parameters* μ *is equal to* $\mu = (\mathbb{E}(Y), \mathbb{E}(X), \mathbb{V}(Y), \mathbb{V}(X))$. *Under assumption [AD], the sample analog of* μ, $\hat{\mu} = \left(n_1^{-1} \sum_{i=1}^{n_1} y_i, n_2^{-1} \sum_{j=1}^{n_2} x_j, etc\right)$, *satisfies [E.3].*□

We continue by imposing assumption on the map $\mu, \boldsymbol{\xi} \mapsto \mathbb{E}\left[c(Y, \boldsymbol{\xi}(Y, Z), \mu)\right]$. We first assume that $\mu, \boldsymbol{\xi} \mapsto \mathbb{E}\left[c(Y, \boldsymbol{\xi}(Y, Z), \mu)\right]$ is pointwise Lipschitz continuous with respect to the nuisance parameters μ and $\boldsymbol{\xi}$:

Assumption - Continuity Conditions on $\mu, \boldsymbol{\xi} \mapsto \mathbb{E}\left[c(Y, \boldsymbol{\xi}(Y, Z), \mu)\right]$.

[E.13] There exists a finite constant $\bar{C}$ *such that*

$$||\mathbb{E}\left[c(Y, \boldsymbol{\xi}(Y, Z), \mu)\right] - \mathbb{E}\left[c(Y, \boldsymbol{\xi}_o(Y, Z), \mu_o)\right]|| < \bar{C} \max\left\{||\mu - \mu_o|| \, , ||\boldsymbol{\xi} - \boldsymbol{\xi}_o||\right\}$$

Assumption [E.13] is convenient because it enables us to preserve properties on estimators of the nuisance while estimating the parameters of interest. In particular it allows us to bound from above the bracketing number of the class of functions $\{c(Y, \boldsymbol{\xi}(Y, Z), \mu) : \mu \in \mathcal{M}, \boldsymbol{\xi} \in \Xi\}$ by the bracketing number of the parameter class $\{\mu \in \mathcal{M}, \boldsymbol{\xi} \in \Xi : ||\boldsymbol{\xi} - \boldsymbol{\xi}_o||_{\Xi} \leq \delta_n\}$ (see Theorem 2.7.11 of van der Vaart and Wellner, 1996).[8]

[8] Assumption [E.13] can be relaxed to the case where the function is not pointwise Lipschitz continuous at the expense of a more tediuous proof using the Theorem 3 by Chen et. al. (2003).

To capture the effect of the estimation of $(\mu_o, \boldsymbol{\xi}_o)$ via $(\hat{\mu}, \hat{\boldsymbol{\xi}})$ on the variability of $\hat{\boldsymbol{\theta}}$ we need to impose assumptions on the derivatives of $\mu, \boldsymbol{\xi} \mapsto c(Y, \boldsymbol{\xi}(Y,Z), \mu)$ with respect to both μ and $\boldsymbol{\xi}$:

Assumption - Differentiability Conditions on $\mu, \boldsymbol{\xi} \mapsto c(y, \boldsymbol{\xi}(Y,Z); \mu)$.

[E.14] The function $\mu \mapsto c(y, \boldsymbol{\xi}; \mu)$ *is twice differentiable. For all* $(\mu, \boldsymbol{\xi}) \in \mathcal{M} \times \Xi$ *and a positive sequence* $\delta_n = o(1)$, *the first derivative of* $\mu \mapsto c(Y, \boldsymbol{\xi}(Y,Z); \mu)$ *with respect to* μ, *denoted* $\mu \mapsto c_\mu(Y, \boldsymbol{\xi}(Y,Z); \mu)$, *satisfies:*

$$||c_\mu(Y, \boldsymbol{\xi}(Y,Z); \mu) - c_\mu(Y, \boldsymbol{\xi}(Y,Z); \mu_o)|| \leq \delta_n \, ||\hat{\mu} - \mu_o||$$

[E.15] There exists a function $(y,z) \mapsto \psi(y,z)$ *such that* $E[\psi(Y_i, Z_i)] = 0$, $V[\psi(Y_i, Z_i)] < \infty$,

$$\left|\left| c(Y, \boldsymbol{\xi}, \mu) - c(Y, \boldsymbol{\xi}_o, \mu) - \psi(Y_i, Z_i) \times [\hat{\boldsymbol{\xi}}(Y,Z) - \boldsymbol{\xi}_o(Y,Z)] \right|\right| \leq ||\hat{\boldsymbol{\xi}}(Y,Z) - \boldsymbol{\xi}_o(Y,Z)||$$

$$n_1^{-1/2} \sum_{i=1}^{n_1} \psi(Y_i, Z_i) \to_P n_1^{1/2} \mathbb{E}[\psi(Y_i, Z_i) \times [\hat{\boldsymbol{\xi}}(Y_i, Z_i) - \boldsymbol{\xi}_o(Y_i, Z_i)]]$$

Newey (1994) discuss how to find $\psi(Y_i, Z_i)$ satisfying [E.15]. We use our examples to illustrate these later assumptions.

Pilot Example 1. For simplicity, suppose that Y *and* X *are zero mean variables with unit variance, so* $c(Y, \boldsymbol{\xi}, \mu) = Y \times \boldsymbol{\xi}$. *Since the product is a continuous function assumption [E.13] is satisfied. Using the results by Newey (1994) it is possible to show that the function* $\psi(Y_i, Z_i) = \partial c(Y, \boldsymbol{\xi}, \mu)/\partial \boldsymbol{\xi} = (Y_i, Y_i)'$ *satisfies [E.15].*□

Consistency and Asymptotic Normality. With assumptions [E.1]-[E-15] in place and the regularity conditions introduced at the end of section 2, we have the following proposition,

Proposition 3 *Let assumptions [AD], [R] and [E.1]-[E.15] holds. Define the random variables* $R^l := c(Y, \xi_o^l(Y,Z); \mu_o) + \psi(Y,Z) + \phi(Y,Z)$ *and* $R^u = c(Y, \xi_o^u(Y,Z); \mu_o) + \psi(Y_i, Z_i) + \phi(Y_i, Z_i)$. *Let* $\sigma_l^{asy} := \mathbb{V}(R^l)$, $\sigma_u^{asy} := \mathbb{V}(R^u)$ *and* $\rho_{l,u}^{asy} := \mathbb{C}(R^l, R^u)$ *denote,* *respectively, the variance of R^l, the variance of R^u and the covariance between R^l and R^u. Then,*

(i) $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o| = o_P(1)$

(ii) $\sqrt{n}|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o| \rightsquigarrow N(0, \Sigma)$*, where* $\Sigma = \begin{pmatrix} \sigma_l^{asy} & \rho_{l,u}^{asy} \\ \rho_{l,u}^{asy} & \sigma_u^{asy} \end{pmatrix}$

(iii) There exists a consistent estimator for Σ.

(iv) $P(\hat{\theta}_l \leq \hat{\theta}_u) = 1$

Proof. See Appendix A. Claims (ii)-(iv) still need to be verified. ■

Claim (i) establishes consistency of our proposed estimator.[9] Claim (ii) establishes asymptotic normality and it basically indicates that the distribution of $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o|$ inflated by $\sqrt{n}$ can be approximated by a bivariate normal distribution with zero mean and variance Σ. There are three components to the asymptotic approximation of the variance of the estimator of the extreme points of the identified set. The first term corrects for averaging $\mathbf{m}(Y_i, \theta, \boldsymbol{\xi}_i)$ while the terms $\psi_Q(y)$ and $\psi_H(y)$ correct, respectively, for the presence of $\hat{Q}_{X|Z}(\tau, z)$ and $\hat{G}_{Y|Z}(y, z)$.

Claims (ii)-(iv) in Proposition 2 enable us to construct confidence intervals for the true value of the parameter of interest, θ_o, of the type proposed by Imbens and Manski (2004), and later refined by Stoye (2009).

4.3 Monte Carlo Experiments

This section reports the results of Monte-Carlo experiments. The aim here is to assess how the proposed estimator for the identified set behaves in small and medium-sized samples. These results complements the asymptotic properties derived previously. The parameter of interest θ in these Monte-Carlo experiments is the Pearson's correlation coefficient between Y and X. This is a prominent example of a parameter belonging to

[9]Notice that Proposition (i) implies that the sequence of non-negative numbers $d_n = \max\{\sup_{\tilde{\theta}\in\Theta_I} \inf_{\breve{\theta}\in\hat{\Theta}_I} |\tilde{\theta} - \breve{\theta}|, \sup_{\breve{\theta}\in\hat{\Theta}_I} \inf_{\tilde{\theta}\in\Theta_I} |\tilde{\theta} - \breve{\theta}|\}$ converges to zero as n goes to infinity. That is, Proposition (i) implies that the Hausdorff distance between $\hat{\Theta}_I$ and Θ_I converge to zero.

the class studied here and it has been the object of interest in the econometric literature on program evaluation. The experiments also illustrate two additional issues. First, they illustrate the implementation of the estimator. Second, they illustrate how much dependence between Y and Z, and between X and Z, is needed to obtain informative bounds on the correlation between Y and X.

Preliminaries. For the sake of simplicity, we work with a scalar common variable Z. We explore two designs for the experiments. In design A, we simulate Z from a standard normal distribution and (Y, X) from a bivariate normal distribution with mean λ and covariance matrix Σ_{MC} given by

$$\lambda = \begin{pmatrix} r_{YZ} \times z \\ r_{XZ} \times z \end{pmatrix} \quad ; \quad \Sigma_{MC} = \begin{pmatrix} (1 - r_{YZ}^2) & \frac{\theta - r_{YZ} r_{XZ}}{\sqrt{(1-r_{YZ}^2)(1-r_{XZ}^2)}} \\ \frac{\theta - r_{YZ} r_{XZ}}{\sqrt{(1-r_{YZ}^2)(1-r_{XZ}^2)}} & (1 - r_{XZ}^2) \end{pmatrix}$$

In this design, the conditional distribution $y \to G_{Y|Z}(y|z)$ is normal with mean $r_{YZ} \times z$ and variance $(1 - r_{YZ}^2)$. Similarly for $x \to G_{X|Z}(x|z)$ with mean $r_{XZ} \times z$ and variance $(1 - r_{XZ}^2)$. In design B, we simulate Z from a standard lognormal distribution and (Y, X) from a bivariate lognormal distribution with mean λ and covariance matrix Σ. These designs are motivated by computational simplicity.

For each design and for each Monte Carlo replication, we simulate from the corresponding population n_1 independent realizations for (Y, Z) and n_2 independent realizations for (Y, Z). The number of replications is equal to 100. Each simulation experiment depends on five design variables: (i) the size of the first sample, n_1; (ii) the size of the second sample, n_2; (iii) the correlation between Y and Z, r_{YZ}, that is, between the two variable observed in the first sample and (iv) the correlation between X and Z, r_{XZ}, that is between the two variables observed in the second sample; and (v) the correlation between Y and X, denoted θ, which is the unobserved parameter of interest.

Before going to the results, we mention some issues related to the implementation of our estimator. All the experiment were carried out in R using the libraries "mvtnorm" (to generate bivariate normal random numbers), "splines" (to generate B-spline basis) and "quantreg" (to solve programming problem (7)).[10] We choose cubic B-splines so $m_1 = m_2 = 3$. We evaluate the robustness of our results with respect to different choices of numbers of knots, K_{n_1} and L_{n_2}. Knots are placed at quantiles of Z.

Results. Tables in appendix B contain the results of our Monte Carlo experiments regarding the properties of our estimator when the parameter of interest is the correlation between Y and X. Tables I and III correspond to data generated according to design A, while Table II and IV to design B. Within each table we make vary: (i) the numbers of knots K_{n_1} and L_{n_2} in the series nonparametric estimators of the distribution function of Y given Z, and of the quantile function of X given Z; (ii) the correlation between Y and Z; and (iii) the correlation between X and Z. Comparisons within a table are aimed to evaluate the sensitivity of the results to different choices for K_{n_1} and L_{n_2}, and to different correlations between Y and Z, and X and Z. Comparison between Tables I and III, and between II and IV are aimed to evaluate the sensitivity of the results to different sample sizes. Comparisons between tables I and II, and between III and IV are aimed to evaluate the sensitivity of the results to normality.

