Identification region of the potential outcome distributions under instrument independence

Toru Kitagawa

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP30/09

Identification Region of the Potential Outcome Distributions under Instrument Independence

Toru Kitagawa[*][†]
University College London and CeMMAP

October 2009.

Abstract

This paper examines identification power of the instrument exogeneity assumption in the treatment effect model. We derive the identification region: The set of potential outcome distributions that are compatible with data and the model restriction. The model restrictions whose identifying power is investigated are (i) instrument independence of each of the potential outcome (*marginal independence*), (ii) instrument *joint independence* of the potential outcomes and the selection heterogeneity, and (iii) instrument monotonicity in addition to (ii) (the *LATE restriction* of Imbens and Angrist (1994)), where these restrictions become stronger in the order of listing. By comparing the size of the identification region under each restriction, we show that the joint independence restriction can provide further identifying information for the potential outcome distributions than marginal independence, but the LATE restriction never does since it solely constrains the distribution of data. We also derive the tightest possible bounds for the average treatment effects under each restriction. Our analysis covers both the discrete and continuous outcome case, and extends the treatment effect bounds of Balke and Pearl (1997) that are available only for the binary outcome case to a wider range of settings including the continuous outcome case.

Keywords: Partial Identification, Program Evaluation, Treatment Effects, Instrumental Variables

JEL Classification: C14, C21.

[*] Email: t.kitagawa@ucl.ac.uk
Webpage: http://www.homepages.ucl.ac.uk/~uctptk0/

[†] An earlier version of this paper appears in a chapter of my Ph.D. dissertation at Brown University. I thank Guido Imbens, Frank Kleibergen, and the seminar and conference participants at Harvard Econometrics Lunch and the 2009 SETA/CeMMAP conference in Kyoto for valuable comments. Financial Support from CeMMAP and the Brown University Merit Dissertation Fellowship are gratefully acknowledged.

1 Introduction

The method of instrumental variables is one of the most important inventions in econometrics, and the instrumental variable accompanied with the instrument exogeneity restriction plays a key role in extracting identifying information for the causal effects in many contexts. This is also the case in a more recent development of the nonseparable triangular simultaneous equation model (Chesher (2003, 2005), Imbens and Newey (2008)). Another important class of models where identification hinges on the instrument exogeneity restriction is the heterogeneous treatment effect model with selection (Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996), Heckman and Vytlacil (2001a, 2005)). In the latter model, if the instrument is randomized and every unit in the population has weakly monotonic selection response to the instrument, then the potential outcome distributions are identified for the subpopulation of those, so called compliers, whose treatment selection response is affected by the instrument (Imbens and Rubin (1997), Abadie, Angrist, and Imbens (2002)).

In this paper, we analyze identification of the potential outcome distributions for the entire population instead of the subpopulation of compliers. In particular, the population potential outcome distributions and the corresponding average treatment effects become the objects of interest when the policy analyst is interested in predicting the impact of policy intervention or making a statistical decision on the treatment choice (Manski (2005, 2007)). If the individual treatment effects are heterogeneous, however, the population distribution of the potential outcomes Y_1 and Y_0 is in general not identified by the instrument exogeneity restrictions. We therefore analyze the problem in the set-identification framework and our focus is on the identification region of the potential outcome distributions: The set of potential outcome distributions that are compatible with empirical evidence (data) and the model restrictions. The model restrictions analyzed in this paper are the following three types of the instrument exogeneity restriction, (i) the instrument is independent of each of the potential outcome (*marginal independence*), (ii) the instrument is *jointly independent* of the potential outcomes and the selection heterogeneity, and (iii) the instrument is jointly independent and selection response is monotonic with respect to the instrument (the *LATE restriction*). These restrictions are nested and become stronger in the order of listing. Although these restrictions are mathematically distinct, they all involve the researcher's belief that the instrument is assigned randomly irrespective of individual unobserved heterogeneities that can influence the outcomes. The use of instrument in the program evaluation context always relies on some of these restrictions, while less research has been done for clarifying what is the maximal identification information for the potential outcome distributions under each of the instrument independence restriction. The main goal of this paper is therefore to provide a rigorous identification analysis for the instrument independence restriction from the perspective of set-identification.

This paper defines, formulates, and derives the identification region of the marginal distributions of the potential outcomes under each of the instrument independence restrictions. We derive a closed-form expression of the identification region, which is represented as a correspondence from the distribution of data to a pair of the marginal distributions of the potential outcomes. Our definition of the identification region does not *a priori* constrain the distribution of data and, therefore constructing the identification region can be viewed as an inductive learning process for the potential outcome distributions. We investigate identification power of each of the instrument independence restriction by comparing the size of the identification region among these restrictions.

We also clarify for which data distribution the identification region can or cannot shrink further by strengthening one restriction to another. We show that for some data generating processes, the instrument joint independence restriction can yield a narrower identification region than the weaker restriction of marginal independence. This result contrasts the role of instrument independence restriction with the one in the missing data context since such identification gain never arises there (Kitagawa (2009a)). Another important finding is that the LATE restriction never provides further identification gain compared with the joint independence restriction, because it only constrains the distribution of data. In this sense, we demonstrate that instrument monotonicity of the LATE restriction is redundant for the purpose of nonparametrically identifying the population potential outcome distributions.

Once the identification region of the potential outcome distributions is obtained, the *sharp* bounds of the parameter θ defined on the potential outcome distributions are constructed by the range of θ with its domain given by the identification region. This implies that the comparative size relationship of the identification region is preserved as it is for the width of the sharp bounds for θ. As an application of this bounding scheme, this paper derives a closed-form expression of the bounds for the average treatment effects $E(Y_1) - E(Y_0)$ under each of the instrument independence restriction. The obtained bounds not only unifies the existing results in the literature, but also generalizes the existing analysis available only for the binary outcome case to the wide range of setting including the continuous outcome case. Manski (1990, 1994, 2003) derive its bounds under the restriction of instrument *mean* independence, $E(Y_1|Z) = E(Y_1)$ and $E(Y_0|Z) = E(Y_1)$. For the binary outcome case, his bounds can be interpreted as the bounds under the instrument marginal independence restriction. Balke and Pearl (1997) considers bounding the average treatment effects in the binary outcome case under the instrument joint independence restriction, and shows that their bounds can be strictly narrower than the Manski's mean independence bounds.[1] In the analysis of Balke and Pearl (1997), the bounds are obtained by solving a finite-dimensional linear programming, and it is not straightforward to extend their procedure to the case where potential outcomes have continuous variation. This paper, in contrast, provides a closed-form expression of the bounds for the average treatment effects that covers the continuous outcome case. Moreover, our derivation does not rely on the machinery of linear optimization, and this will help us develop an intuition behind the construction of the bounds. Our identification analysis also differs from the analysis of Heckman and Vytlacil (2001a, 2001b, 2005) since they assume that the population satisfies the selection equation with the threshold crossing with an additive error. Consequently, their analysis imposes an assumption on the distribution of data, and the tightest possible property of the bounds is limited to a certain class of data distributions. Our analysis ,in contrast, does not restrict the distribution of data so that the identification results presented in this paper are valid for *every* data we potentially encounter.

Since this paper exclusively focuses on identifying the marginal distributions of the potential outcomes, our analysis is free from the structural outcome equation accompanied by some assumptions on the unobserved outcome heterogeneity. That is, validity of our results does not rely on any assumptions on the dimension of the unobservable heterogeneity and the functional form specification

[1] Chen and Small (2006) extend the Balke and Pearl's bound analysis of the binary outcome to the case with three treatment status.

of the structural outcome equation. In this sense, the identification results of this paper provides a benchmark relative to which we can investigate what type of restrictions on the structure provides identifying power for the causal effects. In particular, for the continuous outcome variable case, the set-identification results of this paper provides a vivid contrast with the point-identification results under *outcome monotonicity in a scalar unobservable, or* equivalently *rank invariance* restriction (Chernozhukov and Hansen (2005)). This comparison suggests that the unobservable heterogeneity and the functional form specification for the structural outcome equation often introduced in the nonseparable structural equation model plays an essential role in identifying the potential outcome distributions.

The remainder of the paper is organized as follows. Section 2 introduces the setup and notation of this paper and provides the formal definition of the identification region. In Section 3, we derive the identification region of the potential outcome distributions under each of the instrument independence restriction. In Section 4, we compare the obtained identification regions and also present the sharp bounds for the average treatment effects. Section 5 discusses the link with the nonseparable structural equation model with a binary endogenous variable, and Section 6 concludes. Proofs are provided in Appendix A.

2 Analytical Framework

2.1 Data Generating Process and Population

Consider identifying the causal effect of a binary treatment to a measure of outcome. We use $D \in \{1, 0\}$ as the treatment indicator: $D = 1$ indicates a treated unit and $D = 0$ indicates a non-treated unit. Following the Neyman-Rubin potential outcome framework, Let Y_1 denote the outcome that would be observed if the individual receives treatment and let Y_0 denote the outcome that would be observed if the individual does not receive the treatment. The observed outcome in data is written as $Y \equiv DY_1 + (1 - D)Y_0$. We let the support of Y_1 and Y_0 be a subset of $\mathcal{Y}$.[2] This paper focuses on the situation where the treatment status is not randomized. In this case, we are typically concerned about the selection problem, i.e., the realized treatment status can depend on the underlying potential outcomes. We suppose that a nondegenerate binary variable $Z \in \{1, 0\}$ is available in data, and we consider to use the binary variable Z as an instrumental variable (Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996)). In the experimental setting with incompliance, for instance, the initial treatment assignment is often used as an instrumental variable. Throughout the paper, data is a random sample of (Y, D, Z).

We denote a conditional distribution of (Y, D) given Z by

$$\begin{aligned}
P_{Y_1}(B) &\equiv \Pr(Y \in B, D = 1|Z = 1) = \Pr(Y_1 \in B, D = 1|Z = 1),\\
P_{Y_0}(B) &\equiv \Pr(Y \in B, D = 0|Z = 1) = \Pr(Y_0 \in B, D = 0|Z = 1),\\
Q_{Y_1}(B) &\equiv \Pr(Y \in B, D = 1|Z = 0) = \Pr(Y_1 \in B, D = 1|Z = 0),\\
Q_{Y_0}(B) &\equiv \Pr(Y \in B, D = 0|Z = 0) = \Pr(Y_0 \in B, D = 0|Z = 0).
\end{aligned} \tag{1}$$

[2] The analytical framework considered in this paper is not restricted to a scalar outcome. We can take $\mathcal{Y}$ as an arbtrary space equipped with the Borel σ-algebra $B(\mathcal{Y})$ and a measure μ.

where B is an arbitrary subset of $\mathcal{Y}$. Since $P=(P_{Y_1}(\cdot),P_{Y_0}(\cdot))$ and $Q=(Q_{Y_1}(\cdot),Q_{Y_0}(\cdot))$ uniquely characterize the distribution of data except for the marginal distribution of Z, we represent the *data generating process* by $(P,Q)\in\mathcal{P}$ where $\mathcal{P}$ is the class of data generating processes. We assume that the researcher has knowledge on the dominating measure μ on $\mathcal{Y}$, and we denote the density functions of $P_{Y_j}(\cdot)$ and $Q_{Y_j}(\cdot)$, $j=1,0$, with respect to μ by $p_{Y_j}(\cdot)$ and $q_{Y_j}(\cdot)$. That is, for every subset $B\subset\mathcal{Y}$, we have

$$\begin{aligned} P_{Y_1}(B) &= \int_B p_{Y_1}(y_1)d\mu, \qquad P_{Y_0}(B)=\int_B p_{Y_0}(y_0)d\mu, \\ Q_{Y_1}(B) &= \int_B q_{Y_1}(y_1)d\mu, \qquad Q_{Y_0}(B)=\int_B q_{Y_0}(y_0)d\mu. \end{aligned}$$

It is important to keep in mind that the integration of the density functions $p_{Y_j}(\cdot)$ and $q_{Y_j}(\cdot)$ over $\mathcal{Y}$ yield the conditional probabilities of the observed treatment status given Z, $\Pr(D=j|Z=1)$ and $\Pr(D=j|Z=0)$, that are can be strictly smaller than one. Throughout the analysis, we do not restrict the class of data generating processes $\mathcal{P}$ other than existence of the density functions with respect to μ.

Our identification framework has the selection equation with the unobserved selection heterogeneity V,

$$D=I\{u(Z,V)\geq 0\},$$

where $u(Z,V)$ is the latent utility to rationalize individual's selection on treatment status, and V is the unobserved heterogeneities that affect one's selection response and is possibly dependent on the potential outcomes. We interpret this equation as structural in the sense that, with V fixed, $u(z,V)$ gives the counterfactual selection response for each $z=1,0$. Provided that D and Z are binary, the number of distinct selection responses called *type* are at most four as defined below, and the role of the unobserved heterogeneity V is to randomly categorize the individuals into one of these four types. A random category variable T is used to indicate the type,

$$T=\begin{cases} c: & \text{complier} \quad \text{if } u(1,V)=1,\ u(0,V)=0, \\ n: & \text{never-taker} \quad \text{if } u(1,V)=u(0,V)=0, \\ a: & \text{always-taker} \quad \text{if } u(1,V)=u(0,V)=1, \\ d: & \text{defier} \quad \text{if } u(1,V)=0,\ u(0,V)=1. \end{cases}$$

If we do not impose any restrictions on the distribution of T, we are free from any assumptions on the functional form of the latent utility as well as dimensionality of the unobserved heterogeneity V (Pearl (1994a)).

Every unit in the population of interest has a nonrandom value of (Y_1,Y_0,T,Z) and the parameter of interest is defined on the distribution of (Y_1,Y_0,T,Z). In this sense, we define *population* as a joint probability distribution of $(Y_1,Y_0,T,Z)\in\mathcal{Y}\times\mathcal{Y}\times\{c,n,a,d\}\times\{1,0\}$. Hereafter, f denotes the probability density function of the population variables indicated by the subscripts such as f_{Y_1}, $f_{Y_1,T|Z}$, etc. We use $\mathcal{F}$ to denote the class of populations.

2.2 Defining the Identification Region

Model restrictions are imposed in order to extract identifying information for the potential outcome distributions. Like the instrument exogeneity restriction introduced below, they have the form of statistical relationships among the population random variables (Y_1, Y_0, T, Z). Let A be the model restriction(s) and $\mathcal{F}_A \subset \mathcal{F}$ be the subclass of populations constrained by the imposed restrictions A.

For each data generating process (P, Q), the class of *observationally equivalent* populations $\mathcal{F}^o(P, Q) \subset \mathcal{F}$ is defined as the collection of the distribution of (Y_1, Y_0, T, Z) that generates the empirical evidence (P, Q). Given a data generating process $(P, Q) \in \mathcal{P}$, the identification region under restriction A denoted by $IR(P, Q|A)$, is defined as *the set of populations that are compatible with the empirical evidence* (P, Q) *and restriction* A. That is, $IR(P, Q|A)$ is formulated as the intersection of $\mathcal{F}_A$ and $\mathcal{F}^o(P, Q)$,

$$IR(P, Q|A) \equiv \mathcal{F}_A \cap \mathcal{F}^o(P, Q), \qquad (P, Q) \in \mathcal{P}$$

When $IR(P, Q|A)$ thus defined becomes empty, it implies that restriction A is not compatible with the observed data, and therefore we can refute the model restriction A. Since this refuting rule is based on the emptiness of the identification region, no other testable implications can have more refuting power than this.

