Three Essays in Instrumental Variables

by
Toru Kitagawa
B.A., University of Tokyo, 2002
M.A., Brown University, 2005

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in the
Department of Economics at Brown University

Providence, Rhode Island
May 2009

This dissertation by Toru Kitagawa is accepted in its present form by the Department of Economics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date 4/21/09

Frank Kleibergen, Director

Recommended to the Graduate Council

Date 4/21/09

Stefan Hoderlein, Reader

Date 4/21/09

Sophocles Mavroeidis, Reader

Date 4/21/09

Blaise Melly, Reader

Approved by the Graduate Council

Date 4/21/09

Sheila Bonde
Dean of the Graduate School

Vita

The author was born in September 18th, 1979 in Japan. He received his B.A. in Engineering from the University of Tokyo in 2002. He entered Brown University in 2004 to pursue his degree in Economics. He received his M.A. in 2005 and Ph.D. in 2009.

Acknowledgements

I am deeply indebted to my advisor Frank Kleibergen for thoughtful discussions and continuous encouragement. I also thank Guido Imbens for inspiration and insightful comments. I am grateful to Tony Lancaster without whom I may never have pursued research in econometrics. I am also grateful to Stefan Hoderlein for his constructive comments and encouragement. Thanks to Sophocles Mavroeidis and Blaise Melly for reading the manuscript. I also owe much to my colleagues, Sung Jae Jun, Leandro Magnusson, Yuya Sasaki, and Zhang Zhaoguo, who spent time discussing econometrics with me and who made helpful suggestions throughout my graduate study. I thank my friends at the Robinson basement, Jim Campbell, Tiago De Abreu Freire, Ruben Durante, Emilio Gutierrez-Fernandez, Martin Goetz, Juan-Carlos Gozzi, Young Chul Kim, Mariko Klasing, Petros Milionis, Nathan Schiff, Pablo Suarez Becerra, Norovsambuu Tumennasan, and Yori Zwols, and more, with whom I had a joyful time in Providence. I would like to give special thanks to Yu-lin Lin for her support and consideration. Lastly, I am most indebted to my family in Japan for their unceasing patience, consideration, and support.

Contents

Chapter 1

Partial Identification with an Independent Instrument

1.1 Introduction

A partially identified model is a model for which the parameters of interest cannot be uniquely determined by the observed data. In a sequence of seminal work, Manski (1989, 1990, 1994, 2003, 2007) analyzes the missing data model with selection where some observations of outcome Y can be missing in a nonrandom way, and stimulated research in partial identification analysis (see Manski (2003, 2007) for an overview and economic applications). Manski (1990, 1994) introduces the use of an instrumental variable for partial identification analysis, and analyzes the identification region for the parameters, or for the distribution of outcomes, under various restrictions on the statistical relationship between the instrument and outcome. While the literature has analyzed the identification region of the parameters such as the mean of Y under moment-type restrictions,[1] less is known about the identification region of the outcome distribution under a distributional restriction of statistical independence between instrument and outcome.

In this paper, we focus on the instrument exclusion restriction; that is, an instrument Z that is specified to be *statistically independent* of the underlying outcome. The selection problems that this paper considers are the missing data problem and the counterfactual causal model with a binary treatment.

In the missing data problem, the outcome Y is observed if the selection indicator D is one while it is missing if D is zero. The researcher has a random sample of $(Y \cdot D, D, Z)$ and the object of interest is f_Y, the population distribution of Y. For example, Y could be potential wages that are observed only for those who are employed, and the instrument Z is a variable that is specified to be independent of one's potential wage but may affect one's employment status. A list of instruments

[1] The tight bounds for $E(Y)$ under the mean independence, $E(Y|Z) = E(Y)$, is analyzed by Manski (1994). Manski and Pepper (2000) derive the tight bounds for $E(Y)$ under the restriction of monotonic outcome response: $E(Y|Z = z)$ is increasing with respect to z.

that has been used in this potential wage example includes, for example, the number of children, marital status, and a measure of out-of-work income.

In the counterfactual causal model with a binary treatment, outcome variables are a pair of treatment outcome Y_1 and control outcome Y_0. Since for each individual we can observe only one of the potential outcomes, the counterfactual causal model always involves missing data. Interpreting the selection indicator D as an indicator for treatment status, we observe Y_1 if $D = 1$ and Y_0 if $D = 0$, and data is a random sample of (Y_{obs}, D, Z) where $Y_{obs} = DY_1 + (1 - D)Y_0$. Here, the object of interest is the distribution of potential outcomes (Y_1, Y_0). In particular, we focus on a pair of two marginal distributions of each potential outcome, f_{Y_1} and f_{Y_0}, since the causal effects are defined as a functional of the potential outcome distributions. If individuals self-select to receive the treatment taking into account their potential outcomes, D is not independent of (Y_1, Y_0) and then point-identification of the potential outcome distributions f_{Y_1} and f_{Y_0} fails.

This paper analyzes identification of outcome distributions in these models without imposing point-identifying restrictions. That is, our object of interest is the *identification region*: the set of outcome distributions that are compatible with the empirical evidence and the model restrictions. In the missing data problem, Manski (2003) analyzes the identification region for the outcome distribution f_Y under the independence restriction between Y and Z. The resulting expression there has a rather abstract form and a closed form expression is limited to the discrete outcome case. One of the contributions of this paper is therefore to provide a closed form expression of the identification region that is applicable to a wider range of settings, in particular, a continuous outcome. We use the expression for the identification region of f_Y under the exclusion restriction to examine the possibility of obtaining a narrower identification region by introducing the selection mechanism with latent utility (threshold crossing selection). We consider strengthening the exclusion restriction to the restriction that the instrument Z is *jointly* independent of Y and the selection heterogeneities. We show that this joint independence restriction does not further narrow the identification region of f_Y. We also consider the identification gain of specifying the latent utility to be additively separable (threshold crossing selection with an additive error). We show that threshold crossing selection with an additive error, which is often imposed in the structural selection model, constrains the data generating process in a certain way but does *not* narrow the identification region further than instrument independence. These results imply that once instrument independence is imposed, threshold crossing selection is a redundant restriction in the sense that it does not further contribute to identifying f_Y.

We extend the identification framework of the missing data model to the counterfactual causal model. Since we observe either one of the potential outcomes and the other is missing, we cannot avoid the missing data problem in identifying the causal effects. In particular, if the selection mechanism is nonrandom, that is, individual's participation to treatment depends on his underlying potential outcomes, then the potential outcome distributions cannot be identified. We derive the identification region of the distribution of the potential outcomes (Y_1, Y_0) under the restriction that

Z is jointly independent of (Y_1, Y_0) and the selection heterogeneities. We show that, in the counterfactual causal model, the stronger restriction that Z is jointly independent of (Y_1, Y_0) and the selection heterogeneities can yield a strictly narrower identification region than the independence restriction between Z and (Y_1, Y_0). This finding implies that adding independence of instrument and the selection heterogeneities provides additional identifying information for the potential outcome distributions. This result contrasts the role of the instrument independence restriction in the counterfactual model with the one in the single missing outcome model, since, as we have mentioned above, such identification gain never arises in the single missing outcome model. Our identification analysis clarifies the source of this identification gain and characterizes the condition for the distribution of data under which this identification gain is available.

One advantage of focusing on the identification region (the set of feasible outcome distributions) is that it enables us to derive *tight* bounds for the parameters of the potential outcome distributions.[2] As an application of this way of constructing the tight parameter bounds, we provide the tight bounds for average treatment effects under instrument independence. For the case of binary potential outcomes,[3] Balke and Pearl (1997) consider bounding the causal effects under the same independence restriction within the framework of causal networks. Their derivation of the bounds relies on a certain linear optimization procedure and hence it seems hard to obtain a closed-form expression of the bounds when potential outcome distributions have a large support such as continuous. In contrast, our closed-form expression for the bounds covers the continuous outcome case and its derivation does not use a linear optimization procedure.

The remainder of the paper is organized as follows. Section 1.2 considers the single missing outcome model and derives the identification region of f_Y under instrument independence. It also provides a refutability result of instrument independence based on the emptiness of the identification region. Section 1.3 extends the identification framework developed in Section 1.2 to the counterfactual causal model. We derive the identification region of the potential outcome distributions and the tight bounds for the average treatment effects. For simplicity of exposition, our analytical framework is limited to the case with a binary instrument in the main text. Appendix 1.A.7 discusses the case with a multi-valued discrete instrument. Section 1.4 concludes. Proofs are provided in Appendices.

[2] To the best of our knowledge, there is no consensus on the definition for the tightness (sharpness) of the bounds. In this paper, we define tightness of the parameter bounds as the range for the parameter functional where the domain is given by the identification region.

[3] Chen and Small (2006) derived the tight bounds for average treatment effects for the model with three-arm treatment using linear optimization procedure.

1.2 Single Missing Outcome Model: The Identification Region

1.2.1 Setup and notation

The random variable Y represents a scalar outcome and its support is denoted by $\mathcal{Y} \subset \mathbb{R}$. The marginal distribution of Y is our main interest. We assume that the distribution of Y has a probability density function with respect to a dominating measure μ and we represent the distribution of Y in terms of the probability density function f_Y.[4] Note that Y need not be continuous and we can interpret $f_Y(y)$ to be a probability mass at y when μ is the point mass measure. The reason to focus on the density rather than the cdf is that the identification region for the outcome distribution has a simpler expression when the data generating process and the outcome distributions are represented in terms of densities.

The main text of this paper focuses on a binary instrument $Z \in \{1, 0\}$ since this simplifies the illustration of our main results without losing any essentials of the problem. Our analysis for the binary instrument case can be extended to the case of a multi-valued discrete instrument with finite points of support, which is covered in Appendix 1.A.6.

We do not introduce covariates X into our analysis. When the exclusion restriction of the instrument is specified in terms of conditional independence of Z and Y given X, then the identification analysis for f_Y shown below can be interpreted as the identification analysis for the outcome distribution conditional on each covariate value. Although this approach would be less practical in cases where some of the covariates are continuous, we do not discuss how to control for these covariates here.[5]

The model has missing data. The outcome Y is randomly sampled from f_Y but we do not observe all the realizations of the sampled Y. We use D to denote the selection indicator: $D = 1$ indicates Y is observed and $D = 0$ indicates Y is missing. The data is given as a random sample of $(Y \cdot D, D, Z)$.

We represent the conditional distribution of $(Y \cdot D, D)$ given $Z = 1$ by $P = (P(\cdot), P_{mis})$,

$$\begin{aligned} P(A) &\equiv \Pr(Y \in A | D = 1, Z = 1) \cdot \Pr(D = 1 | Z = 1), \quad A \subset \mathcal{Y}, \\ P_{mis} &\equiv \Pr(D = 0 | Z = 1). \end{aligned}$$

Analogously, we represent the conditional distribution of $(Y \cdot D, D)$ given $Z = 0$ by $Q = (Q(\cdot), Q_{mis})$,

$$\begin{aligned} Q(A) &\equiv \Pr(Y \in A | D = 1, Z = 0) \cdot \Pr(D = 1 | Z = 0), \quad A \subset \mathcal{Y}, \\ Q_{mis} &\equiv \Pr(D = 0 | Z = 0). \end{aligned}$$

[4] We assume that μ is known. In other words, we know the support of Y to be continuous or discrete with known points of support.

[5] When Z is presumed to be generated through a randomized mechanism, we do not need any covariate information for the purpose of identifying f_Y.

$P(\cdot)$ and $Q(\cdot)$ are the conditional distributions of the observed outcomes given Z multiplied by the selection probabilities $\Pr(D = 1|Z)$. P_{mis} and Q_{mis} are simply the missing probabilities given Z: Note that a pair of P and Q uniquely characterizes the distribution of the data except for the marginal distribution of Z, which will not play an important role for identifying f_Y. Thus, we represent the *data generating process* of our model by a pair of P and Q. On the other hand, $\Pr(\cdot)$ and $f.$ each refers to the probability distribution and the probability density of the *population* that is characterized by a value of (Y, D, Z).

We denote the density function of $P(\cdot)$ and $Q(\cdot)$ on $\mathcal{Y}$ by $p(y)$ and $q(y)$, which are linked to the population density via the following identities,

$$\begin{aligned} p(y) &= f_{Y|D,Z}(y|D=1,Z=1)\Pr(D=1|Z=1) = f_{Y,D|Z}(y,D=1|Z=1), \\ q(y) &= f_{Y|D,Z}(y|D=1,Z=0)\Pr(D=1|Z=0) = f_{Y,D|Z}(y,D=1|Z=0). \end{aligned}$$

It is important to keep in mind that the density functions $p(y)$ and $q(y)$ integrate to the selection probabilities $\Pr(D = 1|Z = 1)$ that are smaller than one. Note that without further assumptions P and Q do not reveal any information for the shape of the missing outcome distributions, $f_{Y,D|Z}(y, D = 0|Z = 1)$ and $f_{Y,D|Z}(y, D = 0|Z = 0)$, except for their integral,

$$P_{mis} = \int_{\mathcal{Y}} f_{Y,D|Z}(y, D=0|Z=1)d\mu, \qquad Q_{mis} = \int_{\mathcal{Y}} f_{Y,D|Z}(y, D=0|Z=0)d\mu.$$

The model restrictions given below are restrictions for the population distribution of (Y, D, Z).

Restriction-ER

Exclusion Restriction in the single missing outcome model (ER): Y is statistically independent of Z.

ER is a distributional restriction and cannot be represented by a finite number of moment restrictions if Y is continuous. A weaker version of instrument exogeneity common in econometrics is the mean independence restriction (MI, hereafter).

Restriction-MI

Mean Independence Restriction in the single missing outcome model (MI): Y is mean independent of Z, $E(Y|Z) = E(Y)$.

When we are mainly interested in point-identifying the mean of Y in the selection model, MI is typically sufficient and we do not require the full statistical independence (see, e.g., Andrews and Schafgans (1998)). However, in the partial identification context, these restrictions are different in terms of the identifying information for the mean since the bounds for $E(Y)$ under ER can be strictly narrower than the bounds for $E(Y)$ under MI (see Appendix 1.A.3 for further details).

ER is a *stable* restriction between the instrument and outcome while MI is not (Pearl (2000)). In other words, ER would persist for every distributional parametrization for the outcome and instrument, while MI is not preserved, for example, with respect to a nonlinear transformation of Y. Since we are often not sure about the right measure of Y so as to validate MI, it is hard to argue that an instrument satisfies MI but does not satisfy ER (e.g., can we justify the instrument with respect to which the log wage is mean independent while the raw wage is not?).

1.2.2 The identification region of f_Y under the exclusion restriction

We present the identification region of f_Y under ER. ER implies that the conditional distribution of Y given Z does not depend on Z, $f_Y = f_{Y|Z}$. By applying the law of total probability to the conditional distribution $f_{Y|Z}$, we can decompose f_Y into the conditional density of the observed Y given Z and that of the missing outcomes. Using the notation introduced above, we have

$$\begin{aligned} f_Y(y) &= f_{Y|Z}(y|Z=1) = p(y) + f_{Y,D|Z}(y, D=0|Z=1), \\ f_Y(y) &= f_{Y|Z}(y|Z=0) = q(y) + f_{Y,D|Z}(y, D=0|Z=0). \end{aligned} \tag{1.2.2.1}$$

ER allows us to interpret that the observed outcome distributions $p(y)$ and $q(y)$ provide distinct identifying information for the common f_Y. We aggregate these identifying information for f_Y by taking the envelope,

$$\underline{f}(y) \equiv \max\{p(y), q(y)\}.$$

We refer to $\underline{f}(y)$ as the *envelope density* and the area below the envelope density as the *integrated envelope* $\delta(P,Q) = \int_{\mathcal{Y}} \underline{f}(y) d\mu$.[6]

The formal definition of the identification region under ER is stated as follows.

Definition 1.2.1 (the identification region under ER) *Given a data generating process P and Q, the identification region for f_Y under ER, $IR_{f_Y}(P,Q)$, is the set of f_Y for each of which we can find a joint probability distribution of (Y, D, Z) that is compatible with the data generating process and ER.*

This definition for the identification region under ER is equivalent to *the set of f_Y that yields nonnegative missing outcome distributions $f_{Y,D|Z}(y, D=0|Z=1)$ and $f_{Y,D|Z}(y, D=0|Z=0)$ through (1.2.2.1)* (see the proof of Proposition 2.1 in Appendix A). This implies, without any restrictions on the missing outcome distribution, the conditions for f_Y to be contained in $IR_{f_Y}(P,Q)$ are $f_Y(y) \geq p(y)$ and $f_Y(y) \geq q(y)$ μ-a.e. Hence, $IR_{f_Y}(P,Q)$ is obtained as

$$IR_{f_Y}(P,Q) = \mathcal{F}^{env}_{f_Y}(P,Q) \equiv \left\{ f_Y : \int_{\mathcal{Y}} f_Y(y) d\mu = 1,\ f_Y(y) \geq \underline{f}(y)\ \mu\text{-a.e.} \right\}. \tag{1.2.2.2}$$

[6] Note that the envelope density is not a probability density function on $\mathcal{Y}$ since it does not necessarily integrate to unity.

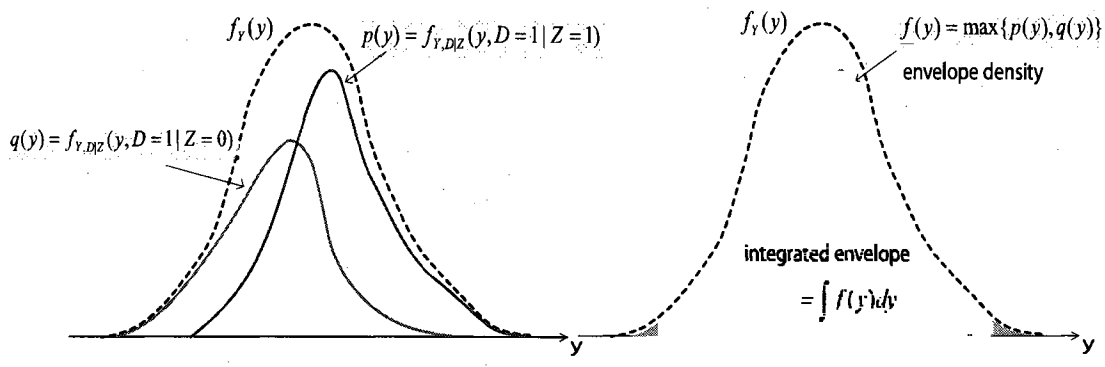

Figure 1.1: *Consider the case with a continuous Y and a binary Z. The dotted curve represents f_Y the probability density of the outcome Y. The identities (1.2.2.1) and the nonnegativity of the missing outcome densities require that the two densities $p(y)$ and $q(y)$ must lie below f_Y. This implies that any f_Y which cover both $p(y)$ and $q(y)$ are compatible with ER and the empirical evidence $p(y)$ and $q(y)$. Hence, the identification region of f_Y is obtained as the collection of the probability distributions such that the individual densities each cover both $p(y)$ and $q(y)$. The right-hand side figure shows the envelope density $\underline{f}(y) = \max\{p(y), q(y)\}$. The integrated envelope $\delta(P,Q) = \int \underline{f}(y)dy$ is the area below the envelope density (shaded area). If $\delta(P,Q)$ exceeds one, then, no probability density function can cover the entire envelope density and we obtain the empty identification region.*

Figure 1.1 provides a graphical illustration for the identification region.

Notice that $\mathcal{F}^{env}_{f_Y}(P,Q)$ becomes empty if and only if the integrated envelope $\delta(P,Q)$ exceeds one. This is because the probability density function f_Y must integrate to one by definition and there do not exist any probability distributions that cover the entire envelope if $\delta(P,Q) > 1$. Thus, refutability of ER depends only on the integrated envelope $\delta(P,Q)$ and testing the emptiness of the identification region is reduced to inferring $\delta(P,Q)$ from data.

The next proposition summarizes the identification region of f_Y and the refutability property for ER in the single missing outcome model. If Y is discrete, this proposition is reduced to Corollary 2.3 of Manski (2003).

Proposition 1.2.1 (the identification region under ER) *Assume that the probability distribution of Y has a density f_Y with respect to a dominating measure μ. Let $\underline{f}(y)$ be the envelope density and $\delta(P,Q)$ be the integrated envelope defined by*

$$\underline{f}(y) \equiv \max\{p(y), q(y)\}, \quad \delta(P,Q) \equiv \int_{\mathcal{Y}} \underline{f}(y)d\mu. \tag{1.2.2.3}$$

(i) The identification region of f_Y under ER, $IR_{f_Y}(P,Q)$, is

$$\mathcal{F}^{env}_{f_Y}(P,Q) = \left\{ f_Y : \int_{\mathcal{Y}} f_Y(y)d\mu = 1,\ f_Y(y) \geq \underline{f}(y)\ \mu\text{-a.e.} \right\}.$$

(ii) $IR_{f_Y}(P,Q)$ *is empty if and only if* $\delta(P,Q) > 1$.

When $IR_{f_Y}(P,Q)$ is nonempty, each $f_Y \in IR_{f_Y}(P,Q)$ has the representation of a mixture of two probability densities weighted by $\delta = \delta(P,Q)$,

$$f_Y(y) = \delta\left(\underline{f}(y)/\delta\right) + (1-\delta)\gamma(y), \tag{1.2.2.4}$$

where $\underline{f}(y)/\delta$ is the normalized envelope density depending only on the data generating process and $\gamma(y)$ is a probability density function that can be arbitrarily chosen to span the identification region. Thus, another way to view $IR_{f_Y}(P,Q)$ is the set of probability distributions generated from (1.2.2.4) by choosing an arbitrary probability density $\gamma(y)$.

By this way of representing $IR_{f_Y}(P,Q)$, F_Y the cdf of Y whose density belongs to $IR_{f_Y}(P,Q)$ is written as

$$F_Y(y) = \int_{(-\infty,y]} \underline{f}(t)d\mu + (1-\delta)\Gamma(y),$$

where $\Gamma(\cdot)$ is the cdf of $\gamma(\cdot)$. Since we can choose any values between zero and one for $\Gamma(y)$, the tight cdf bounds of Y are obtained as

$$\int_{(-\infty,y]} \underline{f}(t)d\mu \leq F_Y(y) \leq \int_{(-\infty,y]} \underline{f}(t)d\mu + 1 - \delta. \tag{1.2.2.5}$$

Note that these cdf bounds can be strictly narrower than the cdf bounds constructed in Blundell et al. (2007) (see Appendix 1.A.2).

The tight bounds for the mean $E(Y)$ also follow from (1.2.2.4). Let Y have a compact support $\mathcal{Y} = [y_l, y_u]$. By specifying $\gamma(y)$ as the degenerate distribution at the lower or upper bound of the outcome support, we obtain the tight bounds for $E(Y)$ under ER,

$$(1-\delta)y_l + \int_{\mathcal{Y}} y\underline{f}(y)d\mu \leq E(Y) \leq \int_{\mathcal{Y}} y\underline{f}(y)d\mu + (1-\delta)y_u. \tag{1.2.2.6}$$

Since ER is stronger than MI, these mean bounds are equally or strictly narrower than the tight mean bounds under MI constructed in Manski (1994). In Appendix 1.A.3, we compare the tight bounds of $E(Y)$ obtained from the exclusion restriction with the ones obtained from the mean independence restriction. A sufficient condition for these two bounds for $E(Y)$ to be identical is that the data generating process reveals either $p(y) \geq q(y)$ μ-a.e. or $q(y) \geq p(y)$ μ-a.e., that is, one of the observed densities covers the other.

1.2.3 Does selection equation help to identify f_Y?

The structural selection model formulates the selection mechanism as

$$D = I\{v(Z,U) \geq 0\}, \tag{1.2.3.7}$$

where $v(Z,U)$ is the latent utility to rationalize the individual selection process, and U represents the unobserved individual heterogeneities that affect one's selection response and are possibly dependent on the outcome Y. Recall that ER only implies independence between the outcome Y and instrument Z, while it is silent about a statistical relationship between the selection heterogeneity U and instrument Z. In the case where we believe Z to be independent of any individual heterogeneities, we might want to explicitly impose joint independence between Z and (Y,U). In that case, can we further narrow the identification region by strengthening ER to joint independence?

As we will discuss further in Section 1.3.3, an importance of this question can be highlighted by a comparison with the counterfactual causal model with endogenous treatment choice (Imbens and Angrist (1994) and Angrist et al. (1996)). Given a pair of treated and control outcomes (Y_1, Y_0) with the nonseparable selection equation (1.2.3.7), the joint independence restriction between Z and (Y_1, Y_0, U) yields a narrower identification region than marginal independence of Z and (Y_1, Y_0) for the distribution of the potential outcomes.[7] The main focus of this section is to investigate whether or not the single missing outcome model can enjoy a similar identification gain from the joint independence restriction.

When we introduce latent utility with unobserved heterogeneities U into the model, we characterize the population by a joint distribution of (Y,D,U,Z) rather than (Y,D,Z). In particular, if the instrument Z is binary, the population random variables (Y,D,U,Z) can be replaced with (Y,T,Z), where T is the individual type that indicates one's selection response to each value of the instrument as defined in Imbens and Angrist (1994) (see also Pearl (1994a)). Define the *potential selection indicator* D_z, $z = 1,0$, representing one's selection response when the instrument was set to $Z = z$, i.e., $D_z = I\{v(z,U) \geq 0\}$. The category variable of *individual type* T is defined as[8]

$$T = \begin{cases} c: & \text{complier} & \text{if } D_1 = 1,\ D_0 = 0, \\ n: & \text{never-taker} & \text{if } D_1 = D_0 = 0, \\ a: & \text{always-taker} & \text{if } D_1 = D_0 = 1, \\ d: & \text{defier} & \text{if } D_1 = 0,\ D_0 = 1. \end{cases}$$

and joint independence of Z and (Y,U) is equivalently stated as joint independence of Z and (Y,T). We call this joint independence restriction Random Assignment Restriction (RA).