5 Conclusions

This paper has considered a situation where there are a sample on variables (Y, Z) and another sample on (X, Z), neither of which has joint information on the variables (Y, X). Knowledge has been sought about a strictly superadditive parameter defined in terms of the joint distribution of (Y, X). This framework has permitted a unified treatment of several applications from different fields, ranging from media planning studies in marketing, the ecological correlation problem in political science, and the measurement of impact heterogeneity in the econometric evaluation of social programs. The issue has been to determine to which extend the common variable

We first have shown that strictly superadditive parameters are not uniquely defined by the available data free of sample variation and the maintained assumptions. Then,

[10] R is a free software which can be dowloaded from http://www.r-project.org/.

we have characterized all the feasible values of the parameter of interest and proposed a consistent and asymptotically estimator for such values. Depending on the particular parameter of interest and the dependence between the observed variables, the common variable Z may help to shrink the identified set. We have illustrated the implementation of the inference procedures we propose through Monte Carlo experiments.

There are many open questions related to the identification problem. In order to make plain the limitations of the available data, throughout the paper we have tried to keep at minimum the assumptions made about the population P. We have also assumed that the parameter of interest is not overidentified, although this may not be the case for some applications such as the linear regression model or those using more than two samples. Research to determine how overidentification and/or ancillary assumptions on P, other than normality or independence of Y, X conditionally on Z, can shrink the identified set are two of a number of topics of current research. There are as well many open questions related to the solution of the inference problem we propose. One of them relates to the optimal choice of the smoothing parameter. Here we have use an ad hoc rule-of-thumb to choose the number of knots for the first step estimators. The derivation of rules providing a fully data-dependent method for choosing such number is an open issue. Another open question relates to efficiency properties of the estimator.

Acknowlegments. *I wish to thank my advisor Thierry Magnac for invaluable help during this project. C. Bontemps, N. Buehren, F. Gomez, G. Ridder, C. Rothe, seminar participants at ENTER Jamboree 2010 and the Group on Asymptotic Statistics at TSE have offered useful comments and suggestions for which I am very grateful. All remaining errors are my responsibility.*

Appendix A. Lemmas and Proofs

Identification Results

Proof of Proposition 1. Let $\mathcal{F}_{Y,X,Z}$ be the class of trivariate distribution functions for which the marginals $G_{Y,Z}$ and $G_{X,Z}$ are given. First, notice that $\mathcal{F}_{Y,X,Z}$ is non-empty since it contains the trivariate distribution which is such that Y and X are conditionally independent given Z -i.e., the distribution $F(y,x,z) = \int_{\mathbb{Z}}^{z} G_{Y|Z}(y,s)G_{X|Z}(x,s)dG_Z(s)$. Second, notice that $\mathcal{F}_{Y,X,Z}$ is convex because of the regularity assumption [R.1]. Consider now the linear map $\mathbb{E}[c(Y,X)] : \mathcal{F}_{Y,X,Z} \mapsto \Theta$. The image of this map is the identified set Θ_I. The identified set then is non-empty because $\mathcal{F}_{Y,X,Z}$ is non-empty. The identified set is bounded because is a subset of the bounded set Θ (see assumption [F]). The Θ_I is convex because it is a linear transformation of the convex set $\mathcal{F}_{Y,X,Z}$. We prove that Θ_I is closed in the in the proof of Theorem 1.■

Proof of Theorem 1. The proof has three steps. In the first step we show $\Theta_I = [\theta_l, \theta_u]$. Second, we show $[\theta_l, \theta_u] = \mathbb{A}[\mathbb{A}(R|Z)]$ and finally we show

$$\mathbb{A}[\mathbb{A}(R|Z)] = \left\{\theta \in \Theta : \mathbb{E}[c(Y,\xi^l)] \leq \theta \leq \mathbb{E}[c(Y,\xi^u)]\right\}$$

Step 1. The identified set Θ_I is a subset of its convex hull because it is convex (see proposition 1 and Rockafellar, 1970 pp. 167, Corollary 18.5.1). In other words $\Theta_I \subseteq [\theta_l, \theta_u]$ where the extreme points θ_l and θ_u are defined by:

$$\begin{aligned} \theta_l &:= \inf_{F_{Y,X,Z} \in \mathcal{F}_{Y,X,Z}} \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y,x) dF_{Y,X,Z}(y,x,z) \\ \theta_u &:= \sup_{F_{Y,X,Z} \in \mathcal{F}_{Y,X,Z}} \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y,x) dF_{Y,X,Z}(y,x,z) \end{aligned}$$

We now look for explicit expressions for θ_l and θ_u in terms of the distributions $G_{Y|Z}$, $G_{X|Z}$, G_Z and show that Θ_I is closed, so $\Theta_I = [\theta_l, \theta_u]$. For each $P \in \mathcal{P}$ we define the conditional density function $y,x \mapsto dF_{Y,X|Z}(y,x,z)$ by $dF_{Y,X|Z}(y,x,z) := \frac{f_{Y,X,Z}(y,x,z)}{g_{\mathbf{Z}}(z)}$, where $y,x,z \mapsto f_{Y,X,Z}(y,x,z)$ denotes the density of (Y,X,Z). Division by 0 in $\frac{f_{Y,X,Z}(y,x,z)}{g_{\mathbf{Z}}(z)}$ is a null probability event by assumption [R.2]. Then $y,x \mapsto dF_{Y,X|Z}(y,x,z)$ is well-defined.

Since $\int_{\mathcal{Y}\times\mathcal{X}} dF_{Y,X}(y,x) = \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} \frac{f_{Y,X,Z}(y,x,z)}{g_{\mathbf{Z}}(z)} g_Z(z)dz$ the objective function in the programming problems defining the extreme points of the identified set can be written as:

$$\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y,x)dF_{Y,X,Z}(y,x,z) = \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y,x)dF_{Y,X|Z}(y,x,z)dG_Z(z)$$

Since $y,x \mapsto c(y,x)$ is an integrable function by assumption [R.1] we can apply Fubini's Theorem in the last expression to get:

$$\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} c(y,x)dF_{Y,X,Z}(y,x,z) = \int_{\mathcal{Z}} \left[\int_{\mathcal{Y}\times\mathcal{X}} c(y,x)dF_{Y,X|Z}(y,x,z)\right] dG_Z(z)$$

Because of the last expression the extreme points θ_l, θ_u are equal to:

$$\begin{aligned}
\theta_l &= \int_{\mathcal{Z}} \left[\inf_{F_{Y,X|Z}\in\mathcal{F}_{Y,X,Z}} \int_{\mathcal{Y}\times\mathcal{X}} c(y,x)dF_{Y,X|Z}(y,x,z)\right] dG_Z(z) \\
\theta_u &= \int_{\mathcal{Z}} \left[\sup_{F_{Y,X|Z}\in\mathcal{F}_{Y,X,Z}} \int_{\mathcal{Y}\times\mathcal{X}} c(y,x)dF_{Y,X|Z}(y,x,z)\right] dG_Z(z)
\end{aligned}$$

Programming problems in square brackets in the last two expressions are Kantarovich mass transportation problems (see Rachev and Ruschendorf, 1998). Since $y,x \mapsto c(y,x)$ satisfies the so-called Monge Condition (see assumption [S]), it follows from Theorem 3.1.2 in Rachev and Ruschendorf (1998, pp. 109) that the objective function $F_{Y,X|Z} \mapsto \int_{\mathcal{Y}\times\mathcal{X}} c(y,x)dF_{Y,X|Z}(y,x,z)$ attains its minimum and maximum value at:

$$\begin{aligned}
F^l_{Y,X|Z}(y,x,z) &= \max\{0, G_{Y|Z}(y,z) + G_{X|Z}(x,z) - 1\} \\
F^u_{Y,X|Z}(y,x,z) &= \min\{G_{Y|Z}(y,z), G_{X|Z}(x,z)\}
\end{aligned}$$

The functions $F^l_{Y,X|Z}(y,x,z)$ and $F^u_{Y,X|Z}(y,x,z)$ are the so-called Frechet distributions of the class $F_{Y,X|Z}$ (Joe, 1997). We have then

$$\begin{aligned}
\theta_l &= \int_{\mathcal{Z}}\int_{\mathcal{Y}\times\mathcal{X}} c(y,x)dF^l_{Y,X|Z}(y,x,z)dG_Z(z) \\
\theta_u &= \int_{\mathcal{Z}}\int_{\mathcal{Y}\times\mathcal{X}} c(y,x)dF^u_{Y,X|Z}(y,x,z)dG_Z(z)
\end{aligned}$$

Let $Q_{Y|Z}(\tau, z)$ and $Q_{X|Z}(\upsilon, z)$ denote, respectively, the τ-quantile of Y given $Z = z$ and the υ-quantile of X given $Z = z$. By using the quantile substitution $y = Q_{Y|Z}(\tau, z)$ and $x = Q_{X|Z}(\upsilon, z)$ we get,

$$\begin{aligned}
\theta_l &= \int_{\mathcal{Z}} \int_{[0,1]\times[0,1]} c\left(Q_{Y|Z}(\tau, z), Q_{X|Z}(\upsilon, z)\right) d\max\{0, \tau + \upsilon - 1\} dG_Z(z) \\
\theta_u &= \int_{\mathcal{Z}} \int_{[0,1]\times[0,1]} c\left(Q_{Y|Z}(\tau, z), Q_{X|Z}(\upsilon, z)\right) d\min\{\tau, \upsilon\} dG_Z(z)
\end{aligned}$$

Since $d\max\{0, \tau + \upsilon - 1\}$ is different from zero only at $\tau + \upsilon - 1 = 0$ and $d\min\{\tau, \upsilon\}$ is different from zero only at $\tau = \upsilon$, we have the following analytical expressions for θ_l and θ_u:

$$\begin{aligned}
\theta_l &= \int_{\mathcal{Z}} \int_{[0,1]} c\left(Q_{Y|Z}(\tau, z), Q_{X|Z}(1 - \tau, z)\right) d\tau dG_Z(z) \\
\theta_u &= \int_{\mathcal{Z}} \int_{[0,1]} c\left(Q_{Y|Z}(\tau, z), Q_{X|Z}(\tau, z)\right) d\tau dG_Z(z)
\end{aligned}$$

Since $F^l_{Y,X|Z}(\cdot, \cdot, z)$ and $F^u_{Y,X|Z}(\cdot, \cdot, z)$ belong to $\mathcal{F}_{Y,X,Z}$ (see Joe, 1997 pp. 65), we have that θ_l and θ_u belong to Θ_I. Therefore, Θ_I is closed.

Step 2. We first prove three useful lemmas. They concern the existence of the random set R and its Aumann expectation. Let ω be an element belonging to Ω. For $z = Z(\omega)$ and $y = Y(\omega)$, define the random variables $\xi^l : \Omega \mapsto \mathcal{X}$ and $\xi^u : \Omega \mapsto \mathcal{X}$ by $\xi^l(y, z) := Q_{X|Z}[1 - G_{Y|Z}(y, z), z]$ and $\xi^u(y, z) := Q_{X|Z}[G_{Y|Z}(y, z), z]$. By construction ξ^l and ξ^u have the same distribution $G_{X|Z}$. Let $\xi : \Omega \mapsto \mathcal{X}$ denote a random variable. Define the set

$$S := \{\xi : P\left(\xi(y, z) \in [\xi^l(y, z), \xi^u(y, z)]\right) = 1\}$$

We have the following lemma proving that the random variable X belongs to S, hence S is non-empty.