If our interest lies in $\theta : \mathcal{F} \rightarrow \Theta$, a feature or parameters of the population, the identification region of θ under A denoted by $IR_\theta(P, Q|A)$ is defined as the range of $\theta(\cdot)$ with its domain given by $IR(P, Q|A)$. When $IR(P, Q|A)$ is empty, we define $IR_\theta(P, Q|A)$ as empty so as to reflect the refutability property of the identification region. So, the identification region of θ under A is defined as

$$IR_\theta(P, Q|A) \equiv \begin{cases} \{\theta(F) : F \in IR(P, Q|A)\} \cap \Theta & \text{if } IR(P, Q|A) \neq \emptyset, \\ \emptyset & \text{if } IR(P, Q|A) = \emptyset. \end{cases} \tag{2}$$

In words, $IR_\theta(P, Q|A)$ is defined as *the set of* θ *for each of which we can construct a population* F *that is compatible with data* P, Q *and the imposed restriction* A.

The identification region defined here does *not* assume that the true population satisfies the imposed restrictions. In this regard, our definition differs from the one of Heckman and Vytlacil (2007). This difference in fact matters when the imposed restriction A is *observationally restrictive* (Koopmans and Reiersøl (1951)), i.e., the restrictions constrains the data generating process. Specifically, when A is observationally restrictive, by assuming that the true population satisfies restriction A, we *a priori* exclude the possibility of having empty $IR(P, Q|A)$, even though data is potentially informative about it. As an unfavorable consequence of assuming the correct specification, we may encounter the *misspecification problem* of the bounds, meaning that the bound *formula* for θ and its tightest-possible property justified under the correct specification are no longer valid if $IR(P, Q|A) = \emptyset$. As we discuss further in Section 4, the bounds of the average treatment effects under the instrument independence restriction provides an example of this type of misspecification problem. In order to avoid the potential misspecification problem, we do keep the class of data generating processes $\mathcal{P}$ invariant no matter which restriction we impose, and construct the bounds by explicitly applying the above definition (2).

2.3 Instrumental Variable Restrictions

Regarding the model restrictions, we consider the following three restrictions in all of which the notion of instrument exogeneity is represented in terms of its statistical independence of the potential outcomes.

Restriction MSI:

Marginal Statistical Independence Restriction: Z is statistically independent of each Y_1 and Y_0, i.e., $Z \perp Y_1$ and $Z \perp Y_0$.

Restriction RA:

Random Assignment Restriction: Z is jointly independent of (Y_1, Y_0, T).

Restriction LATE:

LATE Restriction: Z is jointly independent of (Y_1, Y_0, T), and $f_T(T = d) = 0$ or $f_T(T = c) = 0$.

Obviously, these model restrictions are nested and become stronger in the order that they are listed. The first restriction MSI only imposes marginal independence between the instrument and each of the potential outcome. Since the model restriction has nothing to do with the selection heterogeneity T, the analysis corresponding to this case is robust to dependence between instrument and the selection heterogeneity T.[3] The second restriction RA, in contrast, embodies a stronger version of instrument exogeneity such that the instrument is jointly independent of both the outcome heterogeneities and the selection heterogeneities. RA will be a reasonable restriction if the researcher believes that the instrument is generated through some randomization mechanism as is often the case in the (quasi-)experimental setting. The final restriction $LATE$ is due to Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996), and it plays the fundamental role for identifying the potential outcome distributions for the subpopulation of compliers.

Our primary interest lies in identifying f_{Y_1} and f_{Y_0} the marginal distributions of Y_1 and Y_0. This is often the case if the goal of analysis is to assess the effect of intervention by comparing the marginal distributions of the potential outcomes. For example, the *average treatment effects* is defined as the difference between the mean of f_{Y_1} and f_{Y_0}. Alternatively, we may be interested in the τ-th *quantile differences* defined as the difference of the τ-th quantiles between the two potential outcome distributions. In case where we are interested in the effect of intervention to the inequality of outcomes, the variances of f_{Y_1} and f_{Y_0} may be of our interest. For all these cases, the parameters of interest are defined in terms of the marginal distributions of Y_1 and Y_0, and therefore we shall focus on constructing $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|\cdot)$ the identification region of f_{Y_1} and f_{Y_0}. Note that if our interest lies in a parameter that is defined on the distribution of the individual causal effects $Y_1 - Y_0$, $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|\cdot)$ is less useful since the distribution of $Y_1 - Y_0$ is sensitive not only to the marginals of Y_1 and Y_0 but also to dependence between Y_1 and Y_0. Identification of the distribution of $Y_1 - Y_0$ is out of scope of this paper.[4]

[3] The identification analysis of Cherozhukov and Hansen (2005) and Chesher (2009) is also free from the selection equation. A difference from their analysis is that our analysis do not impose any assumptions on the association between Y_1 and Y_0.

[4] In the situation where the marginal distributions of Y_1 and Y_0 are point-identified, Fan and Park (2008), Firpo

3 Construction of the Identification Region

For the construction of $IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$, our first step is to formulate the conditions for $F \in \mathcal{F}^o(P,Q)$, i.e., compatibility of a distribution of (Y_1,Y_0,T,Z) with the observed data (P,Q). They are obtained by rewriting the right-hand side of the identities (1) in terms of the distribution of (Y_1,Y_0,T,Z).

$$\begin{aligned} p_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1,T=c|Z=1) + f_{Y_1,T|Z}(y_1,T=a|Z=1), \\ q_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1,T=d|Z=0) + f_{Y_1,T|Z}(y_1,T=a|Z=0), \\ p_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0,T=d|Z=1) + f_{Y_0,T|Z}(y_0,T=n|Z=1), \\ q_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0,T=c|Z=0) + f_{Y_0,T|Z}(y_0,T=n|Z=0). \end{aligned} \tag{3}$$

These four equations are interpreted as compatibility of the population with the data generating process $F \in \mathcal{F}^o(P,Q)$.

By the law of total probability, $f_{Y_1|Z}(y_1|Z=z) = \sum_{t\in\{c,n,a,d\}} f_{Y_1,T|Z}(y_1,T=t|Z=z)$ and $f_{Y_0|Z}(y_0|Z=z) = \sum_{t\in\{c,n,a,d\}} f_{Y_0,T|Z}(y_0,T=t|Z=z)$ hold and they imply the difference of $f_{Y_j|Z}$ minus the observed densities p_{Y_j} or q_{Y_j} also has a similar mixture form,

$$\begin{aligned} f_{Y_1|Z}(y_1|Z=1) - p_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1,T=d|Z=1) + f_{Y_1,T|Z}(y_1,T=n|Z=1), \\ f_{Y_1|Z}(y_1|Z=0) - q_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1,T=c|Z=0) + f_{Y_1,T|Z}(y_1,T=n|Z=0), \\ f_{Y_0|Z}(y_0|Z=1) - p_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0,T=c|Z=1) + f_{Y_0,T|Z}(y_0,T=a|Z=1), \\ f_{Y_0|Z}(y_0|Z=0) - q_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0,T=d|Z=0) + f_{Y_0,T|Z}(y_0,T=a|Z=0). \end{aligned} \tag{4}$$

These identities will be used later on to relate the distributions $f_{Y_j|Z}$ to the distribution of $f_{Y_j,T|Z}$ for a given data generating process.

3.1 Identification Region under Marginal Independence (MSI)

If we impose MSI, $f_{Y_1|Z} = f_{Y_1}$ and $f_{Y_0|Z} = f_{Y_0}$ must hold. Therefore $f_{Y_1|Z}$ and $f_{Y_0|Z}$ appearing in the left hand side of (4) are reduced to the unconditional ones, so we have

$$\begin{aligned} f_{Y_1}(y_1) - p_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1,T=d|Z=1) + f_{Y_1,T|Z}(y_1,T=n|Z=1), \\ f_{Y_1}(y_1) - q_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1,T=c|Z=0) + f_{Y_1,T|Z}(y_1,T=n|Z=0), \\ f_{Y_0}(y_0) - p_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0,T=c|Z=1) + f_{Y_0,T|Z}(y_0,T=a|Z=1), \\ f_{Y_0}(y_0) - q_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0,T=d|Z=0) + f_{Y_0,T|Z}(y_0,T=a|Z=0). \end{aligned} \tag{5}$$

Given $(P,Q) \in \mathcal{P}$, any populations contained in $IR(P,Q|MSI)$ satisfy (3) and (5). That is, by noting that the right hand side of (5) has the nonnegative functions, we find necessary conditions for (f_{Y_1},f_{Y_0}) to be contained in $IR_{(f_{Y_1},f_{Y_0})}(P,Q|MSI)$,

$$f_{Y_1}(y_1) \geq \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\} \;\; \mu\text{-a.e.} \quad \text{and} \quad f_{Y_0}(y_0) \geq \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\} \;\; \mu\text{-a.e.}$$

We hereafter call, for each $j = 1,0$, $\max\{p_{Y_j},q_{Y_j}\}$ as the *density envelope* for Y_j and $\delta_{Y_j} \equiv \int_{\mathcal{Y}} \max\{p_{Y_j},q_{Y_j}\}d\mu$ as the integrated envelope for Y_j. The next proposition shows that these conditions are in fact sufficient to build up $IR_{(f_{Y_1},f_{Y_0})}(P,Q|MSI)$, i.e., any f_{Y_1} and f_{Y_0} that each lies

and Ridder (2008), and Heckman, Smith, Clements (1997) analyze identification of the distribution of the individual causal effects $Y_1 - Y_0$.

above the density envelope constitutes the identification region of (f_{Y_1}, f_{Y_0}) under MSI. This result can be seen as a direct extension of Lemma 2.2 of Manski (2003) for the missing data model to the treatment effect model.

Proposition 3.1 (Identification region under marginal independence) *Denote the density envelopes by* $\underline{f_{Y_1}}(y_1) \equiv \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}$ *and* $\underline{f_{Y_0}}(y_0) \equiv \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}$, *and the integrated envelopes by* $\delta_{Y_1} \equiv \int_{\mathcal{Y}} \underline{f_{Y_1}} d\mu$ *and* $\delta_{Y_0} \equiv \int_{\mathcal{Y}} \underline{f_{Y_0}} d\mu$. *Define the sets of probability densities that cover* $\underline{f_{Y_1}}(y_1)$ *and* $\underline{f_{Y_0}}(y_0)$ *respectively by*

$$\begin{aligned}\mathcal{F}^{env}_{f_{Y_1}}(P,Q) &= \left\{ f_{Y_1} : \int_{\mathcal{Y}} f_{Y_1}(y_1) d\mu = 1,\ f_{Y_1}(y_1) \geq \underline{f_{Y_1}}(y_1)\ \mu\text{-}a.e. \right\}, \\ \mathcal{F}^{env}_{f_{Y_0}}(P,Q) &= \left\{ f_{Y_0} : \int_{\mathcal{Y}} f_{Y_0}(y_0) d\mu = 1,\ f_{Y_0}(y_0) \geq \underline{f_{Y_0}}(y_0)\ \mu\text{-}a.e. \right\}.\end{aligned}$$

The identification region under MSI is nonempty if and only if $\delta_{Y_1} \leq 1$ *and* $\delta_{Y_0} \leq 1$, *and it is given by*

$$IR_{(f_{Y_1}, f_{Y_0})}(P, Q|MSI) = \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q).$$

Proof. See Appendix A. ■

It is intuitive that the envelope density $\underline{f_{Y_1}}(y_1)$ provides the maximal identifying information for Y_1's distribution because under MSI each of the observed density $p_{Y_1}(y_1)$ and $q_{Y_1}(y_1)$ must be a part of the common density f_{Y_1} and taking the envelope can be viewed as filling out f_{Y_1} as much as possible with the identified objects $p_{Y_1}(y_1)$ and $q_{Y_1}(y_1)$. The result that $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|MSI)$ takes the form of the Cartesian product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ implies that the two marginal independence restrictions never provide a channel through which the identifying information for f_{Y_1} contributes to identifying f_{Y_0} or vice versa. Therefore, as far as marginal independence is concerned, we can always separate identification analysis of f_{Y_1} from the one of f_{Y_0} without losing any identifying information, and this implication justifies the bounding strategy of *outer bounds* of Manski (2003).

The refutability result of the marginal independence coincides with the testability result for the instrument exclusion restriction analyzed in Pearl (1994b) and is analogous to the missing data case analyzed in Manski (2003). Kitagawa (2009a) considers estimation and inferential aspect of the integrated envelope parameter so as to develop a specification test for instrument independence.

3.2 Identification Region under Random Assignment (RA)

If we strengthen MSI to RA, we can replace the conditional distributions appearing in the right hand side of (3) and (5) with the unconditional ones,

$$\begin{aligned}
p_{Y_1}(y_1) &= f_{Y_1,T}(y_1, T=c) + f_{Y_1,T}(y_1, T=a),\\
q_{Y_1}(y_1) &= f_{Y_1,T}(y_1, T=d) + f_{Y_1,T}(y_1, T=a),\\
p_{Y_0}(y_0) &= f_{Y_0,T}(y_0, T=d) + f_{Y_0,T}(y_0, T=n),\\
q_{Y_0}(y_0) &= f_{Y_0,T}(y_0, T=c) + f_{Y_0,T}(y_0, T=n),\\
f_{Y_1}(y_1) - p_{Y_1}(y_1) &= f_{Y_1,T}(y_1, T=d) + f_{Y_1,T}(y_1, T=n),\\
f_{Y_1}(y_1) - q_{Y_1}(y_1) &= f_{Y_1,T}(y_1, T=c) + f_{Y_1,T}(y_1, T=n),\\
f_{Y_0}(y_0) - p_{Y_0}(y_0) &= f_{Y_0,T}(y_0, T=c) + f_{Y_0,T}(y_0, T=a),\\
f_{Y_0}(y_0) - q_{Y_0}(y_0) &= f_{Y_0,T}(y_0, T=d) + f_{Y_0,T}(y_0, T=a).
\end{aligned} \tag{6}$$

Any population contained in $IR(P,Q|RA)$ must satisfy these equalities so that these consist the necessary condition for the population to belong to $IR(P,Q|RA)$. With the above equations in mind, we can claim that[5] a pair of marginal distributions (f_{Y_1}, f_{Y_0}) belongs to $IR_{(f_{Y_1},f_{Y_0})}(P,Q|RA)$ if and only if we can find four pairs of *nonnegative* functions $(h_{Y_1,t}(y_1), h_{Y_0,t}(y_0))$, $t = c,n,a,d$, that satisfy the *scale constraints*

$$\int h_{Y_1,t}(y_1)d\mu = \int h_{Y_0,t}(y_0)d\mu, \qquad t = c,n,a,d, \tag{7}$$

and the *compatibility constraints*

$$\begin{aligned}
p_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,a}(y_1),\\
q_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,a}(y_1),\\
p_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,n}(y_0),\\
q_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,n}(y_0),\\
f_{Y_1}(y_1) - p_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,n}(y_1),\\
f_{Y_1}(y_1) - q_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,n}(y_1),\\
f_{Y_0}(y_0) - p_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,a}(y_0),\\
f_{Y_0}(y_0) - q_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,a}(y_0).
\end{aligned} \tag{8}$$

In the comparison of (8) with (6), we can observe that each $h_{Y_j,t}$ in (8) corresponds to the unidentified population density $f_{Y_j,T}(y_j, T=t)$ in (6). This tells the rationale behind the above claim, that is, for a fixed (f_{Y_1}, f_{Y_0}), if we can find some nonnegative functions $(h_{Y_1,t}(y_1), h_{Y_0,t}(y_0))$ satisfying all the above constraints (7) and (8), we can impute $f_{Y_j,T}(y_j, T=t)$ by $h_{Y_j,t}$, and propose a compatible population as,[6] for $t = c,n,a,d$,

$$\begin{aligned}
f_{Y_1,Y_0,T}(y_1,y_0,T=t) &= f_{Y_1,Y_0,T|Z}(y_1,y_0,T=t|Z=1) = f_{Y_1,Y_0,T|Z}(y_1,y_0,T=t|Z=0)\\
&= \begin{cases} \left[\int_{\mathcal{Y}} h_{Y_1,t}(y_1)d\mu\right]^{-1} h_{Y_1,t}(y_1)h_{Y_0,t}(y_0) & \text{if } \int_{\mathcal{Y}} h_{Y_1,t}(y_1)d\mu > 0,\\ 0 & \text{if } \int_{\mathcal{Y}} h_{Y_1,t}(y_1)d\mu = 0. \end{cases}
\end{aligned}$$

[5] See Lemma A.1 in Appendix for a formal justification of this claim.