Restriction-RA

[7] Balke and Pearl (1997) analyzed bounding the average treatment effect when Y_1 and Y_0 are binary under the restriction of joint independence $(Y_1, Y_0, U) \perp Z$. They show that the bounds for the average treatment effect can be further narrowed than the average treatment effect bounds of Manski (1994) under mean independence. Note that, when Y_1 and Y_0 are binary, the latter bounds for the average treatment effect are interpreted as the treatment effect bounds under $Y_1 \perp Z$ and $Y_0 \perp Z$.

[8] Although the single missing outcome model is not the counterfactual causal model, we name each type as in Imbens and Angrist.

Random Assignment Restriction in the single missing outcome model (RA): Z is jointly independent of (Y,T).

The definition of the identification region under RA is provided as follows.

Definition 1.2.2 (the identification region under RA) *Given a data generating process P and Q, the identification region for f_Y under the random assignment restriction (RA) is the set of f_Y for each of which we can find a joint probability distribution of (Y,T,Z) that is compatible with the data generating process and the joint independence restriction of Z and (Y,T).*

Appendix 1.A.4 provides a formal analysis on the construction of the identification region under RA. The main result is stated in the next proposition.

Proposition 1.2.2 (invariance of the identification region) *The identification region under ER, $IR_{f_Y}(P,Q)$, is also the identification region of f_Y under RA.*

This proposition shows that a further identification gain from the joint independence restriction, which exists in the counterfactual causal model with an instrument as we mentioned above, does *not* exist in the selection model with a single missing outcome. This redundancy of the joint independence restriction implies that ER is the only refutable restriction for the instrument exogeneity.

An additional restriction we consider is a functional form specification for latent utility. In the standard structural selection model, we specify the selection equation in the form of threshold crossing selection with an additive error,

$$v(Z,U) = \tilde{v}(Z) - U, \tag{1.2.3.8}$$

where U is a scalar and $\tilde{v}(Z)$ depends only on the instrument. Heckman and Vytlacil (2001a, 2001b) show that the expression of the bounds of $E(Y)$ under mean independence constructed in Manski (1994) provides the tight bounds even under the joint independence between Z and (Y,U) and the specification of the additively separable latent utility. This result is somewhat surprising since the tight $E(Y)$ bounds under ER can be strictly narrower than the $E(Y)$ bounds under MI, but the latter becomes the tightest once we impose the joint independence of Z and (Y,U) and threshold crossing with an additive error. We disentangle this puzzle using the expression of the identification region obtained through the envelope density.

By noting the equivalence result of Vytlacil (2002), the selection process with additively separable latent utility can be equivalently analyzed by imposing the monotonicity of Imbens and Angrist (1994). Hence, the definition of the tight identification region in this case is defined as follows.

Definition 1.2.3 (the identification region under separable utility) *Given a data generating process P and Q, the identification region for f_Y under RA and the specification of threshold crossing selection with an additive error is the set of f_Y for each of which we can find a joint probability distribution of (Y,T,Z) that is compatible with the data generating process and satisfies RA with either* $\Pr(T=d)=0$ *or* $\Pr(T=c)=0$.

In Appendix 1.A.5, we derive the identification region for f_Y under these two restrictions. The resulting identification region for f_Y is given in the next proposition.

Proposition 1.2.3 (the identification region under separable utility) *The identification region under RA and the specification of threshold crossing selection with an additive error is*

$$\begin{cases} \mathcal{F}^{env}_{f_Y}(P,Q) & \text{if } p(y) \geq q(y) \ \mu\text{-a.e. or } q(y) \geq p(y) \ \mu\text{-a.e.} \\ \emptyset & \text{otherwise.} \end{cases} \tag{1.2.3.9}$$

This result says that if the data generating process reveals either $p(y) \geq q(y)$ μ-a.e. or $q(y) \geq p(y)$ μ-a.e., the identification region under ER is also the identification region under the restrictions of joint independence and additively separable latent utility. In this sense, threshold crossing selection with an additive error *does not contribute to identifying f_Y further than ER.* This result supports the aforementioned Heckman and Vytlacil's result on the $E(Y)$ bounds since, as already mentioned in Section 2.2, given we observe either $p(y) \geq q(y)$ μ-a.e. or $q(y) \geq p(y)$ μ-a.e., the $E(Y)$ bounds constructed from $\mathcal{F}^{env}_{f_Y}(P,Q)$ coincide with the Manski's $E(Y)$ bounds under MI.

The empty identification region in (1.2.3.9) implies that if joint independence and threshold crossing selection with an additive error hold in the population, we must observe either $p(y) \geq q(y)$ μ-a.e. or $q(y) \geq p(y)$ μ-a.e. In other words, the structural selection model with additively separable latent utility *constrains the data generating process in such a way that either $p(y)$ or $q(y)$ covers the other on the entire $\mathcal{Y}$* (see Figure 1.2 for a visual illustration of the observed densities for this case). Note that the condition of $p(y) \geq q(y)$ μ-a.e. or $q(y) \geq p(y)$ μ-a.e. provides a testable implication for the joint restriction of joint independence and additively separable latent utility. That is, we can refute it by checking whether or not one of the observable densities $p(y)$ or $q(y)$ nests the other.[9]

The envelope density provides the maximal identifying information for f_Y based only on the empirical evidence, and optimality of this aggregating scheme is free from the assumptions that only constrain the data generating process.

[9] Chapter 3 proposes a test procedure for whether the density $p(y)$ nests $q(y)$ in the context of the counterfactual causal model with a binary instrument. This is interpreted as a test for point-identifiability of the local average treatment effect.

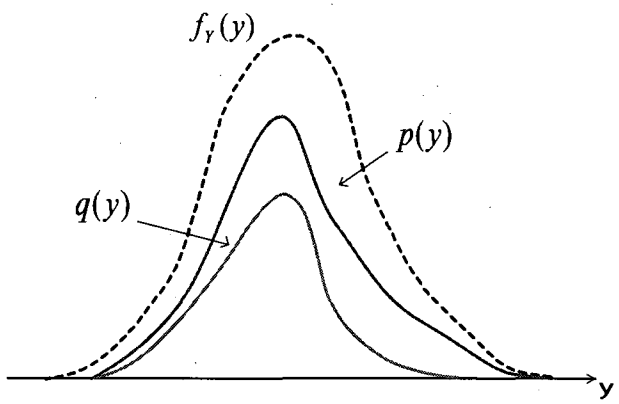

Figure 1.2: *If the instrument is jointly independent of Y and the unobserved heterogeneities in the latent utility, and threshold crossing selection with an additive error holds in the population, then we must observe that either $p(y)$ or $q(y)$ covers the other on the entire $\mathcal{Y}$, as drawn above. Note that this figure also shows the case where the tight mean bounds under ER are identical to the tight mean bounds under MI (see Appendix 1.A.3).*

1.3 Counterfactual Causal Model: The Identification Region

The previous section focused on the selection problem in the missing data context. This section considers extending the use of envelope density in constructing the identification region to the heterogeneous treatment effect model with a binary treatment. In the Rubin-causal model (Rubin (1974)), causal effects are defined in terms of a parameter of the potential outcome distributions. In this section, we construct the identification region for the distribution of potential outcomes when the researcher has an instrumental variable and he is willing to impose the restriction of its exogeneity to the potential outcomes.

1.3.1 Setup and Notation

Let $Y_1 \in \mathbb{R}$ be treatment outcome and $Y_0 \in \mathbb{R}$ be control outcome. In data, we observe Y_1 if one receives a treatment while we observe Y_0 if one does not receive the treatment. In this section, we read D as the treatment indicator: $D = 1$ if one receives the treatment and $D = 0$ otherwise. We denote the observed outcome by $Y_{obs} \equiv DY_1 + (1 - D)Y_0$ and we consider a nondegenerate binary instrumental variable $Z \in \{1, 0\}$. The data is given as a random sample of (Y_{obs}, D, Z).[10]

Our object of interest is a distribution of the potential outcomes. Since it is common to evaluate causal effects by comparing the two marginal distributions of potential outcomes, our interest lies in f_{Y_1} and f_{Y_0} the marginal distributions of Y_1 and Y_0. For example, the *average treatment effect,*

[10]The results presented in this section are limited to the case with a binary instrument. With a binary instrument, we can obtain a closed representation of the identification region for potential outcome distributions. Although it is possible to extend the framework to the case with a milti-valued discrete instrument, a closed form representation of the identification region will be more complex.

which is one of a common measure of treatment effect, is defined as the difference between the mean of f_{Y_1} and f_{Y_0}.[11] Each f_{Y_1} and f_{Y_0} represents the probability density on $\mathcal{Y} \subset \mathbb{R}$ and f_{Y_1} and f_{Y_0} are assumed to have a dominating measure μ. Given that our interest is to bound these causal effect parameters, we focus on constructing the identification region of a pair of f_{Y_1} and f_{Y_0} (the formal definition is given below).

We denote a conditional distribution of (Y_{obs}, D) given Z by

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

Note that the conditional distribution of $(Y_{obs}, D = d)$ given Z provides the probabilities of the event $\{Y_d \in A, D = d\}$ given Z for $d = 1, 0$. Since $P = (P_{Y_1}(\cdot), P_{Y_0}(\cdot))$ and $Q = (Q_{Y_1}(\cdot), Q_{Y_0}(\cdot))$ uniquely characterizes the distribution of data except for the marginal distribution of Z, we represent the data generating process in the counterfactual causal model by (P, Q). The density functions of $P_{Y_d}(\cdot)$ and $Q_{Y_d}(\cdot)$ are denoted by $p_{Y_d}(\cdot)$ and $q_{Y_d}(\cdot)$, $d = 1, 0$,

$$\begin{aligned}
p_{Y_1}(y_1) &\equiv f_{Y_1, D|Z}(y_1, D = 1 | Z = 1), \\
p_{Y_0}(y_0) &\equiv f_{Y_0, D|Z}(y_0, D = 0 | Z = 1), \\
q_{Y_1}(y_1) &\equiv f_{Y_1, D|Z}(y_1, D = 1 | Z = 0), \\
q_{Y_0}(y_0) &\equiv f_{Y_0, D|Z}(y_0, D = 0 | Z = 0).
\end{aligned} \tag{1.3.1.10}$$

Our analysis of this section is an extension of the one in Section 1.2.3, where we explicitly introduce the heterogeneities in the selection response. That is, the model has the structural selection equation with nonseparable latent utility,

$$D = I\{v(Z, U) \geq 0\}.$$

By the same argument as in Section 1.2.3, the population is characterized by a joint distribution of (Y_1, Y_0, T, Z) where T is the individual type defined above.

We define exogeneity of the instrument in this context in terms the statistical independence of the instrument and the two potential outcomes.

Restriction-RA-causal

Random Assignment Restriction in the causal model (RA-causal): Z is jointly independent of (Y_1, Y_0, T).

[11] The *quantile differences* between the two potential outcome distributions can be also a parameter of interest. When we impose the assumption of perfect ranking, that is, the ranking of individuals based on Y_1 is the same as the ranking based on Y_0, then the τ-th quantile difference can be interpreted as the causal effect for the individual whose ranking in terms of potential outcomes is τ.

The restriction RA-causal can be seen as an analogue of the random assignment restriction in the single missing outcome model. RA-causal states that the instrument is randomized regardless of one's potential outcomes and one's selection response and this restriction is standard in the literature of heterogeneous treatment effect model with self-selection (Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1997), Heckman and Vytlacil (2001, 2005, 2008)).

1.3.2 Identification Region of the Potential Outcome Distributions under Instrument Independence

Let $\underline{f_{Y_1}}(y_1)$ be the envelope density for the treated outcomes, $\underline{f_{Y_1}}(y_1) = \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}$ and δ_{Y_1} be its integrated envelope $\delta_{Y_1} = \int_{\mathcal{Y}} \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}d\mu$. Similarly, for control outcomes let $\underline{f_{Y_0}}(y_0) = \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}$ and $\delta_{Y_0} = \int_{\mathcal{Y}} \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}d\mu$. If we naively apply the formula of the sharp mean bounds (1.2.2.6) to each potential outcome, the bounds for $E(Y_1)$ and $E(Y_0)$ can be given by

$$(1-\delta_{Y_1})y_l + \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu \leq E(Y_1) \leq \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu + (1-\delta_{Y_1})y_u, \quad (1.3.2.11)$$

$$(1-\delta_{Y_0})y_l + \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0)d\mu \leq E(Y_0) \leq \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0)d\mu + (1-\delta_{Y_0})y_u. \quad (1.3.2.12)$$

Given bounds of each $E(Y_1)$ and $E(Y_0)$, Manski (2003) considers the *outer bounds*; bounding the average treatment effect $E(Y_1) - E(Y_0)$ by taking the difference between the upper or lower bound of $E(Y_1)$ and the lower or upper bound of $E(Y_0)$.

$$\begin{aligned} & (1-\delta_{Y_1})y_l + \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0)d\mu - (1-\delta_{Y_0})y_u \\ \leq\ & E(Y_1) - E(Y_0) \\ \leq\ & \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu + (1-\delta_{Y_1})y_u - (1-\delta_{Y_0})y_l - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0)d\mu. \end{aligned} \quad (1.3.2.13)$$

Our analysis given below shows that this way of constructing the bounds are *not* necessarily tight under the restriction of RA-causal.

In order to derive the tight bounds for the average treatment effect, we first state the definition of the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal.

Definition 1.3.1 (the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal) *Given a data generating process P and Q, the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal is the set of (f_{Y_1}, f_{Y_0}) for each of which we can construct a joint distribution of (Y_1, Y_0, T, Z) that is compatible with the data generating process and satisfies the independence restriction of (Y_1, Y_0, T) and Z .*

The identification region defined here is a collection of a pair of two marginal distributions (f_{Y_1}, f_{Y_0}) rather than a collection of the joint distribution of (Y_1, Y_0). Focusing on the former is

sufficient to build the tight bounds for the aforementioned causal effects since they are written as a functional of solely f_{Y_1} and f_{Y_0}.

For a joint distribution of (Y_1, Y_0, T, Z) to be compatible with the data generating process P and Q, it must satisfy the equalities (1.3.1.10). In terms of the distribution of (Y_1, Y_0, T, Z), the equalities (1.3.1.10) are written as,

$$\begin{aligned} p_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1, T = c|Z = 1) + f_{Y_1,T|Z}(y_1, T = a|Z = 1), \\ q_{Y_1}(y_1) &= f_{Y_1,T|Z}(y_1, T = d|Z = 0) + f_{Y_1,T|Z}(y_1, T = a|Z = 0), \\ p_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0, T = d|Z = 1) + f_{Y_0,T|Z}(y_0, T = n|Z = 1), \\ q_{Y_0}(y_0) &= f_{Y_0,T|Z}(y_0, T = c|Z = 0) + f_{Y_0,T|Z}(y_0, T = n|Z = 0). \end{aligned} \tag{1.3.2.14}$$

We leave a formal development of the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal to Appendix 1.A.7 and the main text primarily focuses on its heuristic construction. Suppose that the integrated envelopes δ_{Y_1} and δ_{Y_0} are both less than one.[12] Consider an arbitrary pair of two marginal distributions (f_{Y_1}, f_{Y_0}) such that each covers the envelope density of Y_1 and Y_0, that is, $f_{Y_1}(y_1) \geq \underline{f_{Y_1}}(y_1) = \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}$ μ-a.e. and $f_{Y_0}(y_0) \geq \underline{f_{Y_0}}(y_0) = \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}$ μ-a.e.

In order to claim that the proposed (f_{Y_1}, f_{Y_0}) is contained in the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal, we have to show that there exists a joint distribution of (Y_1, Y_0, T, Z) that can generate the data, satisfies RA-causal, and its marginal distributions of Y_1 and Y_0 coincide with the proposed (f_{Y_1}, f_{Y_0}). Since RA-causal implies $f_{Y_1,T|Z} = f_{Y_1,T}$ and $f_{Y_0,T|Z} = f_{Y_0,T}$, candidate distributions of (Y_1, Y_0, T, Z) must yield the equalities (1.3.2.14) without the conditioning variable Z,

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

Furthermore. $f_{Y_1}(y_1) = \sum_{t\in\{c,n,a,d\}} f_{Y_1,T}(y_1, T = t)$ and $f_{Y_0}(y_0) = \sum_{t\in\{c,n,a,d\}} f_{Y_0,T}(y_0, T = t)$ imply

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

[12] This is required in order to have a nonempty identification region since otherwise no density can cover the observed part of Y_1 or Y_0's densitiies and this leads to a violation of independence between Y_1 and Z or Y_0 and Z.

These equalities suggest that given (f_{Y_1}, f_{Y_0}) and (P,Q) 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* $\int h_{Y_1,t}(y_1)d\mu = \int h_{Y_0,t}(y_0)d\mu$ and

$$\begin{aligned}
p_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,a}(y_1), \\
q_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,a}(y_1), \\
p_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,n}(y_0), \\
q_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,n}(y_0), \\
f_{Y_1}(y_1) - p_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,n}(y_1), \\
f_{Y_1}(y_1) - q_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,n}(y_1), \\
f_{Y_0}(y_0) - p_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,a}(y_0), \\
f_{Y_0}(y_0) - q_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,a}(y_0),
\end{aligned} \tag{1.3.2.15}$$

then by setting $f_{Y_1,T|Z}(y_1, T = t|Z = z) = h_{Y_1,t}(y_1)$ and $f_{Y_0,T|Z}(y_0, T = t|Z = z) = h_{Y_0,t}(y_1)$, $t = c,n,a,d$, $z = 1,0$, we can construct a population distribution of (Y_1, Y_0, T, Z) without contradicting RA-causal and the data generating process (P,Q). Thus, given data (P,Q) the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal consists of (f_{Y_1}, f_{Y_0}) for each of which we can find these nonnegative functions $(h_{Y_1,t}(y_1), h_{Y_0,t}(y_0))$, $t = c,n,a,d$, that satisfy the scale constraints $\int h_{Y_1,t}(y_1)d\mu = \int h_{Y_0,t}(y_0)d\mu$.

A closed form expression of the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal is given in the next proposition.

Proposition 1.3.1 (Identification region of (f_{Y_1}, f_{Y_0}) under RA-causal) *Let $\underline{f_{Y_1}}(y_1)$ and $\delta_{Y_1}(P,Q)$ be the envelope density and the integrated envelope for the observed treated outcome distributions*

$$\underline{f_{Y_1}}(y_1) = \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}, \qquad \delta_{Y_1} = \int_{\mathcal{Y}} \max\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}dy_1.$$

and $\underline{f_{Y_0}}(y_0)$ and $\delta_{Y_0}(P,Q)$ be the envelope density and the integrated envelope for the control outcome distribution,

$$\underline{f_{Y_0}}(y_0) = \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}, \qquad \delta_{Y_0} = \int_{\mathcal{Y}} \max\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}dy_0.$$

Let $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ be the set of probability densities that cover the envelope densities $\underline{f_{Y_1}}(y_1)$ and $\underline{f_{Y_0}}(y_0)$,

$$\begin{aligned}
\mathcal{F}^{env}_{f_{Y_1}}(P,Q) &= \left\{ f_{Y_1} : f_{Y_1} \geq \underline{f_{Y_1}} \ \mu\text{-a.e. and } \int_{\mathcal{Y}} f_{Y_1}(y_1)dy_1 = 1 \right\}, \\
\mathcal{F}^{env}_{f_{Y_0}}(P,Q) &= \left\{ f_{Y_0} : f_{Y_0} \geq \underline{f_{Y_0}} \ \mu\text{-a.e. and } \int_{\mathcal{Y}} f_{Y_0}(y_0)dy_0 = 1 \right\}.
\end{aligned}$$

Further, define

$$\lambda_{Y_1} = \int_{\mathcal{Y}} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\} dy_1, \;\; \lambda_{Y_0} = \int_{\mathcal{Y}} \min\{p_{Y_0}(y_0), q_{Y_0}(y_0)\} dy_0.$$

The identification region of (f_{Y_1}, f_{Y_0}) *under RA-causal is nonempty if and only if* $\delta_{Y_1} \leq 1$ *and* $\delta_{Y_0} \leq 1$, *and it is given by,*
(i) for $1 - \delta_{Y_0} < \lambda_{Y_1}$,

$$\left\{(f_{Y_1}, f_{Y_0}) : f_{Y_1} \in \mathcal{F}^*_{f_{Y_1}}(P,Q),\; f_{Y_0} \in \mathcal{F}^{env}_{f_{Y_0}}(P,Q)\right\},$$

where

$$\mathcal{F}^*_{f_{Y_1}}(P,Q) = \left\{f_{Y_1} : f_{Y_1} \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q),\; \int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}\right\} dy_1 \geq \lambda_{Y_1} + \delta_{Y_0} - 1\right\},$$

(ii) for $1 - \delta_{Y_0} > \lambda_{Y_1}$,

$$\left\{(f_{Y_1}, f_{Y_0}) : f_{Y_1} \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q),\; f_{Y_0} \in \mathcal{F}^*_{f_{Y_0}}(P,Q)\right\},$$

where

$$\mathcal{F}^*_{f_{Y_0}}(P,Q) = \left\{f_{Y_0} : f_{Y_0} \in \mathcal{F}^{env}_{f_{Y_0}}(P,Q),\; \int \min\left\{f_{Y_0} - \underline{f_{Y_0}}, \min\{p_{Y_0}(y_0), q_{Y_0}(y_0)\}\right\} dy_0 \geq 1 - \delta_{Y_0} - \lambda_{Y_1}\right\},$$

(iii) for $1 - \delta_{Y_0} = \lambda_{Y_1}$,

$$\left\{(f_{Y_1}, f_{Y_0}) : f_{Y_1} \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \text{ and } f_{Y_0} \in \mathcal{F}^{env}_{f_{Y_0}}(P,Q)\right\}.$$

This proposition states that the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal is smaller than the product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. Specifically, if the data generating process reveals $1 - \delta_{Y_0} \neq \lambda_{Y_1}$, then the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal is strictly smaller than the product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. For example, in case of $1 - \delta_{Y_0} < \lambda_{Y_1}$, then the collection of feasible f_{Y_1} is given by

$$\mathcal{F}^*_{f_{Y_1}}(P,Q) = \left\{f_{Y_1} : f_{Y_1} \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q),\; \int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}\right\} dy_1 \geq \lambda_{Y_1} + \delta_{Y_0} - 1\right\},$$

which is strictly smaller than $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ due to the additional constraint

$$\int \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}\right\} dy_1 \geq \lambda_{Y_1} + \delta_{Y_0} - 1.$$

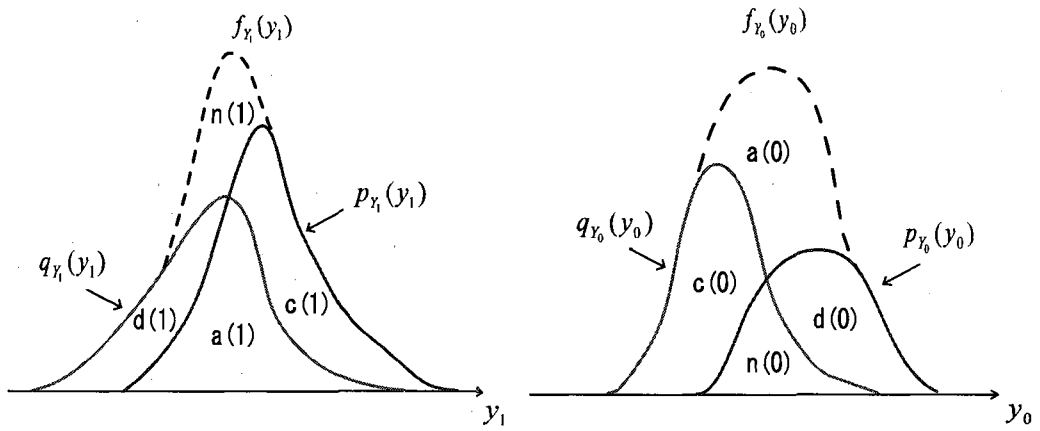

Figure 1.3: This figure depicts the data gerating process with $1-\delta_{Y_0} = \lambda_{Y_1}$ (the area of $a(0)$ is equal to the area of $a(1)$), which corresponds to the case (iii) in Proposition 1.3.1. For each $t = c, n, a, d$, $t(1)$ and $t(0)$ have the same area.

In order to illustrate where this constraint comes from, consider the data generating process that yields the identification region as the product $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ (Figure 1.3). Here, (P,Q) satisfies $1-\delta_{Y_0} = \lambda_{Y_1}$, that is, the area between f_{Y_0} and $\underline{f_{Y_0}}(y_0)$ is equal to the area below $\min\{p_1(y_1), q_1(y_1)\}$. Consider partitioning the subgraph of f_{Y_1} into four denoted as $c(1)$, $a(1)$, $n(1)$, and $d(1)$. Consider also partitioning the subgraph of f_{Y_0} into four denoted as $c(0)$, $a(0)$, $n(0)$, and $d(0)$. The condition $1-\delta_{Y_0} = \lambda_{Y_1}$ implies that the area of $a(1)$ is equal to the area of $a(0)$. In addition, we can show that not only $a(1)$ and $a(0)$ but also $c(1)$ and $c(0)$, $n(1)$ and $n(0)$, and $d(1)$ and $d(0)$ share the same area. This allows us to impute $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$ as the height of the partitions $t(1)$ and $t(0)$ for each $t = c, n, a, d$. Then, the equalities of (1.3.2.15) are all satisfied. 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}) as long as f_{Y_1} and f_{Y_0} each cover the envelope density. Hence the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal is derived as the product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$.