Lemma 1 *Consider the random variables $\xi^l : \Omega \to \mathcal{X}$ and $\xi^u : \Omega \to \mathcal{X}$ as defined previously. Let $x = X(\omega)$. Then, $P\{x \in [\xi^u(y, z), \xi^l(y, z)]\} = 1$ for any ω in Ω.*

Proof. $P\{\xi^l(y,z) \le x \le \xi^u(y,z)\} = P\left[Q_{X|Z}\left(1 - G_{Y|Z}(y,z), z\right) \le x \le Q_{X|Z}\left(G_{Y|Z}(y,z), z\right)\right]$
$P\{\xi^l(y,z) \le x \le \xi^u(y,z)\} = P\left[1 - G_{Y|Z}(y,z) \le G_{X|Z}(x,z) \le G_{Y|Z}(y,z)\right]$
For any ω such that $G_{Y|Z}(y,z) \le 1/2$ we have:
$P\{\xi^l(y,z) \le x \le \xi^u(y,z)\} = 1$
Similarly for any ω such that $1/2 < G_{Y|Z}(y|z)$ we have
$P\{\xi^l(y,z) \le x \le \xi^u(y,z)\} = 1$
Therefore, $P\{x \in [\xi^l(y,z), \xi^u(y,z)]\} = 1$. ■

Let $\mathcal{S}^2(S)$ denote the family of all square integrable selection in S (see Molchanov (2005) for the precise definition of *selection*). Define the map $y, z \mapsto R(y,z)$ with $R(y,z) := \{c(y, \xi(y,z)) : \xi \in \mathcal{S}^2(S)\}$. Here is the second lemma,

Lemma 2 *The mapping $y, z \mapsto R(y,z)$ is a non-empty integrable random closed set defined on the probability space $(\Omega, \mathfrak{F}, P)$.*

Proof. Theorem 1.2.5 in Molchanov (2005) implies that the mapping $y, z \mapsto S(y,z)$ is a set-valued random variable if and only if the support function $q, y, z \mapsto \eta[q, S(y,z)]$,

$$\eta[q, S(y,z)] = \max[q\xi^l(y,z), q\xi^u(y,z)],$$

is a random variable for each $q \in \{-1, 1\}$. The support function $\eta[q, S(y,z)]$ is equal to:

$$\eta[q, S(y,z)] = \begin{cases} \xi^u(y,z) \text{ if } G_{Y|Z}(y,z) \le 1/2 \text{ and } q = 1 \\ -\xi^l(y,z) \text{ if } G_{Y|Z}(y,z) > 1/2 \text{ and } q = -1 \\ \xi^l(y,z) \text{ if } G_{Y|Z}(y,z) \le 1/2 \text{ and } q = 1 \\ -\xi^u(y,z) \text{ if } G_{Y|Z}(y,z) > 1/2 \text{ and } q = -1 \end{cases}$$

From the fact that ξ^l and ξ^u are random variables, also $q, y, z \mapsto \eta[q, S(y,z)]$ is a random variable. From the regularity condition [R.1] and Lemma 1 it follows that $X \in \mathcal{S}^2(S)$, so $\mathcal{S}^2(S)$ is non-empty. Then, $y, z \mapsto R(y,z)$ is an integrable set-valued random variable. Since $y, x \mapsto c(y,x)$ is right-continuous (see assumption [S]) we have that $y, z \mapsto R(y,z)$ is measurable in $(\Omega, \mathfrak{F}, P)$. Hence $y, z \mapsto R(y, z)$ is an integrable set-valued random variable.

■

For an arbitrary set B, let $cl(B)$ denotes its closure. Here is the third lemma:

Lemma 3 *The conditional Aumann expectation* $\mathbb{A}(R|z) := cl\{\mathbb{E}(c(Y, \xi)|z) : \xi \in \mathcal{S}^2(S)\}$ *exists and it is unique.*

Proof. Lemma 3 follows from Molchanov (2005, Theorem 1.46) and the fact that $y, z \mapsto R(y, z)$ is an integrable random closed set (see lemma 2) . ■

We are now in a position to prove that $[\theta_l, \theta_u] = \mathbb{A}[\mathbb{A}(R|Z)]$. We start by showing $[\theta_l, \theta_u] \subset \mathbb{A}[\mathbb{A}(R|Z)]$. Define

$$\begin{aligned} \theta_l(z) \ &: = \int_{[0,1]} c\left(Q_{Y|Z}(\tau, z), Q_{X|Z}(1-\tau, z)\right) d\tau \\ \theta_u(z) \ &: = \int_{[0,1]} c\left(Q_{Y|Z}(\tau, z), Q_{X|Z}(1-\tau, z)\right) d\tau \end{aligned}$$

Pick any $\tilde{\theta}(z) \in [\theta_l(z), \theta_u(z)]$ such that $\tilde{\theta} = \mathbb{E}[\tilde{\theta}(z)]$ with $\tilde{\theta} \in [\theta_l, \theta_u]$. Then there exists a bivariate distribution $\tilde{F} \in \mathcal{F}_{Y,X,Z}$ with support on $\mathcal{Y} \times \mathcal{X}$ and with X-marginal distribution $G_{X|Z}$. Let $(Y, \tilde{X})$ be a random vector with distribution $\tilde{F}$. Since $P\{\tilde{x} \in [\xi^l(\omega), \xi^u(\omega)]\} = 1$ it follows from lemma 2 that $\tilde{X} \in S \ P - a.s.$ and $c(Y, \tilde{X}) \in R \ P - a.z.$ This means that $\tilde{\theta}(z) \in \mathbb{A}(R|Z)$. Since $\tilde{\theta} = \mathbb{E}[\mathbb{E}(c(Y, \tilde{X})|Z)]$, we have $\tilde{\theta} \in cl\{\mathbb{E}[\mathbb{E}(c(Y, \xi)|Z)] : \xi \in S\}$. Therefore, $\tilde{\theta} \in \mathbb{A}[\mathbb{A}(R|Z)]$ and $[\theta_l, \theta_u] \subset \mathbb{A}[\mathbb{A}(R|Z)]$.

We now prove $\mathbb{A}[\mathbb{A}(R|Z)] \subset [\theta_l, \theta_u]$. Pick any $\tilde{\theta}(z) \in \mathbb{A}(R|z)$. Then there exists a random variable $\tilde{X} \in \mathcal{S}(S)$ such that: $\tilde{\theta}(z) = \int_{\mathcal{Y}\times\mathcal{X}} c(y, \tilde{x}) dF_{Y,\tilde{X}|Z}(y, \tilde{x}, z)$. Because $c(Y, \tilde{X})$ is a selection from R, it follows from Beresteanu and Molinari (2008) that for any $y \in \mathcal{Y}$ and $\tilde{x} \in \mathcal{X}$,

$P(Y \leq y, X \leq \tilde{x}|z) \leq P(Y \leq y, S \cap (-\infty, \tilde{x}] \neq \emptyset|z) = P(Y \leq y, \xi^l \leq \tilde{x}|z)$

$P(Y \leq y, X > \tilde{x}|z) > P(Y \leq y_1, S \cap (\tilde{x}] \neq \emptyset|z) = P(Y \leq y, \xi^u \geq \tilde{x}|z)$

Then, the Y-marginal of $F_{Y,\tilde{X}|Z}(y, \tilde{x}|z)$ is $H_{Y|\mathbf{Z}}$. Since by construction ξ^l and ξ^u have marginal distribution $G_{X|\mathbf{Z}}$, the ξ-marginal of $F_{Y,\tilde{X}|Z}(y, \tilde{x}|z)$ it is $G_{X|Z}$. Therefore $F_{Y,\tilde{X}|Z}(y, \tilde{x}|z) \in \mathcal{F}_{Y,X,Z}$ from which $\tilde{\theta}(z) \in [\theta_l(z), \theta_u(z)]$ and $\tilde{\theta} \in [\theta_l, \theta_u]$. This completes step 2.

Step 3. We show $\mathbb{A}[\mathbb{A}(R|Z)] = \{\theta \in \Theta : \theta \leq \mathbb{E}[\mathbb{E}[z(q,R)|Z]]\}$. Since R is integrably bounded (see Lemma 2), from Theorem 1.22 in Molchanov (2005), pp. 157 we have $z(q, \mathbb{A}[\mathbb{A}(R|Z)]) = \mathbb{E}[\mathbb{E}[\eta(q,R)|Z]]$. Then, $\mathbb{A}[\mathbb{A}(R|Z)] = \left\{\theta \in \Theta : \mathbb{E}[c(Y,\xi^l)] \leq \theta \leq \mathbb{E}[c(Y,\xi^u)]\right\}$. which completes the proof of theorem 1. ■

Proof of Corollary 1. The minimum value of θ compatible with hypothetical knowledge of the distribution of Y and of X, θ_l^C, is given by (see Cambanis, et. al. 1976):

$$\theta_l^C = \int_{\mathcal{Y}\times\mathcal{X}} c(y,x) d \max\{0, G_Y(y) + G_X(x) - 1\}$$

by the total law of probability this is equal to:

$$\theta_l^C = \int_{\mathcal{Y}\times\mathcal{X}} c(y,x) d \max\left\{\int_{\mathcal{Z}} G_{Y|Z}(x,z) dG_Z(z) + \int_{\mathcal{Z}} G_{X|Z}(x,z) dG_Z(z) - 1, 0\right\}$$

Since $\max\{a, 0\}$ is a convex function of a, $\theta_l^C \leq \theta_l$ by the conditional Jensen inequality. The maximum value of θ compatible with hypothetical knowledge of the distribution of Y and of X, θ_u^C, is given by:

$$\theta_u^C = \int_{\mathcal{Y}\times\mathcal{X}} c(y,x) d \min\{G_Y(y), G_X(x) - 1\}$$

Since $\min\{a, 0\}$ is concave we have $\theta_u \leq \theta_u^C$. Therefore, $\Theta^C = [\theta_l^C, \theta_u^C] \subseteq \Theta_I$.■

Inference Procedures

We now prove the asymptotic properties of the estimators of the extreme points of the identified set. We begin by proving Proposition 2 and 3. We continue by proving three lemmas we invoke for such proofs.

Let $\bar{C}$ denote a generic constant that may be different in different uses. For any conformable matrices A and B we define the Euclidean norm $||A||_B := trace\,(A'BA)^{1/2}$, where $trace(\cdot)$ denotes the trace operator. For any real-valued function $f(x)$, let $||f(x)||_\infty = \sup_x |f(x)|$ denote the sup-norm. We shall suppress the subscripts in the notations for the norms whenever this can be done without causing confusion. Let $\mathbf{q}(z) = (q_1(z), ..., q_{L_{n_2}}(z))'$ be a $L_{n_2} \times 1$ vector of approximative functions forming the base for $\mathcal{Q}_{n_2}$. Let $\mathbf{p}(z) = (p_1(z), ..., p_{K_{n_1}}(z))'$ be a $K_{n_1} \times 1$ vector of approximative functions forming the base for $\mathcal{H}_{n_1}$. Here is the proof of proposition 2.

Proof of Proposition 2. Let Ξ denote the space of all functions $y, z \mapsto \boldsymbol{\xi}(y, z)$ from $\mathcal{Y} \times \mathcal{Z}$ into $\mathcal{X} \times \mathcal{X}$ defined by $\boldsymbol{\xi}(y, z) := (\xi^l(y, z), \xi^u(y, z))$. For any $\boldsymbol{\xi} \in \Xi$ define the norm:

$$||\boldsymbol{\xi}||_{\Xi} = \max\{||\xi^u||_\infty, ||\xi^l||_\infty\}$$

Then $||\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}||_{\Xi} = o_P(n^{-1/4})$ provided that

$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\hat{\xi}^l(y, z) - \xi^l(y, z)| = o_P(n_1^{-1/4}) \tag{P2.1}$$
$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\hat{\xi}^u(y, z) - \xi^u(y, z)| = o_P(n_1^{-1/4}) \tag{P2.2}$$

We only show that $(P2.1) = o_P(n_1^{-1/4})$. The proof for $(P2.2) = o_P(n_1^{-1/4})$ is similar and thus omitted.