[6] There are many ways possible for combining the density of (Y_1, T) and (Y_0, T) to obtain the joint density of (Y_1, Y_0, T). The one employed here is one of them, so called the conditional independence coupling: The association of Y_1 and Y_0 satisfies $Y_1 \perp Y_0|T$.

The population pinned down in this way by construction satisfies RA, and also it is compatible with the data generating process since it clearly satisfies the constraints (6). Along this line of reasoning, $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA)$ is obtained by *characterizing the conditions for* (f_{Y_1}, f_{Y_0}) *under which we can find such feasible nonnegative functions* $(h_{Y_1,t}(y_1), h_{Y_0,t}(y_0))$, $t = c, n, a, d$.

The next proposition provides $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA)$.

Proposition 3.2 (Identification region under random assignment) *Let λ_{Y_1} be the inner integrated envelope of p_{Y_1} and q_{Y_1} defined by $\lambda_{Y_1} = \int_{\mathcal{Y}} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\} d\mu$.*
(i) The identification region of (f_{Y_1}, f_{Y_0}) under RA is

$$IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA) = \begin{cases} \mathcal{F}^*_{f_{Y_1}}(P, Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P, Q) & \text{if} \quad 1 - \delta_{Y_0} < \lambda_{Y_1} \\ \mathcal{F}^{env}_{f_{Y_1}}(P, Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P, Q) & \text{if} \quad 1 - \delta_{Y_0} = \lambda_{Y_1} \\ \mathcal{F}^{env}_{f_{Y_1}}(P, Q) \times \mathcal{F}^*_{f_{Y_0}}(P, Q) & \text{if} \quad 1 - \delta_{Y_0} > \lambda_{Y_1} \end{cases}$$

where $\mathcal{F}^*_{f_{Y_1}}(P, Q)$ and $\mathcal{F}^*_{f_{Y_0}}(P, Q)$ are proper subsets of $\mathcal{F}^{env}_{f_{Y_1}}(P, Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P, Q)$ respectively defined by

$$\begin{aligned} \mathcal{F}^*_{f_{Y_1}}(P, Q) &= \left\{ f_{Y_1} : f_{Y_1} \in \mathcal{F}^{env}_{f_{Y_1}}(P, Q), \int_{\mathcal{Y}} \min\left\{ f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}, q_{Y_1}\}\right\} d\mu \geq \lambda_{Y_1} + \delta_{Y_0} - 1 \right\}, \\ \mathcal{F}^*_{f_{Y_0}}(P, Q) &= \left\{ f_{Y_0} : f_{Y_0} \in \mathcal{F}^{env}_{f_{Y_0}}(P, Q), \int \min\left\{ f_{Y_0} - \underline{f_{Y_0}}, \min\{p_{Y_0}, q_{Y_0}\}\right\} d\mu \geq 1 - \delta_{Y_0} - \lambda_{Y_1} \right\}. \end{aligned}$$

(ii) $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA)$ is nonempty if and only if $\delta_{Y_1} \leq 1$ and $\delta_{Y_0} \leq 1$.

Proof. See Appendix A. ■

The above proposition clarifies that the identification region under RA can be strictly smaller than the identification region under MSI. In particular, this identification gain arises if the data reveals $1 - \delta_{Y_0} \neq \lambda_{Y_1}$ since $\mathcal{F}^*_{f_{Y_1}}(P, Q)$ and $\mathcal{F}^*_{f_{Y_0}}(P, Q)$ appeared above are strictly smaller than $\mathcal{F}^{env}_{f_{Y_1}}(P, Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P, Q)$ due to the inequality constraints appearing in their definitions.

A proof of this proposition provided in Appendix A proceeds by the method of "guess and verify," so the reader might think an intuition behind this result is rather obscure. Below, for the purpose of providing an intuition of this result, we provide a geometric illustration that clarifies where the additional identification gain of RA relative to MSI comes from.

We first consider the case of $1 - \delta_{Y_0} = \lambda_{Y_1}$ for which Proposition 3.2 says RA does not provide further identification gain than MSI. Figure 1 draws the data generating process and an arbitrary $(f_{Y_1}, f_{Y_0}) \in \mathcal{F}^{env}_{f_{Y_1}}(P, Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P, Q)$ for this case. There, we partition the subgraph of f_{Y_1} into four, $c(1)$, $a(1)$, $n(1)$, and $d(1)$, and similarly partition the subgraph of f_{Y_0} into $c(0)$, $a(0)$, $n(0)$, and $d(0)$. The condition $1 - \delta_{Y_0} = \lambda_{Y_1}$ means that the area of the partition outlined between f_{Y_0} and $\underline{f_{Y_0}}$ is equal to the area of the subgraph of $\min\{p_{Y_1}, q_{Y_1}\}$, i.e., the area of $a(1)$ is equal to the area of $a(0)$. Moreover, it can be shown that, $1 - \delta_{Y_0} = \lambda_{Y_1}$ implies not only $a(1)$ and $a(0)$ but also $c(1)$ and

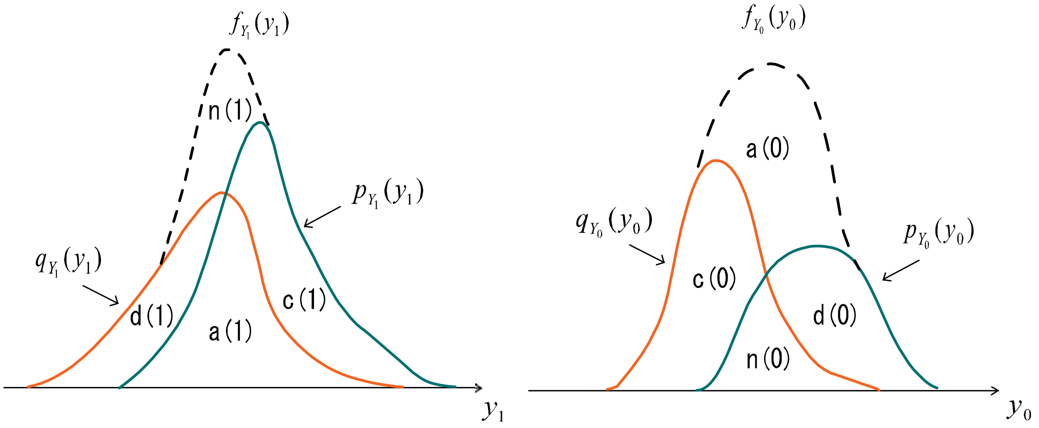

Figure 1: This figure depicts the data gerating process with $1-\delta_{Y_0}=\lambda_{Y_1}$ (the area of $a(0)$ is equal to the area of $a(1)$), which corresponds to the case (iii) in Proposition 1.3.1. For each $t=c,n,a,d$, $t(1)$ and $t(0)$ have the same area.

$c(0)$, $n(1)$ and $n(0)$, and $d(1)$ and $d(0)$ share the same area. This enables us to pin down $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$ to the height of the partitions $t(1)$ and $t(0)$ for each $t=c,n,a,d$, without violating the scale constraints (7). Moreover, this way of pinning down $(h_{Y_1,t}(y_1),h_{Y_0,t}(y_0))$ is compatible with all the constraints of (8) (see also Figure 2). Thus, we can successfully find the feasible nonnegative functions $(h_{Y_1,t},h_{Y_0,t})$ that allow us to construct a population that is compatible with RA and (P,Q). Hence, we conclude that the drawn (f_{Y_1},f_{Y_0}) belongs to $IR_{(f_{Y_1},f_{Y_0})}(P,Q|RA)$. Note that this way of imputing $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$ works for arbitrary $(f_{Y_1},f_{Y_0})\in\mathcal{F}^{env}_{f_{Y_1}}(P,Q)\times\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, so the identification region of (f_{Y_1},f_{Y_0}) under RA is obtained as the Cartesian product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$.

Next, let us consider the case of $1-\delta_{Y_0}<\lambda_{Y_1}$ as drawn in Figure 3, i.e., the area of $a(0)$ is smaller than the area of $a(1)$. The preceding way of pinning down $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$ to $t(1)$ and $t(0)$ will now violate the scale constraints, so we need to come up with a different way of imputing $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$. The following algorithm with graphical assistance of Figure 4 through Figure 7 illustrates a way of imputing $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$ in this case.

Algorithm to impute $(h_{Y_1,t},h_{Y_0,t})$, $t=c,n,a,d$.

Step 1: (Figure 4) Draw an arbitrary $f_{Y_1}\in\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $f_{Y_0}\in\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. We first set $h_{Y_0,a}$ to the height of the partition $a(0)$ and set $h_{Y_1,a}$ to the height of some subset within $\min\{p_{Y_1},q_{Y_1}\}$ such that its area is equal to the area of $a(0)$. Note that the equal area requirement is due to the scale constraint $\int h_{Y_1,a}d\mu=\int h_{Y_0,a}d\mu$. In the top figure, the subset imputed for $h_{Y_1,a}$ is labeled as a. As we pin down $h_{Y_0,a}$ and $h_{Y_1,a}$, we put their copies in the bottom figure for convenience of the later steps. How to choose subset a turns out to be a key for this algorithm and it will be further discussed in Step 4. For now, let us proceed to Step 2 with the drawn subset a.

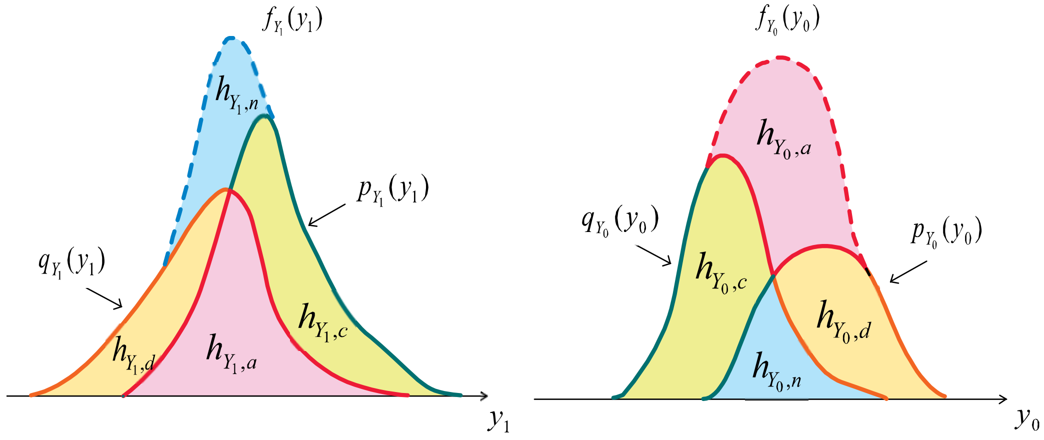

Figure 2: If the data generating process satisfies $1-\delta_{Y_0}=\lambda_{Y_1}$, we can set $h_{Y_1,t}$ and $h_{Y_0,t}$ to the partitions $t(1)$ and $t(0)$ of Figure 1 without contradicting the scale and compatibility constraints.

Step 2: (Figure 5) Impute $h_{Y_1,c}$ and $h_{Y_0,c}$ through the first and seventh constraints of (8). That is, we impute $h_{Y_1,c}$ to the height of subset $c(1)\cup(d\&c)$ and $h_{Y_0,c}$ to the height of subset $c(0)$ as drawn in the top figure. The equal area restriction $\int h_{Y_1,c}d\mu=\int h_{Y_0,c}d\mu$ is automatically satisfied.

Step 3: (Figure 6) Impute $h_{Y_1,d}$ and $h_{Y_0,d}$ via the second and eighth constraints of (8). That is, we impute $h_{Y_1,d}$ to the height of subset $d(1)\cup(d\&c)$ and $h_{Y_0,d}$ to the height of subset $d(0)$ as drawn in the top figure. Note that the equal area restriction $\int h_{Y_1,d}d\mu=\int h_{Y_0,d}d\mu$ is again automatically satisfied. In the bottom figure, the imputed $h_{Y_1,d}$ is piled up on the top of $h_{Y_1,a}$ and $h_{Y_1,c}$.

Step 4: (Figure 7) Since the densities of the other three types have been already imputed, the last piece of the puzzle, $h_{Y_1,n}$ and $h_{Y_0,n}$ must be set at the parts of f_{Y_1} and f_{Y_0} that were left out from the other imputed densities. The imputed $h_{Y_1,n}$ and $h_{Y_0,n}$ are drawn as the shadow areas in the top figure. Algebraically, the imputed $h_{Y_1,n}$ and $h_{Y_0,n}$ are expressed as

$$\begin{aligned} h_{Y_1,n} &= f_{Y_1}-\sum_{t=a,c,n} h_{Y_1,t}=f_{Y_1}-\underline{f_{Y_1}}-[\min\{p_{Y_1},q_{Y_1}\}-h_{Y_1,a}], \\ h_{Y_0,n} &= \min\{p_{Y_0},q_{Y_0}\}. \end{aligned}$$

Since $h_{Y_1,n}$ must be nonnegative, $h_{Y_1,n}\geq 0$ yields the inequality constraint for the possible choices of $h_{Y_1,a}$ (given the proposed f_{Y_1}) that has not been considered in Step 1,

$$h_{Y_1,a}\geq \max\left\{\underline{f_{Y_1}}+\min\{p_{Y_1},q_{Y_1}\}-f_{Y_1},0\right\}. \tag{9}$$

where the maximum operator is needed in the right hand side since $h_{Y_1,a}$ must be nonnegative.

Step 5: As seen in Step 1, the integration of $h_{Y_1,a}$ has been constrained to being equal to $\int h_{Y_0,a}d\mu=$

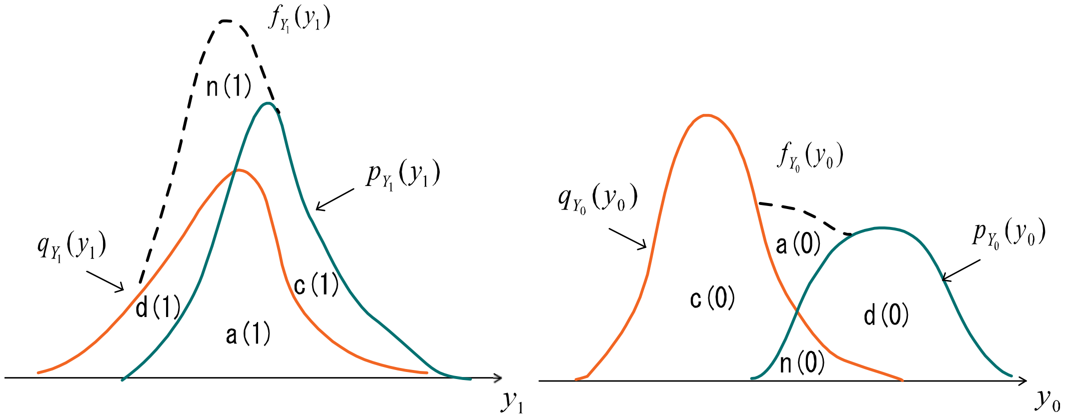

Figure 3: The drawn data generating process satisfies $1 - \delta_{Y_0} < \lambda_{Y_1}$ (the area of $a(0)$ is strictly smaller than the area of $a(1)$). Different from the case drawn in Figure 1, it is not feasible to pin down $(h_{Y_1,t}, h_{Y_0,t})$ to $(t(1), t(0))$ for each $t = c, n, a, d$, because the scale constraints are violated.