On the other hand, consider the case of $1-\delta_{Y_0} \neq \lambda_{Y_1}$. The above way of pinning down $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$ cannot satisfy the scale restriction $\int_{\mathcal{Y}} h_{Y_1,t}(y_1)d\mu = \int_{\mathcal{Y}} h_{Y_1,t}(y_1)d\mu$ and hence we need to develop a different way of constructing $h_{Y_1,t}(y_1)$ and $h_{Y_0,t}(y_0)$. Figure 1.4 draws a data generating process with $1-\delta_{Y_0} < \lambda_{Y_1}$, that is, the area of $a(0)$ is smaller than the area of $a(1)$. Since $1-\delta_{Y_0} < \lambda_{Y_1}$ is equivalent to $\lambda_{Y_0} < 1-\delta_{Y_1}$,[13] the third and fourth constraints of (1.3.2.15) imply that feasible $h_{Y_0,n}$ must satisfy $\int h_{Y_0,n}(y_0)d\mu \leq \lambda_{Y_0} < 1-\delta_{Y_1}$, and this in turn implies that the entire area of $n(1)$ cannot be occupied by $h_{Y_1,n}(y_0)$ since otherwise the equal area restriction is violated. If the identification region for f_{Y_1} under RA-causal was $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$, then, we would be able to allocate the partition $n(1)$ freely so as to span $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$. In order to do this, $n(1)$ must be filled by $h_{Y_1,n}(y_1)$ since $h_{Y_1,n}(y_1)$ is the only density whose shape is completely unrestricted. But

[13]Lemma 1.A.3 in Appendix shows $\lambda_{Y_1}+\lambda_{Y_0}+\delta_{Y_1}+\delta_{Y_0} = 2$. Hence, $1-\delta_{Y_0} < \lambda_{Y_1}$ is equivalent to $\lambda_{Y_0} < 1-\delta_{Y_1}$.

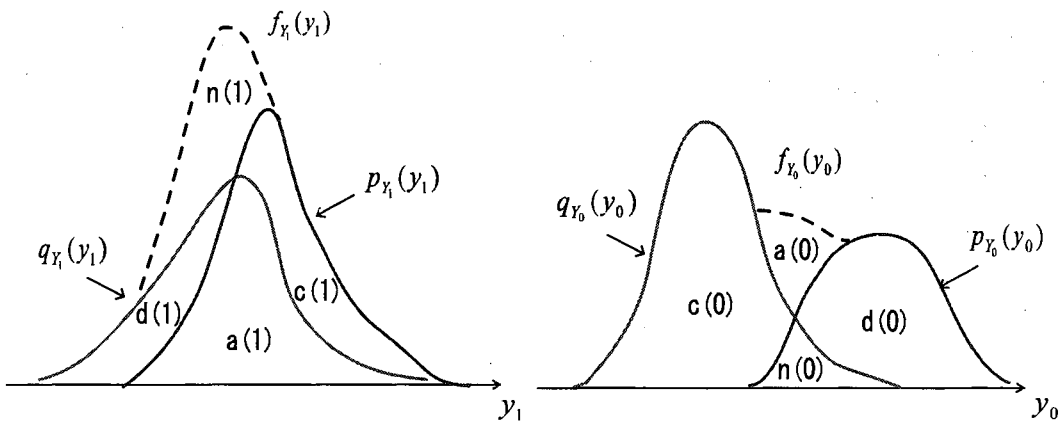

Figure 1.4: The drawn data generating process satisfies $1 - \delta_{Y_0} < \lambda_{Y_1}$ (the area of $a(0)$ is strictly smaller than the area of $a(1)$). Different from the case drawn in Figure 1.3, it is not feasible to pin down $(h_{Y_1,t}, h_{Y_0,t})$ to $(t(1), t(0))$ for each $t = c, n, a, d$.

as was already mentioned, this is not feasible for the given data generating process. Therefore, the identification region for f_{Y_1} must be strictly smaller than $\mathcal{F}^{env}_{f_{Y_1}}(P, Q)$.

Figure 1.5 illustrates the closed-form expression of the identification region obtained in Proposition 1.3.1. Differences from Figure 1.4 are that $a(1)$ is further partitioned into a and $d\&c$, and $n(1)$ is partitioned into n and $d\&c'$. Consider, for each $t \in \{c, n, a, d\}$, we pin down $h_{Y_0,t}(y_0)$ to one of the partitions $t(0)$ in the subgraph of f_{Y_0}. This way of pinning down $h_{Y_0,t}(y_0)$ allows us to span $\mathcal{F}^{env}_{f_{Y_0}}(P, Q)$ since there is no restriction on the shape of $a(0)$. Next, we take a subset a within $a(1)$ so that it shares the same area as $a(0)$, and pin down $h_{Y_1,a}$ to this partition a. Accordingly, the first two constraints of (1.3.2.15) imply $h_{Y_1,c}(y_1)$ is imputed as the sum of $c(1)$ and $d\&c$ and $h_{Y_1,d}(y_1)$ is imputed as the sum of $d(1)$ and $d\&c$. As the types $t = c$ and $t = d$ occupy the area $d\&c$, the area that can be potentially taken by $h_{Y_1,n}$ decreases by the area of $d\&c$ due to the fifth and sixth constraints of (1.3.2.15). This operation can be interpreted as piling up the partition $d\&c$ onto the envelope density, and as a result, the area $d\&c'$ emerges as drawn in Figure 1.5. The subset $d\&c'$ is a copy of $d\&c$ (the height of $d\&c'$ is equal to the height of $d\&c$ at every y_1) and the fifth and sixth constraints of (1.3.2.15) pins down $h_{Y_1,n}(y_1)$ to n, which is the left-over part of $n(1)$ after $d\&c'$. What we can learn from this exercise is that f_{Y_1} in the identification region must spare a enough room for $d\&c'$ above the envelope density. By noting that the possible shape of $d\&c'$ is constrained by $a(1)$, the subset $d\&c'$ can be found for a given f_{Y_1} as long as the area outlined between $\min\{f_{Y_1}, \underline{f_{Y_1}} + \min\{p_{Y_1}, q_{Y_1}\}\}$ and f_{Y_1} is greater than the area of $a(1)$ minus the area of a. This condition is equivalent to $\int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}\right\} dy_1 \geq \lambda_{Y_1} + \delta_{Y_0} - 1$, which appears in the construction of $\mathcal{F}^{*}_{f_{Y_1}}(P, Q)$.

The preceding argument shows that the key constraints that contribute to further narrowing the identification region for f_{Y_1} than $\mathcal{F}^{env}_{f_{Y_1}}(P, Q)$ are the scale constraints, $\int_{\mathcal{Y}} h_{Y_1,t}(y_1) d\mu =$

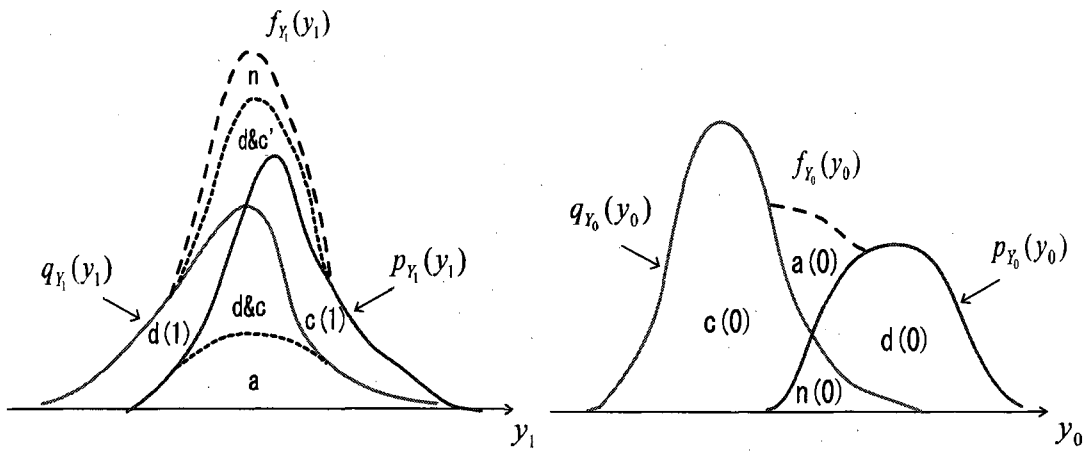

Figure 1.5: The drawn data generating process is the same as Figure 1.4.

$\int_{\mathcal{Y}} h_{Y_1,t}(y_1)d\mu$. Recall that these scale restrictions stem from the independence restriction of T and Z. Thus we can interpret that, when T and Z are assumed to be independent, the observed distributions of Y_0 provide a partial information on the marginal distribution of T and through which we obtain identification gain for the distribution of Y_1. If we only impose independence of (Y_1, Y_0) and Z, which is weaker than RA-causal, then, it can be shown that identification region of (f_{Y_1}, f_{Y_0}) is obtained as the product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ no matter what (P,Q) looks like. This also clarifies the identification power of joint independence between (Y_1, Y_0, T) and Z relative to independence between (Y_1, Y_0) and Z. Recall that in the single missing outcome case, we showed that the restriction of joint independence of (Y,T) and Z does not further narrow the identification region of f_Y than the marginal independence of Y and Z (Proposition 1.2.2). When we consider the causal model, an analogous conclusion is not true and the identification region in general differs between joint independence of (Y_1, Y_0, T) and Z and a weaker restriction of independence between (Y_1, Y_0) and Z.

An important case where $1 - \delta_{Y_0} = \lambda_{Y_1}$ holds is that the data generating process exhibits the nesting structure among densities, i.e., when we observe $p_{Y_1}(y_1) \geq q_{Y_1}(y_1)$ and $q_{Y_0}(y_0) \geq p_{Y_0}(y_0)$, it holds $1 - \delta_{Y_0} = \Pr(D = 1|Z = 0) = \lambda_{Y_1}$.

By applying the same argument as Proposition 1.2.3 to the counterfactual causal model, it can be shown that specifying the selection equation to be additively separable $D = 1\{\tilde{v}(Z) - U \geq 0\}$ restricts the data generating process in such a way that the densities of P and Q exhibit the nesting configuration. Consequently, the specification of additively separable latent utility does not contribute to further narrowing the identification region than the one obtained in Proposition 1.3.1. In order to be precise about the identification region under RA-causal and additively separable latent utility, its formal definition is stated.

Definition 1.3.2 (Identification region of (f_{Y_1}, f_{Y_0}) under RA-causal and separable utility) *Without loss of generality, assume* $\Pr(D=1|Z=1) \geq \Pr(D=1|Z=0)$. *Given the data generating process* (P,Q), *the identification region of* (f_{Y_1}, f_{Y_0}) *under RA-causal and the additively separable latent utility is the collection of* (f_{Y_1}, f_{Y_0}) *for each of which we can construct a joint probability distribution of* (Y_1, Y_0, T, Z) *that satisfies joint independence of* (Y_1, Y_0, T) *and* Z *and* $\Pr(T=d)=0$.

The next proposition summarizes the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal and separable utility.

Proposition 1.3.2 *The identification region of* (f_{Y_1}, f_{Y_0}) *under RA-causal and additively separable latent utility is*

$$\begin{cases} \left\{(f_{Y_1}, f_{Y_0}) : f_{Y_1} \in \mathcal{F}^{env}_{f_{Y_1}}(P,Q),\ f_{Y_0} \in \mathcal{F}^{env}_{f_{Y_0}}(P,Q)\right\} \\ \qquad\qquad \text{if } p_{Y_1}(y_1) \geq q_{Y_1}(y_1) \ \ \mu\text{-a.e. and } \ q_{Y_0}(y_0) \geq p_{Y_0}(y_0) \ \ \mu\text{-a.e.}, \\ \emptyset \qquad\qquad \text{otherwise} \end{cases}$$

1.3.3 Bounding Causal Effects

The outer bounds of the average treatment effects (1.3.2.13) becomes the tight bounds under RA-causal if the identification region of (f_{Y_1}, f_{Y_0}) is given as the product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. However, the preceding analysis showed that the identification region is not necessarily the product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ depending on (P,Q). In this section, we derive the tight bounds of the average treatment effects based on the closed-form expression of the identification region derived in Proposition 1.3.1.

If we can find within the identification region the distribution of Y_1 that is first-order stochastically dominated by the other distributions in the identification region, i.e., the distribution of Y_1 whose cdf is larger than the other distributions in the identification region, then the tight lower bound for $E(Y_1)$ is obtained by calculating the mean with respect to that distribution. Symmetrically, in order to find the upper bound of the mean of Y_1, it suffices to find the distribution within the identification region that first-order stochastically dominates the other distributions in the region. For an illustration on how to find these distributions, see Figure 1.6 where Y_1 and Y_0 are assumed to be continuous on the compact support $\mathcal{Y} = [y_l, y_u]$ (with probability masses at y_l and y_u allowed). The data generating process drawn in this figure is the same as in Figure 1.4, which corresponds to the case (i) of Proposition 1.3.1.

Intuitively speaking, the upper bound for the cdf of Y_1 is found by allocating unidentified or partially identified probabilities of Y_1's distribution to lower values of $\mathcal{Y}$. As we have already discussed, the maximal amount of unidentified probabilities in the distribution of Y_1 is the left-out probabilities $1 - \delta_{Y_1}$ = (the area of $n(1)$). The area $n(1)$ is partitioned into two, n and $d\&c'$, and

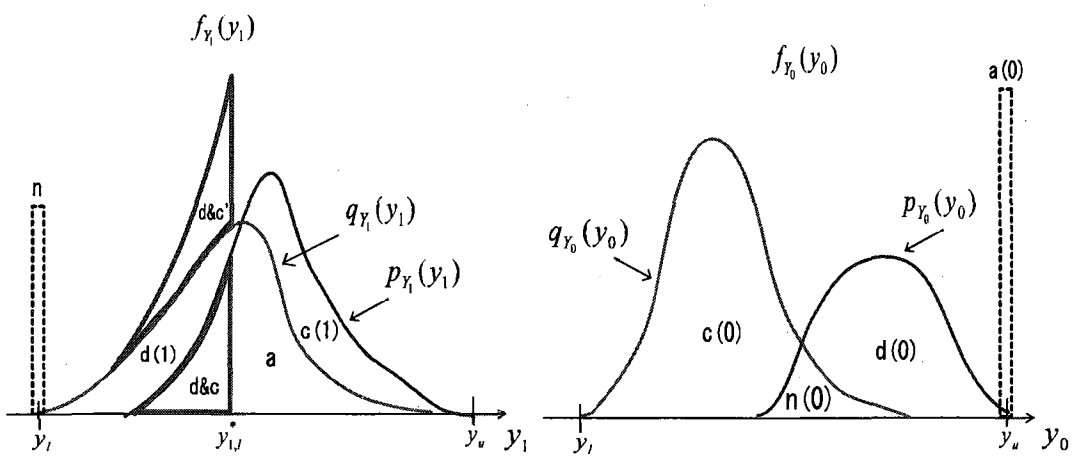

Figure 1.6: This figure shows a data generating process that corresponds to the case (i) in Proposition 1.3.1. The left-hand side figure provides f_{Y_1} in the identification region that achieves the upper bound for Y_1's cdf. The mean of Y_1 with respect to this f_{Y_1} yields the tight lower bound for $E(Y_1)$. The right-hand side figure provides f_{Y_0} that achieves the lower bound for Y_0's cdf, which yields the tight upper bound for $E(Y_0)$.

the shape of n is completely unrestricted so we can assign n to the lower end of the outcome support as drawn in Figure 1.6). On the other hand, the shape of $d\&c'$ is constrained since its copy $d\&c$ must be contained in $a(1)$ (the area below $\min\{p_{Y_1}, q_{Y_1}\}$). Since our goal is to sort out unidentified probabilities to the left, it is graphically obvious that we would like to pinning down $d\&c$ as the left-tail part of $\min\{p_{Y_1}, q_{Y_1}\}$ with keeping its area at $\lambda_{Y_1} + \delta_{Y_1} - 1$. Hence, with $y^*_{1,l}$ chosen so as to satisfy $\int_{y_l}^{y^*_{1,l}} \min\{p_{Y_1}, q_{Y_1}\}\, dy_1 = \lambda_{Y_1} + \delta_{Y_0} - 1$, the cdf of Y_1 that achieves its upper bound is written as

$$F^u_{Y_1}(y) = \lambda_{Y_0} + \int_{y_l}^{y} \underline{f_{Y_1}}(y_1)dy_1 + \int 1\{y \in [y_l, y^*_{1,l}]\} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}dy_1.$$

Symmetrically, by assigning n to the upper end of the outcome support and by taking $d\&c$ as the right tail of $\min\{p_{Y_1}, q_{Y_1}\}$, we obtain the cdf of Y_1 that achieves its lower bound,

$$F^l_{Y_1}(y) = \int_{y_l}^{y} \underline{f_{Y_1}}(y_1)dy_1 + \int 1\{y \in [y^*_{1,u}, y_u]\} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}dy_1 + 1\{y = y_u\}\lambda_{Y_0}.$$

Thus, the tight mean bounds for the distribution of Y_1 are obtained by taking the mean and quantile with respect to these cdfs $F^u_{Y_1}(y)$ and $F^l_{Y_1}(y)$.[14]

Proposition 1.3.3 (The tight average treatment effect bounds under RA-causal) *Assume Y_1 and Y_0 are continuous variables on the interior of the compact support $\mathcal{Y} = [y_l, y_u]$. Let $\delta_{Y_1}, \delta_{Y_0}, \lambda_{Y_1}$, and λ_{Y_0} be the parameters defined in Proposition 1.3.1 and assume $\delta_{Y_1} \leq 1$ and $\delta_{Y_0} \leq 1$.*

[14]The tight bounds for the quantile of f_{Y_1} also follows the tight cdf bounds presented here.

(i) Consider the case where the data generating process exhibits $1-\delta_{Y_0} < \lambda_{Y_1}$. *Let* $y^*_{1,l}$ *and* $y^*_{1,u}$ *be the values in* $\mathcal{Y}$ *that satisfy*

$$\begin{aligned}\int_{y_l}^{y^*_{1,l}} \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 &= \lambda_{Y_1} + \delta_{Y_0} - 1,\\ \int_{y^*_{1,u}}^{y_u} \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 &= \lambda_{Y_1} + \delta_{Y_0} - 1.\end{aligned}$$

Then the tight bounds of the average treatment effects are

$$\begin{aligned}&\int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1)dy_1 + \int_{y_l}^{y^*_{1,l}} y_1 \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 + \lambda_{Y_0} y_l - \int_{y_l}^{y_u} y_0 \underline{f_{Y_0}}(y_0)dy_0 - (1-\delta_{Y_0})y_u\\ \leq\ & E(Y_1) - E(Y_0)\\ \leq\ & \int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1)dy_1 + \int_{y^*_{1,u}}^{y_u} y_1 \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 + \lambda_{Y_0} y_u - \int_{y_l}^{y_u} y_0 \underline{f_{Y_0}}(y_0)dy_0 - (1-\delta_{Y_0})y_l.\end{aligned}$$

(ii) Consider the case where the data generating process exhibits $1-\delta_{Y_0} > \lambda_{Y_1}$. *Define* $y^*_{0,l}$ *and* $y^*_{0,u}$ *as*

$$\begin{aligned}\int_{y_l}^{y^*_{0,l}} \min\left\{p_{Y_0}(y_0), q_{Y_0}(y_0)\right\} dy_0 &= 1 - \delta_{Y_0} - \lambda_{Y_1},\\ \int_{y^*_{0,u}}^{y_u} \min\left\{p_{Y_0}(y_0), q_{Y_0}(y_0)\right\} dy_0 &= 1 - \delta_{Y_0} - \lambda_{Y_1}.\end{aligned}$$

Then the tight bounds of the average treatment effects are

$$\begin{aligned}&\int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1)dy_1 + (1-\delta_{Y_1})y_l - \int y_0 \underline{f_{Y_0}}(y_0)dy_0 - \int_{y^*_{0,u}}^{y_u} y_0 \min\left\{p_{Y_0}(y_0), q_{Y_0}(y_0)\right\} dy_0 - \lambda_{Y_1} y_u\\ \leq\ & E(Y_1) - E(Y_0)\\ \leq\ & \int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1)dy_1 + (1-\delta_{Y_1})y_u - \int y_0 \underline{f_{Y_0}}(y_0)dy_0 - \int_{y_l}^{y^*_{0,l}} y_0 \min\left\{p_{Y_0}(y_0), q_{Y_0}(y_0)\right\} dy_0 - \lambda_{Y_1} y_l.\end{aligned}$$

(iii) If the data generating process exhibits $1-\delta_{Y_0} = \lambda_{Y_1}$, *then the tight bounds of the average treatment effects are*

$$\begin{aligned}&(1-\delta_{Y_1})y_l + \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0)d\mu - (1-\delta_{Y_0})y_u\\ \leq\ & E(Y_1) - E(Y_0)\\ \leq\ & \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu + (1-\delta_{Y_1})y_u - (1-\delta_{Y_0})y_l - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0)d\mu.\end{aligned}$$

This proposition provides the closed form expression of the tight bounds for the average treatment effects under RA-causal. Note that the outer bounds for the average treatment effects considered in

(1.3.2.13) become tight when the data generating process satisfies $1-\delta_{Y_0} = \lambda_{Y_1}$. In particular, when the data generating process satisfies $p_{Y_1}(y_1) \geq q_{Y_1}(y_1)$ μ-a.e. and $q_{Y_0}(y_0) \geq p_{Y_0}(y_0)$ μ-a.e. then, $1-\delta_{Y_0} = \lambda_{Y_1}$ holds and the tight bounds for the average treatment effects take an identical form to the outer bounds. Furthermore, since in the case of $p_{Y_1}(y_1) \geq q_{Y_1}(y_1)$ μ-a.e., the tight $E(Y_1)$ bounds under independence restriction coincides with the tight $E(Y_1)$ bounds under the mean independence restriction $E(Y_1|Z) = E(Y_1)$ (see Appendix 1.A.3), and so is for $E(Y_0)$. Therefore, we can see that in this case imposing RA-causal does not provide further identification gain relative to the mean independence restriction of Y_1 and Y_0 with respect to Z.

If $1-\delta_{Y_0} \neq \lambda_{Y_1}$, then the outer bounds are no longer tight and the tight bounds are the ones given in this proposition. It can be seen that the width of these bounds is shorter than the width of the outer bounds. Specifically, the width of the tight bounds for the case of (i) $1-\delta_{Y_0} < \lambda_{Y_1}$ is

$$(\lambda_{Y_0}+1-\delta_{Y_0})(y_u-y_l) + \int_{y^*_{1,u}}^{y_u} y_1 \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}\, dy_1 - \int_{y_l}^{y^*_{1,l}} y_1 \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}\, dy_1, \tag{1.3.3.16}$$

and if $\min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}$ does not have probability masses on y_l and y_u, this is strictly smaller than

$$\begin{aligned} & (\lambda_{Y_0}+1-\delta_{Y_0})(y_u-y_l) + (\lambda_{Y_1}+\delta_{Y_0}-1)(y_u-y_l) \\ = \;& (\lambda_{Y_0}+\lambda_{Y_1})(y_u-y_l) \\ = \;& (2-\delta_{Y_1}-\delta_{Y_0})(y_u-y_l) \\ = \;& \text{the width of the outer bounds.} \end{aligned}$$

To illustrate a situation where the tight bounds are substantially narrower than the outer bounds, consider the data generating process with $1-\delta_{Y_0} < \lambda_{Y_1}$ and $1-\delta_{Y_0} = 0$, i.e., f_{Y_0} is identified. Then, since $y^*_{1,u} = y_l$ and $y^*_{1,l} = y_u$ hold, the last two terms in (1.3.3.16) cancel out and the width of the tight bounds is $\lambda_{Y_0}(y_u - y_l)$ and it is shorter than the width of the outer bounds by $\lambda_{Y_1}(y_u - y_l)$, which can be substantial if λ_{Y_1} is relatively large, i.e., p_{Y_1} and q_{Y_1} are similar.

Except for the case (iii), the above formulae are valid only for continuous outcome variables. When the support of $\mathcal{Y}$ is allowed to contain discrete points, the closed form expression for the tight bounds needs a slight modification.