By construction $\hat{\xi}^u(y, z) = \mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))$ and $\xi^u(y, z) = Q_{X|Z}(G_{Y|Z}(y, z), z)$ so:

$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\hat{\xi}^u(y, z) - \xi^u(y, z)| = \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)) - Q_{X|Z}(G_{Y|Z}(y, z), z)|$$

Add-and-substract $\mathbf{q}(z)'\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))$, $Q_{X|Z}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y), z)$, $Q_{X|Z}(\mathbf{p}(z)'\boldsymbol{\alpha}(y), z)$ and apply the triangle inequality:

$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\hat{\xi}^u(y, z) - \xi^u(y, z)| \leq$

(A) $\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)) - \mathbf{q}(z)'\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))| +$

(B) $\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\mathbf{q}(z)'\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)) - Q_{X|Z}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y), z)| +$

(C) $\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |Q_{X|Z}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y), z) - Q_{X|Z}(\mathbf{p}(z)'\boldsymbol{\alpha}(y), z)| +$

(D) $\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |Q_{X|Z}(\mathbf{p}(z)'\boldsymbol{\alpha}(y), z) - Q_{X|Z}(G_{Y|Z}(y, z), z)|$

We now find the order for each of the terms in the right hand side of the later expression. Consider the term (A). Take common factor $\mathbf{q}(z)'$:

$$(A){=}\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\left[\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))-\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))\right]\right|$$

By the Cauchy-Schwarz inequality,

$$(A)\leq\sup_{z\in\mathcal{Z}}||\mathbf{q}(z)'||\times\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))-\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))\right|$$

Since $\mathbf{q}(z)'$ is a vector of splines functions and z is a scalar, we have $\sup_{z\in\mathcal{Z}}||\mathbf{q}(z)'||\leq L_{n_2}^{1/2}$ (see Newey, 1997 pp. 160). Then,

$$(A)\leq L_{n_2}^{1/2}\times\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))-\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))\right|$$

By lemma 4 (see below) the term $\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\hat{\boldsymbol{\beta}}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))-\boldsymbol{\beta}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y))\right|$ is $o_P(n^{-1/4})$. Then, $(A)=O(L_{n_2}^{1/2})\times o_P(n_1^{-1/4})$. Since $O(L_{n_2}^{1/2})\times o_P(n_2^{-1/4})=o_P(n_2^{-1/4})$ we have $(A)=o_P(n_2^{-1/4})$.

Consider now the term (B). According to assumption [E.7], we have $(B)=O(L_{n_2}^{-1/2})$. Consider now the term (C). Since $\tau{\mapsto}Q_{X|Z}(\tau,z)$ is continuous (see assumption [E.4]), we can take the following Taylor approximation of such function around $\tau=\mathbf{p}(z)'\boldsymbol{\alpha}(y)$:

$$\begin{aligned}Q_{X|Z}(\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y),z)-Q_{X|Z}(\mathbf{p}(z)'\boldsymbol{\alpha}(y),z) &= \left.\tfrac{\partial Q_{X|Z}(\tau,z)}{\partial\tau}\right|_{\tau=\mathbf{p}(z)'\boldsymbol{\alpha}(y)}\times[\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)-\mathbf{p}(z)'\boldsymbol{\alpha}(y)]\\ &+\tfrac{1}{2}\left.\tfrac{\partial^2 Q_{X|Z}(\tau,z)}{\partial\tau\partial\tau}\right|_{\tau=\mathbf{p}(z)'\tilde{\boldsymbol{\alpha}}(y)}[\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)-\mathbf{p}(z)'\boldsymbol{\alpha}(y)]^2\end{aligned}$$

for some vector $\tilde{\boldsymbol{\alpha}}(y)$ such that $\mathbf{p}(z)'\tilde{\boldsymbol{\alpha}}(\mathbf{y})\in[\mathbf{p}(z)'\boldsymbol{\alpha}(y),\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)]$**.** By the triangle inequality,

$$\begin{aligned}(C)\leq\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}&\left|\left.\tfrac{\partial Q_{X|Z}(\tau,z)}{\partial\tau}\right|_{\tau=\mathbf{p}(z)'\boldsymbol{\alpha}(y)}\times\mathbf{p}(z)'[\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)]\right|\\ +\tfrac{1}{2}\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}&\left|\left.\tfrac{\partial^2 Q_{X|Z}(\tau,z)}{\partial\tau\partial\tau}\right|_{\tau=\mathbf{p}(z)'\tilde{\boldsymbol{\alpha}}(y)}\times[\mathbf{p}(z)'(\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y))]^2\right|\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$(C) \leq \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left| \frac{\partial Q_{X|Z}(\tau,z)}{\partial \tau}\Big|_{\tau=\mathbf{p}(z)'\boldsymbol{\alpha}(y)} \right| \times \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|\mathbf{p}(z)'[\hat{\boldsymbol{\alpha}}(y) - \boldsymbol{\alpha}(y)]\right| +$$
$$\frac{1}{2} \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left| \frac{\partial^2 Q_{X|Z}(\tau,z)}{\partial \tau \partial \tau}\Big|_{\tau=\mathbf{p}(z)'\tilde{\boldsymbol{\alpha}}(y)} \right| \times \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|[\mathbf{p}(z)'[\hat{\boldsymbol{\alpha}}(y) - \boldsymbol{\alpha}(y)]]^2\right|$$

Since $\tau \mapsto Q_{X|Z}(\tau,z)$ has continuous uniformly bounded derivatives (see assumption [E.4]),

$$(C) \leq \bar{C} \times \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y) - \mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| +$$
$$\bar{C} \times \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|[\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y) - \mathbf{p}(z)'\boldsymbol{\alpha}(y)]^2\right|$$

By Lemma 5 (see below), $\left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y) - \mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| = o_P(n_1^{1/4})$, so $(C) = \bar{C} \times o_P(n_1^{1/4}) \times o_P(n_1^{1/4})$. Since $o_P(n_1^{1/4}) \times o_P(n_1^{1/4}) = o_P(n_1^{1/4})$ we have $(C) = o_P(n_1^{1/4})$

Consider now the term (D). We take the following Taylor approximation of $\tau \mapsto Q_{X|Z}(\tau,z)$ around $G_{Y|Z}(y,z)$,

$$Q_{X|Z}(\mathbf{p}(z)'\boldsymbol{\alpha}(y), z) - Q_{X|Z}(G_{Y|Z}(y,z), z) = \frac{\partial Q_{X|Z}(\tau,z)}{\partial \tau}\Big|_{\tau=G_{Y|Z}(y,z)} \times [\mathbf{p}(z)'\boldsymbol{\alpha}(y) - G_{Y|Z}(y,z)]$$
$$+ \frac{1}{2} \frac{\partial^2 Q_{X|Z}(\tau,z)}{\partial \tau \partial \tau}\Big|_{\tau=\tilde{\boldsymbol{\tau}}} [\mathbf{p}(z)'\boldsymbol{\alpha}(y) - G_{Y|Z}(y,z)]^2$$

for some $\tilde{\boldsymbol{\tau}} \in [\mathbf{p}(z)'\boldsymbol{\alpha}(y), G_{Y|Z}(y,z)]$**.** By the triangle inequality, the Cauchy-Schwarz inequalities and the fact that $\tau \mapsto Q_{X|Z}(\tau,z)$ has continuous uniformly bounded derivatives (see assumption [E4]),

$$(D) \leq \bar{C} \times \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|\mathbf{p}(z)'\boldsymbol{\alpha}(y) - G_{Y|Z}(y,z)\right|$$
$$+ \bar{C} \times \sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|[\mathbf{p}(z)'\boldsymbol{\alpha}(y) - G_{Y|Z}(y,z)]^2\right|$$

By assumption [E.6], $(D) = O(K_{n_1}^{-1/2})$. Combining the results for (A), (B), (C) and (D) we have that

$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} |\hat{\xi}^u(y,z) - \xi^u(y,z)| = o_P(n_2^{-1/4}) + O(L_{n_2}^{-1/2}) + o_P(n_1^{-1/4}) + O(K_{n_1}^{-1/2})$$

Under assumptions [E.9] and [E.10] the terms $O(L_{n_2}^{-1/2})$ and $O(K_{n_1}^{-1/2})$ vanish so

$$\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}|\hat{\xi}^u(y,z)-\xi^u(y,z)|=o_P(n^{-1/4})$$

which complets the proof.■

We continue by proving proposition 3.

Proof of Proposition 3. We first prove that the estimator $(\hat{\theta}_l,\hat{\theta}_u)$ of the extreme points of the identified set is consistent. The following lemma, which has been proved by Chen et. al. (2003), provides sufficient conditions for the general class of semiparametric Z-estimators to be consistent.

Lemma 4 *Let data be iid replications* $\{\mathbf{W}_i\}_{i=1}^n$ *of the random vector* $\mathbf{W}$*. Let* Θ *be a finite dimensional parameter set and* $\mathcal{H}$ *a infinite dimensional parameter set. Suppose that there exists a function* $M(\boldsymbol{\theta},h)=\mathbb{E}[m(\mathbf{W},\boldsymbol{\theta},h)]$ *from* $\Theta\times\mathcal{H}$ *into* $\mathbb{R}^d$ *for some positive integer* d*. Let* $M(\boldsymbol{\theta},h)=\sum_{i=1}^n m(\mathbf{W}_i,\boldsymbol{\theta},h)]$ *denote the empirical analog of* $M(\boldsymbol{\theta},h)$*. Suppose that* $\boldsymbol{\theta}_o\in\Theta$ *satisfies* $M(\boldsymbol{\theta}_o,h_o)=0$ *for some* $h_o\in\mathcal{H}$*, and that:*

(1.1) $\left|\left|M_n(\hat{\boldsymbol{\theta}},\hat{h})\right|\right|\leq\inf_{\theta\in\Theta}\left|\left|M_n(\boldsymbol{\theta},\hat{h})\right|\right|+o_P(1)$

For all $\delta>0$ *there exists* $\epsilon(\delta)$ *such that*

(1.2) $\inf_{||\boldsymbol{\theta}-\boldsymbol{\theta}_o||>\delta}\left|\left|M(\boldsymbol{\theta},\hat{h})\right|\right|\geq\epsilon(\delta)$

For all $\delta>0$ *there exists* $\varepsilon>0$ *such that*

(1.3) $||h-h_o||_{\mathcal{H}}<\delta$ *imply* $\sup_{\theta\in\Theta}||M(\boldsymbol{\theta},h)-M(\boldsymbol{\theta},h_o)||<\varepsilon$

(1.4) $\left|\left|\hat{h}-h\right|\right|_{\mathcal{H}}=o_P(1)$

(1.5) For all positive sequences $\delta_n=o(1)$

$$\sup_{\theta\in\Theta,\left|\left|\hat{h}-h\right|\right|_{\mathcal{H}}<\delta_n,}||M_n(\theta,h)-M(\theta,h)||=o_P(1)$$

Then, $\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}=o_P(1)$*.*

Proof of Lemma 4. See Chen et. al. (2003).■

Our application correspond to the case $\hat{\boldsymbol{\theta}} = (\hat{\theta}_l, \hat{\theta}_u)$, $h = (\mu, \boldsymbol{\xi})$, $\mathcal{H} = (\mathcal{M}, \Xi)$, $M(\boldsymbol{\theta}, h) = \mathbb{E}[c(Y, \mu, \boldsymbol{\xi}) - \boldsymbol{\theta}]$ and $M_n(\boldsymbol{\theta}, h) := \sum_{i=1}^{n_1} c(Y_i, \mu, \boldsymbol{\xi}) - \boldsymbol{\theta}$. In order to prove consistency of the estimators of the extreme points of the identified set $(\hat{\theta}_l, \hat{\theta}_u)$ it is enough thus to show that conditions (1.1)-(1.5) are met.