$1 - \delta_{Y_0}$. So, the integration of (9) gives

$$1 - \delta_{Y_0} \geq \int \max\left\{\underline{f_{Y_1}} + \min\left\{p_{Y_1}, q_{Y_1}\right\} - f_{Y_1}, 0\right\} d\mu,$$

and this can be rewritten as

$$\begin{aligned} & 1 - \delta_{Y_0} \geq -\int \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\left\{p_{Y_1}, q_{Y_1}\right\}\right\} d\mu + \lambda_{Y_1} \\ \Longleftrightarrow \quad & \int \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\left\{p_{Y_1}, q_{Y_1}\right\}\right\} d\mu \geq \underbrace{\lambda_{Y_1} - [1 - \delta_{Y_0}]}_{\text{the area of } d\&c}. \end{aligned} \tag{10}$$

This inequality is exactly the one appearing in the definition of $\mathcal{F}^*_{f_{Y_1}}(P, Q)$. If f_{Y_1} proposed in Step 1 meets this inequality, it implies that there exists a choice of $h_{Y_1,a} \geq 0$ based on which Step 2 through Step 4 guarantee the existence of feasible $(h_{Y_1,t}, h_{Y_0,t})$, $t = c, n, d$.

By the implication obtained in Step 5 of the above algorithm, we can claim that the $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA) \supset \mathcal{F}^*_{f_{Y_1}}(P, Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P, Q)$. In fact, it is also possible to show $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA) \subset \mathcal{F}^*_{f_{Y_1}}(P, Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P, Q)$ (See the proof of Proposition 3.2 in Appendix A). An interpretation of inequality (10) is that, any f_{Y_1} contained in $\mathcal{F}^*_{f_{Y_1}}(P, Q)$ must spare enough room on the top of the envelope density $\underline{f_{Y_1}}$ so that the region between f_{Y_1} and $\underline{f_{Y_1}}$ can contain the region of $d\&c'$ shown in the top figure of Figure 7, which is the exact copy of $d\&c$. It provides an intuition of why RA yields the identification gain compared with MSI. Suppose that the overlapping area of p_{Y_1} and q_{Y_1} is large while the overlapping area of p_{Y_0} and q_{Y_0} is less, implying that the area of $n(1)$ is relatively larger than the area of

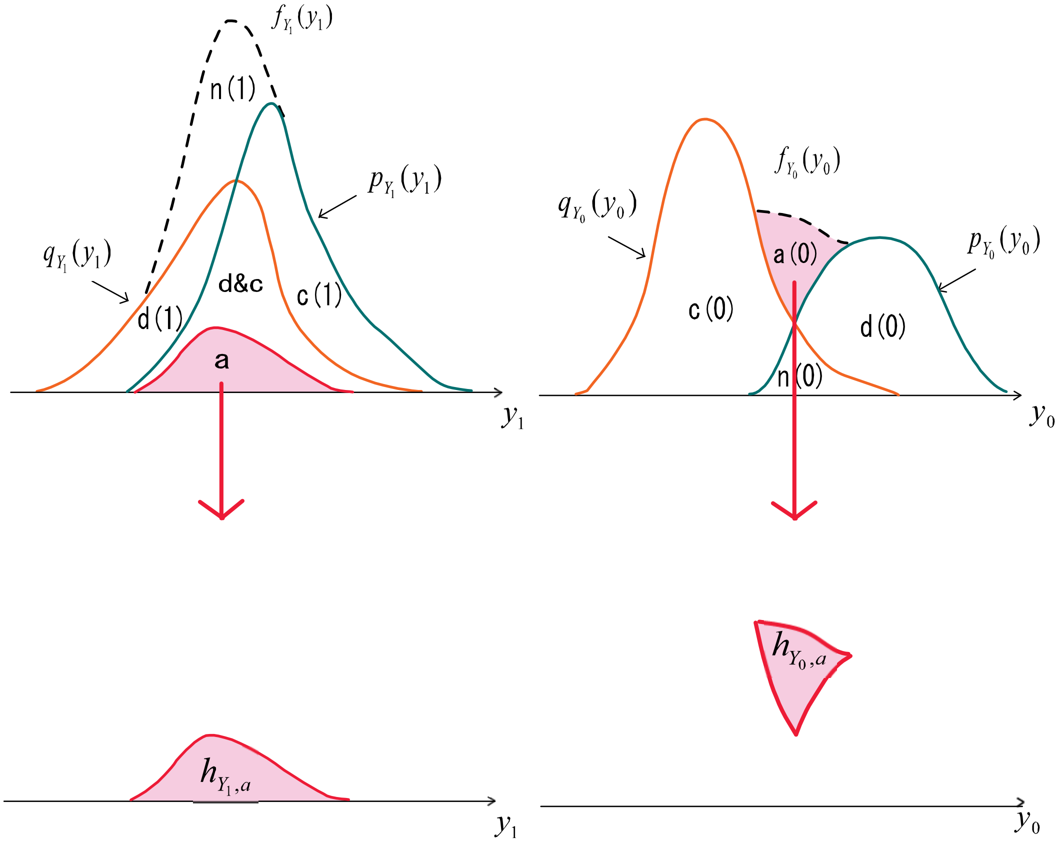

Figure 4: Step 1: Imputation of $h_{Y_1,a}$ and $h_{Y_0,a}$.

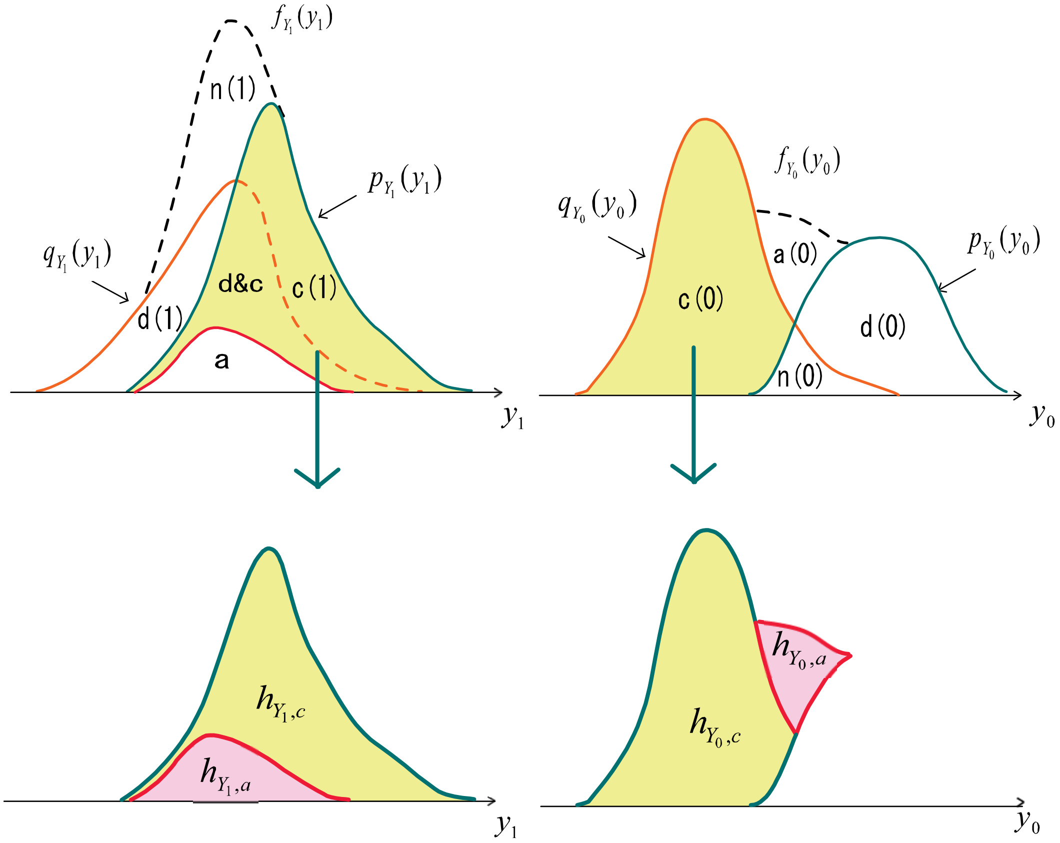

Figure 5: Step 2: Imputation of $h_{Y_1,c}$ and $h_{Y_0,c}$.

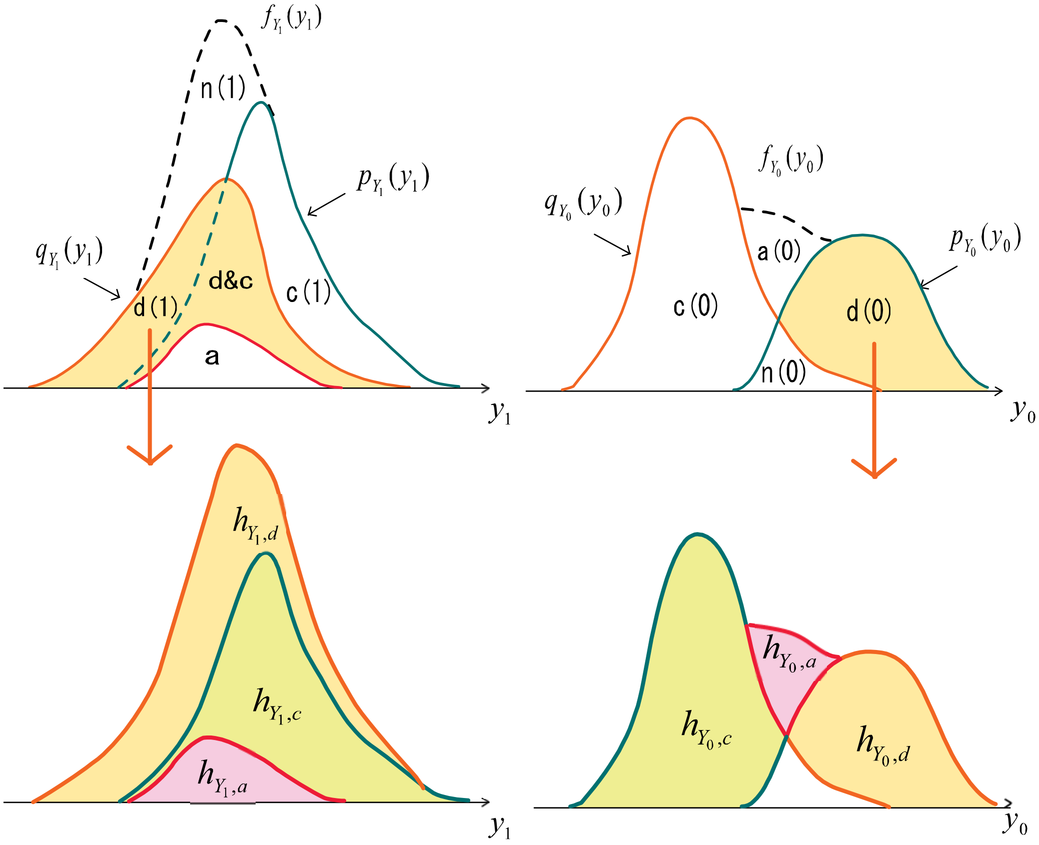

Figure 6: Step 3: Imputation of $h_{Y_1,d}$ and $h_{Y_0,d}$.

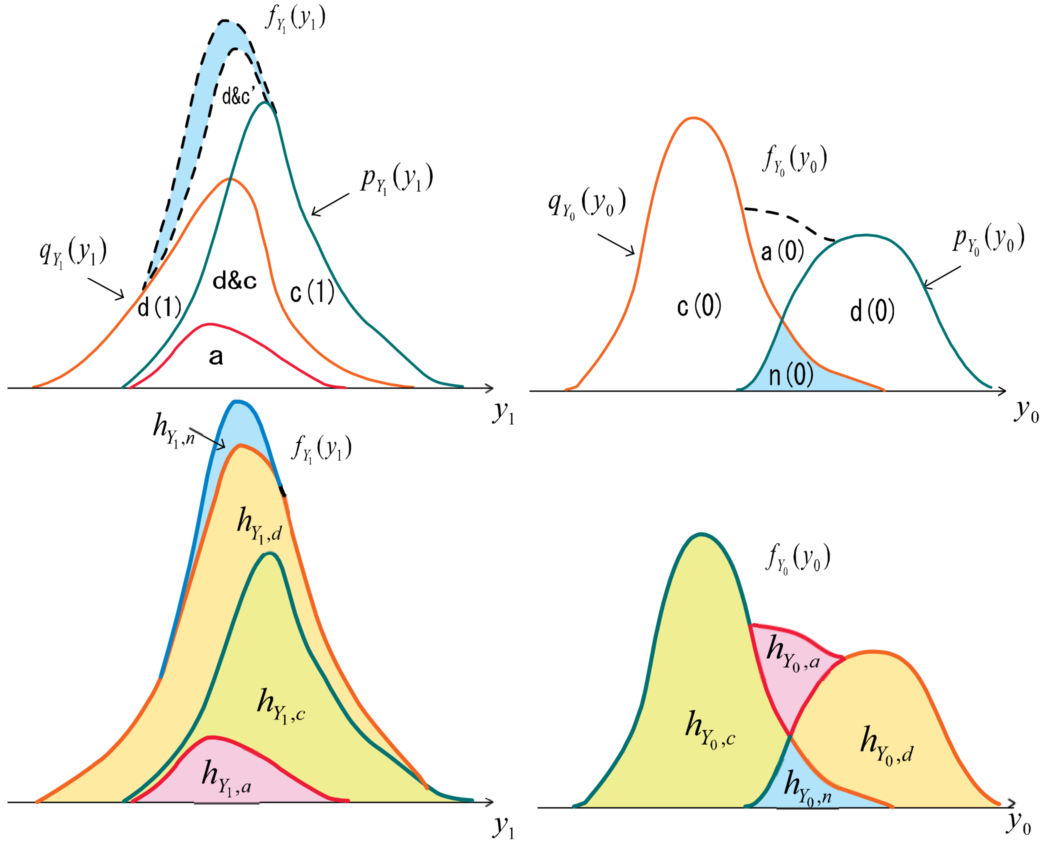

Figure 7: Step 4: The last piece of puzzle. imputation of $h_{Y_1,n}$ and $h_{Y_0,n}$.

$n(0)$. Then, under RA, we are able to learn that there exists only a small fraction of never-takers because the fraction of never-takers is at most the area of $n(0)$. This in turn implies that the entire part of $n(1)$ cannot be imputed as the never-taker's outcome density. If the identification region for f_{Y_1} under RA was $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$, then, it must be the case that the entire $n(1)$ can be imputed by the never-taker's density $h_{Y_1,n}$ because $h_{Y_1,n}$ is the only density whose shape is completely unrestricted. But, we cannot do so since the fraction of the never takers learned from the area of $n(0)$ is not big enough to fill the entire $n(1)$. Therefore the identification region for f_{Y_1} becomes strictly smaller than $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$. Inequality (10) clarifies the channel through which identifying information for f_{Y_0} contributes to identifying f_{Y_1}.

The symmetric argument works for the case of $1-\delta_{Y_0} > \lambda_{Y_1}$. By noting that $1-\delta_{Y_0} > \lambda_{Y_1}$ is equivalent to $1-\delta_{Y_1} < \lambda_{Y_0}$ (see Lemma A.2 in Appendix A), the symmetric analysis can be implemented to construct the identification region. In this case, the identification region for f_{Y_0} becomes smaller than $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, implying that the identifying information for f_{Y_1} contributes to identifying f_{Y_0}

3.3 Identification Region under the LATE restriction

Proposition 3.2 clarifies that if the observed data meets $1-\delta_{Y_0} = \lambda_{Y_1}$, then the difference between MSI and RA does not matter for identifying f_{Y_1} and f_{Y_0}. One situation where this condition is satisfied is the case of *nested densities*:

$$\begin{gathered} p_{Y_1}(y_1) \geq q_{Y_1}(y_1)\ \mu\text{-a.e. and } q_{Y_0}(y_0) \geq p_{Y_0}(y_0)\ \mu\text{-a.e., or} \\ p_{Y_1}(y_1) \leq q_{Y_1}(y_1)\ \mu\text{-a.e. and } q_{Y_0}(y_0) \leq p_{Y_0}(y_0)\ \mu\text{-a.e.} \end{gathered} \tag{11}$$

In this section, we shall show that the configuration of the nested densities is a key for constructing the identification region under the LATE restriction.