In case of (i) $1-\delta_{Y_0} < \lambda_{Y_1}$, we define

$$\begin{aligned} y^*_{1,l} &\equiv \inf\left\{ y : \int_{[y_l,y]} \min\{p_{Y_1}, q_{Y_1}\} d\mu \geq \lambda_{Y_1} + \delta_{Y_0} - 1 \right\}, \\ y^*_{1,u} &\equiv \sup\left\{ y : \int_{[y,y_u]} \min\{p_{Y_1}, q_{Y_1}\} d\mu \geq \lambda_{Y_1} + \delta_{Y_0} - 1 \right\}. \end{aligned}$$

Then, it can be shown that the density of Y_1 that yields the cdf upper bound is

$$\begin{aligned} f_{Y_1}^{upper}(y_1) &= \lambda_{Y_0} 1\{y_1 = y_l\} + \underline{f_{Y_1}}(y_1) + 1\{y_1 \in [y_l, y_{1,l}^*]\} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\} \\ &\quad - \left(\int_{[y_l, y_{1,l}^*]} \min\{p_{Y_1}, q_{Y_1}\} d\mu - [\lambda_{Y_1} + \delta_{Y_0} - 1] \right) 1\{y_1 = y_{1,l}^*\}, \end{aligned}$$

and the one that yields the cdf lower bound is

$$\begin{aligned} f_{Y_1}^{lower}(y_1) &= \lambda_{Y_0} 1\{y_1 = y_u\} + \underline{f_{Y_1}}(y_1) + 1\{y_1 \in [y_{1,u}^*, y_u]\} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\} \\ &\quad - \left(\int_{[y_{1,u}^*, y_u]} \min\{p_{Y_1}, q_{Y_1}\} d\mu - [\lambda_{Y_1} + \delta_{Y_0} - 1] \right) 1\{y_1 = y_{1,u}^*\}. \end{aligned}$$

Hence, the tight bounds for $E(Y_1)$ is obtained as $\int_{\mathcal{Y}} y_1 f_{Y_1}^{upper}(y_1) d\mu \leq E(Y_1) \leq \int_{\mathcal{Y}} y_1 f_{Y_1}^{lower}(y_1) d\mu$. Note that the tight bounds for $E(Y_0)$ do not change even when we allow discrete points in the support. Hence, the closed form expression of the tight bounds for average treatment effect is

$$\begin{aligned} & \int_{\mathcal{Y}} y_1 f_{Y_1}^{upper}(y_1) d\mu - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0) d\mu - (1 - \delta_{Y_0}) y_u \\ \leq\ & E(Y_1) - E(Y_0) \\ \leq\ & \int_{\mathcal{Y}} y_1 f_{Y_1}^{lower}(y_1) d\mu - \int_{\mathcal{Y}} y_0 \underline{f_{Y_0}}(y_0) d\mu - (1 - \delta_{Y_0}) y_l \end{aligned}$$

A similar modification is needed for case (ii) $1 - \delta_{Y_0} > \lambda_{Y_1}$. Define

$$\begin{aligned} y_{0,l}^* &\equiv \inf \left\{ y : \int_{[y_l, y]} \min\{p_{Y_0}, q_{Y_0}\} d\mu \geq 1 - \delta_{Y_0} - \lambda_{Y_1} \right\}, \\ y_{0,u}^* &\equiv \sup \left\{ y : \int_{[y, y_u]} \min\{p_{Y_0}, q_{Y_0}\} d\mu \geq 1 - \delta_{Y_0} - \lambda_{Y_1} \right\}. \end{aligned}$$

Then, it can be shown that the density of Y_0 that yields the cdf upper bound is

$$\begin{aligned} f_{Y_0}^{upper}(y_0) &= \lambda_{Y_1} 1\{y_0 = y_l\} + \underline{f_{Y_0}}(y_0) + 1\{y_0 \in [y_l, y_{0,l}^*]\} \min\{p_{Y_0}(y_0), q_{Y_0}(y_0)\} \\ &\quad - \left(\int_{[y_l, y_{0,l}^*]} \min\{p_{Y_0}, q_{Y_0}\} d\mu - [1 - \delta_{Y_0} - \lambda_{Y_1}] \right) 1\{y_0 = y_{0,l}^*\}, \end{aligned}$$

and the one that yields the cdf lower bound is

$$\begin{aligned} f_{Y_0}^{lower}(y_0) &= \lambda_{Y_1} 1\{y_0 = y_u\} + \underline{f_{Y_0}}(y_0) + 1\{y_0 \in [y_{0,u}^*, y_u]\} \min\{p_{Y_0}(y_0), q_{Y_0}(y_0)\} \\ &\quad - \left(\int_{[y_{0,u}^*, y_u]} \min\{p_{Y_0}, q_{Y_0}\} d\mu - [1 - \delta_{Y_0} - \lambda_{Y_1}] \right) 1\{y_0 = y_{0,u}^*\}. \end{aligned}$$

Hence, the tight bounds for $E(Y_0)$ become $\int_{\mathcal{Y}} y_0 f_{Y_0}^{upper}(y_0)d\mu \leq E(Y_0) \leq \int_{\mathcal{Y}} y_0 f_{Y_0}^{lower}(y_0)d\mu$. As a result, the tight bounds for average treatment effect is

$$\begin{aligned} & \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu + (1-\delta_{Y_1})y_l - \int_{\mathcal{Y}} y_0 f_{Y_0}^{lower}(y_0)d\mu \\ \leq \; & E(Y_1) - E(Y_0) \\ \leq \; & \int_{\mathcal{Y}} y_1 \underline{f_{Y_1}}(y_1)d\mu + (1-\delta_{Y_1})y_u - \int_{\mathcal{Y}} y_0 f_{Y_0}^{upper}(y_0)d\mu. \end{aligned}$$

Note that when outcome variables are binary, these tight bounds coincide with the treatment effect bounds obtained in Balke and Pearl (1997).

1.4 Concluding Remarks

This paper derives the identification region of the outcome distributions in the single missing outcome model and the counterfactual causal model under the restriction of statistical independence of an instrument and outcomes.

For the single missing outcome model, we extend the identification region obtained in Manski (2003) to a general setting where the outcome variable can be continuous. Using the envelope density, we provide an analytically tractable representation of the identification region for the outcome distribution under the restriction of instrument independence. We derive the integrated envelope, which is the key parameter for examining the emptiness of the identification region. Since the empty identification region implies misspecification of the exclusion restriction, this parameter is useful for testing instrument independence in the selection model. Chapter 2 discusses the use of the integrated envelope for the purpose of testing the instrument exclusion restriction.

We analyze the single missing outcome model with heterogeneities in the selection response to an instrument. We show that a stronger exclusion restriction — that the instrument is jointly independent of the outcome and the selection heterogeneities — does not further narrow the identification region. In addition, we show that threshold crossing selection with an additive error constrains the data generating process, but does not further narrow the identification region. These identification results imply that, regardless of we specify the selection equation or whether or not we are explicit about the selection heterogeneities, the envelope density always provides maximal identifying information for the outcome distribution once the instrument is assumed to be independent of the outcome.

We extend this identification framework to the counterfactual causal model with a binary treatment and derive a closed-form expression for the identification region of the potential outcome distributions under the instrument independence restriction. We show that the independence restriction is in terms of independence between instrument and (Y_1, Y_0), the set of potential outcome distributions that cover the envelope densities of treatment and control outcomes provide the identification region for the potential outcome distributions. We show that, as the independence restriction is

strengthened to instrument's *joint* independence of (Y_1, Y_0) and the selection heterogeneities, we can obtain a narrower identification region than the case of the weaker independence restriction. This finding implies that the additional independence restriction of instrument and the selection heterogeneities provide further identifying information of the potential outcome distributions. We characterize the condition for the data generating process under which such identification gain arises. Based on the obtained identification region, we derive tight bounds for the average treatment effect under the instrument exclusion restriction. If we consider the case of binary outcomes, our tight bounds for the average treatment effect produce the same bounds as in Balke and Pearl (1997). In this sense, our results can be seen as one generalization of Balke and Pearl's result to continuous outcomes.

The identification region in the single missing outcome model considered in this paper can be straightforwardly extended up to a multi-valued discrete instrument. On the other hand, the identification region of the counterfactual outcome distributions considered in this paper is limited to the case of a binary treatment and a binary instrument. Under a weaker restriction that instrument is solely independent of (Y_1, Y_0), an extension to a multi-valued instrument case is straightforward. However, under the stronger restriction of instrument independence, it is not clear yet how the identification region looks like when the model contains multiple treatments and a multi-valued instrument.

1.A Appendices

1.A.1 Proof of Proposition 1.2.1

(i) Let P and Q be given by data and assume $\delta(P,Q) \leq 1$. Let $\mathcal{F}_{f_Y}^{env}(P,Q) = \{f_Y : f_Y(y) \geq \underline{f}(y)\ \mu\text{-a.e.}\}$. For an arbitrary $f_Y \in \mathcal{F}_{f_Y}^{env}(P,Q)$, we shall construct a joint probability law of (Y,D,Z) that is compatible with the data generating process P and Q, and ER. Since the marginal distribution of Z is irrelevant to the analysis, we focus on the conditional law of (Y,D) given Z. Let B be an arbitrary Borel set. In order for the conditional law of (Y,D) given Z to be compatible with the data generating process, we must have

$$\begin{aligned} \Pr(Y \in B, D=1|Z=1) &= \int_B p(y)d\mu, \\ \Pr(Y \in B, D=1|Z=0) &= \int_B q(y)d\mu. \end{aligned}$$

Pin down the probability of $\{Y \in B, D=0\}$ given Z to

$$\begin{aligned} \Pr(Y \in B, D=0|Z=1) &= \int_B [f_Y(y) - p(y)]d\mu, \\ \Pr(Y \in B, D=0|Z=0) &= \int_B [f_Y(y) - q(y)]d\mu. \end{aligned}$$

Note that the constructed probabilities are nonnegative by construction and they satisfy ER since $\Pr(Y \in B|Z=1) = \Pr(Y \in B|Z=0) = \int_B f_Y(y)d\mu$. Hence, $\mathcal{F}^{env}_{f_Y}(P,Q)$ is contained in the identification region under ER.

On the other hand, consider a marginal outcome distribution $f_Y \notin \mathcal{F}^*$. Then, there exists a Borel set A with $\mu(A) > 0$ such that

$$\int_A [f_Y(y) - p(y)]d\mu < 0 \quad \text{or} \quad \int_A [f_Y(y) - q(y)]d\mu < 0. \tag{1.1.1.17}$$

Note that the probabilities of $\{Y \in A, D=0\}$ given Z are written as

$$\begin{aligned}
\Pr(Y \in A, D=0|Z=1) &= \Pr(Y \in A|Z=1) - \Pr(Y \in A, D=1|Z=1) \\
&= \int_A [f_{Y|Z}(y|Z=1) - p(y)]d\mu \\
\Pr(Y \in A, D=0|Z=0) &= \Pr(Y \in A|Z=0) - \Pr(Y \in A, D=1|Z=0) \\
&= \int_A [f_{Y|Z}(y|Z=0) - q(y)]d\mu
\end{aligned}$$

If ER is true, $f_{Y|Z} = f_Y$ must hold. Then, by (1.1.1.17) one of the above probabilities are negative, and therefore we cannot construct a conditional law of (Y,D) given Z that is compatible with the data generating process and ER.
Thus, we conclude $\mathcal{F}^{env}_{f_Y}(P,Q)$ is the identification region under ER. (ii) is obvious.

1.A.2 A Comparison with the cdf bounds in Blundell et al. (2007)

In this appendix, we compare the tight cdf bounds based on the envelope density (1.2.2.5) with the cdf bounds used in Blundell et al. (2007). We shall show that the latter do not always yield the tightest bounds.

Based on a moment restriction for the cdf of Y, $F_{Y|Z}(y|z) = E(I\{Y \in (-\infty, y]\}|Z=z) = E(I\{Y \in (-\infty, y]\}) = F_Y(y)$, Blundell et al. (2007) use the mean independence bounds of Manski (1994) to construct the bounds for $F_Y(y)$,

$$\begin{aligned}
&\max\{P((-\infty,y]), Q((-\infty,y])\} \leq F_Y(y) \\
\leq\ &\min\{P((-\infty,y]) + P_{mis}, Q((-\infty,y]) + Q_{mis}\}.
\end{aligned} \tag{1.1.2.18}$$

These bounds, which we call the *naive cdf bounds* hereafter, are not necessarily the tightest possible under ER. The reason is that the naive cdf bounds only utilize the restriction that the probability of the event $\{Y \leq y\}$ does not depend on Z. This restriction is certainly weaker than the statistical independence restriction since the full statistical independence requires that $\Pr(Y \in A|Z)$ for *any* subsets $A \subset \mathcal{Y}$ does not depend on Z.

For stating the main result of this section, we define the dominance relationship between $p(y)$ and $q(y)$.

Definition 1.A.1 (dominance in density) *(i) The density $p(y)$ dominates $q(y)$ on $A \subset \mathcal{Y}$ if $p(y) \geq q(y)$ holds μ-a.e. on A.*
(ii) $p(y)$ is the dominating density if $p(y)$ dominates $q(y)$ on $\mathcal{Y}$.

$p(y)$ is the dominating density if $p(y)$ covers $q(y)$ on the entire outcome support. If this is the case, $q(y)$ does not provide identifying information for f_Y further than $p(y)$ because the maximal area under f_Y is occupied by $p(y)$ alone. The existence of the dominance relationship guarantees the interchangeability between max operation and integration, that is,

$$\int_A \max\{p(y), q(y)\} d\mu = \max\left\{\int_A p(y)d\mu, \int_A q(y)d\mu\right\}.$$

if and only if $p(y)$ dominates $q(y)$ on A

This fundamental identity provides the following tightness result of the naive cdf bounds.

Proposition 1.A.1 (tightness of the naive cdf bounds) *(i) The naive cdf bounds at $y \in \mathcal{Y}$ are tight under ER if and only if either $p(y)$ or $q(y)$ dominates the other on $(-\infty, y]$ and either $p(y)$ or $q(y)$ dominates the other on (y, ∞).*
(ii) The naive cdf bounds are tight under ER for all $y \in \mathcal{Y}$ if and only if the data generating process reveals the dominating density.

Proof. (i) Fix $y \in \mathcal{Y}$. For the lower bound of the naive cdf bounds,

$$\begin{aligned}
\max\left\{\int_{(-\infty,y]} p(y)d\mu, \int_{(-\infty,y]} q(y)d\mu\right\} &\leq \int_{(-\infty,y]} \max\{p(y), q(y)\} d\mu \\
&= \int_{(-\infty,y]} \underline{f}(y)d\mu \\
&= \text{the lower bound of the tight cdf bounds.}
\end{aligned}$$

Note that the inequality holds in equality if and only if either $p(y)$ or $q(y)$ dominates the other on $(-\infty, y]$.

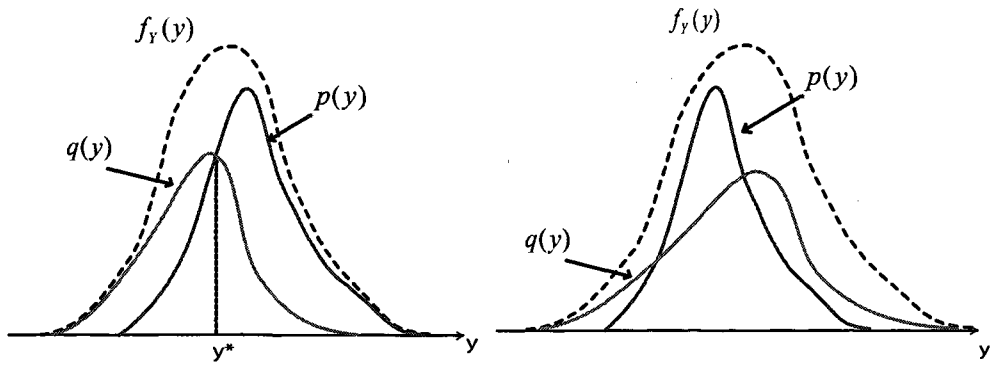

Figure 1.7: *In the left-hand side figure, the naive cdf bounds at y^* are tight. On the other hand, when $p(y)$ and $q(y)$ are drawn as in the right-hand side figure, the naive cdf bounds are not tight at any $y \in \mathcal{Y}$ (Proposition B.1).*

For the upper bound of the naive cdf bounds,

$$\begin{aligned}
& \min\left\{\int_{(-\infty,y]} p(y)d\mu + P_{mis}, \int_{(-\infty,y]} q(y)d\mu + Q_{mis}\right\} \\
= \; & \min\left\{1 - \int_{(y,\infty)} p(y)d\mu, 1 - \int_{(y,\infty)} q(y)d\mu\right\} \\
= \; & 1 - \max\left\{\int_{(y,\infty)} p(y)d\mu, \int_{(y,\infty)} q(y)d\mu\right\} \\
\geq \; & 1 - \int_{(y,\infty)} \underline{f}(y)d\mu \\
= \; & \int_{(-\infty,y]} \underline{f}(y)d\mu + 1 - \delta \\
= \; & \text{the upper bound of the tight cdf bounds,}
\end{aligned}$$

where the inequality holds in equality if and only if either $p(y)$ or $q(y)$ dominates the other on (y, ∞).
The statement (ii) clearly follows from (i). ■

When we employ the naive cdf bounds, we would refute ER if the lower and upper bound of the cdf cross at some y. This refuting rule is as powerful as the one based on the integrated envelope if the condition in Proposition B.1 (i) holds at some y. However, this holds in a rather limited situation where some left unbounded intervals $(-\infty, y]$ or right unbounded intervals (y, ∞) can correctly divide $\mathcal{Y}$ into $\{y : p(y) \geq q(y)\}$ and $\{y : p(y) < q(y)\}$ (see Figure 1.7).

1.A.3 Identification gain of ER relative to MI

Consider the bounded outcome support $\mathcal{Y} = [y_l, y_u]$. Manski (1994) derives the tight $E(Y)$ bounds under MI,

$$\begin{aligned} & \max\left\{\int_{\mathcal{Y}} yp(y)d\mu + y_l P_{mis},\ \int_{\mathcal{Y}} yq(y)d\mu + y_l Q_{mis}\right\} \\ \leq\ & E(Y) \leq \min\left\{\int_{\mathcal{Y}} yp(y)d\mu + y_u P_{mis},\ \int_{\mathcal{Y}} yq(y)d\mu + y_u Q_{mis}\right\}. \end{aligned} \tag{C.1}$$

The next proposition shows the necessary and sufficient condition for the MI mean bounds (C.1) to coincide with the ER mean bounds (1.2.2.6).

Proposition 1.A.2 (Identification power of ER relative to MI) *The MI mean bounds (C.1) coincide with the ER mean bounds (1.2.2.6) if and only if the data generating process reveals a dominating density on $(y_l, y_u]$ and $[y_l, y_u)$.*

Proof. The lower bound of the MI mean bounds is written as

$$\begin{aligned} & \max\left\{\int_{\mathcal{Y}} yp(y)d\mu + y_l\left(1 - \int_{\mathcal{Y}} p(y)d\mu\right),\ \int_{\mathcal{Y}} yq(y)d\mu + y_l\left(1 - \int_{\mathcal{Y}} q(y)d\mu\right)\right\} \\ =\ & \max\left\{\int_{\mathcal{Y}} (y - y_l)p(y)d\mu, \int_{\mathcal{Y}} (y - y_l)q(y)d\mu\right\} + y_l \\ \leq\ & \int_{\mathcal{Y}} (y - y_l)\underline{f}(y)d\mu + y_l \\ =\ & \int_{\mathcal{Y}} y\underline{f}(y)d\mu + (1 - \delta)y_l \\ =\ & \text{the lower bound of the ER mean bounds,} \end{aligned}$$

where the inequality holds in equality if and only if either $(y - y_l)p(y) \geq (y - y_l)q(y)$, μ-a.e. on $[y_l, y_u]$ or $(y - y_l)p(y) \leq (y - y_l)q(y)$, μ-a.e. on $[y_l, y_u]$ holds. This condition is equivalently stated as the existence of the dominating density on $(y_l, y_u]$ since the necessary and sufficient condition for $(y - y_l)p(y) \geq (y - y_l)q(y)$, μ-a.e. on $[y_l, y_u]$ is $p(y) \geq q(y)$ μ-a.e. on $(y_l, y_u]$.
Similarly, for the upper bound of the MI mean bounds, we have

$$\begin{aligned} & \min\left\{\int_{\mathcal{Y}} yp(y)d\mu + y_u\int_{\mathcal{Y}} (1 - p(y))d\mu,\ \int_{\mathcal{Y}} yq(y)d\mu + y_u\int_{\mathcal{Y}} (1 - q(y))d\mu\right\} \\ =\ & y_u - \max\left\{\int_{\mathcal{Y}} (y_u - y)p(y)d\mu, \int_{\mathcal{Y}} (y_u - y)q(y)d\mu\right\} \\ \geq\ & y_u - \int_{\mathcal{Y}} (y_u - y)\underline{f}(y)d\mu \\ =\ & \int_{\mathcal{Y}} y\underline{f}(y)d\mu + (1 - \delta)y_u \\ =\ & \text{the upper bound of the ER mean bounds,} \end{aligned}$$

where the inequality holds in equality if and only if either $(y_u - y)p(y) \geq (y_u - y)q(y)$ μ-a.e. on $[y_l, y_u]$ or $(y_u - y)p(y) \leq (y_u - y)q(y)$ μ-a.e. on $[y_l, y_u]$ is true. Similarly to the lower bound case, this is equivalent to the existence of the dominating density on $[y_l, y_u)$.
By combining the results for the lower and upper bound, we conclude that the MI mean bounds coincide with the ER mean bounds if and only if *the data generating process reveals a dominating density on* $(y_l, y_u]$ and $[y_l, y_u)$. ■

This proposition demonstrates that when we observe the dominating density, that is, either $p(y)$ or $q(y)$ covers the other on the entire $\mathcal{Y}$, ER does not provide narrower bounds for $E(Y)$ than MI. The intuition of this proposition is given as follows. When we construct the ER mean bounds, we allocate the amount of unidentified probability, which is given by one minus the integrated envelope $1 - \delta$, to the worst-case or best-case outcome. Consequently, the width of the mean bounds is determined by the amount of unidentified probability, $(y_u - y_l)(1 - \delta)$. On the other hand, when we construct the MI mean bounds, we first construct the bounds for $E(Y)$ from P and Q separately and then, we take the intersection of these. The width of these two bounds are therefore determined by P_{mis} and Q_{mis}. If one of them is equal to $1 - \delta$, it implies that we cannot reduce the amount of unidentified probability by strengthening MI to ER. Therefore the ER mean bounds coincide with the MI mean bounds if $\min\{P_{mis}, Q_{mis}\} = 1 - \delta$. A sufficient condition for this is the presence of the dominating density on $\mathcal{Y}$. Note that, when Y is binary, ER mean bounds and MI mean bounds always coincide since these restrictions are equivalent.

1.A.4 Proof of Proposition 1.2.2.

The next lemma, which will be used in the proof of Proposition 1.2.2, summarizes the implication of imposing RA.

Lemma 1.A.1 *(i) If a joint probability distribution on (Y, T, Z) satisfies RA, then, the following identities hold μ-a.e.,*

$$\begin{aligned} p(y) &= h_c(y) + h_a(y), \\ q(y) &= h_d(y) + h_a(y), \\ f_Y(y) - p(y) &= h_d(y) + h_n(y), \\ f_Y(y) - q(y) &= h_c(y) + h_n(y), \end{aligned} \tag{*}$$

where $h_t(y) = f_{Y,T}(y, T = t)$, $t = c, n, a, d$.
(ii) Conversely, given a data generating process P and Q, and a marginal distribution of outcome f_Y, if there exist nonnegative functions $h_t(y)$, $t = c, n, a, d$, that satisfy (*) μ-a.e., then we can construct a joint probability law on (Y, T, Z) that is consistent with the data generating process and RA.

Proof. Assume that a population distribution of (Y,T,Z) satisfies RA. Then, for $B \in \mathcal{B}(\mathcal{Y})$,

$$\begin{aligned} P(B) &= \Pr(Y \in B, D = 1 | Z = 1) \\ &= \Pr(Y \in B, T \in \{c,a\} | Z = 1) \\ &= \Pr(Y \in B, T = c | Z = 1) + \Pr(Y \in B, T = a | Z = 1) \\ &= \Pr(Y \in B, T = c) + \Pr(Y \in B, T = a). \end{aligned}$$

The second line follows since the event $\{Y \in B, D = 1 | Z = 1\}$ is equivalent to $\{Y \in B, T \in \{c,a\} | Z = 1\}$ and the fourth line follows by RA. As the density expression of the above, we obtain

$$p(y) = f_{Y,T}(y, T = c) + f_{Y,T}(y, T = a),$$

which corresponds to the first identity of (*). We obtain the second constraint in a similar manner and we omit its derivation for brevity. As for the third constraint in (*),

$$\begin{aligned} \Pr(Y \in B) - P(B) &= \Pr(Y \in B | Z = 1) - \Pr(Y \in B, D = 1 | Z = 1) \\ &= \Pr(Y \in B, D = 0 | Z = 1) \\ &= \Pr(Y \in B, T \in \{n,d\} | Z = 1) \\ &= \Pr(Y \in B, T = n) + \Pr(Y \in B, T = d) \end{aligned}$$

We obtain the fourth constraint in a similar manner. This completes the proof of the first part of the lemma.

To prove the converse statement of the proposition, suppose that, for a given data generating process P and Q and a marginal distribution of f_Y, we have nonnegative functions $h_t(\cdot)$ for $t = c,n,a,d$ satisfying the constraints (*). Since the marginal distribution of Z is irrelevant to the analysis, we focus on constructing the conditional law of (Y,T) given Z. Let us specify both $\Pr(Y \in B, T = t | Z = 1)$ and $\Pr(Y \in B, T = t | Z = 0)$ to be equal to $\int_B h_t(y) d\mu \geq 0$, $t = c,n,a,d$. These are valid probability measures since $\sum_t \Pr(Y \in \mathcal{Y}, T = t | Z = z) = \sum_t \int_{\mathcal{Y}} h_t(y) d\mu = \int_{\mathcal{Y}} f_Y(y) d\mu = 1$. This probability law satisfies RA by construction. Furthermore, the constructed joint distribution

is compatible with the data generating process and the proposed f_Y since

$$\begin{aligned}
\Pr(Y \in B, D = 1 | Z = 1) &= \Pr(Y \in B, T = c | Z = 1) + \Pr(Y \in B, T = a | Z = 1) \\
&= \int_B h_c(y) d\mu + \int_B h_a(y) d\mu = P(B), \\
\Pr(Y \in B, D = 1 | Z = 0) &= \Pr(Y \in B, T = d | Z = 0) + \Pr(Y \in B, T = a | Z = 0) \\
&= \int_B h_d(y) d\mu + \int_B h_a(y) d\mu = Q(B), \\
\Pr(Y \in B) &= \sum_{t=c,n,a,d} \Pr(Y \in B, T = t) \\
&= \sum_{t=c,n,a,d} \int_B h_t(y) d\mu = \int_B f_Y(y) d\mu.
\end{aligned}$$

This completes the proof of the converse statement. ■

By the converse part of the above lemma, the identification region of f_Y under RA is formed as the collection of f_Y's for each of which we can find the feasible nonnegative functions $h_t(\cdot)$, $t = c, n, a, d$ satisfying (*). Recall that, when we construct $IR_{f_Y}(P, Q)$, we only concern whether $f_Y(y)$ is greater than or equal to $p(y)$ and $q(y)$. Here, we need to concern the existence of the nonnegative densities $h_t(\cdot)$, $t = c, n, a, d$, compatible with the constraints (*).