Consistency. We start by checking condition (1.1). Recall that $(\hat{\theta}_l, \hat{\theta}_u)$ have been defined by $\hat{\theta}_l := n^{-1} \sum_{i=1}^{n} c(y_i, \hat{\xi}_i^l; \hat{\mu})$ and $\hat{\theta}_u := n^{-1} \sum_{i=1}^{n} c(y_i, \hat{\xi}_i^u; \hat{\mu})$. Thus $\hat{\boldsymbol{\theta}}$ can interpreted as the argument of the minimum of the function $\boldsymbol{\theta} \mapsto \left|\left|M_n(\boldsymbol{\theta}, \hat{\mu}, \hat{\boldsymbol{\xi}})\right|\right|_{I_2}^2$, where I_2 is the identity matrix of dimension 2. From this follows that $\left|\left|M_n(\hat{\boldsymbol{\theta}}, \hat{\mu}, \hat{\boldsymbol{\xi}})\right|\right| = \inf_{\theta \in \Theta^2} \left|\left|M_n(\boldsymbol{\theta}, \hat{\mu}, \hat{\boldsymbol{\xi}})\right|\right|$, so (1.1) is verified. We continue by checking condition (1.2). Since in our case the function $\boldsymbol{\theta} \mapsto M(\boldsymbol{\theta}, \hat{\mu}, \hat{\boldsymbol{\xi}})$ is linear, the true values of the extreme points of the identified set estimator $\boldsymbol{\theta}_o = (\theta_l, \theta_u)$ are the unique minimizers of $\boldsymbol{\theta} \mapsto \left|\left|M(\boldsymbol{\theta}, \hat{\mu}, \hat{\boldsymbol{\xi}})\right|\right|_{I_2}^2$, where I_2 is the identity matrix of dimension 3. Thus, (1.2) is satisfied. We now check condition (1.3). Since $M(\boldsymbol{\theta}, \mu, \boldsymbol{\xi}) = \boldsymbol{\theta} - \mathbb{E}\left[c(Y, \boldsymbol{\xi}; \mu)\right]$ we have:

$$\sup_{\theta \in \Theta^2} ||M(\boldsymbol{\theta}, \mu, \boldsymbol{\xi}) - M(\boldsymbol{\theta}, \mu_o, \boldsymbol{\xi}_o)|| = ||\mathbb{E}\left[c(Y, \boldsymbol{\xi}_o; \mu_o)\right] - \mathbb{E}\left[c(Y, \boldsymbol{\xi}; \mu)\right]||$$

Then, condition (1.3) is satisfied because of the Lipschitz condition [E.13] in [EC]. We now check condition (1.4). For $h = (\mu, \boldsymbol{\xi}) \in \mathcal{H} = (\mathcal{M}, \Xi)$ define the norm

$$||h - h||_{\mathcal{H}} = \max\{||\mu||, ||\xi^u||_\infty, ||\xi^l||_\infty\}$$

Assumption [E.11] in [EI] and Proposition 2 imply that condition (1.4) is satisfied by the case at hand. Finally, we check condition (1.5). According to Chen et. al. (2003) condition (1.5) will be satisfied whenever the class of functions $\{M(\theta, h), \theta, h\}$ is P-Glivenko-Cantelli. Notice that.

$$\sup_{\theta \in \Theta, \left|\left|\hat{h} - h\right|\right|_{\mathcal{H}} < \delta_n,} ||M_n(\theta, h) - M(\theta, h)||$$

$$= \sup_{\theta \in \Theta^2, \left|\left|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right|\right| < \delta_n, ||\hat{\mu} - \boldsymbol{\mu}|| < \delta_n} ||\theta - \sum c(Y_i, \boldsymbol{\xi}, \mu) - \theta + \mathbb{E}\left[c(Y_i, \boldsymbol{\xi}, \mu)\right]||$$

$$= \sup_{||\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}||<\delta_n, ||\hat{\mu}-\boldsymbol{\mu}||<\delta_n} ||\sum c(Y_i, \boldsymbol{\xi},\mu) - \mathbb{E}\left[c(Y_i, \boldsymbol{\xi},\mu)\right]||$$

For the case at hand then, Condition (1.5) will be satisfied when the class of functions $\{\mathbb{E}\left[c(Y_i, \boldsymbol{\xi},\mu)\right], \boldsymbol{\xi} \in\Xi, \mu\in\mathcal{M}\}$ is P-Glivenko-Cantelli. Since $\boldsymbol{\xi},\mu \mapsto \mathbb{E}\left[c(Y_i, \boldsymbol{\xi},\mu)\right]$ is Lipschitz continuous (see assumption [E.13]), it follows from Theorem 2.10.6 by van der Vaart and Wellner (1996) that $\{\mathbb{E}\left[c(Y_i, \boldsymbol{\xi},\mu)\right], \boldsymbol{\xi} \in\Xi, \mu\in\mathcal{M}\}$ is P-Glivenko-Cantelli whenever $\mathcal{M}$ and Ξ are P-Glivenko-Cantelli. Since $\mathcal{M}$ is a compact subset of $\mathbb{R}^d$, the bracketing number of $\mathcal{M}$ is known and thus $\mathcal{M}$ is P-Glivenko-Cantelli. Because of lemma 8 (see below), the space Ξ is also P-Glivenko-Cantelli and then condition (1.5) is met. Therefore, the estimator of the extreme points of the identified set is consistent.

Asymptotic Normality. We now prove that $\sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converge in distribution to a random variable with normal distribution. Our starting point is the definition of the estimator

$$\hat{\theta} = n_1^{-1} \sum_{i=1}^{n_1} c(Y_i, \hat{\xi}(Y_i, Z_i))$$

Add-and-substract $\theta_o = \mathbb{E}[c(Y_i, \xi_o(Y_i, Z_i))]$, $\mathbb{E}[c(Y_i, \hat{\xi}(Y_i, Z_i))]$, $c(Y_i, \xi_o(Y_i, Z_i))$, and multiply both sides by $n_1^{1/2}$

$$\begin{aligned} n_1^{1/2}\left(\hat{\theta} - \theta_o\right) &= n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\left(c(Y_i, \hat{\xi}(Y_i, Z_i)) - \mathbb{E}[c(Y_i, \hat{\xi}(Y_i, Z_i))]\right)\right] \\ &+ n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\left(c(Y_i, \xi_o(Y_i, Z_i)) - \mathbb{E}[c(Y_i, \xi_o(Y_i, Z_i))]\right)\right] \\ &- n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\left(c(Y_i, \xi_o(Y_i, Z_i)) - \mathbb{E}[c(Y_i, \hat{\xi}(Y_i, Z_i))]\right)\right] \end{aligned}$$

Define the empirical process $\nu_{n_1}(\xi, \mu) := n^{-1/2}\left[\sum_{i=1}^{n_1}[c(Y_i, \xi_i; \mu) - \mathbb{E}(c(Y_i, \xi_i; \mu))]\right]$. Stochastic equicontinuity of $\nu_{n_1}(\xi, \mu)$ at (ξ_o, μ_o) (see below), consistency of $(\hat{\xi}, \hat{\mu})$ for (ξ_o, μ_o) (see proposition 3) and $P(\hat{\xi} \in \Xi) \to 1$ yield $\nu_{n_1}(\hat{\xi}, \hat{\mu}) - \nu_{n_1}(\xi_o, \mu_o) = o_P(1)$,so we have

$$n_1^{1/2}\left(\hat{\theta} - \theta_o\right) = n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\left(\mathbb{E}[c(Y_i, \hat{\xi}(Y_i, Z_i))] - c(Y_i, \xi_o(Y_i, Z_i))\right)\right] + o_P(1)$$

Approximate $c(Y_i, \hat{\xi}(Y_i, Z_i))$ inside the expectation by $c(Y_i, \xi_o(Y_i, Z_i)) + \psi(Y_i, Z_i)[\hat{\xi}(Y_i, Z_i) -$

$\xi_o(Y_i, Z_i)]$

$$\begin{aligned} n_1^{1/2}\left(\hat{\theta}-\theta_o\right) &= n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}(\mathbb{E}[c(Y_i,\xi_o(Y_i,Z_i))]-c(Y_i,\xi_o(Y_i,Z_i)))\right] \\ &+n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\mathbb{E}[\psi(Y_i,Z_i)[\hat{\xi}(Y_i,Z_i)-\xi_o(Y_i,Z_i)]]\right] \\ &+n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\mathbb{E}[rem(\hat{\xi}(Y_i,Z_i)-\xi_o(Y_i,Z_i))]\right] \\ &+o_P(1) \end{aligned}$$

where $rem(\hat{\xi}(Y_i, Z_i) - \xi_o(Y_i, Z_i))$ denote the difference

$$rem(\hat{\xi}(Y_i,Z_i)-\xi_o(Y_i,Z_i)) := c(Y_i,\hat{\xi}(Y_i,Z_i))-c(Y_i,\xi_o(Y_i,Z_i))-\psi(Y_i,Z_i)[\hat{\xi}(Y_i,Z_i)-\xi_o(Y_i,Z_i)]$$

By assumption [E.15.(i)] we have this difference is $o_P(1)$. It then follows that

$$\begin{aligned} n_1^{1/2}\left(\hat{\theta}-\theta_o\right) &= n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}(\mathbb{E}[c(Y_i,\xi_o(Y_i,Z_i))]-c(Y_i,\xi_o(Y_i,Z_i)))\right] \\ &+n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}\mathbb{E}[\psi(Y_i,Z_i)[\hat{\xi}(Y_i,Z_i)-\xi_o(Y_i,Z_i)]]\right] \\ &+o_P(1) \end{aligned}$$

Since (Y_i, Z_i) are identically distributed (see assumption [AD])

$$\begin{aligned} n_1^{1/2}\left(\hat{\theta}-\theta_o\right) &= n^{1/2}\left[n_1^{-1}\sum_{i=1}^{n_1}(\mathbb{E}[c(Y_i,\xi_o(Y_i,Z_i))]-c(Y_i,\xi_o(Y_i,Z_i)))\right] \\ &+n^{1/2}\mathbb{E}[\psi(Y_i,Z_i)[\hat{\xi}(Y_i,Z_i)-\xi_o(Y_i,Z_i)]] \\ &+o_P(1) \end{aligned}$$

By assumption [E15.(ii)]

$$n_1^{1/2}\left(\hat{\theta}-\theta_o\right) = n_1^{-1/2}\sum_{i=1}^{n_1}(\mathbb{E}[c(Y_i,\xi_o(Y_i,Z_i))]-c(Y_i,\xi_o(Y_i,Z_i))+\psi(Y_i,Z_i))+o_P(1)$$

Under assumptions [R.1], $n_1^{1/2}\left(\hat{\theta}-\theta_o\right)$ is asymptotically normal, say $N(0,\Sigma)$, by the central limit theorem, since it is mean zero sample average normalized by $n^{1/2}$. The expression for the variance covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_l^{asy} & \rho_{l,u}^{asy} \\ \rho_{l,u}^{asy} & \sigma_u^{asy} \end{pmatrix}$$

where , $\sigma_l^{asy} := \mathbb{V}[c(Y,\xi_l(Y,Z);\mu)+\psi(Y,Z)]$, $\sigma_u^{asy} := \mathbb{V}[c(Y,\xi_u(Y,Z);\mu)]$ and

$$\rho_{l,u}^{asy} := \mathbb{C}[c(Y,\xi_l(Y,Z);\mu)+\psi(Y,Z), c(Y,\xi_u(Y,Z);\mu)+\psi(Y,Z)]$$

which completes the proof. ■

We now prove the two lemmas we have invoked in the proof of proposition 2. Here is the first of these lemmas,