The LATE restriction further constrains the population by deleting one of the selection types. Specifically, in case of $\Pr(D=1|Z=1) \geq \Pr(D=1|Z=0)$, it implies the no-defier condition $f_T(T=d)=0$. Since the analysis of no-compliers case and the no-defiers case is symmetric, we without loss of generality consider the case of $\Pr(D=1|Z=1) \geq \Pr(D=1|Z=0)$.

Under the LATE restriction (RA and the no-defier condition), the equations in the previous section (6) are simplified to

$$\begin{aligned} &p_{Y_1}(y_1) = f_{Y_1,T}(y_1,T=c) + f_{Y_1,T}(y_1,T=a), \\ &q_{Y_1}(y_1) = f_{Y_1,T}(y_1,T=a), \\ &p_{Y_0}(y_0) = f_{Y_0,T}(y_0,T=n), \\ &q_{Y_0}(y_0) = f_{Y_0,T}(y_0,T=c) + f_{Y_0,T}(y_0,T=n), \\ &f_{Y_1}(y_1) - p_{Y_1}(y_1) = f_{Y_1,T}(y_1,T=n), \\ &f_{Y_1}(y_1) - q_{Y_1}(y_1) = f_{Y_1,T}(y_1,T=c) + f_{Y_1,T}(y_1,T=n), \\ &f_{Y_0}(y_0) - p_{Y_0}(y_0) = f_{Y_0,T}(y_0,T=c) + f_{Y_0,T}(y_0,T=a), \\ &f_{Y_0}(y_0) - q_{Y_0}(y_0) = f_{Y_0,T}(y_0,T=a). \end{aligned}$$

The first four of the above constraints imply that when the population satisfies the LATE restriction, the data generating process must reveal the nested densities since $p_{Y_1}(y_1) - q_{Y_1}(y_1) = f_{Y_1,T}(y_1, T =$

$c) \geq 0$ and $q_{Y_0}(y_0) - p_{Y_0}(y_0) = f_{Y_0,T}(y_0, T = c) \geq 0$. This is equivalent to saying that observing the non-nested densities must yield the empty identification region under LATE. On the other hand, when data reveals the nested densities, then for every $(f_{Y_1}, f_{Y_0}) \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, we can uniquely solve the above constraints to obtain the nonnegative densities of (Y_1, T) and (Y_0, T), and they can be combined to obtained the distribution of (Y_1, Y_0, T) independent of Z. Accordingly, the next proposition follows.

Proposition 3.3 (Identification region under the LATE restriction) *The identification region of (f_{Y_1}, f_{Y_0}) under the LATE restriction is*

$$IR_{(f_{Y_1}, f_{Y_0})}(P,Q|LATE) = \begin{cases} \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q) & \text{for nested densities (11),} \\ \emptyset & \text{otherwise.} \end{cases}$$

Proof. A proof is given in the preceding paragraphs of this section. ■

This proposition states that if the data generating process reveals the nested densities, the identification region under LATE coincides with the identification region under MSI. Moreover, the fact that the nested densities satisfy $1 - \delta_{Y_0} = \lambda_{Y_1}$ implies that the identification region under LATE also coincides with the identification region under RA (Proposition 3.2 (i)). If the nested densities are not observed, then LATE restriction is refuted while the identification region under RA or MSI can yield the nonempty identification region. Put another way, as far as the population distributions of the potential outcomes are concerned, adding instrument monotonicity, or equivalently threshold crossing selection with an additive error, to the instrument independence restriction *only constrains the data generating process without helping us learn about (f_{Y_1}, f_{Y_0}) further than MSI or RA.* In this sense, we can safely drop the instrument monotonicity restriction from the analysis if the goal of analysis is to acquire the maximal identifying information for the potential outcome distributions. Note that the refutability result of the LATE restriction is not new in the literature. Heckman and Vytlacil (2005, Theorem 1 in Appendix A) demonstrates a testable implication for the LATE restriction, which is equivalent to the nested density condition given here.[7]

4 Bounding Causal Parameters

By appropriately defining the outcome support and its dominating measure μ, the identification regions obtained in the previous section can be applied to the wide range of settings including discrete, unbounded, and even multi-dimensional outcomes. Moreover, for a parameter (vector) θ that maps (f_{Y_1}, f_{Y_0}) to Θ, we can make a comparison of the size of the sharp bounds of θ without explicitly computing the bounds.

[7] The emptiness result of Proposition 3.3 implies that the configuration of the nested densities is interpreted as a necesarry testable implication for the LATE restriction. Since the LATE assumption plays a crucial role in validating the Wald type instrumental variable estimator to be consistent to the local average treatment effects, we can use this testable implication to develop a necessary specification test for the instrument validity in the context of the LATE estimation. See Kitagawa and Hoderlein (2009) for a test procedure for the nested configuration of the densities.

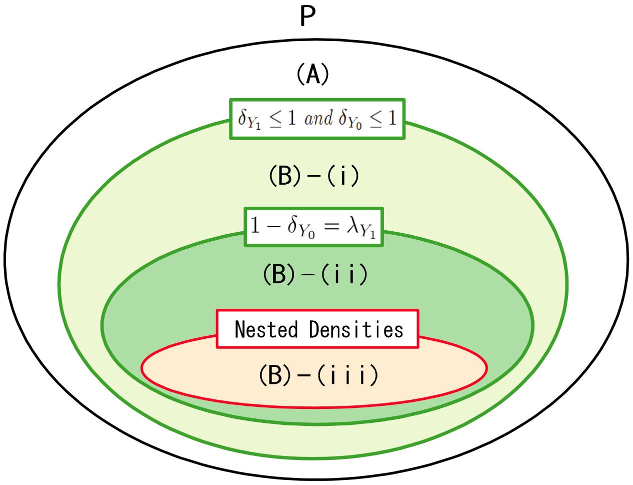
P
(A)
$\delta_{Y_1} \leq 1$ and $\delta_{Y_0} \leq 1$
(B) – (i)
$1 - \delta_{Y_0} = \lambda_{Y_1}$
(B) – (ii)
Nested Densities
(B) – (iii)

Figure 8: The classification of the data generating processes in Theorem 1.

Theorem 1 *Let θ be a parameter (vector) that maps (f_{Y_1}, f_{Y_0}) to Θ. Then, for each layer of the data generating process (see Figure 8), the sharp bounds of θ under MSI, RA, and LATE meet the following relationships.*
(A) If $\delta_{Y_1} > 1$ or $\delta_{Y_0} > 1$, then

$$IR_\theta(P, Q|\cdot) = \emptyset \text{ for all of MSI, RA, and LATE.}$$

(B) If $\delta_{Y_1} \leq 1$ and $\delta_{Y_0} \leq 1$, and
(i) if $1 - \delta_{Y_0} \neq \lambda_{Y_1}$, then,

$$IR_\theta(P, Q|MSI) \supset IR_\theta(P, Q|RA) \neq \emptyset, \qquad IR_\theta(P, Q|LATE) = \emptyset.$$

(ii) if $1 - \delta_{Y_0} = \lambda_{Y_1}$ and the data generating process does not reveal the nested densities, then

$$IR_\theta(P, Q|MSI) = IR_\theta(P, Q|RA) \neq \emptyset, \qquad IR_\theta(P, Q|LATE) = \emptyset.$$

(iii) if the data generating process reveals the nested densities,

$$IR_\theta(P, Q|MSI) = IR_\theta(P, Q|RA) = IR_\theta(P, Q|LATE) \neq \emptyset.$$

Proof. By the definition of $IR_\theta(P, Q|\cdot)$ given in (2) these results are implied by Proposition 3.1, 3.2, and 3.3. ■

Provided that the outcome is a scalar with compact support $\mathcal{Y} = [y_l, y_u]$, this theorem clearly applies to the sharp bounds of the average treatment effects (ATE) $\theta = E(Y_1) - E(Y_0)$. Below, we shall present the formula for the sharp ATE bounds under each restriction.

In order to simplify the expression of the sharp ATE bounds, we introduce the α-th left- or right-trimming of a nonnegative integrable function $g : \mathcal{Y} \to \mathbb{R}$. For $\alpha < \int_{\mathcal{Y}} g d\mu$, let $q_\alpha^{left} \equiv \inf \left\{ t : \int_{(-\infty,t]} g d\mu \geq \alpha \right\}$, and define the *$\alpha$-th left-trimming* of g by

$$[g]_\alpha^{ltrim}(y) \equiv g(y) 1\left\{ y > q_\alpha^{left} \right\} + \left(\int_{(-\infty, q_\alpha^{left}]} g(y) d\mu - \alpha \right) 1 \left\{ y = q_\alpha^{left} \right\}.$$

Similarly, with $q_\alpha^{right} \equiv \sup \left\{ t : \int_{[t,\infty)} g d\mu \geq \alpha \right\}$, we define the *$\alpha$-th right-trimming* of g by

$$[g]_\alpha^{rtrim}(y) \equiv g(y) 1\left\{ y < q_\alpha^{right} \right\} + \left(\int_{[q_\alpha^{right}, \infty)} g(y) d\mu - \alpha \right) 1 \left\{ y = q_\alpha^{right} \right\}.$$

In words, the α-th (right-) left-trimming is obtained by trimming the (right-) left-tail part of the function g with the trimmed masses equal to α. Note that if the underlying measure has point masses the second terms in the right-hand side of the above definitions can be nonzero, and these adjustment terms are needed to make the trimmed area exactly equal to α.

Proposition 4.1 (The sharp ATE bounds) *Assume Y_1 and Y_0 have the compact support $\mathcal{Y} = [y_l, y_u]$ and their distributions are absolutely continuous with respect to the measure μ that allows the point masses at y_l and y_u. We also assume that the data generating process meets $\delta_{Y_1} \leq 1$ and $\delta_{Y_0} \leq 1$ so as to exclude Case (A) of Theorem 1.*
(i) The sharp ATE bounds under MSI are

$$\begin{aligned} IR_{ATE}(P,Q|MSI) \quad = \quad & \left[(1-\delta_{Y_1}) y_l + \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}} d\mu - (1-\delta_{Y_0}) y_u, \right. \\ & \left. \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu + (1-\delta_{Y_1}) y_u - (1-\delta_{Y_0}) y_l - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}} d\mu \right]. \end{aligned} \tag{12}$$

(ii) The sharp ATE bounds under RA are, for $1 - \delta_{Y_0} = \lambda_{Y_1}$,

$$IR_{ATE}(P,Q|RA) = IR_{ATE}(P,Q|MSI),$$

for $1 - \delta_{Y_0} < \lambda_{Y_1}$,

$$\begin{aligned} & IR_{ATE}(P,Q|RA) \\ = \quad & \left[\int_{\mathcal{Y}} y_1 \left(\underline{f_{Y_1}} + [\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{rtrim} \right) d\mu + \lambda_{Y_0} y_l - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}} d\mu - (1-\delta_{Y_0}) y_u, \right. \\ & \left. \int_{\mathcal{Y}} y_1 \left(\underline{f_{Y_1}} + [\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{ltrim} \right) d\mu + \lambda_{Y_0} y_u - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0) d\mu - (1-\delta_{Y_0}) y_l \right], \end{aligned} \tag{13}$$

and, for $1 - \delta_{Y_0} > \lambda_{Y_1}$,

$$\begin{aligned} & IR_{ATE}(P,Q|RA) \\ = \quad & \left[\int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu + (1-\delta_{Y_1}) y_l - \int_{\mathcal{Y}} y_0 \left(\underline{f_{Y_0}} + [\min\{p_{Y_0}, q_{Y_0}\}]_{1-\delta_{Y_1}}^{ltrim} \right) d\mu - \lambda_{Y_1} y_u, \right. \\ & \left. \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu + (1-\delta_{Y_1}) y_u - \int_{\mathcal{Y}} y_0 \left(\underline{f_{Y_0}} + [\min\{p_{Y_0}, q_{Y_0}\}]_{1-\delta_{Y_1}}^{rtrim} \right) d\mu - \lambda_{Y_1} y_l \right], \end{aligned} \tag{14}$$

(iii) The sharp ATE bounds under LATE are

$$IR_{ATE}(P,Q|LATE)$$
$$= \begin{cases} \left[\begin{array}{c} \max_z\{E(Y|D=1,Z=z)\Pr(D=1|Z=z)+y_l\Pr(D=0|Z=z)\} \\ -\min_z\{E(Y|D=0,Z=z)\Pr(D=0|Z=z)+y_u\Pr(D=1|Z=z)\}, \end{array} \right. \\ \left. \begin{array}{c} \min_z\{E(Y|D=1,Z=z)\Pr(D=1|Z=z)+y_u\Pr(D=0|Z=z)\} \\ -\max_z\{E(Y|D=0,Z=z)\Pr(D=0|Z=z)+y_l\Pr(D=1|Z=z)\} \end{array} \right] \\ \hfill \textit{for nested densities,} \\ \emptyset \hfill \textit{otherwise.} \end{cases}$$

Proof. See Appendix A. ■

When the data generating process reveals $1-\delta_{Y_0} \neq \lambda_{Y_1}$, the ATE bounds under RA is strictly narrower than the bounds under MSI. For instance, in case of $1-\delta_{Y_0} < \lambda_{Y_1}$, the comparison of the lower bounds of (13) and (12) shows that the former is larger than the latter by

$$[\lambda_{Y_1}-(1-\delta_{Y_0})] \times \int_{y_1}(y_1-y_l)\frac{[\min\{p_{Y_1},q_{Y_1}\}]^{rtrim}_{1-\delta_{Y_0}}}{[\lambda_{Y_1}-(1-\delta_{Y_0})]}d\mu.$$

By noting $[\min\{p_{Y_1},q_{Y_1}\}]^{rtrim}_{1-\delta_{Y_0}}/[\lambda_{Y_1}-(1-\delta_{Y_0})]$ to be a probability measure, this expression implies that the identification gain for ATE becomes more as $[\lambda_{Y_1}-(1-\delta_{Y_0})]$ gets bigger and/or $[\min\{p_{Y_1},q_{Y_1}\}]^{rtrim}_{1-\delta_{Y_0}}$ becomes more spread than the degenerate function at the lower bound y_l.

On the other hand, when (P,Q) reveals the nesting configuration, the ATE bounds are given by (12) irrespective of the imposed restrictions as claimed in Theorem 1. Moreover, it can be shown that the configuration of the nested densities reduces the bound formula (12) to the ATE bounds of Manski (1994) under the mean independence restriction, $E(Y_1|Z)=E(Y_1)$ and $E(Y_0|Z)=E(Y_0)$. This observation supports the result of Heckman and Vytlacil (2001a, 2001b, 2007), which says that the sharp ATE bounds under the LATE restriction coincides with Manski's mean independence bounds. Validity of this statement, however, relies on the situation where the data reveals the nested densities. When the data does not, then, the LATE restriction is misspecified and a naive implementation of the formula of the Manski's mean independence bounds no longer yields the tightest possible bounds. Furthermore, the formula of Manski's mean independence bounds do not necessarily become empty even though $IR_{(f_{Y_1},f_{Y_0})}(P,Q|LATE)$ is empty. These phenomena raise some concern about the misspecification problem of the bound formula justified under the observationally restrictive assumptions, and also highlight the advantage of constructing the sharp bounds with being explicit about its definition given in Section 2.

In the special case where the outcome variables are binary, the sharp ATE bounds under RA presented above coincide with the treatment effect bounds of Balke and Pearl (1997) (see Appendix B for details). Since the analysis of Balke and Pearl (1997) relies on a linear optimization procedure with the finite number of choice variables, their approach cannot be straightforwardly applied to the case in which the outcome variables have continuous variation. Thus the bound formula obtained here can be seen as a nontrivial generalization of the Balke and Pearl's bounds to the continuous outcome case.