Proof of Proposition 1.2.2. Given a data generating process, $p(y)$ and $q(y)$, pick an arbitrary $f_Y \in \mathcal{F}^{env}_{f_Y}(P, Q)$. Figure 1.8 illustrates the proof of this redundancy result. We can find four partitions in the subgraph of $f_Y(y)$, which are labeled as C, N, A, and D. Consider imputing the type-specific density $h_t(y)$ as the height of one of the proposed partitions,

$$\begin{aligned}
\mathsf{C} &: \quad h_c(y) = \underline{f}(y) - q(y), \\
\mathsf{N} &: \quad h_n(y) = f_Y(y) - \underline{f}(y), \\
\mathsf{A} &: \quad h_a(y) = \min\{p(y), q(y)\}, \\
\mathsf{D} &: \quad h_d(y) = \underline{f}(y) - p(y).
\end{aligned}$$

Note that the obtained $h_t(y)$, $t = c, n, a, d$, satisfy the constraints (*) and they are nonnegative by construction. This way of imputing the four densities is feasible for any $f_Y \in \mathcal{F}^{env}_{f_Y}(P, Q)$. By Lemma 1.A.1 (ii), we conclude $\mathcal{F}^{env}_{f_Y}(P, Q)$ is contained in the identification region of f_Y under RA. For $f_Y \notin \mathcal{F}^{env}_{f_Y}(P, Q)$, if there exists a population compatible with (P, Q) and RA, then the third and fourth constraints of Lemma 1.A.1 (i) imply $f_Y(y) - p(y) \geq 0$ and $f_Y(y) - q_Y(y) \geq 0$ μ-a.e. and this contradicts $f_Y \notin \mathcal{F}^{env}_{f_Y}(P, Q)$. Hence, the identification region of f_Y under RA is contained in $\mathcal{F}^{env}_{f_Y}(P, Q)$. This completes the proof of the invariance result. ■

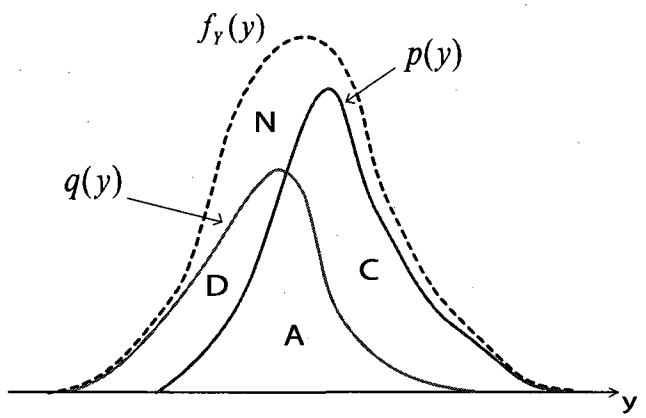

Figure 1.8: *A graphical illustration of the invariance result of the identification region under RA (Proposition 2.2).*

1.A.5 Proof of Proposition 1.2.3.

Provided that the population distribution satisfies RA, threshold crossing selection with an additive error is equivalent to the monotonicity of Imbens and Angrist (1994) (Vytlacil (2002)). Thus, the identification gain of imposing the additively separable threshold crossing formulation is examined by adding Imbens and Angrist's monotonicity to our analysis.[15] In this appendix, we refer to the monotonicity of Imbens and Angrist, or equivalently, threshold crossing selection with an additive error, as the *monotonic selection response to an instrument (*MSR, hereafter*).* Throughout the analysis, we assume $\Pr(D=1|Z=1) \geq \Pr(D=1|Z=0)$. This is equivalent to assuming that the observed selection probability is nondecreasing with respect to Z. Since we can always redefine the value of Z compatible with this assumption, we do not lose any generality by restricting our analysis to this case.

Restriction-MSR

Monotonic Selection Response to an Instrument (MSR): Without loss of generality, assume $\Pr(D=1|Z=1) \geq \Pr(D=1|Z=0)$. The selection process satisfies MSR if no defiers exist in the population $\Pr(T=d)=0$.

From the partial identification point of view, the implication of MSR is summarized in the next proposition, which covers Proposition 1.2.3 in the main text in the statement (iii).

Proposition 1.A.3 (Existence of the dominating density under RA and MSR) *Suppose that a population distribution of (Y,T,Z) satisfies RA and MSR.*
(i) Then, $p(y)$ is the dominating density.

[15]Note that the monotonicity of Imbens and Angrist is discussed in the context of the counterfactual causal model. Although our analysis is on the missing data model with a single outcome, we can consider an analogous restriction to the monotonicity since the monotonicity only concerns the population distribution of the potential selection indicators.

(ii) The MI mean bounds (1.2.2.6) coincide with the ER mean bounds (C.1).
Conversely, for a given data generating process, P and Q,
(iii) The identification region under RA and MSR is given by

$$\begin{cases} \mathcal{F}_{f_Y}^{env}(P,Q) & \text{if } p(y) \text{ is the dominating density} \\ \emptyset & \text{if } p(y) \text{ is not the dominating density} \end{cases}$$

Proof of Prposition 1.A.3 . (i) From the first two constraints in (*), $h_t(y) = 0$ implies $p(y) - q(y) = h_c(y) \geq 0$. (ii) This follows from Proposition 1.A.2. (iii) Suppose that $p(y)$ is the dominating density. For an arbitrary $f_Y \in \mathcal{F}_{f_Y}^{env}(P,Q)$, we want to show that there exists type specific nonnegative functions $h_t(y)$, $t = c, n, a, d$, that are compatible with the constraints (*) and MSR, i.e., the defier's density $h_d(y)$ is zero. Consider the following way of imputing the type specific densities,

$$\begin{aligned} h_c(y) &= p(y) - q(y), \\ h_n(y) &= f_Y(y) - p(y), \\ h_a(y) &= q(y), \\ h_d(y) &= 0. \end{aligned} \tag{1.1.5.19}$$

These densities satisfy the constraints (*) and as in the proof of the converse statement of Lemma 1.A.1, they yield a joint distribution of (Y,T,Z) that meets RA and MSR. Since this way of constructing $h_t(y)$ is feasible for any $f_Y \in \mathcal{F}_{f_Y}^{env}(P,Q)$, we claim that $\mathcal{F}_{f_Y}^{env}(P,Q)$ is contained in the identification region under RA and MSR. For $f_Y \notin \mathcal{F}_{f_Y}^{env}(P,Q)$, f_Y is not contained in the identification region because it is not compatible with RA as we showed in the proof of Proposition 1.2.2. Hence, we conclude that $\mathcal{F}_{f_Y}^{env}(P,Q)$ is the identification region of f_Y under RA and MSR. The emptiness of the identification region when $p(y)$ is not the dominating density is implied by (i) of this proposition. ■

This proposition shows that when RA and MSR hold in the population distribution of (Y,T,Z), then the data generating process must reveal the dominating density. The presence of the dominating density makes ER redundant relative to MI in terms of the width of $E(Y)$ bounds (Proposition 1.A.2).

If the data generating process reveals the dominating density, then, imposing MSR does not further narrow $IR_{f_Y}(P,Q)$. This is because MSR does not constrain how to impute the missing outcomes. To see why, consider the configuration of $p(y)$ and $q(y)$ and an arbitrary $f_Y(y)$ as shown in Figure 1.9. In (1.1.5.19), we pin down the type-specific densities, $h_c(y)$, $h_a(y)$, and $h_n(y)$ to the height of the area C, A, and N of Figure 1.9. This implies that each $f_Y \in \mathcal{F}_{f_Y}^{env}(P,Q)$ is obtained by the unique imputation of the never-taker's density without violating MSR. Hence, we obtain the identification region under RA and MSR as $\mathcal{F}_{f_Y}^{env}(P,Q)$.

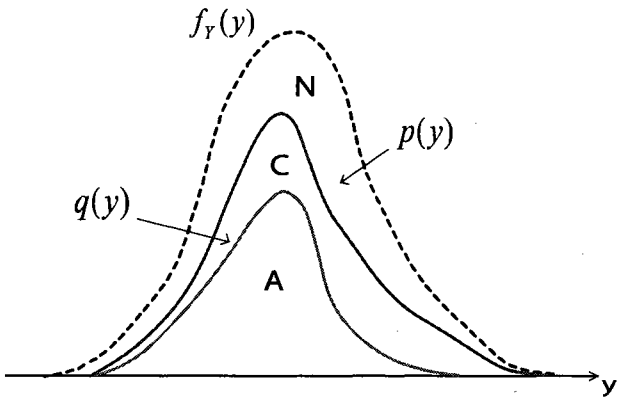

Figure 1.9: *If RA and MSR are satisfied, we must observe the above configuration of the densities (Proposition 1.A.3).* A *indicates the subgraph of* $q(y)$. *The subgraph of* $p(y)$ *minus that of* $q(y)$ *and the subgraph of* $f_Y(y)$ *minus that of* $p(y)$ *are labeled as* C *and* N, *respectively.*

1.A.6 Extension to a multi-valued discrete instrument

This appendix provides a framework that covers the single missing outcome model with a multi-valued instrument.

Assume that the support of Z consists of K points denoted by $\mathcal{Z} = \{z_1, \ldots, z_K\}$. Denote the probability distribution of Y_{data} conditional on $Z = z_k$ by $P_k = (P_k(\cdot), P_{k,mis})$,

$$\begin{aligned} P_k(A) &= \Pr(Y \in A | D = 1, Z = z_k) \Pr(D = 1 | Z = z_k), \\ P_{k,mis} &= \Pr(D = 0 | Z = z_k). \end{aligned}$$

We represent the data generating process by $\mathcal{P} = (P_1, \ldots, P_K)$. We use the lowercase letter p_k to stand for the density of P_k on $\mathcal{Y}$. The envelope density is defined as

$$\underline{f}(y) = \max_k \{p_k(y)\}.$$

Analogous to the binary instrument case, we say $p_k(y)$ is the dominating density on A if for all $l \neq k$, $p_k(y) \geq p_l(y)$ holds μ-a.e. on A.

Results similar to Proposition 2.1, B.1, and C.1 are obtained even when Z is multi-valued. Proofs proceed in the same way as in the binary instrument case and are therefore omitted for brevity. We notate the identification region of f_Y, $\{f_Y : f_Y(y) \geq \underline{f}(y)\ \mu\text{-a.e.}\}$, by $IR_{f_Y}(\mathcal{P})$.

In order to demonstrate a generalization of Proposition 2.2 (invariance of $IR_{f_Y}(\mathcal{P})$ under RA) and D.2 (existence of the dominating density under RA and MSR), we construct the type indicator T in the following manner. For the K-valued instrument, individual's selection response is uniquely characterized by an array of K potential selection indicators D_k, $k = 1, \ldots, K$. D_k indicates whether the individual is selected when Z is exogenously set at z_k. In total, there are 2^K number of types

in the population and we interpret T as a random variable indicating one of the 2^K types. Let $\mathcal{T}$ be the set of all types and define $\mathcal{T}_k \subset \mathcal{T}$ be the set of types with $D_k = 1$, $\mathcal{T}_k = \{t \in \mathcal{T} : D_k = 1\}$. $\mathcal{T}_k$ is interpreted as the subpopulation of those who are selected when $Z = z_k$.

Similarly to the binary Z case, RA is stated that Z is jointly independent of (Y, T). We keep the notation $\pi_t = \Pr(T = t)$ and $g_t(y) = f_{Y|T}(y|T = t)$. Analogous to the equations (*), if the population satisfies RA, then, for all $k = 1, \ldots, K$, we have

$$\begin{aligned} p_k(y) &= \textstyle\sum_{t \in \mathcal{T}_k} \pi_t g_t(y), \\ f_Y(y) - p_k(y) &= \textstyle\sum_{t \in \mathcal{T} \setminus \mathcal{T}_k} \pi_t g_t(y), \end{aligned}$$

The converse statement in Lemma D.1 holds as well for the multi-valued instrument case. That is, for a given data generating process $\mathcal{P}$ and a marginal outcome distribution f_Y, if we can find the nonnegative functions $\{h_t(y) : t \in \mathcal{T}\}$ that satisfy, for all $k = 1, \ldots, K$,

$$\begin{aligned} p_k(y) &= \textstyle\sum_{t \in \mathcal{T}_k} h_t(y), \\ f_Y(y) - p_k(y) &= \textstyle\sum_{t \in \mathcal{T} \setminus \mathcal{T}_k} h_t(y), \end{aligned} \tag{**}$$

then we can construct a joint distribution of (Y, T, Z) that is compatible with $\mathcal{P}$ and RA. A proof of this follows in a similar manner to the proof of Lemma D.1 and we do not present it here.

The redundancy of RA holds even when Z is multi-valued.

Proposition 2.2'. *For a multi-valued instrument, $IR_{f_Y}(\mathcal{P})$ is the identification region under RA.*

Proof. When $IR_{f_Y}(\mathcal{P})$ is empty, it is obvious that the identification region under RA is empty. Hence, assume $IR_{f_Y}(\mathcal{P})$ is nonempty.

Pick an arbitrary $f_Y \in IR_{f_Y}(\mathcal{P})$. Our goal is to find the set of nonnegative functions $\{h_t(y)\}_{t \in \mathcal{T}}$ that are compatible with the constraints (**).

Let $\mathcal{S}_k$ be the subgraph of $p_k(y)$ and $\mathcal{S}_k^c$ the supgraph of $p_k(y)$, i.e., $\mathcal{S}_k = \{(y, f) \in \mathcal{Y} \times \mathbb{R}_+ : 0 \leq f \leq p_k(y)\}$ and $\mathcal{S}_k^c = \{(y, f) \in \mathcal{Y} \times \mathbb{R}_+ : f > p_k(y)\}$. We denote the subgraph of f_Y by $\mathcal{S}_{f_Y}$. Note that, by the construction of $IR_{f_Y}(\mathcal{P})$, $\mathcal{S}_k \subset \mathcal{S}_{f_Y}$ holds for all k. Using the K subgraphs $\{S_k, k = 1, \ldots, K\}$, $\mathcal{S}_{f_Y}$ is partitioned into 2^K disjoint subsets. Each of these is represented by the K intersection of the subgraphs or supgraphs of $p_k(y)$ such as $\mathcal{S}_1 \cap \mathcal{S}_2^c \cap \cdots \cap \mathcal{S}_K \cap \mathcal{S}_{f_Y}$.

By noting that each t is one-to-one corresponding to a unique binary array of $\{D_k : k = 1, \ldots, K\}$, we define a subset $A(t) \subset \mathcal{S}_{f_Y}$ by assigning one of the disjoint subsets formed within $\mathcal{S}_{f_Y}$,

$$A(t) = \left(\bigcap_{l : D_l = 1} \mathcal{S}_l \right) \cap \left(\bigcap_{l : D_l = 0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y}.$$

Let us fix k. Note that the set of types $\mathcal{T}_k = \{t \in \mathcal{T} : D_k = 1\}$ and $\mathcal{T} \setminus \mathcal{T}_k = \{t \in \mathcal{T} : D_k = 0\}$ both

contain 2^{K-1} distinct types. Consider taking the union of $A(t)$ over $t \in \mathcal{T}_k$ and $t \in \mathcal{T} \setminus \mathcal{T}_k$,

$$\bigcup_{t\in\mathcal{T}_k} A(t) = \bigcup_{t\in\mathcal{T}_k} \left(\mathcal{S}_k \cap \left(\bigcap_{l\neq k: D_l=1} \mathcal{S}_l \right) \cap \left(\bigcap_{l\neq k: D_l=0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y} \right), \tag{E.1}$$

$$\bigcup_{t\in\mathcal{T}\setminus\mathcal{T}_k} A(t) = \bigcup_{t\in\mathcal{T}\setminus\mathcal{T}_k} \left(\mathcal{S}_k^c \cap \left(\bigcap_{l\neq k: D_l=1} \mathcal{S}_l \right) \cap \left(\bigcap_{l\neq k: D_l=0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y} \right). \tag{E.2}$$

In the above expressions, the subset $\left(\bigcap_{l\neq k: D_l=1} \mathcal{S}_l \right) \cap \left(\bigcap_{l\neq k: D_l=0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y}$ can be seen as one of the disjoint subsets within S_{f_Y} partitioned by the $(K-1)$ subgraphs $S_1, \ldots, S_{k-1}, S_{k+1}, \ldots, S_K$. Since each $t \in \mathcal{T}_k$ one-to-one corresponds to one of the partitioned subsets $\left(\bigcap_{l\neq k: D_l=1} \mathcal{S}_l \right) \cap \left(\bigcap_{l\neq k: D_l=0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y}$ and each $t \in \mathcal{T}\setminus\mathcal{T}_k$ also one-to-one corresponds to one of them, the union in the right hand side of (E.1) is the union of mutually disjoint and exhaustive partitions of $S_k \cap S_{f_Y}$. Therefore, the identities (E.1) and (E.2) are reduced to

$$\begin{aligned} \bigcup_{t\in\mathcal{T}_k} A(t) &= \mathcal{S}_k \cap \mathcal{S}_{f_Y} = \mathcal{S}_k, \\ \bigcup_{t\in\mathcal{T}\setminus\mathcal{T}_k} A(t) &= \mathcal{S}_k^c \cap \mathcal{S}_{f_Y}. \end{aligned}$$

For a set $A \in \mathcal{Y} \times \mathbb{R}_+$, define the coordinate projection on $\mathbb{R}_+$ by $\Pi_y(A) = \{f \in \mathbb{R}_+ : (y, f) \in A\}$. Since $A(t)$'s are mutually disjoint, applying the coordinate projection to the above identities yields

$$\begin{aligned} \bigcup_{t\in\mathcal{T}_k} \Pi_y(A(t)) &= \Pi_y(S_k), \\ \bigcup_{t\in\mathcal{T}\setminus\mathcal{T}_k} \Pi_y(A(t)) &= \Pi_y(S_k^c \cap \mathcal{S}_{f_Y}). \end{aligned}$$

We take the Lebesgue measure $Leb(\cdot)$ to the above identities. By noting $\Pi_y(A(t))$ are disjoint over t, $Leb\left[\Pi_y(S_k)\right] = p_k(y)$, and $Leb\left[\Pi_y(S_k^c \cap \mathcal{S}_{f_Y})\right] = f_Y(y) - p_k(y)$, we have

$$\begin{aligned} \sum_{t\in\mathcal{T}_k} Leb\left[\Pi_y(A(t))\right] &= p_k(y), \\ \sum_{t\in\mathcal{T}\setminus\mathcal{T}_k} Leb\left[\Pi_y(A(t))\right] &= f_Y(y) - p_k(y). \end{aligned}$$

These equations suggest us to pin down each $h_t(y)$ to $Leb\left[\Pi_y(A(t))\right]$. Each $h_t(y)$ is by construction nonnegative and we can see they agree with the constraints (**). Since k is arbitrary, this completes the proof. ■

For a generalization of Proposition D.2, we without loss of generality assume that $k < l$ implies $\Pr(D_k = 1) \leq \Pr(D_l = 1)$.

Restriction-MSR (Multivariate Z)

Without loss of generality, assume $\Pr(D_k = 1) \leq \Pr(D_{k+1} = 1)$ for all $k = 1, \ldots, (K-1)$. The selection process satisfies MSR if $D_k \leq D_{k+1}$ for all $k = 1, \ldots, (K-1)$ over the entire population.

Proposition D.2'. *Suppose that a population distribution of (Y, T, Z) satisfies RA and MSR.*

(i) Then, the data generating process $\mathcal{P}$ satisfies

$$p_1(y) \leq p_2(y) \leq, \ldots, \leq p_K(y) \quad \mu\text{-a.e.}$$

(ii) The MI mean bounds

$$\max_k \left\{ \int_{\mathcal{Y}} y p_k(y) d\mu + y_l P_k(\{mis\}) \right\} \leq E(Y) \leq \min_k \left\{ \int_{\mathcal{Y}} y p_k(y) d\mu + y_u P_k(\{mis\}) \right\}$$

are identical to the ER mean bounds (1.2.2.6).

Conversely, given the data generating process $\mathcal{P} = (P_1, \ldots, P_K)$, the identification region under RA and MSR is given by

$$\begin{cases} IR_{f_Y}(\mathcal{P}) & \text{if } p_1(y) \leq p_2(y) \leq, \ldots, \leq p_K(y) \quad \mu\text{-a.e.} \\ \emptyset & \text{otherwise.} \end{cases}$$

Proof. (i) From (**), we have

$$\begin{aligned} p_k(y) &= \sum_{t \in \mathcal{T}_k \cap \mathcal{T}_{k+1}} \pi_t g_t(y) + \sum_{t \in \mathcal{T}_k \cap (\mathcal{T} \setminus \mathcal{T}_{k+1})} \pi_t g_t(y), \\ p_{k+1}(y) &= \sum_{t \in \mathcal{T}_{k+1} \cap \mathcal{T}_k} \pi_t g_t(y) + \sum_{t \in \mathcal{T}_{k+1} \cap (\mathcal{T} \setminus \mathcal{T}_k)} \pi_t g_t(y). \end{aligned}$$

Note that the types in $\mathcal{T}_k \cap (\mathcal{T} \setminus \mathcal{T}_{k+1})$ have $D_k = 1$ and $D_{k+1} = 0$ and they do not exist in the population by MSR. Therefore, $\sum_{t \in \mathcal{T}_k \cap (\mathcal{T} \setminus \mathcal{T}_{k+1})} \pi_t g_t(y) = 0$ holds and we conclude

$$p_{k+1}(y) - p_k(y) = \sum_{t \in \mathcal{T}_{k+1} \cap (\mathcal{T} \setminus \mathcal{T}_k)} \pi_t g_t(y) \geq 0.$$

This proposition implies the existence of the dominating density. An application of Proposition C.1 yields (ii).

For the converse statement, we assume that the data generating process reveals $p_1(y) \leq p_2(y) \leq , \ldots, \leq p_K(y)$ μ-a.e. Let us pick an arbitrary $f_Y \in IR_{f_Y}(\mathcal{P})$. We construct a joint distribution of (Y, T, Z) that is compatible with RA and MSR. Note that under MSR the possible types in the population are characterized by a nondecreasing sequence of K binary variables $\{D_k\}_{k=1}^K$. Hence, there are at most $(K+1)$ types allowed to exist in the population. We use t_l^*, $l = 1, \ldots, K$, to indicate the type whose $\{D_k\}_{k=1}^K$ is zero up to the l-th element and one afterwards. We denote the type whose $\{D_k\}_{k=1}^K$ is one for all k by t_0^*. Note that $\mathcal{T}_{l+1} \cap (\mathcal{T} \setminus \mathcal{T}_l)$ the set of types with $D_l = 0$

and $D_{l+1} = 1$ consists of only t_l^* under MSR. Let

$$\begin{aligned} &h_{t_0^*}(y) = p_1(y), \\ &h_{t_l^*}(y) = p_{l+1}(y) - p_l(y), \quad \text{for } l = 1, \ldots, (K-1), \\ &h_{t_K^*}(y) = f_Y(y) - p_K(y), \\ &h_t(y) = 0, \qquad\qquad \text{for the rest of } t \in \mathcal{T}. \end{aligned}$$

This construction provides nonnegative $h_t(y)$'s. The constructed $h_t(y)$'s satisfy (**) since for each $k = 1, \ldots, K$, we have

$$\begin{aligned} \sum_{t \in \mathcal{T}_k} h_t(y) &= \sum_{l=0}^{k-1} h_{t_l^*}(y) = p_k(y), \\ \sum_{t \in \mathcal{T} \setminus \mathcal{T}_k} h_t(y) &= \sum_{l=k}^{K} h_{t_l^*}(y) = f_Y(y) - p_k(y). \end{aligned}$$

Thus, we conclude that there exists a joint probability law of (Y, T, Z) that is compatible with the data generating process and satisfies RA and MSR. Since this way of constructing $h_t(y)$'s is feasible for any $f_Y \in IR_{f_Y}(\mathcal{P})$, we conclude that $IR_{f_Y}(\mathcal{P})$ is the identification under RA and MSR. The emptiness of the identification region follows immediately from (i). ■

1.A.7 Proof of Proposition 1.3.1 and 1.3.2

The following lemma are used for the proof of 1.3.1 and 1.3.2.

Lemma 1.A.2 *Let the data generating process P and Q be given: Fix f_{Y_1} and f_{Y_0} the marginal distributions of Y_1 and Y_0. We can construct a joint distribution of (Y_1, Y_0, T, Z) that is compatible with the data generating process, satisfies RA-causal, 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.*

$$
\begin{aligned}
p_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,a}(y_1), && (1.1.7.20)\\
q_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,a}(y_1), && (1.1.7.21)\\
p_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,n}(y_0), && (1.1.7.22)\\
q_{Y_0}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,n}(y_0), && (1.1.7.23)\\
f_{Y_1}(y_1) - p_{Y_1}(y_1) &= h_{Y_1,d}(y_1) + h_{Y_1,n}(y_1), && (1.1.7.24)\\
f_{Y_1}(y_1) - q_{Y_1}(y_1) &= h_{Y_1,c}(y_1) + h_{Y_1,n}(y_1), && (1.1.7.25)\\
f_{Y_0}(y_0) - p_{Y_1}(y_0) &= h_{Y_0,c}(y_0) + h_{Y_0,a}(y_0), && (1.1.7.26)\\
f_{Y_0}(y_0) - q_{Y_0}(y_0) &= h_{Y_0,d}(y_0) + h_{Y_0,a}(y_0), && (1.1.7.27)\\
\int_{\mathcal{Y}} h_{Y_1,c}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,c}(y_0)d\mu, && (1.1.7.28)\\
\int_{\mathcal{Y}} h_{Y_1,n}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,n}(y_0)d\mu, && (1.1.7.29)\\
\int_{\mathcal{Y}} h_{Y_1,a}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,a}(y_0)d\mu, && (1.1.7.30)\\
\int_{\mathcal{Y}} h_{Y_1,d}(y_1)d\mu &= \int_{\mathcal{Y}} h_{Y_0,d}(y_0)d\mu. && (1.1.7.31)
\end{aligned}
$$

Proof. First, we prove "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 $\pi_t = \int_{\mathcal{Y}} h_{Y_1,t}d\mu = \int_{\mathcal{Y}} h_{Y_1,t}d\mu \geq 0$ for $t \in \{c, n, a, d\}$. We claim that the conditional densities of (Y_1, Y_0, T) given Z can be proposed as

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

By construction ER-causal is satisfied. The scale constraints (1.1.7.28) through (1.1.7.31) imply $f_{Y_1,T|Z}(y_1, T = t|Z = 1) = f_{Y_1,T|Z}(y_1, T = t|Z = 0) = h_{Y_1,t}(y_1)$ and $f_{Y_0,T|Z}(y_0, T = t|Z = 1) = f_{Y_0,T|Z}(y_0, T = t|Z = 0) = h_{Y_0,t}(y_0)$. The constructed probability distributions are compatible with the data generating process. For example, the constraint (1.1.7.20) 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}. Lastly, this way of constructing the population distribution gives the marginal distribution of Y_1 as $\sum_{t=c,n,a,d} f_{Y_1,T}(y_1, t) = \sum_{t=c,n,a,d} h_{Y_1,t}(y_1)$, which by the constraint (1.1.7.20) and (1.1.7.24) coincides with the proposed f_{Y_1}. A similar reasoning also holds for f_{Y_0}

Now, we prove "only if" part. Assume that there exists a population distribution of (Y_1, Y_0, T, Z) that is compatible with the data generating process and RA-causal. Then the constraints (1.1.7.20) through (1.1.7.31) must hold with $h_{Y_1,t}(y_1) = f_{Y_1,T}(y_1, T = t)$ and $h_{Y_0,t}(y_0) = f_{Y_0,T}(y_0, T = t)$ for each t as we discussed in the main text. Since $f_{Y_1,T}(y_1, T = t)$ and $f_{Y_0,T}(y_0, T = t)$ are nonnegative functions with satisfying the scale constraints, the conclusion holds. ■

Lemma 1.A.3 *Let δ_{Y_1}, δ_{Y_0}, λ_{Y_1}, and λ_{Y_0} be the parameters defined in the statement of Proposition 1.3.1.*

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

Proof.