Lemma 5 *Let assumption [E] and [AD] holds. Define* $\hat{\boldsymbol{\beta}}(\tau)$ *as in (5). Then,*

$$\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\tau)-\mathbf{q}(z)'\boldsymbol{\beta}(\tau)\right| = o_P\left(n_2^{-1/4}\right)$$

Proof of Lemma 5. Take common factor $\mathbf{q}(z)'$

$$\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\tau)-\mathbf{q}(z)'\boldsymbol{\beta}(\tau)\right| = \sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\left[\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)\right]\right|$$

By the Cauchy-Schwarz inequality,

$$\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\tau)-\mathbf{q}(z)'\boldsymbol{\beta}(\tau)\right| \leq \sup_{z\in\mathcal{Z}}||\mathbf{q}(z)'|| \times \sup_{z\in\mathcal{Z}}\sup_{\tau\in[0,1]}|\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)|$$

Since $\mathbf{p}(z)'$ are splines functions we have $\sup_{z\in\mathcal{Z}}||\mathbf{q}(z)'|| \leq L_{n_2}^{1/2}$ (see Newey, 1997 pp. 160). Then,

$$\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\tau)-\mathbf{q}(z)'\boldsymbol{\beta}(\tau)\right| \leq L_{n_2}^{1/2}\sup_{z\in\mathcal{Z}}\sup_{\tau\in[0,1]}|\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)|$$

It follows from theorem 4.1 in Horowitz and Lee (2004) that $\sup_{z\in\mathcal{Z}}|\hat{\boldsymbol{\beta}}(\tau)-\boldsymbol{\beta}(\tau)| = O_p\left(L_{n_2}^{1/2}n_2^{-1/2}\right)$ (it remains to extent this result uniformly over τ). Then,

$$\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\tau)-\mathbf{q}(z)'\boldsymbol{\beta}(\tau)\right| = O_p\left(L_{n_2}n_2^{-1/2}\right)$$

Therefore under assumption [E.9] in [ES]

$$\sup_{\tau\in[0,1]}\sup_{z\in\mathcal{Z}}\left|\mathbf{q}(z)'\hat{\boldsymbol{\beta}}(\tau)-\mathbf{q}(z)'\boldsymbol{\beta}(\tau)\right| = o_P(n_2^{-1/4})$$

which completes the proof. ■

Define the $n_1\times K_{n_1}$ matrix $\mathbf{P} := (\mathbf{p}(Z_1)',...,\mathbf{p}(Z_{n_1})')'$. Using this notation we can rewrite the series estimator of the conditional distribution $G_{Y|Z}(y|z)$, $\hat{G}_{Y|Z}(y,z)$, as $\hat{G}_{Y|Z}(y,z) = \mathbf{p}(z)'(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{w}(y)$ where $\mathbf{w}(y) = (\mathbb{I}(Y_1\le y),..,\mathbb{I}(Y_{n_1}\le y))$ is a $n_1\times 1$ vector indexed by $y\in\mathcal{Y}$. Here is the second lemma:

Lemma 6 *Let assumption [E] and [AD] holds. Define $\hat{G}_{Y|Z}(y,z)$ as in (5). Then,*

$$\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)-\mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| = o_P\left(n_1^{-1/4}\right)$$

Proof of Lemma 6. Take common factor $\mathbf{p}(z)'$

$$\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)-\mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| = \sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'[\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)]\right|$$

By the Cauchy-Schwarz inequality,

$$\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)-\mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| \le \sup_{z\in\mathcal{Z}}||\mathbf{p}(z)'|| \times \sup_{z\in\mathcal{Z}}\sup_{y\in\mathcal{Y}}|\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)|$$

Since $\mathbf{p}(z)'$ are splines functions we have $\sup_{z\in\mathcal{Z}}||\mathbf{p}(z)'|| \le K_{n_1}^{1/2}$ (see Newey, 1997 pp.

160). Then,

$$\sup_{y\in\mathcal{Y}}\sup_{z\in\mathcal{Z}}\left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y)-\mathbf{p}(z)'\boldsymbol{\alpha}(y)\right|\leq K_{n_1}^{1/2}\times\sup_{z\in\mathcal{Z}}\sup_{y\in\mathcal{Y}}\left|\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)\right|$$

Now we prove $\sup_{y\in\mathcal{Y}}\left|\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)\right| = O_P(K_{n_1}^{1/2}n_1^{-1/2})$. Define the $n_1\times 1$ vector $\mathbf{g}:=(g_0(y_1,z_1),..,g_0(y_{n_1},z_{n_1}))'$ and $\boldsymbol{\varepsilon}(y):=\mathbf{w}(y)-\mathbf{g}$. By construction

$$\hat{\boldsymbol{\alpha}}(y)=(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{w}(y)$$

Add-and-substract $(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{g}$

$$\hat{\boldsymbol{\alpha}}(y)=(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{w}(y)-(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{g}+(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{g}$$

Take common factor $(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$ in the first two terms and use the definition of $\boldsymbol{\varepsilon}(y)$

$$\hat{\boldsymbol{\alpha}}(y)=(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)+(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{g}$$

Add-and-substract $(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{P}\boldsymbol{\alpha}(y)$

$$\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)=(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)+(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{g}-(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{P}\boldsymbol{\alpha}(y)$$

Take common factor $(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$ in the last two terms

$$\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)=(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)+(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'(\mathbf{g}-\mathbf{P}\boldsymbol{\alpha}(y))$$

By the triangle inequality,

$$\sup\nolimits_{z\in\mathcal{Z}}\sup\nolimits_{y\in\mathcal{Y}}||\hat{\boldsymbol{\alpha}}(y)-\boldsymbol{\alpha}(y)||\leq$$

$$(A)\ \sup\nolimits_{z\in\mathcal{Z}}\sup\nolimits_{y\in\mathcal{Y}}||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)/n_1||+$$

$$(B)\ \sup\nolimits_{z\in\mathcal{Z}}\sup\nolimits_{y\in\mathcal{Y}}||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'(\mathbf{g}-\mathbf{P}\boldsymbol{\alpha}(y))/n_1||$$

Consider expression (A). Let 1_n be the indicator function for the smallest eigenvalue of $(\mathbf{P}'\mathbf{P}/n_1)^{-1}$ being greater than 1/2, so $P(1_n = 1) \to 1$ as n_1 goes to infinity. Multiplying both sides of the later inequality by 1_n

$$\begin{aligned}\sup_{y\in\mathcal{Y}} 1_n \times ||\hat{\boldsymbol{\alpha}}(y) - \boldsymbol{\alpha}(y)|| \leq 1_n \times \sup_{y\in\mathcal{Y}} ||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y_1)||/n_1 \\ +1_n \times \sup_{y\in\mathcal{Y}} ||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))||/n_1\end{aligned}$$

Consider the expectation of the squared of the first term in the right hand side of the inequality

$$\begin{aligned}&\mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)/n_1||^2 \,|Z\right] \\ &= 1_n \times \mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} trace \left(\mathbf{P}(\mathbf{P}'\mathbf{P}/n_1)^{-1}(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)\boldsymbol{\varepsilon}(y)'\right) |Z\right] /n_1\end{aligned}$$

by the cycle property of the trace operator $trace(\cdot)$. Since $trace(\cdot)$ is linear

$$\begin{aligned}&\mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)/n_1||^2 \,|Z\right] \\ &= 1_n \times \mathbb{E}^* \left[trace \left(\sup_{y\in\mathcal{Y}} \mathbf{P}(\mathbf{P}'\mathbf{P}/n_1)^{-1}(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)\boldsymbol{\varepsilon}(y)'\right) |Z\right] /n_1\end{aligned}$$

Since $\mathbf{P}(\mathbf{P}'\mathbf{P}/n_1)^{-1}(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'$ does not depend on y and we are conditioning on Z

$$\begin{aligned}&\mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)/n_1||^2 \,|Z\right] \\ &1_n \times trace \left(\mathbf{P}(\mathbf{P}'\mathbf{P}/n_1)^{-1}(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} \boldsymbol{\varepsilon}(y)\boldsymbol{\varepsilon}(y)' \,|Z\right]\right) /n_1\end{aligned}$$

Boundedness of $z \mapsto \mathbb{E}[\mathbb{I}(Y < y)^q|z]$ uniformly over y (see assumption E.8 in [EA]) and independence of the observations (see assumption AD) imply $\mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} \boldsymbol{\varepsilon}(y)\boldsymbol{\varepsilon}(y)' \,|Z\right] \leq \bar{C}\mathbf{I}$, where $\leq$ denotes the usual positive semi-definite order and $\mathbf{I}$ is the identity matrix. Then,

$$\mathbb{E}^* \left[\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\boldsymbol{\varepsilon}(y)/n_1||^2 \,|Z\right] \leq 1_n \times trace \left(\mathbf{P}(\mathbf{P}'\mathbf{P}/n_1)^{-1}(\mathbf{P}'\mathbf{P}/n_1)^{-1}\mathbf{P}'\bar{C}\mathbf{I}\right) /n_1$$

Taking expectations over Z,

$$\mathbb{E}^*\left[\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P'P}/n_1)^{-\mathbf{1}}\mathbf{P'}\boldsymbol{\varepsilon}(y)/n_1||^2\right] \leq \bar{C}\times 1_n \times trace\left(\mathbb{E}^*\left[\mathbf{P}(\mathbf{P'P}/n_1)^{-1}(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}\right]\right)/n_1$$

Since $trace\left(\mathbb{E}^*\left[\mathbf{P}(\mathbf{P'P}/n_1)^{-1}(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}\right]\right) = K_{n_1}$

$$\mathbb{E}^*\left[\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P'P}/n_1)^{-\mathbf{1}}\mathbf{P'}\boldsymbol{\varepsilon}(y)/n_1||^2\right] \leq \bar{C} \times 1_n \times K_{n_1}/n_1$$

By the Markov inequality,

$$P\left(\sup_{y\in\mathcal{Y}} 1_n \times ||(\mathbf{P'P}/n_1)^{-\mathbf{1}}\mathbf{P'}\boldsymbol{\varepsilon}(y)/n_1|| > \bar{C} \times K_{n_1}^{1/2} \times n_1^{-1/2}\right) \leq \bar{C} \times K_{n_1}/n_1$$

That is$(A) = O_P(K_{n_1}^{1/2} n_1^{-1/2})$.