5 Discussion: The Source of Identification Gain in the Structural Equation Model

In this section, we shall discuss the link between the counterfactual causal model analyzed in this paper and the nonseparable structural equation model with a binary endogenous variable. Specifically, we compare our identification regions with some identification results known in the literature of the nonseparable structural equations.

Let $Y = \phi(D, U)$ be the structural outcome equation where U represents the unobserved heterogeneity that affects one's outcome response. The structural outcome equation and the counterfactual outcomes are linked by $Y_1 = \phi(1, U)$ and $Y_0 = \phi(1, U)$ (see, e.g., Athey and Imbens (2006), Chernozhukov and Hansen (2005), and Pearl (2000)). Therefore, the potential outcome distributions f_{Y_1} and f_{Y_0} correspond to the marginal distributions of the transformed random variables $\phi(1, U)$ and $\phi(0, U)$ where U's distribution is the unconditional one.

Structure is absent in our identification analysis. That is, validity of our results does not rely on any type of assumptions including the dimension of U, distribution of U, and the functional form specification of $\phi(j, U)$, $j = 1, 0$. In this sense, the identification results of this paper provide a benchmark compared with which we are able to analyze what type of restrictions on the structure in addition to instrument independence plays a crucial role for identifying the causal effects. One insightful comparison we shall make for this purpose is with the restriction of *outcome monotonicity in unobservable.*

Monotonicity in unobservable assumes that $\phi(1, U)$ and $\phi(0, U)$ are increasing with respect to a *scalar* unobservable term U following uniform distribution on the unit interval. In other words, we interpret $\phi(1, \tau)$ and $\phi(0, \tau)$ as the τ-th quantile of the distributions of Y_1 and Y_0.[8] When the outcome is binary, Chesher (2009) obtains the bounds of the average treatment effects that can be substantially narrower than the one presented in this paper (see Hahn (2009) for the comparison between these bounds).[9] Moreover, in the continuous outcome case, Chernozhukov and Hansen (2005) shows that rank invariance and independence of Z and U can point-identify $\phi(1, \cdot)$ and $\phi(0, \cdot)$, implying point-identification of the potential outcome distributions. This transition from the set-identification result of Proposition 3.2 of this paper to the point-identification result of Chernozhukov and Hansen (2005) highlights strong identification power of outcome monotonicity in unobservable (or equivalently rank invariance between Y_1 and Y_0). This in turn implies that identification in this

[8] This restriction has been employed in Chesher (2003, 2005) and it is also refered to as the rank invariance restriction considered in Chernozhukov and Hansen (2005), i.e., the individual ranked at τ in terms of the value of Y_1 and the one ranked at τ in terms of the value of Y_0 share the same unobservable characteristics U.

[9] If the outcome is binary, monotonicity in unobservable implies *monotonic outcome response to treatment*, which means that $\phi(j, U)$ is weaky monotonic with respect to j. That is, $\phi(1, U) \geq \phi(0, U)$ for every U or $\phi(1, U) \leq \phi(0, U)$ for every U, and, in terms of the potential outcome notation, it is equivalently stated as $Y(1) \geq Y(0)$ with probability one or $Y(1) \leq Y(0)$ with probability one. In a more general setting, Heckman and Vytlacil (2007) and Bhattacharya, Shaikh, and Vytlacil (2008) demonstrate that imposing monotonic outcome response to treatment in addition to the LATE restriction can further narrow the ATE bounds. Although it is not as powerful as point-identifying the causal effects, it indicates that the monotonic outcome response to instrument contributes to identifying the potential outcome distributions. Note that they assume the specification of threshold crossing with an additive error for the selection equation so that their analysis is restricted to the case where the data generating process reveals the nested densities.

case largely relies on the assumptions on the association between the individual potential outcomes, and this point should be acknowledged if the researcher imposes it without a convincing economic theory or sound background knowledge for it.

6 Concluding Remarks

From the perspective of partial identification, this paper clarifies identification power of the instrument independence assumptions in the heterogeneous treatment effect model. We derive the identification region of the marginal distributions of the potential outcomes under each restriction, and compare the size of the identification region among them. We clarify for which data generating process the identification region can be further tightened or not. We show that for some data generating processes the instrument joint independence restriction can provide further identification gain than instrument marginal independence. Another important finding is that adding the instrument monotonicity restriction to instrument independence restriction is redundant for identifying the potential outcome distributions because it only constrains the data generating process without further identifying the potential outcome distributions. We also present the sharp bounds for the average treatment effects under each restriction. Our analysis covers binary, discrete, and continuous outcome support, and our bounds under joint independence extend the bounds of Balke and Pearl (1997) for the binary outcome case to a more general setting including the continuous outcome case.

Our identification framework exclusively focuses on the causal effects defined in terms of the population distribution of the potential outcomes, and our analysis does not impose any assumptions that constrain the association of the potential outcomes. This would be a reasonable approach if the researcher has little knowledge on the association of the potential outcomes, but a disadvantage is that we may have to give up drawing an informative conclusion out of data. If one can justify the association of Y_1 and Y_0 based on an economic theory or some causal knowledge, then it is possible to increase informativeness of the conclusion. For example, in case of the continuous outcome, adding the rank invariance assumption of Chernozhukov and Hansen (2005), i.e., individual's rank of the outcome does not vary with treatment status, gives point-identification of the potential outcome distributions (Chernozhukov and Hansen (2005)). The comparison of their result with our identification result highlights strong identification power of the rank invariance assumption, which is worth attention if the researcher imposes it without a convincing economic theory or sound background knowledge about it.

Appendix A: Proofs

The proofs of constructing $IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$ proceed in the manner of "guess and verify." We first propose a guess for $IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$, say, $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$. In order to verify that the guess $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$ is correct, we need to show the two things. First, for an arbitrary $(f_{Y_1},f_{Y_0})\in IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$, we shall show that there exists a distribution of (Y_1,Y_0,T,Z) that is compatible with (P,Q) and the imposed restrictions. This first step proves $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)\subset$

$IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$. Next, in order to prove $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot) \supset IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$, it suffices to show that every $(f_{Y_1}, f_{Y_0}) \in IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$ meet a necessary condition for $(f_{Y_1}, f_{Y_0}) \in IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$ (e.g., Proof for Proposition 3.1 and 3.3). Alternatively, we may demonstrate that any $(f_{Y_1}, f_{Y_0}) \notin IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$ bring up contradiction to some of the imposed restrictions (e.g., Proof of Proposition 3.2.). In either way, we can conclude $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot) \supset IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$. By combining them, we conclude that the guess is correct, $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot) = IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$.

Proof of Proposition 3.1. Fix $(P,Q) \in \mathcal{P}$, and guess the identification region under MSI to be $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|MSI) = \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. Clearly, $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|MSI)$ is nonempty if and only if $\delta_{Y_1} \leq 1$ and $\delta_{Y_1} \leq 1$, since otherwise no probability densities can cover the entire density envelopes. Let us pick an arbitrary $(f_{Y_1}, f_{Y_0}) \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. Consider the distribution of (Y_1, Y_0, T) given Z as follows.

$$\begin{aligned}
f_{Y_1,Y_0,T|Z}(y_1,y_0,T=a|Z=1) &= \frac{1}{\Pr(D=1|Z=1)} p_{Y_1}(y_1)[f_{Y_0}(y_0) - p_{Y_0}(y_0)],\\
f_{Y_1,Y_0,T|Z}(y_1,y_0,T=a|Z=0) &= \frac{1}{\Pr(D=1|Z=0)} q_{Y_1}(y_1)[f_{Y_0}(y_0) - q_{Y_0}(y_0)],\\
f_{Y_1,Y_0,T|Z}(y_1,y_0,T=n|Z=1) &= \frac{1}{\Pr(D=0|Z=1)} [f_{Y_1}(y_1) - p_{Y_1}(y_1)] p_{Y_0}(y_0),\\
f_{Y_1,Y_0,T|Z}(y_1,y_0,T=n|Z=0) &= \frac{1}{\Pr(D=0|Z=0)} [f_{Y_1}(y_1) - q_{Y_1}(y_1)] q_{Y_0}(y_0),\\
f_{Y_1,Y_0,T|Z}(y_1,y_0,T=c|Z=z) &= 0 \quad \text{for } z=1,0,\\
f_{Y_1,Y_0,T|Z}(y_1,y_0,T=d|Z=z) &= 0 \quad \text{for } z=1,0.
\end{aligned}$$

By noting $\int p_{Y_1} d\mu = \Pr(D=1|Z=1)$, $\int p_{Y_0} d\mu = \Pr(D=0|Z=1)$, $\int q_{Y_1} d\mu = \Pr(D=1|Z=0)$, and $\int q_{Y_0} d\mu = \Pr(D=0|Z=0)$, we can see that the constructed population meets the constraints (3). Furthermore, by plugging the constructed population densities into the identities, $f_{Y_1|Z} = \sum_{t\in\{c,n,a,d\}} \int_{y_0} f_{Y_1,Y_0,T|Z} d\mu$ and $f_{Y_0|Z} = \sum_{t\in\{c,n,a,d\}} \int_{y_1} f_{Y_1,Y_0,T|Z} d\mu$, we claim that the constructed population meets MSI. Therefore, $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot) \subset IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$. The other direction is straightforward since if $(f_{Y_1}, f_{Y_0}) \in IR_{(f_{Y_1},f_{Y_0})}(P,Q|MSI)$ $f_{Y_1} \geq \underline{f_{Y_1}}$ and $f_{Y_0} \geq \underline{f_{Y_0}}$ must hold as discussed in the main text. Hence, $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot) \supset IR_{(f_{Y_1},f_{Y_0})}(P,Q|\cdot)$, and the conclusion follows. ■

The following lemma are used for the proof of Proposition 3.2, and 4.1

Lemma A.1. *Let the data generating process $(P,Q) \in \mathcal{P}$ be given. Fix f_{Y_1} and f_{Y_0} the marginal probability densities of Y_1 and Y_0. There exists a joint distribution of (Y_1, Y_0, T, Z) that is compatible with the data generating process, satisfies RA, and whose marginal distributions of Y_1 and Y_0 coincide with the provided f_{Y_1} and f_{Y_0} if and only if we can find nonnegative functions*

$\{(h_{Y_1,t}, h_{Y_0,t}), t = c, n, a, d\}$ *that satisfy the following constraints μ-a.e.*

$$
\begin{aligned}
p_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,a}(y_1), && (15)\\
q_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,a}(y_1), && (16)\\
p_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,n}(y_0), && (17)\\
q_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,n}(y_0), && (18)\\
f_{Y_1}(y_1) - p_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,n}(y_1), && (19)\\
f_{Y_1}(y_1) - q_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,n}(y_1), && (20)\\
f_{Y_0}(y_0) - p_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,a}(y_0), && (21)\\
f_{Y_0}(y_0) - q_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,a}(y_0), && (22)\\
\int_{\mathcal{Y}} h_{Y_1,c}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,c}(y_0)d\mu, && (23)\\
\int_{\mathcal{Y}} h_{Y_1,n}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,n}(y_0)d\mu, && (24)\\
\int_{\mathcal{Y}} h_{Y_1,a}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,a}(y_0)d\mu, && (25)\\
\int_{\mathcal{Y}} h_{Y_1,d}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,d}(y_0)d\mu. && (26)
\end{aligned}
$$

Proof of Lemma A.1. The "only if" part is implied by the equations (6) in the main text. So, we focus on proving the "if" part of the lemma. Given the nonnegative functions $\{(h_{Y_1,t}, h_{Y_0,t}), t = c, n, a, d\}$ satisfying the above constraints, let $\pi_t = \int_{\mathcal{Y}} h_{Y_1,t}d\mu = \int_{\mathcal{Y}} h_{Y_0,t}d\mu \geq 0$ for $t \in \{c, n, a, d\}$. Consider the conditional densities of (Y_1, Y_0, T) given Z constructed by

$$
\begin{aligned}
& f_{Y_1,Y_0,T|Z}(y_1, y_0, T = t|Z = 1) = f_{Y_1,Y_0,T|Z}(y_1, y_0, T = t|Z = 0)\\
= & \begin{cases} \pi_t^{-1} h_{Y_1,t}(y_1) h_{Y_0,t}(y_0) & \text{if} \quad \pi_t > 0, \\ 0 & \text{if} \quad \pi_t = 0. \end{cases}
\end{aligned}
$$

By construction the constructed population satisfies RA. Also, the constraint (15) implies

$$
\begin{aligned}
p_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,a}(y_1)\\
&= f_{Y_1,T|Z}(y_1, T = c|Z = 1) + f_{Y_1,T|Z}(y_1, T = a|Z = 1),
\end{aligned}
$$

and a similar result holds for p_{Y_0}, q_{Y_1}, and q_{Y_0}. Hence, the constructed population is compatible with the data generating process. Lastly, this way of constructing the population distribution gives the proposed distribution of Y_1 since $\sum_{t=c,n,a,d} \int_{y_0 \in \mathcal{Y}} f_{Y_1,Y_0,T}(y_1, y_0, t)d\mu = \sum_{t=c,n,a,d} h_{Y_1,t}(y_1) = f_{Y_1}$ as implied by the constraints (15) and (19). This is also the case for f_{Y_0}. Thus, the given (f_{Y_1}, f_{Y_0}) belongs to $IR_{(f_{Y_1}, f_{Y_0})}(P, Q|RA)$. This completes the proof. ■

Lemma A.2. Let δ_{Y_1}, δ_{Y_0}, λ_{Y_1}, be the parameters defined in the statement of Proposition 3.1 and 3.2. In addition, define $\lambda_{Y_0} \equiv \int_{\mathcal{Y}} \min\{p_{Y_0}, q_{Y_0}\}d\mu$.

$\delta_{Y_1} + \delta_{Y_0} + \lambda_{Y_1} + \lambda_{Y_0} = 2.$

Proof of Lemma A.2.

$$\begin{aligned}
\Pr(D=1|Z=1)+\Pr(D=1|Z=0) &= \int_{\mathcal{Y}}[p_{Y_1}+q_{Y_1}]d\mu \\
&= \int_{\mathcal{Y}}[\max\{p_{Y_1},q_{Y_1}\}+\min\{p_{Y_1},q_{Y_1}\}]d\mu \\
&= \delta_{Y_1}+\lambda_{Y_1}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\Pr(D=1|Z=1)+\Pr(D=1|Z=0) &= 2-\Pr(D=0|Z=1)+\Pr(D=0|Z=0) \\
&= 2-\int_{\mathcal{Y}}[p_{Y_0}+q_{Y_0}]d\mu \\
&= 2-\int_{\mathcal{Y}}[\max\{p_{Y_0},q_{Y_0}\}+\min\{p_{Y_0},q_{Y_0}\}]d\mu \\
&= 2-\delta_{Y_0}-\lambda_{Y_0}.
\end{aligned}$$

Hence, $\delta_{Y_1}+\lambda_{Y_1}=2-\delta_{Y_0}-\lambda_{Y_0}$ holds. ■

Proof of Proposition 3.2. As shown in Proposition 3.1, if the data generating process reveals $\delta_{Y_1}>1$ or $\delta_{Y_0}>1$, no population is compatible with MSI, and this clearly implies $IR_{(f_{Y_1},f_{Y_0})}(P,Q|RA)$ is empty. So, we precludes this trivial case from the proof and focus on the data generating process with $\delta_{Y_1}\leq 1$ and $\delta_{Y_0}\leq 1$.

First, let us consider the data generating process with $1-\delta_{Y_0}<\lambda_{Y_1}$, and guess the identification region to be $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|RA)=\mathcal{F}^{*}_{f_{Y_1}}(P,Q)\times\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. Note that $\mathcal{F}^{*}_{f_{Y_1}}(P,Q)$ is nonempty since it always contains $f_{Y_1}=\underline{f_{Y_1}}+\frac{1-\delta_{Y_1}}{\lambda_{Y_1}}\min\{p_{Y_1},q_{Y_1}\}$.