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

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 1.3.1. By Lemma 1.A.3, the identification region of (f_{Y_1}, f_{Y_0}) under ER-causal is obtained by identifying the set of a pair of probability densities (f_{Y_1}, f_{Y_0}) for each of which we can find the nonnegative functions $\{(h_{Y_1,t}, h_{Y_0,t}), t = c, n, a, d\}$ satisfying all the constraints. Consider the case where (P, Q) reveals $\delta_{Y_1} > 1$. Then, for an arbitrary probability density f_{Y_1}, there must exist $A \subset \mathcal{Y}$ with $\mu(A) > 0$ on which $f_{Y_1} - p_{Y_1} < 0$or $f_{Y_1} - q_{Y_1} < 0$. This precludes the possibility that we can find nonnegative h functions satisfying the constraints (1.1.7.20) and (1.1.7.21). Hence, $\delta_{Y_1} \leq 1$ is necessary for the identification region to be nonempty. By the same reasoning, $\delta_{Y_0} \leq 1$ is also necessary for the identification region to be nonempty.

From now on, we assume that (P, Q) reveals $\delta_{Y_1} \leq 1$ and $\delta_{Y_0} \leq 1$. Consider the case of (i) $1 - \delta_{Y_0} < \lambda_{Y_1}$. Choose an arbitrary f_{Y_1} from

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

and choose an arbitrary f_{Y_0} from $F^{env}_{f_{Y_0}}(P,Q)$. Note that $\mathcal{F}^*_{f_{Y_1}}(P,Q)$ is nonempty since it always contains $f_{Y_1} = \underline{f_{Y_1}} + \frac{1-\delta_{Y_1}}{\lambda_{Y_1}} \min\{p_{Y_1}, q_{Y_1}\}$. Define a nonnegative function

$$g_{Y_1}(y_1) = \frac{\lambda_{Y_1} + \delta_{Y_0} - 1}{\int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}, q_{Y_1}\}\right\} d\mu} \min\left\{f_{Y_1}(y_1) - \underline{f_{Y_1}}(y_1), \min\{p_{Y_1}(y_1), q_{Y_1}(u_1)\}\right\},$$

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

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

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 easily seen that the constraints (1.1.7.20) through (1.1.7.27) are satisfied. Also, by utilizing Lemma 1.A.3, we can check the scale constraints (1.1.7.28) through (1.1.7.31) are valid. Hence, we conclude that $\mathcal{F}^*_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ is contained in the identification region under RA-causal.

Next, consider f_{Y_1} that does not satisfy $\int_{\mathcal{Y}} \min\left\{f_{Y_1} - \underline{f_{Y_1}}, \min\{p_{Y_1}, q_{Y_1}\}\right\} d\mu \geq \lambda_{Y_1} + \delta_{Y_0} - 1$. Suppose that the nonnegative functions $\{(h_{Y_1,t}, h_{Y_0,t}), t = c, n, a, d\}$ satisfying the constraints (1.1.7.20) through (1.1.7.27) exist. Then, the constraints (1.1.7.26) and (1.1.7.27) imply that $\int_{\mathcal{Y}} h_{Y_0,a} d\mu \leq 1 - \delta_{Y_0}$. On the other hand, since

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

it follows that

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

Hence, $\int h_{Y_1,a} d\mu > 1 - \delta_{Y_0}$. Thus, the scale constraint for $t = a$ does not hold and 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 1.A.2.

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-causal.

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

Lastly, consider the case of (iii). $1 - \delta_{Y_0} = \lambda_{Y_1}$. As we argued in the main text, 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 meets all the constraints. 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-causal. ■

Proof of Proposition 1.3.2. If the data generating process reveals $p_{Y_1} \geq q_{Y_1}$ μ-a.e. and $q_{Y_0} \geq p_{Y_0}$ μ-a.e., then $1 - \delta_{Y_0} = \lambda_{Y_1}$ holds, and Proposition 1.3.1 (iii) implies that 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 find the distribution of (Y_1, Y_0, T, Z) that satisfies RA-causal. In the proof of Proposition 1.3.1 (iii), we propose one way to find compatible h functions. If we apply it to the current case, we obtain $h_{Y_1,d} = h_{Y_0,d} = 0$. This implies that the imputed population meets $\Pr(T = d) = 0$ and therefore the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal and separable utility is contained in $\mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$. For $(f_{Y_1}, f_{Y_0}) \notin \mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, RA-causal is violated. Hence, we conclude that $\mathcal{F}^{env}_{f_{Y_1}}(P,Q) \times \mathcal{F}^{env}_{f_{Y_0}}(P,Q)$ is the identification region of (f_{Y_1}, f_{Y_0}) under RA-causal and separable utility.

If the data generating process does not reveal $p_{Y_1} \geq q_{Y_1}$ μ-a.e. and $q_{Y_0} \geq p_{Y_0}$ μ-a.e., this implies that under RA there exist a subset A with positive measure on which at least one of the following inequalities hold,

$$\begin{aligned}
\int_A q_{Y_1} d\mu - \int_A p_{Y_1} d\mu &= \Pr(Y_1 \in A, T = d) - \Pr(Y_1 \in A, T = c) \geq 0, \\
\int_A p_{Y_0} d\mu - \int_A q_{Y_0} d\mu &= \Pr(Y_0 \in A, T = d) - \Pr(Y_0 \in A, T = c) \geq 0.
\end{aligned}$$

These inequalities imply the population must allow $T = d$ with positive probability and no population can satisfies RA-causal and separable utility. Hence the identification region is empty. ■

1.A.8 Proof of Proposition 1.3.3

Proof. We first consider bounding the mean of Y_1 when the data generating process reveals (i) $1-\delta_{Y_0} < \lambda_{Y_1}$. Since $\int_{y_l}^{y_u} y_1 dF \leq \int_{y_l}^{y_u} y_1 dF'$ whenever F is first-order stochastically dominated by F', the lower bound of $E(Y_1)$ is obtained if we can find the density $f_{Y_1}^{upper}$ whose cdf satisfies $\int_{y_l}^{y} f_{Y_1}^{upper} dy_1 \geq \int_{y_l}^{y} \tilde{f}_{Y_1} dy_1$ at every $y \in \mathcal{Y}$ for all $\tilde{f}_{Y_1} \in \mathcal{F}^*_{f_{Y_1}}(P,Q)$. Let us guess such $f_{Y_1}^{upper}$ to be

$$f_{Y_1}^{upper}(y_1) = 1\{y_1 = y_l\}\lambda_{Y_0} + \underline{f_{Y_1}}(y_1) + 1\{y_1 \in [y_l, y^*_{1,l}]\}\min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\}$$

and verify $\int_{y_l}^{y} f_{Y_1}^{upper} dy_1 \geq \int_{y_l}^{y} \tilde{f}_{Y_1} dy_1$ at every $y \in \mathcal{Y}$ for all $\tilde{f}_{Y_1} \in \mathcal{F}^*_{f_{Y_1}}(P,Q)$. By Lemma 1.A.2 and Proposition 1.3.1, there exist nonnegative functions $\{\tilde{h}_{Y_1,t}, t = c,n,a,d\}$ by which $\tilde{f}_{Y_1}$ can be represented as

$$\begin{aligned}
\tilde{f}_{Y_1} &= \sum_t \tilde{h}_{Y_1,t} \\
&= p_{Y_1} + q_{Y_1} - \tilde{h}_{Y_1,a} + \tilde{h}_{Y_1,n} \\
&= \underline{f_{Y_1}} + \min\{p_{Y_1}, q_{Y_1}\} - \tilde{h}_{Y_1,a} + \tilde{h}_{Y_1,n}, \qquad (1.1.8.32)
\end{aligned}$$

where in the second line we use the constraints (1.1.7.20) and (1.1.7.21). The difference between $\int_{y_l}^{y} f_{Y_1}^{upper} dy_1$ and $\int_{y_l}^{y} \tilde{f}_{Y_1} dy_1$ is written as

$$\begin{aligned}
&\int_{y_l}^{y} f_{Y_1}^{upper} dy_1 - \int_{y_l}^{y} \tilde{f}_{Y_1} dy_1 \\
= \; &\lambda_{Y_0} + \int_{y_l}^{y} \tilde{h}_{Y_1,a} dy_1 - \int_{y_l}^{y} \tilde{h}_{Y_1,n} dy_1 - \int_{y_l}^{y} 1\{y_1 \in (y^*_{1,l}, y_u]\}\min\{p_{Y_1}, q_{Y_1}\} dy_1. \qquad (1.1.8.33)
\end{aligned}$$

Note that $\int_{y_l}^{y} \tilde{h}_{Y_1,n} dy_1$ is bounded by

$$\begin{aligned}
\int_{y_l}^{y} \tilde{h}_{Y_1,n} dy_1 &\leq \int_{\mathcal{Y}} \tilde{h}_{Y_1,n} dy_1 \\
&= 1 - \delta_{Y_1} - \lambda_{Y_1} + \int_{\mathcal{Y}} \tilde{h}_{Y_1,a} dy_1 \qquad (1.1.8.34)
\end{aligned}$$

where the last line follows from the integral of (1.1.8.32). The last term in (1.1.8.33) is rewritten as

$$\begin{aligned}
&\int_{y_l}^{y} 1\{y_1 \in (y^*_{1,l}, y_u]\}\min\{p_{Y_1}, q_{Y_1}\} dy_1 \\
= \; &\int_{\mathcal{Y}} 1\{y_1 \in (y^*_{1,l}, y_u]\}\min\{p_{Y_1}, q_{Y_1}\} dy_1 - \int_{y}^{y_u} 1\{y_1 \in (y^*_{1,l}, y_u]\}\min\{p_{Y_1}, q_{Y_1}\} dy_1 \\
= \; &1 - \delta_{Y_0} - \int_{y}^{y_u} 1\{y_1 \in (y^*_{1,l}, y_u]\}\min\{p_{Y_1}, q_{Y_1}\} dy_1 \qquad (1.1.8.35)
\end{aligned}$$

where the last line follows by the definition of $y^*_{1,l}$. By plugging (1.1.8.34) and (1.1.8.35) into (1.1.8.33),

$$\begin{aligned}
&\int_{y_l}^{y} f_{Y_1}^{upper} dy_1 - \int_{y_l}^{y} \tilde{f}_{Y_1} dy_1 \\
\geq\ & \delta_{Y_1} + \delta_{Y_0} + \lambda_{Y_1} + \lambda_{Y_0} - 2 + \int_{y}^{y_u} \left[1\{y_1 \in (y^*_{1,l}, y_u]\} \min\{p_{Y_1}, q_{Y_1}\} - \tilde{h}_{Y_1,a} \right] dy_1 \\
=\ & \int_{y}^{y_u} \left[1\{y_1 \in (y^*_{1,l}, y_u]\} \min\{p_{Y_1}, q_{Y_1}\} - \tilde{h}_{Y_1,a} \right] dy_1 \qquad (1.1.8.36)
\end{aligned}$$

where we use Lemma 1.A.3 to obtain the equality. For $y > y^*_{1,l}$, this integral is nonnegative since the constraints (1.1.7.20) and (1.1.7.21) imply $\tilde{h}_{Y_1,a} \leq \min\{p_{Y_1}, q_{Y_1}\}$. For $y \leq y^*_{1,l}$,

$$\begin{aligned}
&\int_{y}^{y_u} \left[1\{y_1 \in (y^*_{1,l}, y_u]\} \min\{p_{Y_1}, q_{Y_1}\} - \tilde{h}_{Y_1,a} \right] dy_1 \\
=\ & 1 - \delta_{Y_0} - \int_{y}^{y_l} \tilde{h}_{Y_1,a} dy_1 \\
\geq\ & 1 - \delta_{Y_0} - \int_{\mathcal{Y}} \tilde{h}_{Y_1,a} dy_1 \\
=\ & 1 - \delta_{Y_0} - \int_{\mathcal{Y}} \tilde{h}_{Y_0,a} dy_1 \\
\geq\ & 0 \qquad (1.1.8.37)
\end{aligned}$$

where the fourth line follows from the scale constraint $\int_{\mathcal{Y}} \tilde{h}_{Y_1,a} dy_1 = \int_{\mathcal{Y}} \tilde{h}_{Y_0,a} dy_1$ and the last inequality follows since $\int_{\mathcal{Y}} \tilde{h}_{Y_o,a} dy_1 \leq 1 - \delta_{Y_0}$ by the constraints (1.1.7.26) and (1.1.7.27). Thus, the cdf upper bound of Y_1 under RA-causal is given by the cdf of $f_{Y_1}^{upper}$.

Next, we derive the cdf lower bound of Y_1 under RA-causal. Consider

$$f_{Y_1}^{lower}(y_1) = \underline{f_{Y_1}}(y_1) + 1\{y_1 \in [y_l, y^*_{1,l}]\} \min\{p_{Y_1}(y_1), q_{Y_1}(y_1)\} + 1\{y_1 = y_u\}\lambda_{Y_0}$$

and $\int_y^{y_u} f_{Y_1}^{upper} dy_1 - \int_y^{y_u} \tilde{f}_{Y_1} dy_1$. By repeating a similar procedure used to derive (1.1.8.36), we obtain

$$\begin{aligned}
&\int_{y}^{y_u} f_{Y_1}^{lower} dy_1 - \int_{y}^{y_u} \tilde{f}_{Y_1} dy_1 \\
\geq\ & \int_{y_l}^{y} \left[1\{y_1 \in [y_l, y^*_{1,u})\} \min\{p_{Y_1}, q_{Y_1}\} - \tilde{h}_{Y_1,a} \right] dy_1.
\end{aligned}$$

By the construction of $y^*_{1,u}$ and the same reasoning made in deriving (1.1.8.37), $\int_y^{y_u} f_{Y_1}^{upper} dy_1 - \int_y^{y_u} \tilde{f}_{Y_1} dy_1 \geq 0$. Therefore, the cdf of $f_{Y_1}^{lower}$ first-order stochastically dominates the cdf of $\tilde{f}_{Y_1} \in \mathcal{F}^*_{\tilde{f}_{Y_1}}(P, Q)$.

By taking the mean of Y_1 with respect to $f_{Y_1}^{upper}$ and $f_{Y_1}^{lower}$, we obtain the tight bounds of $E(Y_1)$

under RA-causal.

$$\begin{aligned} & \int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1) dy_1 + \int_{y_l}^{y^*_{1,l}} y_1 \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 + \lambda_{Y_0} y_l \\ \leq \; & E(Y_1) \\ \leq \; & \int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1) dy_1 + \int_{y^*_{1,u}}^{y_u} y_1 \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 + \lambda_{Y_0} y_u. \end{aligned}$$

Since the identification region of f_{Y_0} in this case is $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, the tight bounds for $E(Y_0)$ are

$$\int_{y_l}^{y_u} y_0 \underline{f_{Y_0}}(y_0) dy_0 + (1-\delta_{Y_0}) y_l \leq E(Y_0) \leq \int_{y_l}^{y_u} y_0 \underline{f_{Y_0}}(y_0) dy_0 + (1-\delta_{Y_0}) y_u.$$

Given that the identification region of under RA-causal is the Cartesian product of $\mathcal{F}^{*}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, the tight bounds for $E(Y_1) - E(Y_0)$ under ER-causal become

$$\begin{aligned} & \int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1) dy_1 + \int_{y_l}^{y^*_{1,l}} y_1 \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 + \lambda_{Y_0} y_l - \int_{y_l}^{y_u} y_0 \underline{f_{Y_0}}(y_0) dy_0 - (1-\delta_{Y_0}) y_u \\ \leq \; & E(Y_1) - E(Y_0) \\ \leq \; & \int_{y_l}^{y_u} y_1 \underline{f_{Y_1}}(y_1) dy_1 + \int_{y^*_{1,u}}^{y_u} y_1 \min\left\{p_{Y_1}(y_1), q_{Y_1}(y_1)\right\} dy_1 + \lambda_{Y_0} y_u - \int_{y_l}^{y_u} y_0 \underline{f_{Y_0}}(y_0) dy_0 - (1-\delta_{Y_0}) y_l. \end{aligned}$$

For the case of (ii) $1 - \delta_{Y_0} > \lambda_{Y_1}$, the tight average treatment effect bounds are derived by a symmetric argument to the case of (i) and we omit a proof for brevity.

For the case (iii) $1 - \delta_{Y_0} = \lambda_{Y_1}$, since the identification region is given by the Cartesian product of $\mathcal{F}^{env}_{f_{Y_1}}(P,Q)$ and $\mathcal{F}^{env}_{f_{Y_0}}(P,Q)$, the tight bounds coincide with the outer bounds. ■

Bibliography

[1] Andrews, D. W. K. and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies,* 65, 497-517.

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

[4] Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323-363.

[5] Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189-218.

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

[7] Heckman, J. J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313-318.

[8] 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.

[9] 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.

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

[11] Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845-1857.

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

[13] Manski, C. F. (1989): "Anatomy of the Selection Problem." *Journal of Human Resources*, 24, 343-360.

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

[15] 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.

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

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

[18] Manski, C. F. and J. Pepper (2000): "Monotone Instrument Variables: With Application to the Returns to Schooling," *Econometrica*, 68, 997-1010.

[19] Pakes, A., J. Porter, K, Ho, and J. Ishii (2006): "Moment Inequalities and Their Application," manuscript, Harvard University.

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

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

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

[23] Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66: 688-701.

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

Chapter 2

Testing for Instrument Independence in the Selection Model

2.1 Introduction

This paper develops a nonparametric specification test for the instrument exclusion restriction in the sample selection model. In order to obtain a testable implication for the instrument exclusion restriction in a nonparametric way, this paper focuses on the *identification region* for f_Y considered in Chapter 2. Since the identification region under exclusion restriction is the set of outcome distributions that are compatible with the empirical evidence and the exclusion restriction restrictions, an empty identification region implies a misspecification of the exclusion restriction. Hence, our specification test infers from data the emptiness of the identification region.

Specification tests based on the emptiness of the identification region for the partially identified parameters have been studied in the literature of the moment inequality model.[1] Our analysis, however, differs from the moment inequality model since the independence restriction we consider is a distributional restriction rather than a moment restriction, and, especially for continuous Y, the identification region for the outcome distribution cannot be expressed by a finite number of moment inequalities. In Chapter 1, we showed that the size of the identification region for the outcome distribution is characterized by a scalar parameter, the *integrated envelope*: the integral of the envelope over the conditional densities of the observed Y given Z. In particular, the identification region

[1] In the partially identified model with moment inequalities, a specification test for moment restrictions is obtained as a by-product of the confidence sets for the partially identified parameters, that is, we reject the null restriction if the confidence set is empty. A list of the literature that analyses the confidence sets in the moment inequality model contains Andrews, Berry and Jia (2004), Andrews and Guggenberger (2008), Andrews and Soares (2007), Bugni (2008), Canay (2007), Chernozhukov, Hong, and Tamer (2007), Guggenberger, Hahn, and Kim (2008), Imbens and Manski (2004), Pakes, Porter, Ho, and Ishii (2006), Romano and Shaikh (2008a, 2008b), and Rosen (2008).

is empty if and only if the integrated envelope exceeds one. We therefore obtain a nonparametric specification test for the instrument exclusion restriction by developing an inferential procedure for whether the integrated envelope exceeds one. We propose an estimator for the integrated envelope and derive its asymptotic distribution. An asymptotically size correct specification test for instrument independence is obtained by inverting the one-sided confidence intervals for the integrated envelope. A parameter similar to the integrated envelope is considered in Manski (2003) and Pearl (1994b), but its estimation and inference have not been analyzed. Hence, this paper is the first that provides a formal asymptotic analysis for the integrated envelope.

Another contribution of the paper is the implementation of the test procedure. The asymptotic distribution of the integrated envelope estimator is given by a supremum functional of Gaussian processes and it is difficult to obtain the critical values analytically. Furthermore, due to a non-pivotal feature of the asymptotic distribution, the standard nonparametric bootstrap fails to yield asymptotically valid critical values (Andrews (2000)). We therefore develop a bootstrap procedure for the integrated envelope estimator and verify its asymptotic validity. Similarly to the bootstrap procedure for the moment inequality model (Bugni (2008) and Canay (2007)), we first select the asymptotic distribution for which the bootstrap approximation is targeted. Given the targeted asymptotic distribution, we bootstrap the empirical processes so as to approximate the Gaussian processes (van der Vaart and Wellner (1996)).

Blundell, Gosling, Ichimura, and Meghir (2007) consider testing the instrument independence by inferring whether the bounds for the cumulative distribution function (cdf) of f_Y intersects or not. Our specification test, however, differs from their method in the following ways. First, their procedure tests the emptiness of potentially non-tight cdf bounds for f_Y while our procedure always tests the emptiness of the *tightest* cdf bounds. Therefore, our procedure have more refuting power for the instrument exclusion than theirs. Second, the asymptotic validity of their bootstrap procedure is not formally investigated and its asymptotic property is not known. Our bootstrap algorithm in contrast has an asymptotic justification in terms of correct size.

Monte Carlo simulations illustrate the finite sample performance of our bootstrap test procedure. While the standard subsampling procedure by Politis and Romano (1994) is shown to be valid, we present simulation evidence that our bootstrap has better finite sample performance. We apply the proposed test procedure to the classical model of self-selection into the labor market using data from Blundell et al. (2007). We test whether the measure of out-of-work income constructed in Blundell et al. (2003) is independent of the potential wage. Our test results provide an evidence that the exclusion restriction for the out-of-work income is misspecified. Since our procedure tests the emptiness of the identification region, this conclusion is based on the empirical evidence alone and free from any assumptions about the potential wage distribution and the selection mechanism.

The remainder of the paper is organized as follows. Section 2.2 develops an estimator of the integrated envelope and derive its asymptotic distribution. Based on this asymptotic distribution, Section 2.3 formalizes the test procedure by developing an asymptotically valid bootstrap algorithm. We also demonstrate the validity of subsampling. Section 2.4 provides simulation results and

compares the finite sample performance of the bootstrap with subsampling. For simplicity, our analysis is limited to the case of a binary instrument up to Section 2.4. In Section 2.5, we cover the model with a multi-valued discrete instrument. In order to illustrate the use of testing procedure, Section 2.6 tests whether the out-of-work income constructed in Blundell et al. (2003) is independent of the potential wage. Section 2.7 concludes. Proofs are provided in Appendices.

2.2 Estimation of the integrated envelope and a specification test of the exclusion restriction

The identification analysis in Chapter 1 clarified that the emptiness of the identification region under ER $IR_{f_Y}(P,Q)$ is summarized by the integrated envelope $\delta(P,Q)$. We also showed that in the single missing outcome model the stronger restriction RA does not narrow $IR_{f_Y}(P,Q)$. These results imply that $\delta(P,Q)$ is the only relevant parameter for the purpose of refuting the instrument exogeneity. This paper focuses on estimation and inference for $\delta(P,Q)$ so as to develop a specification test for the instrument independence assumption.

Without losing any distributional information of data $(Y \cdot D, D, Z)$, we define an outcome observation recorded in data by $Y_{data} \equiv DY + (1-D)\{mis\}$ and express data as i.i.d observations of $(Y_{data,i}, Z_i)$, $i = 1, \ldots, N$, where $\{mis\}$ indicates that the observation of Y is missing. Clearly, the data generating process $P = (P(\cdot), P_{mis})$ and $Q = (Q(\cdot), Q_{mis})$ are interpreted as the conditional distributions of the random variable Y_{data} given Z, which have the support $\mathcal{Y} \cup \{mis\}$. We divide the full sample into two subsamples based on the assigned value of $Z \in \{1,0\}$. We denote the size of these subsamples by $m = \sum_{i=1}^{N} Z_i$ and $n = \sum_{i=1}^{N}(1-Z_i)$. We assume Z_i is Bernoulli with mean $\lambda \equiv \Pr(Z=1) \in [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$ and define $\lambda_N \equiv m/N$. We adopt the two-sample problem with nonrandom sample size, i.e., our asymptotic analysis is conditional on the sequence $\{Z_i : i = 1, 2, ...\}$. Since $\lambda_N \to \lambda$, $m \to \infty$, and $n \to \infty$ as $N \to \infty$, we interpret the stochastic limit with respect to $N \to \infty$ equivalent to the limit with respect to $m \to \infty$, $n \to \infty$, and $\lambda_N \to \lambda$.