Consider now expression (B). Take the expectation of $\sup_{y\in\mathcal{Y}} ||(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))||/n_1$ conditional on Z

$$\begin{aligned}
&\mathbb{E}^*\left[\sup_{y\in\mathcal{Y}} ||(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))/n_1||^2 \,|Z\right] \\
&= \mathbb{E}^*\left[\sup_{y\in\mathcal{Y}} tr\left((\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))'\mathbf{P}(\mathbf{P'P}/n_1)^{-1}(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))\right)/n_1 \,|Z\right] \\
&= tr\left(\mathbb{E}^*\left[\sup_{y\in\mathcal{Y}}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))'\mathbf{P}(\mathbf{P'P}/n_1)^{-1}(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y)) \,|Z\right]\right)/n_1 \\
&= tr\left(\mathbf{P}(\mathbf{P'P}/n_1)^{-1}(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}\mathbb{E}^*\left[\sup_{y\in\mathcal{Y}}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))'(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y)) \,|Z\right]\right)/n_1
\end{aligned}$$

Because of assumption [E.6] in [EA], and taking expectations over Z

$$\begin{aligned}
\mathbb{E}^*\left[\sup_{y\in\mathcal{Y}} ||(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))/n_1||^2\right] &\leq K_{n_1}^{-1/2} \times K_{n_1}/n_1 \\
&\leq K_{n_1}^{1/2}/n_1
\end{aligned}$$

By the Markov inequality,

$$P\left(\sup_{y\in\mathcal{Y}} ||(\mathbf{P'P}/n_1)^{-1}\mathbf{P'}(\mathbf{g} - \mathbf{P}\boldsymbol{\alpha}(y))/n_1|| > \bar{C} \times K_{n_1}^{1/4} n_1^{-1/2}\right) \leq \bar{C} \times K_{n_1}^{1/2}/n_1$$

That is, $(B) = O_P(K_{n_1}^{1/4} n_1^{-1/2})$. Collecting results for (A) and (B) we have $\sup_{y\in\mathcal{Y}} ||\hat{\boldsymbol{\alpha}}(y) - \boldsymbol{\alpha}(y)|| = O_P(K_{n_1}^{1/2} n_1^{-1/2}) + O_P(K_{n_1}^{1/4} n_1^{-1/2})$. Furthermore,

$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y) - \mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| = O(K_{n_1}^{1/2}) \times O_P(K_{n_1}^{1/2} n_1^{-1/2})$$

and then

$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y) - \mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| = O_P\left(K_{n_1} n_1^{-1/2}\right)$$

Notice that $n_1^{-1/4} K_{n_1} n_1^{-1/2} = n_1^{-3/4} K_{n_1}$. Therefore under assumption [E.9] in [ES]

$$\sup_{y\in\mathcal{Y}} \sup_{z\in\mathcal{Z}} \left|\mathbf{p}(z)'\hat{\boldsymbol{\alpha}}(y) - \mathbf{p}(z)'\boldsymbol{\alpha}(y)\right| = o_P(n_1^{-1/4})$$

which completes the proof. ■

We now prove the lemma concerning the size of Ξ we have invoked in the proof of proposition 3. Let $L_2(P)$ denote the collection of square integrable functions with respect to the population P.

Lemma 7 ***(Bounded Bracketing Integral).*** *Let assumption [E], [R] and [AD] holds. Then,*

$$\int_0^\infty \sqrt{\ln N_{[]}\left(\varepsilon, \Xi, L_2(P)\right)} d\varepsilon < \infty$$

where $N_{[]}$ denotes the bracketing number of Ξ.

Proof of Lemma 7. For any vector of positive integers $a = (a_y, .., a_{d_Z})$, define the differential operator $D^a = \partial^{|a|}/\partial y^{a_y} ... \partial z^{a_{d_Z}}$, where $|a| = a_y + ... + a_{d_Z}$ and d_Z is the dimension of z. Let $\lfloor w \rfloor$ be the largest integer v such that $v \leq w$ for any $w \in \mathbb{R}$. For the function $y, z \mapsto \xi^l(y, z)$ define the norm

$$||\xi^l||_{\infty,w} := \max_{a:|a|\leq\lfloor w\rfloor} \sup_{(y,z)\in]0,1[^{1+d_z}} |D^a\xi^l(y,z)| + \max_{a:|a|=\lfloor w\rfloor} \sup_{(y,y',z,z')\in]0,1[^{(1+d_z)^2}} \frac{|D^a\xi^l(y,z) - D^a\xi^l(y',z')|}{||(y,z)-(y',z')||}$$

and similarly for $||\xi^u||_{\infty,w}$. Define $||\xi||_{\infty,w} := \max\{||\xi^l||_{\infty,w}, ||\xi^u||_{\infty,w}\}$. Because of assumption [EC] and the chain rule we have $||\xi||_{\infty,w} \leq \bar{C}$ for any $w \leq 2(1 + d_Z)$. Then, Ξ is a subset of the popular space of functions studied by var der Vaart and Wellner (1996, pp. 154) (see also Chen et. al. 2003, pp. 1598). It follows then from corollary 2.7.2 of van der Vaart and Wellner (1996) that the entropy with bracketing of the class Ξ is bounded by

$$\log N_{[]}\left(\epsilon, \Xi, L_2(P)\right) \leq \bar{C}\left(\frac{1}{\epsilon}\right)^{\frac{(1+d_Z)}{2(1+d_Z)}}$$

In our case $d_Z = 1$, hence $\int_0^\infty \sqrt{\ln N_{[]}\left(\varepsilon, \Xi, L_2(P)\right)} d\varepsilon < \infty$.■

Appendix B: Tables

TABLE I. *Simulation Results: Finite Sample Properties of the Estimator $\hat{\Theta}_I$.*

Design A - Normal Overlapping Marginals.

Design Variables							Lower Bound				Upper Bound			
K_{n_1}	L_{n_2}	n_1	n_2	$r_{1,Z}$	$r_{2,Z}$	θ_0	θ_l	SE	RMSE	AAD	θ_u	SE	MSE	AAD
2	2	100	100	.8	.9	.5	.458	.110	.110	.098	.981	.112	.259	.234
3	3	100	100	.8	.9	.5	.458	.125	.144	.123	.981	.127	.307	.283
4	4	100	100	.8	.9	.5	.458	.110	.121	.097	.981	.119	.277	.251
5	5	100	100	.8	.9	.5	.458	.115	.126	.099	.981	.125	.300	.275
6	6	100	100	.8	.9	.5	.458	.109	.119	.099	.981	.102	.297	.279
7	7	100	100	.8	.9	.5	.458	.126	.130	.108	.981	.108	.287	.266
2	2	100	100	.5	.5	.5	-.500	.102	.225	.203	1.00	.111	.299	.278
3	3	100	100	.5	.5	.5	-.500	.084	.226	.210	1.00	.093	.339	.326
4	4	100	100	.5	.5	.5	-.500	.093	.242	.225	1.00	.101	.373	.359
5	5	100	100	.5	.5	.5	-.500	.106	.279	.259	1.00	.094	.360	.348
6	6	100	100	.5	.5	.5	-.500	.107	.271	.249	1.00	.096	.409	.398
7	7	100	100	.5	.5	.5	-.500	.091	.292	.278	1.00	.089	.416	.406
2	2	100	100	.2	.1	.5	-.954	.100	.330	.315	.994	.102	.336	.321
3	3	100	100	.2	.1	.5	.954	.097	.361	.348	.994	.106	.366	.351
4	4	100	100	.2	.1	.5	-.954	.106	.380	.365	.994	.098	.398	.385
5	5	100	100	.2	.1	.5	-.954	.103	.392	.378	.994	.107	.398	.383
6	6	100	100	.2	.1	.5	-.954	.102	.408	.395	.994	.108	.426	.412
7	7	100	100	.2	.1	.5	-.954	.104	.429	.416	.994	.105	.438	.425

This table reports small sample properties for the estimator of the identified set when the parameter of interest is the Person's correlation between Y_1 and Y_2 (θ_o), and the distribution of Y_1, Y_2, Z is trivariate normal. The labels in the columns stand for: K_{n_1}:number of knots in the estimation of $H_{Y_1|Z}$ |L_{n_2}: number of knots in the estimation of $Q_{Y_1|Z}$ |n_1: size of sample 1 | n_2: size of sample 2 | r_{1Z}: correlation between Y_1, Z | r_{2Z}: correlation between Y_2, Z | θ_l: Identified set lower bound | SE: Standard Errors | RMSE: root mean squared error | AAD: Average Absolute Deviation| θ_u: Identified Set upper bound.

TABLE II. *Simulation Results: Finite Sample Properties of the Estimator* $\hat{\Theta}_I$.

Design B - LogNormal Overlapping Marginals.

			Design Variables					Lower Bound				Upper Bound		
K_{n_1}	L_{n_2}	n_1	n_2	$r_{1,Z}$	$r_{2,Z}$	θ_0	θ_l	SE	RMSE	AAD	θ_u	SE	MSE	AAD
2	2	100	100	.8	.9	.35	.338	.160	.162	.131	.971	.156	.271	.222
3	3	100	100	.8	.9	.35	.338	.134	.137	.100	.971	.129	.236	.198
4	4	100	100	.8	.9	.35	.338	.149	.148	.120	.971	.150	.298	.258
5	5	100	100	.8	.9	.35	.338	.162	.165	.133	.971	.122	.255	.224
6	6	100	100	.8	.9	.35	.338	.162	.166	.130	.971	.158	.313	.271
7	7	100	100	.8	.9	.35	.338	.141	.140	.113	.971	.172	.336	.288
2	2	100	100	.5	.5	.35	-.228	.122	.125	.096	1.00	.145	.326	.292
3	3	100	100	.5	.5	.35	-.228	.099	.109	.082	1.00	.135	.325	.296
4	4	100	100	.5	.5	.35	-.228	.125	.129	.099	1.00	.125	.339	.315
5	5	100	100	.5	.5	.35	-.228	.107	.110	.082	1.00	.131	.371	.347
6	6	100	100	.5	.5	.35	-.228	.110	.131	.095	1.00	.128	.377	.355
7	7	100	100	.5	.5	.35	-.228	.129	.145	.111	1.00	.131	.388	.365
2	2	100	100	.2	.1	.35	-.357	.083	.092	.074	.991	.131	.313	.285
3	3	100	100	.2	.1	.35	-.357	.090	.096	.076	.991	.135	.369	.344
4	4	100	100	.2	.1	.35	-.357	.090	.090	.072	.991	.116	.383	.364
5	5	100	100	.2	.1	.35	-.357	.094	.094	.075	.991	.123	.394	.375
6	6	100	100	.2	.1	.35	-.357	.079	.079	.061	.991	.132	.435	.414
7	7	100	100	.2	.1	.35	-.357	.078	.077	.061	.991	.146	.452	.428

This table reports small sample properties for the estimator of the identified set when the parameter of interest is the Person's correlation coefficient between Y_1 and Y_2 (θ_o), and the distribution of Y_1, Y_2, Z is trivariate lognormal. The labels in the columns stand for: K_{n_1}:number of knots in the estimation of $H_{Y_1|Z}$ $|L_{n_2}$: number of knots in the estimation of $Q_{Y_1|Z}$ $|n_1$: size of sample 1 | n_2: size of sample 2 | r_{1Z}: correlation between Y_1, Z | r_{2Z}: correlation between Y_2, Z | θ_l: Identified set lower bound | SE: Standard Errors | RMSE: root mean squared error | AAD: Average Absolute Deviation| θ_l: Identified Set upper bound.

TABLE III. *Simulation Results: Finite Sample Properties of the Estimator $\hat{\Theta}_I$.*

Design A - Normal Overlapping Marginals.

Design Variables							Lower Bound				Upper Bound			
K_{n_1}	L_{n_2}	n_1	n_2	$r_{1,Z}$	$r_{2,Z}$	θ_0	θ_l	SE	RMSE	AAD	θ_u	SE	MSE	AAD
2	2	500	500	.8	.9	.5	.458	.038	.069	.061	.981	.020	.078	.076
3	3	500	500	.8	.9	.5	.458	.047	.048	.038	.981	.049	.158	.150
4	4	500	500	.8	.9	.5	.458	.037	.056	.047	.981	.020	.090	.088
5	5	500	500	.8	.9	.5	.458	.057	.057	.046	.981	.056	.160	.150
6	6	500	500	.8	.9	.5	.458	.034	.051	.043	.981	.021	.093	.091
7	7	500	500	.8	.9	.5	.458	.050	.050	.040	.981	.054	.157	.148
2	2	500	500	.5	.5	.5	-.500	.046	.168	.162	1.00	.029	.156	.154
3	3	500	500	.5	.5	.5	-.500	.044	.202	.197	1.00	.048	.250	.245
4	4	500	500	.5	.5	.5	-.500	.054	.159	.150	1.00	.028	.160	.157
5	5	500	500	.5	.5	.5	-.500	.039	.196	.192	1.00	.049	.240	.235
6	6	500	500	.5	.5	.5	-.500	.053	.162	.155	1.00	.029	.166	.163
7	7	500	500	.5	.5	.5	-.500	.044	.208	.199	1.00	.045	.246	.242
2	2	500	500	.2	.1	.5	-.954	.032	.184	.181	.994	.030	.184	.182
3	3	500	500	.2	.1	.5	-.954	.049	.285	.281	.994	.042	.284	.281
4	4	500	500	.2	.1	.5	-.954	.038	.199	.195	.994	.033	.198	.195
5	5	500	500	.2	.1	.5	-.954	.044	.289	.286	.994	.041	.295	.292
6	6	500	500	.2	.1	.5	-.954	.042	.287	.284	.994	.041	.287	.281
7	7	500	500	.2	.1	.5	-.954	.039	.283	.280	.994	.046	.287	.283

This table reports small sample properties for the estimator of the identified set when the parameter of interest is the Person's correlation coefficient between Y_1 and Y_2 (θ_o), and the distribution of Y_1, Y_2, Z is trivariate normal. The labels in the columns stand for: K_{n_1}:number of knots in the estimation of $H_{Y_1|Z}$ $|L_{n_2}$: number of knots in the estimation of $Q_{Y_1|Z}$ $|n_1$: size of sample 1 | n_2: size of sample 2 | r_{1Z}: correlation between Y_1, Z | r_{2Z}: correlation between Y_2, Z | θ_l: Identified set lower bound | SE: Standard Errors | RMSE: root mean squared error | AAD: Average Absolute Deviation| θ_l: Identified Set upper bound.