Pick an arbitrary f_{Y_1} from $\mathcal{F}^{*}_{f_{Y_1}}(P,Q)$ and an arbitrary f_{Y_0} from $F^{env}_{f_{Y_0}}(P,Q)$. Define a nonnegative function

$$g_{Y_1}=\frac{\lambda_{Y_1}+\delta_{Y_0}-1}{\int_{\mathcal{Y}}\min\left\{f_{Y_1}-\underline{f_{Y_1}},\min\{p_{Y_1},q_{Y_1}\}\right\}d\mu}\min\left\{f_{Y_1}-\underline{f_{Y_1}},\min\{p_{Y_1},q_{Y_1}\}\right\}, \tag{27}$$

and consider the following choice of $\{(h_{Y_1,t},h_{Y_0,t}),t=c,n,a,d\}$,

$$\begin{aligned}
&h_{Y_1,c}=p_{Y_1}-\min\{p_{Y_1},q_{Y_1}\}+g_{Y_1},\\
&h_{Y_1,n}=f_{Y_1}-\underline{f_{Y_1}}-g_{Y_1},\\
&h_{Y_1,a}=\min\{p_{Y_1},q_{Y_1}\}-g_{Y_1},\\
&h_{Y_1,d}=q_{Y_1}-\min\{p_{Y_1},q_{Y_1}\}+g_{Y_1},\\
&h_{Y_0,c}=q_{Y_0}-\min\{p_{Y_0},q_{Y_0}\},\\
&h_{Y_0,n}=\min\{p_{Y_0},q_{Y_0}\},\\
&h_{Y_0,a}=f_{Y_0}-\underline{f_{Y_0}},\\
&h_{Y_0,d}=p_{Y_0}-\min\{p_{Y_0},q_{Y_0}\}.
\end{aligned} \tag{28}$$

Since $g_{Y_1}\leq\min\{p_{Y_1},q_{Y_1}\}$ and $g_{Y_1}\leq f_{Y_1}-\underline{f_{Y_1}}$ by construction, $\{h_{Y_1,t}(y_1),t=c,n,a,d\}$ are all nonnegative functions. It can be seen that the constraints (15) through (22) are satisfied. Also, by utilizing Lemma A.2, we can check the scale constraints (23) through (26) are satisfied. Hence, by Lemma A.1, we conclude that the proposed (f_{Y_1}, f_{Y_0}) belongs to $IR_{(f_{Y_1},f_{Y_0})}(P,Q|RA)$ so that $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|RA) \subset IR_{(f_{Y_1},f_{Y_0})}(P,Q|RA)$.

Next, consider f_{Y_1} that does not satisfy $\int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}, q_{Y_1}\}\right\} d\mu \geq \lambda_{Y_1} + \delta_{Y_0} - 1$ and $f_{Y_0} \in \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. Suppose that the nonnegative functions $\{(h_{Y_1,t}, h_{Y_0,t}), t = c,n,a,d\}$ satisfying the constraints (15) through (22) exist. Then, the constraints (21) and (22) imply that $\int_{\mathcal{Y}} h_{Y_0,a} d\mu \leq 1 - \delta_{Y_0}$. Moreover,

$$\begin{aligned} f_{Y_1} &= \sum_{t=c,n,a,d} h_{Y_1,t} \\ &\geq p_{Y_1} + q_{Y_1} - h_{Y_1,a} \\ &= \underline{f_{Y_1}} + \min\{p_{Y_1}, q_{Y_1}\} - h_{Y_1,a}, \end{aligned}$$

implies

$$f_{Y_1} - \underline{f_{Y_1}} \geq \min\{p_{Y_1}, q_{Y_1}\} - h_{Y_1,a}. \tag{29}$$

Now, since $f_{Y_1} \notin \mathcal{F}^*_{f_{Y_1}}(P,Q)$, it follows that

$$\begin{aligned} \lambda_{Y_1} + \delta_{Y_0} - 1 &> \int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}, q_{Y_1}\}\right\} d\mu \\ &\geq \int_{\mathcal{Y}} \min\left\{\min\{p_{Y_1}, q_{Y_1}\} - h_{Y_1,a}, \min\{p_{Y_1}, q_{Y_1}\}\right\} d\mu \\ &= \int_{\mathcal{Y}} [\min\{p_{Y_1}, q_{Y_1}\} - h_{Y_1,a}] d\mu \\ &= \lambda_{Y_1} - \int h_{Y_1,a} d\mu. \end{aligned}$$

where the second line follows by the inequality (29). Hence, $\int h_{Y_1,a} d\mu > 1 - \delta_{Y_0}$. This and $\int_{\mathcal{Y}} h_{Y_0,a} d\mu \leq 1 - \delta_{Y_0}$ violates the scale constraint for $t = a$. So, we conclude that there are no feasible $\{(h_{Y_1,t}, h_{Y_0,t}), t = c,n,a,d\}$ that meets the constraints of Lemma A.1. Note $f_{Y_0} \notin \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ immediately implies violation of $Y_0 \perp Z$. Therefore, we conclude $IR^{guess}_{(f_{Y_1},f_{Y_0})}(P,Q|RA) \supset IR_{(f_{Y_1},f_{Y_0})}(P,Q|RA)$.

By combining these results, we conclude that $\mathcal{F}^*_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ is the identification region of (f_{Y_1}, f_{Y_0}) under RA.

For the case of $1 - \delta_{Y_0} > \lambda_{Y_1}$, the identification region is derived by a symmetric argument to the case of $1 - \delta_{Y_0} < \lambda_{Y_1}$. So, for the sake of brevity we omit a proof.

Lastly, consider the case of $1 - \delta_{Y_0} = \lambda_{Y_1}$. As we presented in the main text and Figure 3, for every $f_{y_1} \in F^{env}_{f_{Y_1}}(P,Q)$ and $f_{Y_0} \in F^{env}_{f_{Y_0}}(P,Q)$, we can find $\{(h_{Y_1,t}, h_{Y_0,t}), t = c,n,a,d\}$ that satisfies all the constraints of Lemma A.1. Hence, $\mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ is the identification region of (f_{Y_1}, f_{Y_0}) under RA. ■

Proof of Proposition 4.1. The mean parameter respects stochastic dominance (Manski (2003)). So, the sharp upper (lower) bounds of $E(Y_1)$ are obtained by finding f_{Y_1} within the identification region that (is) first-order stochastically dominates (dominated by) the others in the identification region. Consider bounding the mean of Y_1 when the density f_{Y_1} belongs to the class of densities $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{*}_{f_{Y_1}}(P,Q)$ respectively. For the former, it is known that the bounds of $E(Y_1)$ is given by

$$(1-\delta_{Y_1})y_l + \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu \leq E(Y_1) \leq (1-\delta_{Y_1})y_u + \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu.$$

See Lemma 2.2.2 in Manski (2003) for the discrete outcome case and Kitagawa (2009) for the continuous outcome case. For the latter, deriving the bounds is slightly more involved. Consider the density

$$f^{lower}_{Y_1}(y_1) = \lambda_{Y_0} 1\{y_1 = y_l\} + \underline{f_{Y_1}}(y_1) + [\min\{p_{Y_1}, q_{Y_1}\}]^{rtrim}_{1-\delta_{Y_0}}(y_1).$$

Note $f^{lower}_{Y_1} \geq \underline{f_{Y_1}}$ and

$$\int_{\mathcal{Y}} \min\left\{ f^{lower}_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}, q_{Y_1}\} \right\} d\mu = \lambda_{Y_1} - (1-\delta_{Y_0}),$$

so $f^{lower}_{Y_1} \in \mathcal{F}^{*}_{f_{Y_1}}(P,Q)$. By applying the decomposition trick (28) proposed in the proof of Proposition 3.2, we can decompose $f^{lower}_{Y_1}$ into the nonnegative functions $\{h^{lower}_{Y_1,t}\}$. That is, for $t = a$ and $t = n$, we obtain

$$\begin{aligned} h^{lower}_{Y_1,a} &= \min\{p_{Y_1}, q_{Y_1}\} - [\min\{p_{Y_1}, q_{Y_1}\}]^{rtrim}_{1-\delta_{Y_0}}, \\ h^{lower}_{Y_1,n} &= \lambda_{Y_0} 1\{y_1 = y_l\}. \end{aligned}$$

Furthermore, $f^{lower}_{Y_1}$ is expressed as

$$\begin{aligned} f^{lower}_{Y_1} &= \sum_t h^{lower}_{Y_1,t} \\ &= p_{Y_1} + q_{Y_1} - h^{lower}_{Y_1,a} + h^{lower}_{Y_1,n}. \qquad (30) \end{aligned}$$

where in the second line we use the constraints (15) and (16). Let $\tilde{f}_{Y_1}$ be an arbitrary element in $\mathcal{F}^{*}_{f_{Y_1}}(P,Q)$. By Lemma A.1 and Proposition 3.2, there exist nonnegative functions $\{\tilde{h}_{Y_1,t}, t = c, n, a, d\}$ by which $\tilde{f}_{Y_1}$ can be represented as

$$\begin{aligned} \tilde{f}_{Y_1} &= \sum_t \tilde{h}_{Y_1,t} \\ &= p_{Y_1} + q_{Y_1} - \tilde{h}_{Y_1,a} + \tilde{h}_{Y_1,n}, \qquad (31) \end{aligned}$$

and, again, by applying the decomposition trick (28) of the proof of Proposition 3.2, $\tilde{h}_{Y_1,a}$ and $\tilde{h}_{Y_1,n}$ can be expressed as

$$\begin{aligned} \tilde{h}_{Y_1,a} &= \min\{p_{Y_1}, q_{Y_1}\} - \tilde{g}_{Y_1}, \\ \tilde{h}_{Y_1,n} &= \tilde{f}_{Y_1} - \underline{f_{Y_1}} - \tilde{g}_{Y_1}. \end{aligned}$$

where $\tilde{g}_{Y_1}$ is obtained by plugging $\tilde{f}_{Y_1}$ into (27). From (30) and (31), for $t \in [y_l, y_u]$, the difference between $\int_{[y_l,t]} f_{Y_1}^{lower} d\mu$ and $\int_{[y_l,t]} \tilde{f}_{Y_1} d\mu$ is written as

$$\begin{aligned}
& \int_{[y_l,t]} f_{Y_1}^{lower} d\mu - \int_{[y_l,t]} \tilde{f}_{Y_1} d\mu \\
= & \int_{[y_l,t]} [h_{Y_1,n}^{lower} - \tilde{h}_{Y_1,n}] d\mu + \int_{[y_l,t]} [\tilde{h}_{Y_1,a} - h_{Y_1,a}^{lower}] d\mu \\
= & \lambda_{Y_0} - \int_{[y_l,t]} \left(\tilde{f}_{Y_1} - \underline{f_{Y_1}} - \tilde{g}_{Y_1} \right) d\mu + \int_{[y_l,t]} \left([\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{rtrim} - \tilde{g}_{Y_1} \right) d\mu. \qquad (32)
\end{aligned}$$

Regarding the second term of (32), since $\tilde{f}_{Y_1} - \underline{f_{Y_1}} - \tilde{g}_{Y_1} \geq 0$, it can be bounded above by

$$\begin{aligned}
\int_{[y_l,t]} \left(\tilde{f}_{Y_1} - \underline{f_{Y_1}} - \tilde{g}_{Y_1} \right) d\mu \quad & \leq \quad \int_{\mathcal{Y}} \left(\tilde{f}_{Y_1} - \underline{f_{Y_1}} - \tilde{g}_{Y_1} \right) \\
& = \quad 1 - \delta_{Y_1} - \lambda_{Y_1} - 1 + \delta_{Y_0}.
\end{aligned}$$

Regarding the third term of (32), if t is strictly less than the $(1-\delta_{Y_0})$-th right-trimming point $q_{1-\delta_{Y_0}}^{right} = \sup\left\{ s : \int_{[s,y_u]} \min\{p_{Y_1}, q_{Y_1}\} d\mu \geq 1 - \delta_{Y_0} \right\}$, then $[\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{rtrim} = \min\{p_{Y_1}, q_{Y_1}\} \geq \tilde{g}_{Y_1}$ holds on $y_1 \in [y_l, t]$. So the integral is nonnegative. On the other hand, if $t \geq q_{1-\delta_{Y_0}}^{right}$,

$$\begin{aligned}
& \int_{[y_l,t]} \left([\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{rtrim} - \tilde{g}_{Y_1} \right) \\
= \quad & \lambda_{Y_1} - (1-\delta_{Y_0}) - \int_{[y_l,t]} \tilde{g}_{Y_1} d\mu \\
\geq \quad & \lambda_{Y_1} - (1-\delta_{Y_0}) - \int_{\mathcal{Y}} \tilde{g}_{Y_1} d\mu \\
= \quad & \lambda_{Y_1} - (1-\delta_{Y_0}) - [\lambda_{Y_1} - (1-\delta_{Y_0})] = 0.
\end{aligned}$$

By combining them, for each $t \in [y_l, y_u]$, (32) is bounded below by $\lambda_{Y_0} + \lambda_{Y_1} + \delta_{Y_1} + \delta_{Y_0} - 2$, and this is zero by Lemma A.2. Therefore, we conclude that $f_{Y_1}^{lower}$ first order stochastically dominates $\tilde{f}_{Y_1}$, and the mean of Y_1 with respect to $f_{Y_1}^{lower}$ minimizes $E(Y_1)$ over $f_{Y_1} \in \mathcal{F}_{f_{Y_1}}^*(P,Q)$.
Next, we shall find the upper bound of $E(Y_1)$ by essentially repeating the same procedure as above. Define

$$f_{Y_1}^{upper}(y_1) = \underline{f_{Y_1}}(y_1) + [\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{ltrim}(y_1) + \lambda_{Y_0} 1\{y_1 = y_u\},$$

which is shown to belong to $\mathcal{F}_{f_{Y_1}}^*(P,Q)$. Similarly to the lower bound case (31), represent $f_{Y_1}^{upper}$ by

$$\begin{aligned}
f_{Y_1}^{upper} \quad & = \quad \sum_t h_{Y_1,t}^{upper} \\
& = \quad p_{Y_1} + q_{Y_1} - h_{Y_1,a}^{upper} + h_{Y_1,n}^{upper},
\end{aligned}$$

$$\begin{aligned}
h_{Y_1,a}^{upper} \quad & = \quad \min\{p_{Y_1}, q_{Y_1}\} - [\min\{p_{Y_1}, q_{Y_1}\}]_{1-\delta_{Y_0}}^{ltrim}, \\
h_{Y_1,n}^{upper} \quad & = \quad \lambda_{Y_0} 1\{y_1 = y_u\}.
\end{aligned}$$

For an arbitrary $\tilde{f}_{Y_1} \in \mathcal{F}^*_{f_{Y_1}}(P,Q)$, consider the difference between $\int_{(t,y_u]} f_{Y_1}^{upper} d\mu$ and $\int_{(t,y_u]} \tilde{f}_{Y_1} d\mu$. Analogous to (32), we obtain

$$\begin{aligned} & \int_{(t,y_u]} f_{Y_1}^{upper} d\mu - \int_{(t,y_u]} \tilde{f}_{Y_1} d\mu \\ = \; & \lambda_{Y_0} - \int_{(t,y_u]} \left(\tilde{f}_{Y_1} - \underline{f_{Y_1}} - \tilde{g}_{Y_1} \right) d\mu + \int_{(t,y_u]} \left([\min \{ p_{Y_1}, q_{Y_1} \}]^{ltrim}_{1-\delta_{Y_0}} - \tilde{g}_{Y_1} \right) d\mu. \end{aligned}$$

Now, by repeating the same procedure as above, the right hand side is bounded below by $\lambda_{Y_0} + \lambda_{Y_1} + \delta_{Y_1} + \delta_{Y_0} - 2 = 0$. Hence, we conclude that $f_{Y_1}^{upper}$ is first order stochastically dominated by $\tilde{f}_{Y_1}$, and the mean of Y_1 with respect to $f_{Y_1}^{upper}$ maximizes $E(Y_1)$ over $f_{Y_1} \in \mathcal{F}^*_{f_{Y_1}}(P,Q)$.