The test strategy considered in this paper is as follows. The null hypothesis is that $IR_{f_Y}(P,Q)$ is nonempty, that is, $\delta(P,Q) \leq 1$. Let $\hat{\delta}$ be the point estimator of $\delta(P,Q)$ such that $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ has an asymptotic distribution,

$$\sqrt{N}(\hat{\delta} - \delta(P,Q)) \rightsquigarrow J(\cdot; P, Q, \lambda),$$

where "$\rightsquigarrow$" denotes weak convergence and $J(\cdot; P, Q, \lambda)$ represents the cdf of the asymptotic distribution which can depend on P, Q, and λ. We infer whether or not $\delta(P,Q) \leq 1$ with a prespecified maximal false rejection rate α by inverting the one-sided confidence intervals with coverage $1-\alpha$. That is, our goal is to obtain $\hat{c}_{1-\alpha}$, a consistent estimator of the $(1-\alpha)$-th quantile of $J(\cdot; P, Q, \lambda)$, and to check whether the one-sided confidence intervals $[\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}}, \infty)$ contain 1 or not. We reject the null hypothesis if we observe $\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} > 1$. This procedure provides a *pointwise* asymptotically

size correct test[2] since for every (P,Q) satisfying the null $\delta(P,Q) \leq 1$, we have

$$\begin{aligned} Prob_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} > 1\right) &\leq Prob_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} > \delta(P,Q)\right) \\ &= Prob_{P,Q,\lambda_N}\left(\sqrt{N}(\hat{\delta} - \delta(P,Q)) > \hat{c}_{1-\alpha}\right) \\ &\xrightarrow{N\to\infty} 1 - J(c_{1-\alpha}; P, Q, \lambda) = \alpha. \end{aligned}$$

We decompose our theoretical development into two parts. First, we develop an estimator of $\delta(P,Q)$ and derive the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$. Second, we focus on how to consistently estimate quantiles of the asymptotic distribution $J(\cdot; P, Q, \lambda)$.

2.2.1 An illuminating example: binary Y

To motivate our estimation and inference procedure for $\delta(P,Q)$, we consider a simple example in which Y is binary. The main focus of this section is to illuminate the non-pivotal asymptotic distribution for the estimation of $\delta(P,Q)$. We also illustrate how our bootstrap strategy resolves the problem.

Estimation of δ

When Y is binary, P and Q are represented by the three probabilities, (p_1, p_0, p_{mis}) and (q_1, q_0, q_{mis}), where p_y and q_y, $y = 1, 0, \{mis\}$, are the probabilities of $Y_{data} = y$ given $Z = 1$ and $Z = 0$ respectively. Here, the integrated envelope $\delta = \delta(P,Q)$ is defined as

$$\delta \equiv \max\{p_1, q_1\} + \max\{p_0, q_0\}. \tag{2.2.1.1}$$

A sample analogue estimator for δ is constructed as

$$\hat{\delta} = \max\{\hat{p}_1, \hat{q}_1\} + \max\{\hat{p}_0, \hat{q}_0\},$$

where $(\hat{p}_1, \hat{p}_0)$ and $(\hat{q}_1, \hat{q}_0)$ are the maximum likelihood estimators of (p_1, p_0) and (q_1, q_0). Here, the maximum likelihood estimators are the sample fractions of the observations classified in the corresponding category conditional on Z. The standard central limit theorem yields

$$\sqrt{N}\begin{pmatrix} \hat{p}_1 - p_1 \\ \hat{p}_0 - p_0 \\ \hat{q}_1 - q_1 \\ \hat{q}_0 - q_0 \end{pmatrix} \rightsquigarrow \begin{pmatrix} X_1 \\ X_0 \\ W_1 \\ W_0 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_{P,\lambda} & O \\ O & \Sigma_{Q,\lambda} \end{pmatrix}\right),$$

[2] Andrews and Guggenberger (2008), Canay (2007), Imbens and Manski (2004), and Romano and Shaikh (2008) analyze the uniform asymptotic validity of the confidence regions for partially identified parameters in the moment inequality model. In this paper, we establish the pointwise asymptotic valdity of our inferential procedure for the integrated envelope. It is not yet known whether our inferential procedure for the integrated envelope is uniformly asymptotically valid.

where

$$\Sigma_{P,\lambda} = \lambda^{-1} \begin{pmatrix} p_1(1-p_1) & -p_1 p_0 \\ -p_1 p_0 & p_0(1-p_0) \end{pmatrix} \text{ and}$$
$$\Sigma_{Q,\lambda} = (1-\lambda)^{-1} \begin{pmatrix} q_1(1-q_1) & -q_1 q_0 \\ -q_1 q_0 & q_0(1-q_0) \end{pmatrix}.$$

Although the maximum likelihood estimators for p and q are asymptotically normal, $\hat{\delta}$ is not necessarily normal due to the max operator. Specifically, asymptotic normality fails when the data generating process has *ties* in the max operator in (2.2.1.1), meaning $p_1 = q_1$ and/or $p_0 = q_0$. For example, consider the case of $p_1 = q_1$ and $p_0 > q_0$. Then, it follows that

$$\begin{aligned} \sqrt{N}(\hat{\delta} - \delta) &= \max \left\{ \begin{array}{c} \sqrt{N}(\hat{p}_1 - p_1) \\ \sqrt{N}(\hat{q}_1 - q_1) \end{array} \right\} + \max \left\{ \begin{array}{c} \sqrt{N}(\hat{p}_0 - p_0) \\ \sqrt{N}(\hat{q}_0 - q_0) + \sqrt{N}(q_0 - p_0) \end{array} \right\} \\ &\rightsquigarrow \max \left\{ \begin{array}{c} X_1 \\ W_1 \end{array} \right\} + X_0, \end{aligned}$$

where the second max operation in the first line converges in distribution to X_0 since $\sqrt{N}(q_0 - p_0) \to -\infty$. In contrast, when there are no ties ($p_1 \neq q_1$ and $p_0 \neq q_0$), $\sqrt{N}(\hat{\delta} - \delta)$ is asymptotically normal since it converges to the sum of the two normal random variables.

In order to summarize all the possible asymptotic distributions, we introduce

$$\begin{aligned} \delta_1 &= p_1 + p_0, \quad G_1 = X_1 + X_0, \\ \delta_2 &= p_1 + q_0, \quad G_2 = X_1 + W_0, \\ \delta_3 &= q_1 + p_0, \quad G_3 = W_1 + X_0, \\ \delta_4 &= q_1 + q_0, \quad G_4 = W_1 + W_0, \end{aligned}$$

where δ_j, $j = 1, \ldots, 4$, are the candidates of δ and at least one of them achieves the true integrated envelope. G_j each represents the Gaussian random variable that is obtained from the asymptotic distribution of $\sqrt{N}(\hat{\delta}_j - \delta_j)$, where $\hat{\delta}_j$ is the sample analogue estimator of δ_j. Using this notation, the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is expressed as

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \max_{\{j:\delta_j = \delta\}} \{G_j\}. \tag{2.2.1.2}$$

The index set of the max operator $\{j : \delta_j = \delta\}$ indicates whether there are ties between P and Q. For instance, in case of $p_1 = q_1$ and $p_0 > q_0$, we have $\{j : \delta_j = \delta\} = \{1, 3\}$. If $\{j : \delta_j = \delta\}$ is a singleton, we obtain asymptotic normality, while if it contains more than one element, asymptotic normality fails and the asymptotic distribution is given by the extremum value among the normal random variables $\{G_j : \delta_j = \delta\}$. Thus, $\sqrt{N}(\hat{\delta} - \delta)$ is not uniformly asymptotically normal over the

data generating process.

The failure of uniform asymptotic normality of a statistic is known as discontinuity of the asymptotic distribution and it arises in many contexts in econometrics (e.g., weak instruments, unit root, etc.). The integrated envelope also has this issue. This raises difficulties in conducting inference on δ since we do not know which asymptotic distribution gives a better approximation for the sampling distribution of $\sqrt{N}(\hat{\delta} - \delta)$.

Inconsistency of the nonparametric bootstrap

The issue of discontinuity of the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ cannot be bypassed by standard implementation of the nonparametric bootstrap. By following an argument similar to Andrews (2000), it can be shown that the nonparametric bootstrap fails to consistently estimate the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$. The case of binary Y provides a canonical example for this.

In the standard nonparametric bootstrap, we form a bootstrap sample using m i.i.d. draws from the subsample $\{Y_{data,i} : Z_i = 1\}$ and n i.i.d. draws from the subsample $\{Y_{data,i} : Z_i = 0\}$. Let $\hat{\delta}^* = \max\{\hat{p}_1^*, \hat{q}_1^*\} + \max\{\hat{p}_0^*, \hat{q}_0^*\}$ be the bootstrap estimator of δ where $(\hat{p}_1^*, \hat{p}_0^*)$ and $(\hat{q}_1^*, \hat{q}_0^*)$ are the maximum likelihood estimators computed from the bootstrap sample. If the standard nonparametric bootstrap were consistent, then, for almost every sequence of the original sample, we could replicate the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ by that of $\sqrt{N}(\hat{\delta}^* - \hat{\delta})$. This is, however, not the case when there are ties between P and Q.

Consider again the case of $p_1 = q_1$ and $p_0 > q_0$ where the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is given by $\max\{X_1, W_1\} + X_0$. The bootstrap statistic $\sqrt{N}(\hat{\delta}^* - \hat{\delta})$ is written as

$$\begin{aligned}
\sqrt{N}(\hat{\delta}^* - \hat{\delta}) &= \sqrt{N}(\max\{\hat{p}_1^*, \hat{q}_1^*\} + \max\{\hat{p}_0^*, \hat{q}_0^*\}) - \sqrt{N}(\max\{\hat{p}_1, \hat{q}_1\} + \max\{\hat{p}_0, \hat{q}_0\}) \\
&= \underbrace{\max\left\{\sqrt{N}(\hat{p}_1^* - \hat{q}_1^*), 0\right\} - \max\left\{\sqrt{N}(\hat{p}_1 - \hat{q}_1), 0\right\}}_{(i)} \\
&\quad \underbrace{+\max\left\{\sqrt{N}(\hat{q}_0^* - \hat{p}_0^*), 0\right\} - \max\left\{\sqrt{N}(\hat{q}_0 - \hat{p}_0), 0\right\}}_{(ii)} \\
&\quad \underbrace{+\sqrt{N}(\hat{q}_1^* - \hat{q}_1) + \sqrt{N}(\hat{p}_0^* - \hat{p}_0)}_{(iii)}. \qquad (2.2.1.3)
\end{aligned}$$

We denote the probability distribution for the bootstrap sample with size N by $\{\mathbb{P}_N : N \geq 1\}$. Let ω be an element of the sample space Ω. Since $\sqrt{N}(\hat{p}_1 - \hat{q}_1)$ weakly converges to the Gaussian random variable $G = X_1 - W_1$, we can find an Ω on which $\hat{p}_1$, $\hat{q}_1$, and G are defined and $\sqrt{N}(\hat{p}_1(\omega) - \hat{q}_1(\omega)) \to_{N\to\infty} G(\omega)$ for almost all $\omega \in \Omega$ (the Almost Sure Representation Theorem, see, e.g., Pollard (1984)). The central limit theorem of triangular arrays and the strong law of large numbers

imply, for almost every $\omega \in \Omega$,

$$\sqrt{N}\begin{pmatrix} \hat{p}_1^* - \hat{p}_1(\omega) \\ \hat{p}_0^* - \hat{p}_0(\omega) \\ \hat{q}_1^* - \hat{q}_1(\omega) \\ \hat{q}_0^* - \hat{q}_0(\omega) \end{pmatrix} \rightsquigarrow \begin{pmatrix} X_1 \\ X_0 \\ W_1 \\ W_0 \end{pmatrix}, \qquad (2.2.1.4)$$
$$\hat{q}_0(\omega) - \hat{p}_0(\omega) \rightarrow q_0 - p_0 < 0.$$

Let us consider the event $B_c = \{\omega \in \Omega : G(\omega) < -c\}$ for a constant $c > 0$. Clearly, $\Pr(B_c) > 0$ holds. For $\omega \in B_c$, the stochastic limit of each term in (2.2.1.3) is obtained as

$$\begin{aligned}
(i) &= \max\left\{\sqrt{N}(\hat{p}_1^* - \hat{p}_1(\omega)) - \sqrt{N}(\hat{q}_1^* - \hat{q}_1(\omega)) + \sqrt{N}(\hat{p}_1(\omega) - \hat{q}_1(\omega)), 0\right\} \\
&\quad - \max\left\{\sqrt{N}(\hat{p}_1(\omega) - \hat{q}_1(\omega)), 0\right\} \\
&\leq \max\left\{\sqrt{N}(\hat{p}_1^* - \hat{p}_1(\omega)) - \sqrt{N}(\hat{q}_1^* - \hat{q}_1(\omega)) - c, 0\right\} \quad \text{for sufficiently large } N, \\
&\rightsquigarrow \max\{X_1 - W_1 - c, 0\}, \\
(ii) &= \max\left\{\sqrt{N}(\hat{q}_0^* - \hat{q}_0(\omega)) - \sqrt{N}(\hat{p}_0^* - \hat{p}_0(\omega)) + \sqrt{N}(\hat{q}_0(\omega) - \hat{p}_0(\omega)), 0\right\} \\
&\quad - \max\left\{\sqrt{N}(\hat{q}_0(\omega) - \hat{p}_0(\omega)), 0\right\} \\
&\rightarrow 0 \text{ in probability with respect to } \{\mathbb{P}_N : N \geq 1\},
\end{aligned}$$

and the term (iii) weakly converges to $W_1 + X_0$ by (2.2.1.4). To sum up, we have for large N

$$\sqrt{N}(\hat{\delta}^* - \hat{\delta}(\omega)) \leq \max\{X_1 - c, W_1\} + X_0 \leq \max\{X_1, W_1\} + X_0, \qquad (2.2.1.5)$$

where the second inequality is strict with positive probability in terms of the randomness in drawing a bootstrap sample. Note that the last terms in (2.2.1.5) have the same probability law as the limiting distribution of $\sqrt{N}(\hat{\delta} - \delta)$. Therefore, along the sampling sequence of $\omega \in B_c$, the asymptotic distribution of the bootstrap statistic $\sqrt{N}(\hat{\delta}^* - \hat{\delta}(\omega))$ fails to coincide with that of $\sqrt{N}(\hat{\delta} - \delta)$. Provided that $\Pr(B_c) > 0$, this refutes the consistency of the nonparametric bootstrap.

Asymptotically valid inference

We provide two procedures for asymptotically valid inference on δ. The first approach estimates the asymptotic distribution $\max_{\{j:\delta_j=\delta\}}\{G_j\}$ in two steps. In the first step, we estimate the index set $\mathbb{V}^{\max} \equiv \{j : \delta_j = \delta\}$. In the second step, we estimate the joint distribution of G_j's. The latter part is straightforward in this example since the G_j's are Gaussian and their covariance matrix can be consistently estimated. For the former part, we estimate $\mathbb{V}^{\max}$ using the sequence of *slackness variables* $\{\eta_N : N \geq 1\}$,

$$\hat{\mathbb{V}}^{\max}(\eta_N) = \{j \in \{1,2,3,4\} : \sqrt{N}(\hat{\delta} - \hat{\delta}_j) \leq \eta_N\}.$$

In this construction of $\hat{\mathbb{V}}^{\max}(\eta_N)$, we determine which δ_j achieves the population δ in terms of whether the estimator of δ_j is close to $\hat{\delta} = \max_j\{\hat{\delta}_j\}$ or not. The value of $\eta_N/\sqrt{N}$ gives the cut-off value for how small $(\hat{\delta} - \hat{\delta}_j)$ should be in order for such j to be included in the estimator of $\mathbb{V}^{\max}$. This estimator for $\mathbb{V}^{\max}$ is asymptotically valid[3] if the slackness sequence $\{\eta_N : N \geq 1\}$ meets the following conditions,

$$\frac{\eta_N}{\sqrt{N}} \to 0 \text{ and } \frac{\eta_N}{\sqrt{\log\log N}} \to \infty.$$

That is, η_N diverges to positive infinity faster than $\sqrt{\log\log N}$, but not as fast as $\sqrt{N}$. This speed of divergence is implied by the law of iterated logarithm (see, e.g., Shiryaev (1996)).

By combining these two estimations, we are able to consistently estimate the asymptotic distribution $\max_{j\in\mathbb{V}^{\max}}\{G_j\}$ by

$$\max_{j\in\hat{\mathbb{V}}^{\max}(\eta_N)}\{\hat{G}_j\}$$

where the $\hat{G}_j$'s are Gaussian and their covariance matrix is estimated from the sample.

Instead of plugging in $\hat{G}_j$'s, we can incorporate the nonparametric bootstrap for estimating the asymptotic distribution; given the estimator $\hat{\mathbb{V}}^{\max}(\eta_N)$, we resample,

$$\max_{j\in\hat{\mathbb{V}}^{\max}(\eta_N)}\{\sqrt{N}(\hat{\delta}_j^* - \hat{\delta}_j)\}$$

where $\hat{\delta}_j^*$ is the bootstrapped $\hat{\delta}_j$. Since the standard argument of the bootstrap consistency shows $\sqrt{N}(\hat{\delta}_j^* - \hat{\delta}_j) \rightsquigarrow G_j$, we can build in the nonparametric bootstrap inside the max operator so as to obtain the consistent estimator for the asymptotic distribution. In Section 4, we extend this approach to a general setting.

As Andrews (2000) points out, another asymptotically valid method is subsampling (Politis and Romano (1994)). In subsampling, we resample fewer observations than the original sample randomly *without* replacement, i.e., we resample $b_m(< m)$ observations from $\{Y_{data,i} : Z_i = 1\}$ and $b_n(< n)$ observations from $\{Y_{data,i} : Z_i = 0\}$. By tuning the blocksizes to $(b_m, b_n) \to \infty$, $(b_m/m, b_n/n) \to 0$, and $b_m/(b_m + b_n) \to \lambda$, the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is consistently estimated by the repeated sampling of

$$\sqrt{B}(\hat{\delta}^*_{b_m,b_n} - \hat{\delta}),$$

where $B = b_m + b_n$ and $\hat{\delta}^*_{b_m,b_n} = \max\{\hat{p}^*_{1,b_m}, \hat{q}^*_{1,b_n}\} + \max\{\hat{p}^*_{0,b_m}, \hat{q}^*_{0,b_n}\}$ is the estimator of δ obtained from the subsamples of size b_m and b_n. To see why subsampling works, consider the same setup

[3]For the formal statement of the consistency of $\hat{\mathbb{V}}^{\max}(\eta_N)$, see Lemma A.2 and the proof of Proposition 4.1 in Appendix A.

$p_1 = q_1$, $p_0 > q_0$, and

$$\begin{aligned}\sqrt{B}(\hat{\delta}^*_{b_m,b_n} - \hat{\delta}) &= \underbrace{\max\left\{\sqrt{B}(\hat{p}^*_{1,b_m} - \hat{q}^*_{1,b_n}), 0\right\} - \max\left\{\sqrt{B}(\hat{p}^*_{1,b_m} - \hat{q}^*_{1,b_n}), 0\right\}}_{(i)'} \\ &\quad \underbrace{+\max\left\{\sqrt{B}(\hat{q}^*_{0,b_m} - \hat{p}^*_{0,b_n}), 0\right\} - \max\left\{\sqrt{B}(\hat{q}_0 - \hat{p}_0), 0\right\}}_{(ii)'} \\ &\quad \underbrace{+\sqrt{B}(\hat{q}^*_{1,b_n} - \hat{q}_1) + \sqrt{B}(\hat{p}^*_{0,b_m} - \hat{p}_0)}_{(iii)'}.\end{aligned}$$

Given the above choice of blocksizes, we can see that the asymptotic distributions of $(ii)'$ and $(iii)'$ are the same as (ii) and (iii). While, for $(i)'$, we obtain

$$\begin{aligned}(i)' &= \max\left\{\sqrt{B}(\hat{p}^*_{1,b_m} - \hat{p}_1) - \sqrt{B}(\hat{q}^*_{1,b_n} - \hat{q}_1) + \sqrt{B}(\hat{p}_1 - \hat{q}_1), 0\right\} \\ &\quad - \max\left\{\sqrt{B}(\hat{p}_1 - \hat{q}_1), 0\right\} \\ &\rightsquigarrow \max\{X_1 - W_1, 0\}\end{aligned}$$

since $\sqrt{B}(\hat{p}_1 - \hat{q}_1) = \sqrt{B/N}\sqrt{N}(\hat{p}_1 - \hat{q}_1) \to 0$ in probability (with respect to the randomness in the original sampling sequence). Thus, the resampling distribution of the statistic $\sqrt{B}(\hat{\delta}^*_{b_m,b_n} - \hat{\delta})$ correctly replicates the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$.

2.2.2 Generalization to an arbitrary Y

The framework of this section allows Y to be an arbitrary scalar random variable. We keep the instrument binary for simplicity. With additional notation, we can extend our analysis to the case with a multi-valued discrete instrument with finite points of support (see Section 2.5 and Appendix 2.A.2).

An estimator of δ

In the binary Y example, we write the true integrated envelope by

$$\begin{aligned}\delta(P,Q) &= \max_j\{\delta_j\} = \max\left\{\begin{array}{c} p_1 + p_0 \\ p_1 + q_0 \\ p_0 + q_1 \\ q_1 + q_0 \end{array}\right\} \\ &= \max\left\{\begin{array}{c} P(\{1,0\}) + Q(\emptyset) \\ P(\{1\}) + Q(\{0\}) \\ P(\{0\}) + Q(\{1\}) \\ P(\emptyset) + Q(\{1,0\}) \end{array}\right\}.\end{aligned}$$

Note that the last expression is further rewritten as

$$\delta(P,Q) = \max_{V \in \mathcal{B}(\{1,0\})} \{P(V) + Q(V^c)\}, \tag{2.2.2.6}$$

where $\mathcal{B}(\{1,0\})$ is the power set of $\{1,0\}$, $\mathcal{B}(\{1,0\}) = \{\{1,0\},\{1\},\{0\},\emptyset\}$, and $V^c = \{1,0\} \setminus V$, the complement of V. Here, $P(V)+Q(V^c)$ is seen as a function from the power set of $\mathcal{Y} = \{1,0\}$ to $\mathbb{R}_+$ and the integrated envelope is defined as its maximum over the possible subsets of $\mathcal{Y} = \{1,0\}$. A generalization to an arbitrary Y utilizes this representation of $\delta(P,Q)$.

Let $\mathcal{B}(\mathcal{Y})$ be the Borel σ-algebra on $\mathcal{Y}$. We define a set function $\delta(\cdot) : \mathcal{B}(\mathcal{Y}) \to \mathbb{R}_+$,

$$\delta(V) = P(V) + Q(V^c), \tag{2.2.2.7}$$

where V^c is the complement of V, $\mathcal{Y} \backslash V$. The function $\delta(V)$ returns the sum of the probability on V with respect to P and the probability on V^c with respect to Q. Note that the integrated envelope $\delta(P,Q)$ is given by the value of $\delta(\cdot)$ evaluated at $E = \{y \in \mathcal{Y} : p(y) \geq q(y)\}$ since

$$\begin{aligned} \delta(P,Q) &= \int_{\mathcal{Y}} \max\{p(y), q(y)\} d\mu \\ &= \int_{\{y:p(y)\geq q(y)\}} p(y) d\mu + \int_{\{y:p(y)<q(y)\}} q(y) d\mu \\ &= P(E) + Q(E^c). \end{aligned}$$

It can be shown that for an arbitrary $V \in \mathcal{B}(\mathcal{Y})$, $\delta(E) - \delta(V) \geq 0$, and therefore E is a maximizer of $\delta(\cdot)$ over $\mathcal{B}(\mathcal{Y})$.[4] Hence, an alternative expression for the integrated envelope $\delta(P,Q)$ is the supremum of $\delta(\cdot)$ over $\mathcal{B}(\mathcal{Y})$,

$$\delta(P,Q) = \sup_{V \in \mathcal{B}(\mathcal{Y})} \{\delta(V)\}. \tag{2.2.2.8}$$

We can see this expression of $\delta(P,Q)$ as a direct analogue of (2.2.2.6) for a more complex $\mathcal{Y}$, and the only complication appears in the class of subsets in $\mathcal{Y}$ on which the supremum operates.

Let P_m and Q_n be the empirical probability measures for $\{Y_{data,i} : Z_i = 1\}$ and $\{Y_{data,i} : Z_i = 0\}$, i.e., for $V \in \mathcal{B}(\mathcal{Y})$,

$$P_m(V) \equiv \frac{1}{m} \sum_{i:Z_i=1} I\{Y_{data,i} \in V\}, \quad Q_n(V) \equiv \frac{1}{n} \sum_{i:Z_i=0} I\{Y_{data,i} \in V\}.$$

We define a sample analogue of $\delta(\cdot)$ by replacing the population distribution of $P(\cdot)$ and $Q(\cdot)$ in

[4] Let $(P-Q)(B) = P(B) - Q(B)$ and $(Q-P)(B) = Q(B) - P(B)$. For an arbitray $B \in \mathcal{B}(\mathcal{Y})$, we have

$$\delta(E) - \delta(B) = (P-Q)(E \cap B^c) + (Q-P)(E^c \cap B).$$

Since $(P-Q)(\cdot)$ and $(Q-P)(\cdot)$ are nonnegative on any subsets contained in E and E^c, $\delta(E) - \delta(B) \geq 0$ holds.