TABLE IV. *Simulation Results: Finite Sample Properties of the Estimator* $\hat{\Theta}_I$*.*

Design B - LogNormal Overlapping Marginals.

Design Variables							Lower Bound				Upper Bound			
K_{n_1}	L_{n_2}	n_1	n_2	$r_{1,Z}$	$r_{2,Z}$	θ_0	θ_l	SE	RMSE	AAD	θ_u	SE	MSE	AAD
2	2	500	500	.8	.9	.35	.338	.084	.114	.091	.971	.077	.144	.121
3	3	500	500	.8	.9	.35	.338	.083	.113	.092	.971	.085	.147	.120
4	4	500	500	.8	.9	.35	.338	.076	.109	.092	.971	.090	.165	.138
5	5	500	500	.8	.9	.35	.338	.077	.108	.085	.971	.069	.142	.124
6	6	500	500	.8	.9	.35	.338	.098	.108	.077	.971	.098	.191	.164
7	7	500	500	.8	.9	.35	.338	.091	.109	.084	.971	.092	.176	.149
2	2	500	500	.5	.5	.35	-.228	.054	.077	.060	1.00	.069	.286	.278
3	3	500	500	.5	.5	.35	-.228	.059	.084	.066	1.00	.088	.295	.282
4	4	500	500	.5	.5	.35	-.228	.069	.094	.075	1.00	.096	.292	.276
5	5	500	500	.5	.5	.35	-.228	.059	.079	.062	1.00	.091	.295	.280
6	6	500	500	.5	.5	.35	-.228	.059	.076	.058	1.00	.083	.286	.274
7	7	500	500	.5	.5	.35	-.228	.053	.076	.059	1.00	.082	.289	.277
2	2	500	500	.2	.1	.35	-.357	.067	.067	.052	.991	.096	.375	.362
3	3	500	500	.2	.1	.35	-.357	.059	.059	.042	.991	.085	.373	.363
4	4	500	500	.2	.1	.35	-.357	.073	.073	.053	.991	.110	.391	.375
5	5	500	500	.2	.1	.35	-.357	.058	.060	.049	.991	.078	.363	.355
6	6	500	500	.2	.1	.35	-.357	.061	.061	.047	.991	.077	.367	.359
7	7	500	500	.2	.1	.35	-.357	.069	.069	.053	.991	.089	.386	.376

This table reports small sample properties for the estimator of the identified set when the parameter of interest is the Person's correlation coefficient between Y_1 and Y_2 (θ_o), and the distribution of Y_1, Y_2, Z is trivariate lognormal. The labels in the columns stand for: K_{n_1}:number of knots in the estimation of $H_{Y_1|Z}$ $|L_{n_2}$: number of knots in the estimation of $Q_{Y_1|Z}$ $|n_1$: size of sample 1 | n_2: size of sample 2 | r_{1Z}: correlation between Y_1, Z | r_{2Z}: correlation between Y_2, Z | θ_l: Identified set lower bound | SE: Standard Errors | RMSE: root mean squared error | AAD: Average Absolute Deviation| θ_l: Identified Set upper bound.

References

[1] ANDREWS, Donald W.K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity", *Econometrica*, 62(1), pp. 43-72.

[2] ATHEY, Susan and Guido IMBENS (2006): "Identification and Inference in Nonlinear Difference-in-Difference Models", *Econometrica*, 74(2), pp. 431-497.

[3] BERESTEANU, Arie and Francesca MOLINARI: "Asymptotic Properties for a Class of Partially Identified Models", *Econometrica*, 76(4), pp. 763-814 .

[4] CAMBANIS, Stamatis, Gordon SIMONS and William STOUT (1976): "Inequalities for Ek(X,Y) when the marginals are fixed", *Probability Theory and Related Fields*, 36 (4), pp. 285-294.

[5] CHEN, Xiaohong (2007): "Large-Sample Sieve Estimation of Semi-Nonparameteric Models", Chapter 76 in J. Heckman and E. Leamer (eds), *Handbook of Econometrics. Volume 6B*, Elsevier.

[6] CHEN, Xiaohong, Olivier LINTON and Ingrid van KEILEGOM (2003): "Estimation of Semiparametric Models when the Criterion Function Is not Smooth", *Econometrica*, 71(5), 1591-1608.

[7] CROSS, Phillip and Charles MANSKI (2002): "Regressions, Short and Long", *Econometrica*, 70(2), pp. 357-368

[8] DJEBBARI, Habiba and Jeffrey SMITH (2008): "Heterogenous Impacts in Progresa", *Journal of Econometrics*, 145(1), pp. 64-80.

[9] FAN, Yanquin and Sang Soo PARK (2010): "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference", *Econometric Theory*, forthcoming.

[10] FAN, Yanquin and Jisong WU (2010): "Partial Identification of the Distribution of Treatment Effects in Switching Regimes Models and its Confidence Sets", *Review of Economic Studies*, forthcoming.

[11] FAN, Yanquin and Dongming ZHU (2010): "Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes", mimeo.

[12] GENTZKOW, (2006): "Television and Voter Turnout", *Quarterly Journal of Economics*, 121(3), 971

[13] GILULA, Zvi, Robert McCULLOCH and Peter ROSSI (2006): "A Direct Approach to Data Fusion", *Journal of Marketing Research*, 43 pp. 73-83

[14] HECKMAN, James, Jeffry SMITH AND Nancy CLEMENTS (1997): "Making the Most Out of Programme Evaluation and Social Experiments: Accounting for Heterogeneity in Program Impacts", *Review of Economic Studies*, 64(4), pp. 487-435.

[15] HECKMAN, James, Edward VYTLACIL (2007): "Econometric Evaluation of Social Programs", Chapter 70 in J. Heckman and E. Leamer (eds), *Handbook of Econometrics. Volume 6B*, Elsevier.

[16] HOROWITZ, Joel and Sokbae LEE (2005): "Nonparametric Estimation of an Additive Quantile Regression Model", *Journal of the American Statistical Association*, 100(472) pp. 1238-1249.

[17] HOROWITZ, Joel and Charles MANSKI (1995): "Identification and Robustness with Contaminated and Corrupted Data", *Econometrica*, 63(2) pp. 281-302.

[18] HOROWITZ, Joel and Charles MANSKI (1998): "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations", *Journal of Econometrics*, 84 pp 37-58.

[19] HOROWITZ, Joel and Charles MANSKI (2006): "Identification and Estimation of Statistical Functionals Using Incomplete Data", *Journal of Econometrics*, 132, pp. 445-459

[20] IMBENS, Guido and Charles MANSKI (2004): "Confidence Intervals for Partially Identified Parameters", *Econometrica*, 74 pp. 1845-1857

[21] IMBENS, Guido and Whitney Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", *Econometrica*, 77(5) pp. 1481-1512.

[22] JOE, Harry (1997): *Multivariate Models and Dependence Concepts*, Monographs on Statistics and Applied Probability 73, Chapman & Hall/CRC.

[23] KADANE, J. (2001): "Some Statistical Problems in Merging Data Files", *Journal of Official Statistics*, 17, pp. 423-433.

[24] KOENKER, Roger (2005): *Quantile Regression*, Econometric Society Monographs Series, Cambridge University Press.

[25] MANSKI, Charles (1997): "Monotone Treatment Response", *Econometrica*, 65(6), pp. 1311-1334.

[26] MANSKI, Charles (2005): Partial Identification in Econometrics, in Durlauf and Blume (eds.), *New Palgrave Dictionary of Economics*, Vol 6 pp. 300-306.

[27] MOLCHANOV, Ilya (2005): *Theory of Random Sets*, Springer Verlag.

[28] MOLINARI, Francesca and Marcin PESKI (2006): "Generalization of a Result on Regressions, Short and Long", *Econometric Theory*, 22(1) pp. 159-163.

[29] MORIARTY, Chris and Fritz SCHEUREN (2003): "Statistical Matching: A Paradigm for Assesing the Uncertainty in the Procedure", *Journal of Official Statistics*, 17, pp. 407-422.

[30] NEWEY, Whitney (1994): "The Asymptotic Variance of Semiparametric Estimators", *Econometrica*, 62(6), pp. 1349-1382.

[31] NEWEY, Whitney (1997): "Convergence Rates and Asymptotic Normality for Series Estimators", *Journal of Econometrics*, 79(1), pp. 147-168.

[32] OGBURN, (1919): "How Women Vote", *Political Science Quarterly*, pp. 413-433.

[33] RACHEV, Svetlozar and Ludger RUSCHENDORF (1998): *Mass Transportation Problems. Volume I: Theory,* Springer Verlag.

[34] RASSLER, Sussane (2002): *Statistical Matching*, Springer Verlag.

[35] RIDDER, Geert and Robert MOFFIT (2007): "The Econometrics of Data Combination", Chapter 75 in J. Heckman and E. Leamer (eds), *Handbook of Econometrics. Volume 6B*, Elsevier.

[36] ROBINSON, W. (1950): "Ecological Correlation and the Behavior of Individuals", *American Sociological Review,* 15(3), pp. 351-357.

[37] ROCKAFELLAR, Tyrrel (1970): *Convex Analysis*, Princeton University Press.

[38] RUSCHENDORF, Ludger (1991): "Bounds for Distributions with Multivariate Marginals", in K. Mosler and M. Scarsini (eds.), *Stochastic Orders and Decision under Risk*, IMS Lecture Notes-Monograph Series, Vol. 19, pp. 285-310.

[39] SONG, Kyungchul (2008): "Uniform Convergence Rates of Series Estimators over Function Spaces", *Econometric Theory*, 24 pp. 1463-1499.

[40] STOYE, Jörg (2009): "More on Confidence Intervals for Partially Identified Parameters", *Econometrica*, 77(4), pp. 1299-1315

[41] TAMER, Elie (2009): "Partial Identification in Econometrics", *Annual Review of Economics*, forthcomming.

[42] TCHEN, André (1980): "Inequalities for Distributions with Given Marginals", *The Annals of Probability*, 8(4), pp. 814-827.

[43] THE NIELSEN COMPANY (2007): *Introduction to Data Fusion*, available at www.nielsen.com.

[44] VAN DER VAART, Aad and Jon WELLNER (1996): *Weak Convergence and Empirical Processes*, Springer-Verlag.

[45] VITALE, Richard (1979): "Regression with Given Marginals", *The Annals of Statistics*, 7(3), pp. 653-658

[46] YU, Keming, Zudi LU and Julian STANDER (2003): "Quantile Regression: Applications and Current Research Areas", *Journal of the Royal Statistical Association. Series D (The Statistician)*, 52(2) pp. 331-350.