The bounds for $E(Y_0)$ when the density f_{Y_0} belongs to the class of densities $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ and $\mathcal{F}^*_{f_{Y_0}}(P,Q)$ are derived by a symmetric argument to the case of $E(Y_1)$, so we do not duplicate the proof here.

In order to combine the bounds of $E(Y_1)$ and $E(Y_0)$, we note that the identification region of (f_{Y_1}, f_{Y_0}) takes the form of the Cartesian product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ or $\mathcal{F}^*_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ or $\mathcal{F}^*_{f_{Y_0}}(P,Q)$. Hence, by applying the argument of the outer bounds of (Manski (2003)), it is valid to bound $E(Y_1) - E(Y_0)$ by subtracting the upper (lower) of $E(Y_0)$ from the lower (upper) bound of $E(Y_1)$ for each corresponding underlying identification regions of f_{Y_1} and f_{Y_0} . This completes the proof of the sharp bounds under MSI and RA.

As for the bounds under LATE, the bounds become empty when the nested densities are not observed because the identification region of (f_{Y_1}, f_{Y_0}) in this case is empty (Proposition 3.3). On the other hand, when the data generating process exhibits the nested densities, the formula of the sharp ATE bounds corresponding to $\mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ is reduced to the presented formula since for $j = 1, 0$, we have $\delta_{Y_j} = \max_z \{\Pr(D = j|Z = z)\}$ and $\int_{\mathcal{Y}} y_1 \underline{f_{Y_1}} d\mu = \max_z \{E(Y|D = j, Z = z) \Pr(D = j|Z = z)\}$.
■

Appendix B: Binary Y: A Comparison with the Balke and Pearl's bounds

In this appendix, we shall show that when the outcome is binary the bound formula for $IR_{ATE}(P, Q|RA)$ presented in Proposition 4.1 coincides with the Balke and Pearl's bounds (Balke and Pearl (1997)). Now, the dominating measure μ puts point masses on $\{1, 0\}$. Accordingly, each $p_{Y_j}(y_j)$ or $q_{Y_j}(y_j)$ for $y_j \in \{1, 0\}$ and $j = 1, 0$, represents probability masses $\Pr(Y = y_j, D = j|Z = 1)$ or $\Pr(Y = y_j, D = j|Z = 0)$.

By solving a linear optimization, Balke and Pearl (1997, pp.1172) derives the following bound formulas for $E(Y_1)$ and $E(Y_0)$.

$$\max\left\{\begin{array}{c} q_{Y_1}(1) \\ p_{Y_1}(1) \\ -q_{Y_0}(0) - q_{Y_1}(0) + p_{Y_0}(0) + p_{Y_1}(1) \\ -q_{Y_1}(0) - q_{Y_0}(1) + p_{Y_0}(1) + p_{Y_1}(1) \end{array}\right\} \leq E(Y_1) \tag{33}$$

$$\leq \quad \min\left\{\begin{array}{c} 1 - p_{Y_1}(0) \\ 1 - q_{Y_1}(0) \\ q_{Y_0}(0) + q_{Y_1}(1) + p_{Y_0}(1) + p_{Y_1}(1) \\ q_{Y_1}(1) + q_{Y_1}(1) + p_{Y_0}(0) + p_{Y_1}(1) \end{array}\right\} \tag{34}$$

and

$$\max\left\{\begin{array}{c} p_{Y_0}(1) \\ q_{Y_0}(1) \\ q_{Y_0}(1) + q_{Y_1}(1) - p_{Y_0}(0) - p_{Y_1}(1) \\ q_{Y_1}(0) + q_{Y_0}(1) - p_{Y_0}(0) - p_{Y_1}(0) \end{array}\right\} \leq E(Y_0) \tag{35}$$

$$\leq \quad \min\left\{\begin{array}{c} 1 - p_{Y_0}(0) \\ 1 - q_{Y_0}(0) \\ q_{Y_1}(0) + q_{Y_0}(1) + p_{Y_0}(1) + p_{Y_1}(1) \\ q_{Y_0}(1) + q_{Y_1}(1) + p_{Y_1}(0) + p_{Y_0}(1) \end{array}\right\}. \tag{36}$$

In addition, Balke and Pearl show that the bounds for $E(Y_1) - E(Y_0)$ is obtained by the difference of these bounds, that is, the lower bound of $E(Y_1) - E(Y_0)$ is equal to the lower bound of $E(Y_1)$ less $E(Y_0)$'s upper bound, and the upper bound of $E(Y_1) - E(Y_0)$ is equal to the upper bound of $E(Y_1)$ less $E(Y_0)$'s lower bound.

In order to make the comparison of our bounds with their bounds easier, we rewrite the above bounds as follows. First, consider the lower bound of $E(Y_1)$. The first two elements in the maximum operator can be combined and written as a single element $\max\{p_{Y_1}(1), q_{Y_1}(1)\}$. As for the third and the forth elements in the maximum operator, we plug in $p_{Y_0}(0) = 1 - p_{Y_1}(0) - p_{Y_1}(1) - p_{Y_0}(1)$ and $p_{Y_1}(1) + q_{Y_1}(1) = \max\{p_{Y_1}(1), q_{Y_1}(1)\} + \min\{p_{Y_1}(1), q_{Y_1}(1)\}$, and take their maximum. As a result, the maximum of the third and the forth elements is written as

$$\begin{aligned} & \max\left\{\begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array}\right\} + \min\left\{\begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array}\right\} - 1 + \max\left\{\begin{array}{c} p_{Y_0}(0) \\ q_{Y_0}(0) \end{array}\right\} + \max\left\{\begin{array}{c} p_{Y_0}(1) \\ q_{Y_0}(1) \end{array}\right\} \\ = \quad & \max\left\{\begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array}\right\} + \min\left\{\begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array}\right\} - (1 - \delta_{Y_0}). \end{aligned} \tag{37}$$

Regarding the $E(Y_1)$'s upper bound (34), the minimum of the first two elements is written as

$$1 - \max\{p_{Y_1}(0), q_{Y_1}(0)\} = \max\{p_{Y_1}(1), q_{Y_1}(1)\} + 1 - \delta_{Y_1}, \tag{38}$$

and, by noting $p_{Y_1}(1) + q_{Y_1}(1) = \max\{p_{Y_1}(1), q_{Y_1}(1)\} + \min\{p_{Y_1}(1), q_{Y_1}(1)\}$, the minimum of the third and fourth elements becomes

$$\begin{aligned} & \max\left\{ \begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array} \right\} + \min\left\{ \begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array} \right\} + \min\left\{ \begin{array}{c} p_{Y_0}(0) \\ q_{Y_0}(0) \end{array} \right\} + \min\left\{ \begin{array}{c} p_{Y_0}(1) \\ q_{Y_0}(1) \end{array} \right\} \\ = & \max\left\{ \begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array} \right\} + \min\left\{ \begin{array}{c} p_{Y_1}(1) \\ q_{Y_1}(1) \end{array} \right\} + \lambda_{Y_0}. \end{aligned} \tag{39}$$

By taking the maximum of $\max\{p_{Y_1}(1), q_{Y_1}(1)\}$ and (37), and the minimum of (38) and (39), we obtain an alternative expression for the Balke and Pearl's bounds of $E(Y_1)$,

$$\begin{aligned} & \max\{p_{Y_1}(1), q_{Y_1}(1)\} + \max\left\{ \begin{array}{c} 0 \\ \min\{p_{Y_1}(1), q_{Y_1}(1)\} - (1-\delta_{Y_0}) \end{array} \right\} \leq E(Y_1) \qquad (40) \\ \leq & \max\{p_{Y_1}(1), q_{Y_1}(1)\} + \min\left\{ \begin{array}{c} (1-\delta_{Y_1}) - \lambda_{Y_0} \\ \min\{p_{Y_1}(1), q_{Y_1}(1)\} \end{array} \right\} + \lambda_{Y_0}. \end{aligned}$$

Similarly, we consider the same type of transformations on the bounds of $E(Y_0)$. That is, we express the maximum over the first two elements and the maximum over the latter two elements in (35). Also, we take the minimum of the first two elements and the latter two elements of (35) separately. Then, Balke and Pearl's $E(Y_0)$ bounds are written as

$$\begin{aligned} & \max\{p_{Y_0}(1), q_{Y_0}(1)\} + \max\left\{ \begin{array}{c} 0 \\ \min\{p_{Y_0}(1), q_{Y_0}(1)\} - (1-\delta_{Y_1}) \end{array} \right\} \leq E(Y_0) \qquad (41) \\ \leq & \max\{p_{Y_0}(1), q_{Y_0}(1)\} + \min\left\{ \begin{array}{c} (1-\delta_{Y_0}) - \lambda_{Y_1} \\ \min\{p_{Y_0}(1), q_{Y_0}(1)\} \end{array} \right\} + \lambda_{Y_1}. \end{aligned}$$

Now, consider the data generating process that satisfies $1-\delta_{Y_0} = \lambda_{Y_1}$, which also implies $1-\delta_{Y_1} = \lambda_{Y_0}$ by Lemma A.2. In this case, the transformed Balke and Pearl's bounds (40) and (41) yield

$$\max\{p_{Y_1}(1), q_{Y_1}(1)\} \leq E(Y_1) \leq \max\{p_{Y_1}(1), q_{Y_1}(1)\} + 1 - \delta_{Y_1}, \tag{42}$$

$$\max\{p_{Y_0}(1), q_{Y_0}(1)\} \leq E(Y_0) \leq \max\{p_{Y_0}(1), q_{Y_0}(1)\} + 1 - \delta_{Y_0}. \tag{43}$$

We can see these bounds yields the bound formula for ATE (12), and therefore, we obtain the consistent result with the first case of Proposition 4.1 (ii).

Next, consider the case for $1-\delta_{Y_0} < \lambda_{Y_1}$, which also implies $1-\delta_{Y_1} > \lambda_{Y_0}$. In this case, the bounds for $E(Y_0)$ is the same as (43), while $E(Y_1)$'s bound can differ from (42) since the second terms of the expression of the lower and the upper bound of (40) can be nonzero. In fact, the second term of the lower bound of (40) is seen as the probability mass on $y_1 = 1$ for $(1-\delta_{Y_0})$-right-trimming of $\min\{p_{Y_1}, q_{Y_1}\}$. Also, the second term of the upper bound of (40) is seen as the probability mass on $y_1 = 1$ for the $(1-\delta_{Y_0})$-left-trimming of $\min\{p_{Y_1}, q_{Y_1}\}$. Therefore, the resulting bounds of $E(Y_1) - E(Y_0)$ coincide with (13), the second case of Proposition 4.1 (ii).

Last, consider the case for $1-\delta_{Y_0} > \lambda_{Y_1}$, which also implies $1-\delta_{Y_1} < \lambda_{Y_0}$. Contrary to the previous case, the bounds for $E(Y_1)$ becomes the same as (42), while $E(Y_0)$'s bound is not always given by (43). Note that the second term of the lower bound of (41) is seen as the probability mass on $y_0 = 1$ for $(1-\delta_{Y_1})$-right-trimming of $\min\{p_{Y_0}, q_{Y_0}\}$. Also, the second term of the upper bound of (41) is seen as the probability mass on $y_1 = 1$ for the $(1-\delta_{Y_0})$-left-trimming of $\min\{p_{Y_1}, q_{Y_1}\}$.

Therefore, the resulting bounds of $E(Y_1) - E(Y_0)$ coincide with (14), the third case of Proposition 4.1 (ii).

Thus, we conclude that, for every possible data generating process with the binary outcome, the bound formula of Proposition 4.1 (ii) yields the same bounds as Balke and Pearl (1997).

References

[1] Abadie, A., J. D. Angrist, and G. W. Imbens (2002) "Instrumental Variables Estimates of the. Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 1,: pp91-117

[2] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association,* 91: 444 - 472.

[3] Athey, S. and G.W. Imbens (2006): "Identification and Inference in Nonlinear Difference-In-Differences Models," *Econometrica* 74, 431-497.

[4] Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171-1176.

[5] Bhattacharya J., A.M. Shaikh, and E. Vytlacil (2008): "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization," *American Economic Review: Papers & Proceedings*, 98, 351-356.

[6] Chen, J. and D. S. Small (2006): "Bounds on Causal Effects in Three-arm Trials with Non-compliance," *Journal of the Royal Statistical Society, Series B*, 68, part 5, 815-836.

[7] Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica* 71, 1405-1441.

[8] Chesher, A. (2005): "Nonparametric Identification Under Discrete Variation," *Econometrica*, 73, 1525-1550.

[9] Chesher, A. (2009): "Instrumental Variable Models for Discrete Outcomes," CeMMAP Working Paper, 30/08.

[10] Fan Y. and S. Park (2007): "Sharp Bounds on the Distribution of the Treatment Effects and Their Statistical Inference," manuscript, Department of Economics, Vanderbilt University.

[11] Firpo, S. and G. Ridder (2008): "Bounds on Functionals of the Distribution of Treatment Effects," IEPR Working Paper 08.09.

[12] Hahn, J. (2009): "Bounds on ATE with Discrete Outcomes," manuscript, Department of Economics, UCLA.

[13] Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most out of Programme Evaluation and Social Experiments Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487-535.

[14] Heckman, J. J. and E. Vytlacil (2001a): "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effects," in Lechner, M., and M. Pfeiffer editors, *Econometric Evaluation of Labour Market Policies.* pp. 1-15, Center for European Economic Research, New York.

[15] Heckman, J. J. and E. Vytlacil (2001b): "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powell editors, *Nonlinear Statistical Model: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya,* pp. 1-46. Cambridge University Press, Cambridge UK.

[16] Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73, 669-738.

[17] Heckman, J. J. and E. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," J.J. Heckman and E. E. Leamer ed. *Handbook of Econometrics Vol. 6B*, Chapter 71. Elsevier B.V., New York.

[18] Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

[19] Imbens, G. W. and Newey, W. (2008): "Identification and Estimation of Triangular Simultaneous Equations Model Without Additivity," *forthcoming in Econometrica.*

[20] Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, 64, 555-574.

[21] Kitagawa, T. (2009a): "Testing for Instrument Independence in the Selection Model," manuscript, Department of Economics, University College London.

[22] Kitagawa, T. (2009b): *Three Essays in Instrumental Variables*, Ph.D. dissertation, Brown University.

[23] Kitagawa, T. and S. Hoderlein (2009): "Testing for Instrument Validity in the Heterogeneous Treatment Effect Model," manuscript, Department of Economics, University College London.

[24] Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Reviews Papers and Proceedings*, 80, 319-323.

[25] Manski, C. F. (1994): "The Selection Problem," In C. Sims, editor, *Advances in Econometrics, Sixth World Congress, Vol 1*, 143-170, Cambridge University Press, Cambridge, UK.

[26] Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer-Verlag, New York.

[27] Manski, C.F. (2005): *Social Choice with Partial Knowledge of Treatment Response*, Princeton University Press, Princeton, New Jersey.

[28] Manski, C. F. (2007): *Identification for Prediction and Decision*, Harvard University Press, Cambridge, Massachusetts.

[29] Pearl, J. (1994a): "From Bayesian Networks to Causal Networks," A. Gammerman ed. *Bayesian Networks and Probabilistic Reasoning*, pp. 1-31. London: Alfred Walter.

[30] Pearl, J. (1994b): "On the Testability of Causal Models with Latent and Instrumental Variables," *Uncertainty in Artificial Intelligence*, 11, 435-443.

[31] Pearl, J. (2000): *Causality*, Cambridge University Press, Cambridge, UK.

[32] Vytlacil, E. J. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result". *Econometrica*, 70, 331-341.