(2.2.2.7) with the empirical distributions $P_m(\cdot)$ and $Q_n(\cdot)$,

$$\hat{\delta}(V) = P_m(V) + Q_n(V^c). \tag{2.2.2.9}$$

Analogous to the construction of the integrated envelope in (2.2.2.8), we propose an estimator of $\delta(P,Q)$ by maximizing $\hat{\delta}(\cdot)$ over a class of subsets $\mathbb{V} \subset \mathcal{B}(\mathcal{Y})$,[5]

$$\hat{\delta} \equiv \sup_{V \in \mathbb{V}} \{\hat{\delta}(V)\}. \tag{2.2.2.10}$$

This estimator for $\delta(P,Q)$ has the class of subsets $\mathbb{V}$ in its construction and the estimation procedure requires specifying $\mathbb{V}$ beforehand. In the next subsection, we discuss how to specify $\mathbb{V}$ in order to guarantee the asymptotic validity of the estimator.

VC-class

When Y is discrete, $\mathbb{V}$ is specified as the power set of $\mathcal{Y}$ as in the binary Y case (2.2.2.6). On the other hand, when Y is continuous, we cannot take $\mathbb{V}$ as large as $\mathcal{B}(\mathcal{Y})$. The reason is that if we specify $\mathbb{V} = \mathcal{B}(\mathcal{Y})$, $\mathbb{V}$ can contain the subset, $V^{\max} = \left\{ \bigcup_{i:Z_i=1, Y_{data,i} \neq \{mis\}} \{Y_{data,i}\} \right\}$ for any sampling sequence of $\{(Y_{data,i}, Z_i)\}_{i=1}^N$, $N = 1, 2, \ldots$. This subset almost surely gives the trivial maximum of $\hat{\delta}(\cdot)$,

$$\hat{\delta}(V^{\max}) = m^{-1} \sum_{i:Z_i=1} D_i + n^{-1} \sum_{i:Z_i=0} D_i,$$

and therefore provides little information on the integrated envelope no matter how large the sample size is because it converges to $\Pr(D=1|Z=1) + \Pr(D=1|Z=0)$. This forces us to restrict the size of $\mathbb{V}$ smaller than $\mathcal{B}(\mathcal{Y})$ in order to guarantee the consistency of $\hat{\delta}$.

An appropriate restriction for this purpose is that $\mathbb{V}$ is the *Vapnik-Červonenkis class (VC-class)* (see, e.g., Dudley (1999) for the definition of VC-class). The class of the right unbounded intervals $\mathbb{V} = \{[y, \infty) : y \in \mathbb{R}\}$ is an example of the VC-class. In Figure 2.1, the function $\delta(\cdot)$ is plotted with respect to this choice of $\mathbb{V}$ and provides a visual illustration for how $\delta(\cdot)$ attains the integrated envelope at its maximum.

By specifying $\mathbb{V}$ as the collection of right and left unbounded intervals, we obtain the *half unbounded interval class* $\mathbb{V}_{half}$,

$$\mathbb{V}_{half} = \{\emptyset, \mathbb{R}\} \cup \{(-\infty, y] : y \in \mathbb{R}\} \cup \{[y, \infty) : y \in \mathbb{R}\}. \tag{2.2.2.11}$$

In order for the estimator $\hat{\delta}$ to be consistent to the true integrated envelope $\delta(P,Q)$, we need to assume that the specified $\mathbb{V}$ contains some V which attain $\delta(V) = \delta(P,Q)$. This assumption, or, for short, the choice of $\mathbb{V}$, may be interpreted as restrictions on the global properties of the densities

[5] Forming an estimator by maximizing a set function with respect to a class of subsets is found in the literature of estimation for the density contours (Hartigan (1988) and Polonik (1995)).

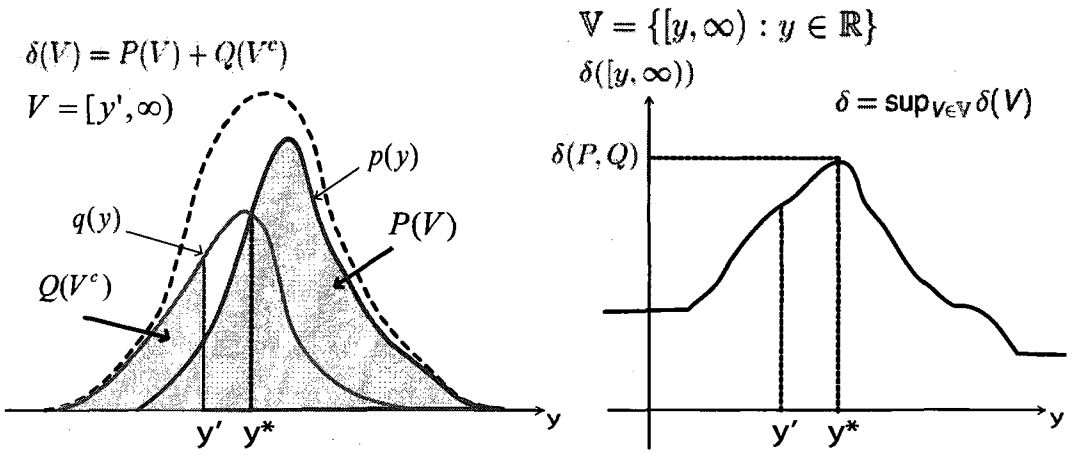

Figure 2.1: *Let Y be a continuous outcome on $\mathbb{R}$. In order to draw $\delta(\cdot)$ in two dimensions, we plot $\delta(\cdot)$ with respect to the collection of right unbounded intervals $\mathbb{V} = \{[y, \infty) : y \in \mathbb{R}\}$. As the left-hand side figure shows, $P(V)$ corresponds to the right tail area of $p(\cdot)$ while $Q(V^c)$ corresponds to the left tail area of $q(\cdot)$. $\delta(V)$ returns the sum of these areas. The right-hand side figure plots $\delta([y, \infty))$ with respect to y. When $p(y)$ and $q(y)$ cross only at y^* as in the left-hand side figure, $\delta([y, \infty))$ achieves its unique maximum at y^* and the maximum corresponds to the integrated envelope $\delta(P, Q)$. Note that the sample analogue $\hat{\delta}([y, \infty))$ is drawn as a random step function centered around the true $\delta([y, \infty))$.*

rather than the local properties such as smoothness. For example, when we specify $\mathbb{V} = \mathbb{V}_{half}$, we are imposing the restriction on the configuration of $p(y)$ and $q(y)$ such that $p(y)$ and $q(y)$ can cross at most once as in the left-hand side panel of Figure 2.1.

An alternative to $\mathbb{V}_{half}$ considered in this paper is the *histogram class* $\mathbb{V}_{hist}$, which is defined as the power set of histogram bins whose breakpoints can float over $\mathbb{R}$. For an illustration for $\mathbb{V}_{hist}$, consider fixed L histogram bins with a prespecified binwidth. Let $(\hat{p}_1, \ldots, \hat{p}_L)$ and $(\hat{q}_1, \ldots, \hat{q}_L)$ be the histogram estimators for the discretized P and Q on $\mathcal{Y}$. Then, analogously to the binary Y case, we can form the estimator of the integrated envelope in terms of the specified bins as $\sum_{l=1}^{L} \max\{\hat{p}_l, \hat{q}_l\}$. When we employ the histogram class, we maximize $\sum_{l=1}^{L} \max\{\hat{p}_l, \hat{q}_l\}$ over the possible choices of histogram bins (with a fixed binwidth).

The algebraic definition of the histogram class is given as follows. Let $h > 0$ be the bin width and L the number of bins. Pick an initial breakpoint $y_0 \in \mathbb{R}$ and consider equally distanced L points $-\infty < y_0 < y_1 < \cdots < y_{L-1} < \infty$ where $y_l = y_0 + lh$, $l = 1, \ldots, (L-1)$. Denote the $(L+1)$ disjoint intervals formed by these L points by $H_0(y_0, h) = (-\infty, y_0]$, $H_l(y_0, h) = (y_{l-1}, y_l]$, $l = 1, \ldots, (L-1)$, and $H_L(y_0, h) = (y_{L-1}, \infty)$. Let $I_j(L)$, $j = 1, \ldots, 2^{L+1}$ indicate all the possible subsets of the indices $\{0, 1, \ldots, L\}$. Given $\mathcal{Y}_0$ a set of the smallest breakpoint y_0, the histogram

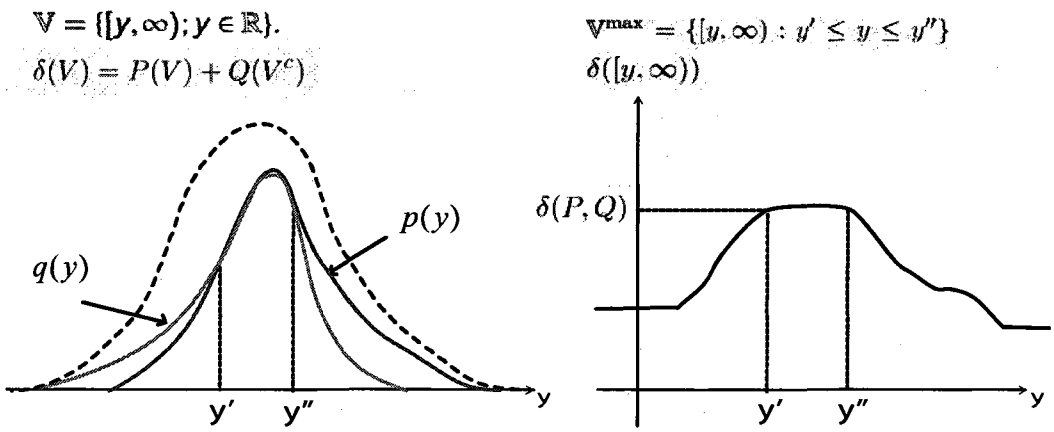

Figure 2.2: *$p(y)$ and $q(y)$ are tied over $[y', y'']$. Given $\mathbb{V}$ as the collection of right unbounded intervals, $\delta([y, \infty))$ is constant over $[y', y'']$ and there is a continuum of maximizers of $\delta(\cdot)$. Here, the maximizer subclass is given by $\mathbb{V}^{\max} = \{[y, \infty) : y \in [y', y'']\}$.*

class with bin width h and the number of bins L is expressed as

$$\mathbb{V}_{hist}(h, L, \mathcal{Y}_0) = \left\{ \bigcup_{l \in I_j(L)} H_l(y_0, h) : y_0 \in \mathcal{Y}_0,\ j = 1, \ldots, 2^{L+1} \right\}. \tag{2.2.2.12}$$

Although the binwidth is a tuning parameter, we obtain a finer VC-class than $\mathbb{V}_{half}$.

As we saw in the binary Y case, ties between P and Q cause the non-pivotal asymptotic distribution for the estimator of $\delta(P, Q)$. In order to consider how the ties between P and Q can be represented in terms of the class of subsets $\mathbb{V}$, let us specify $\mathbb{V}$ as the right unbounded interval class $\{[y, \infty) : y \in \mathbb{R}\}$. If P and Q have ties as in Figure 2.2, the maximizer of $\delta(\cdot)$ over $\mathbb{V}$ is no longer unique and any elements in $\mathbb{V}^{\max} = \{[y, \infty) : y' \leq y \leq y''\}$ can yield the integrated envelope. This example illustrates that we can identify the existence of ties between P and Q with respect to $\mathbb{V}$ by the size of the subclass

$$\mathbb{V}^{\max} = \{V \in \mathbb{V} : \delta(V) = \delta(P, Q)\}.$$

If $\mathbb{V}^{\max}$ consists of a single element $V^{\max}$, this means that $V^{\max}$ is the only subset in $\mathbb{V}$ that divides the outcome support into $\{y : p(y) \geq q(y)\}$ and $\{y : p(y) < q(y)\}$. Hence, there are no ties between P and Q (with respect to the specification of $\mathbb{V}$). On the other hand, if $\mathbb{V}^{\max}$ contains two distinct elements, $V_1^{\max}$ and $V_2^{\max}$ with $\mu(V_1^{\max} \Delta V_2^{\max}) > 0$, it can be shown that $p(y) = q(y)$ on $V_1^{\max} \Delta V_2^{\max}$, and therefore P and Q are tied on the set with positive measure $V_1^{\max} \Delta V_2^{\max}$.

Throughout our asymptotic analysis, we do not explicitly specify $\mathbb{V}$. Provided that the assumptions given below are satisfied, the main asymptotic results of the present paper are valid independent of the choice of $\mathbb{V}$. In practice, however, there is a trade-off between the flexibility of $\mathbb{V}$ (richness of $\mathbb{V}$) and the precision of the estimator $\hat{\delta}$. That is, as we choose a larger $\mathbb{V}$ for a given sample size (e.g., the histogram class with finer bins), we have more upward-biased $\hat{\delta}$ due to data

overfitting. On the other hand, as we choose a smaller $\mathbb{V}$, the assumption that $\mathbb{V}$ contains some V satisfying $\delta(V) = \delta(P, Q)$ becomes less credible. Regardless of its practical importance, we do not discuss how to choose $\mathbb{V}$ in this paper and leave it for future research.

Asymptotic distribution of $\hat{\delta}$

The main assumptions that are needed for our asymptotic results are given as follows.

Assumptions

(A1) *Nondegeneracy*: The data generating process P and Q are nondegenerate probability distributions on $\mathcal{Y} \cup \{mis\}$ and the integrated envelope is positive $\delta(P, Q) > 0$.

(A2) *VC-class*: $\mathbb{V}$ is a VC-class of measurable subsets in $\mathcal{Y}$.

(A3) *Optimal partition*: There exists a nonempty *maximizer subclass* $\mathbb{V}^{\max} \subset \mathbb{V}$ *defined by*

$$\mathbb{V}^{\max} = \{V \in \mathbb{V} : \delta(V) = \delta(P, Q)\}$$

(A4) *Existence of maximizer*: With probability one, there exists a sequence of random sets $\hat{V}_N \in \mathbb{V}$ and $\hat{V}_N^{\max} \in \mathbb{V}^{\max}$ such that for every $N \geq 1$,

$$\hat{\delta}(\hat{V}_N) = \sup_{V \in \mathbb{V}} \{\hat{\delta}(V)\}, \quad \hat{\delta}(\hat{V}_N^{\max}) = \sup_{V \in \mathbb{V}^{\max}} \{\hat{\delta}(V)\}.$$

Assumption (A3) implies that $\mathbb{V}$ contains at least one optimal subset at which the set function $\delta(\cdot)$ achieves the true integrated envelope. Since these subsets maximize $\delta(\cdot)$, we refer to the collection of these subsets as the *maximizer subclass* $\mathbb{V}^{\max}$. We allow $\mathbb{V}^{\max}$ to contain more than one element to handle the aforementioned issue of ties between P and Q. Assumption (A4) is imposed since this simplifies our proof of the asymptotic results.

The consistency of $\hat{\delta}$ follows from the uniform convergence of the empirical probability measure (Glivenko-Cantelli theorem).

For the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P, Q))$, consider

$$\sqrt{N}(\hat{\delta} - \delta(P, Q)) = \sup_{V \in \mathbb{V}} \left\{ \sqrt{N}(\hat{\delta}(V) - \delta(V)) + \sqrt{N}(\delta(V) - \delta(P, Q)) \right\}. \tag{2.2.2.13}$$

The first term in the supremum of (2.2.2.13) can be written as the sum of two independent empirical processes on $\mathbb{V}$,

$$\sqrt{N}(\hat{\delta}(V) - \delta(V)) = \left(\frac{N}{m}\right)^{1/2} \sqrt{m}(P_m(V) - P(V)) + \left(\frac{N}{n}\right)^{1/2} \sqrt{n}(Q_n(V^c) - Q(V^c)).$$

By applying the uniform central limit theorem of empirical processes (the Donsker theorem), $\sqrt{m}(P_m(V) - P(V))$ and $\sqrt{n}(Q_n(V^c) - Q(V^c))$ each converges weakly to mean zero tight Gaussian processes on $\mathbb{V}$ (see, e.g., van der Vaart and Wellner (1996)). Since the sum of independent Gaussian processes also yields Gaussian processes, $\sqrt{N}(\hat{\delta}(V) - \delta(V))$ weakly converges to mean zero tight Gaussian processes on $\mathbb{V}$. On the other hand, the second term in the supremum of (2.2.2.13) vanishes for $V \in \mathbb{V}^{\max}$ and it diverges to negative infinity for $V \notin \mathbb{V}^{\max}$. Therefore, for large N, the supremum is attained at some $V \in \mathbb{V}^{\max}$. This argument implies that the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ is given by the supremum of the set indexed Gaussian processes over the maximizer subclass $\mathbb{V}^{\max}$.

Proposition 2.2.1 (consistency and weak convergence of $\hat{\delta}$) *Assume (A1), (A2), and (A3).*
(i) $\hat{\delta} \to \delta(P,Q)$ as $N \to \infty$ with probability one.
(ii) Assume further (A4). Let $\mathbb{V}^{\max}$ be the maximizer subclass $\{V \in \mathbb{V} : \delta(V) = \delta(P,Q)\}$. Then,

$$\sqrt{N}(\hat{\delta} - \delta(P,Q)) \rightsquigarrow \sup_{V \in \mathbb{V}^{\max}} \{G(V)\}, \tag{2.2.2.14}$$

where $G(V)$ is the set indexed mean zero tight Gaussian process in $l^\infty(\mathbb{V})$ with the covariance function, for V_1, $V_2 \in \mathbb{V}$,

$$\begin{aligned} Cov(G(V_1), G(V_2)) &= \lambda^{-1}\left[P(V_1 \cap V_2) - P(V_1)P(V_2)\right] \\ &\quad + (1-\lambda)^{-1}[Q(V_1^c \cap V_2^c) - Q(V_1^c)Q(V_2^c)]. \end{aligned}$$

The asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ depends not only on the data generating process P, Q, and λ, but also on the maximizer subclass $\mathbb{V}^{\max}$ or, equivalently, on the choice of $\mathbb{V}$. If P and Q do not have ties and Assumption (A3) holds, $\mathbb{V}^{\max}$ has the unique element $V^{\max}$, then, the distribution of (2.2.2.14) is given by the projection of the Gaussian processes onto $V^{\max}$ so $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ is asymptotically normal. We present this special case in the next corollary.

Corollary 2.2.1 (asymptotic normality of $\hat{\delta}$) *Assume (A1) through (A4). If $\mathbb{V}^{\max}$ is a singleton with the unique element $V^{\max}$, then,*

$$\sqrt{N}(\hat{\delta} - \delta(P,Q)) \rightsquigarrow \mathcal{N}(0, \sigma^2(P,Q,\lambda)),$$

where

$$\sigma^2(P,Q,\lambda) = \lambda^{-1}P(V^{\max})(1 - P(V^{\max})) + (1-\lambda)^{-1}Q((V^{\max})^c)(1 - Q((V^{\max})^c)).$$

The asymptotic variance is consistently estimated by

$$\hat{\sigma}^2 = (N/m)P_m(\hat{V}_N)(1 - P_m(\hat{V}_N)) + (N/n)Q_n(\hat{V}_N^c)(1 - Q_n(\hat{V}_N^c)).$$

where $\hat{V}_N$ is a random sequence of sets that satisfy $\hat{\delta}(\hat{V}_N) = \sup_{V \in \mathbb{V}}\{\hat{\delta}(V)\}$ for $N \geq 1$.

Asymptotic normality with the consistently estimable variance makes inference straightforward. In some situations, however, the singleton assumption seems to be too restrictive. For instance, consider the case where the instrument is weak in the sense that $p(y)$ does not differ much from $q(y)$. Then, assuming $p(y) \neq q(y)$ almost everywhere is too restrictive.

2.3 Implementation of resampling methods: bootstrap and subsampling validity

Given the expression of the asymptotic distribution (Proposition 2.2.1), we want to consistently estimate the $(1-\alpha)$-th quantile of the asymptotic distribution. We propose two asymptotically valid resampling methods in this section. The resampling methods are particularly useful since the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ given in Proposition 2.2.1 has the form of a supremum functional of the Gaussian processes, and, especially when $\mathbb{V}^{\max}$ is not a singleton, it is difficult to obtain the critical values analytically (Romano (1988)).

2.3.1 Resampling method I: a modified bootstrap

The asymptotic distribution given in Proposition 2.2.1 can be replicated by the asymptotic distribution of $\sup_{V \in \mathbb{V}^{\max}} \left\{\sqrt{N}(\hat{\delta}(V) - \delta(V))\right\}$. Hence, one method to estimate it is plugging a consistent estimator for $\mathbb{V}^{\max}$ and the bootstrap analogue of $\sqrt{N}(\hat{\delta}(V) - \delta(V))$ into $\sup_{V \in \mathbb{V}^{\max}} \left\{\sqrt{N}(\hat{\delta}(V) - \delta(V))\right\}$. In this section, we validate this approach for approximating the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$.

Let $\mathbf{Y}^1_{data,m}$ represent the original sample of Y_{data} with $Z = 1$ and size m. Similarly, let $\mathbf{Y}^0_{data,n}$ be the original sample of Y_{data} with $Z = 0$ and size n. Our bootstrap algorithm is summarized as follows.

Algorithm: bootstrap for the integrated envelope

1. *Pick a slackness sequence $\{\eta_N : N \geq 1\}$ that satisfies*

$$\frac{\eta_N}{\sqrt{N}} \to 0, \quad \frac{\eta_N}{\sqrt{\log\log N}} \to \infty.$$

2. *Estimate the maximizer subclass by*

$$\hat{\mathbb{V}}^{\max}(\eta_N) = \left\{V \in \mathbb{V} : \sqrt{N}(\hat{\delta} - \hat{\delta}(V)) \leq \eta_N\right\}.$$

3. *Sample m observations from $\mathbf{Y}^1_{data,m}$ and sample n observations from $\mathbf{Y}^0_{data,n}$ randomly with replacement and construct*

$$\hat{\delta}^*(V) = P^*_m(V) + Q^*_n(V^c), \quad V \in \mathbb{V},$$

where P^*_m and Q^*_n are the empirical distributions constructed by the bootstrap sample.

4. *Compute*

$$\sup_{V \in \hat{\mathbb{V}}^{\max}(\eta_N)} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)) \right\}.$$

5. *Iterate Step 3 and 4 many times and obtain $\hat{c}^{boot}_{1-\alpha}$ as the sample $(1-\alpha)$-th quantile of the iterated statistics.*

6. *Reject the null hypothesis $\delta(P,Q) \leq 1$ if $\hat{\delta} - \frac{\hat{c}^{boot}_{1-\alpha}}{\sqrt{N}} > 1$.*

In Step 1, we specify a value of the tuning parameter η_N. Given the choice of η_N, we estimate $\mathbb{V}^{\max}$ in Step 2 and the above rate of divergence for η_N guarantees the estimator $\hat{\mathbb{V}}^{\max}(\eta_N)$ to be consistent to $\mathbb{V}^{\max}$ (see Lemma 2.A.2 in Appendices). Since the asymptotic argument only governs the speed of divergence of η_N, it provides little guidance on how to set its value in practice. We further address this issue in the Monte Carlo study of Section 2.4.

Given $\hat{\mathbb{V}}^{\max}(\eta_N)$, in Step 3 and 4, we bootstrap the function $\hat{\delta}(\cdot)$ and plug in $\sqrt{N}(\hat{\delta}^*(\cdot) - \hat{\delta}(\cdot))$, a bootstrap analogue of $\sqrt{N}(\hat{\delta}(\cdot) - \delta(\cdot))$, to the supremum operator $\sup_{V \in \hat{\mathbb{V}}^{\max}(\eta_N)} \{\cdot\}$. The bootstrap validity for empirical processes guarantees that $\sqrt{N}(\hat{\delta}^*(\cdot) - \hat{\delta}(\cdot))$ approximates the Gaussian process $G(\cdot)$ obtained in Proposition 2.2.1 (see van der Vaart and Wellner (1996) for bootstrap validity for empirical processes). By combining consistency of $\hat{\mathbb{V}}^{\max}(\eta_N)$ and bootstrap validity of $\sqrt{N}(\hat{\delta}^*(\cdot) - \hat{\delta}(\cdot))$, the statistic $\sup_{V \in \hat{\mathbb{V}}^{\max}(\eta_N)} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)) \right\}$ approximates $\sup_{V \in \mathbb{V}^{\max}} \{G(V)\}$.

The next proposition validates our specification test based on the above bootstrap algorithm.

Proposition 2.3.1 (bootstrap validity) *Assume (A1) through (A4). Then, the above bootstrap test procedure yields a pointwise asymptotically size correct test for the null $\delta(P,Q) \leq 1$, that is, for every P and Q satisfying $\delta(P,Q) \leq 1$,*

$$\lim_{N \to \infty} Prob_{P,Q,\lambda_N} \left(\hat{\delta} - \frac{\hat{c}^{boot}_{1-\alpha}}{\sqrt{N}} > 1 \right) \leq \alpha.$$

2.3.2 Resampling method II: subsampling

Subsampling is valid for any statistics that possess the asymptotic distribution (Politis and Romano (1994)). Therefore, subsampling is a valid alternative to the above bootstrap. Our subsampling proceeds in the standard manner as in Politis and Romano (1994) except for the two-sample nature

of our problem. To illustrate our subsampling algorithm, we use the following notation. We divide the full sample into $\mathbf{Y}^1_{data,m}$ and $\mathbf{Y}^0_{data,n}$ as described in Section 2.2.1. Let (b_m, b_n) be a pair of subsample sizes and $B = b_m + b_n$. There exist $N_m = \binom{m}{b_m}$ distinct subsamples from $\mathbf{Y}^1_{data,m}$, and $N_n = \binom{n}{b_n}$ distinct subsamples from $\mathbf{Y}^0_{data,n}$. The subscripts $k = 1, \ldots, N_m$ and $l = 1, \ldots, N_n$ indicate each distinct subsample. We denote the estimator $\hat{\delta}$ evaluated at the k-th subsample of $\mathbf{Y}^1_{data,m}$ and at the l-th subsample of $\mathbf{Y}^0_{data,n}$ by $\hat{\delta}^*_{k,l}$. The subsample estimator of $c_{1-\alpha}$ is defined as

$$\hat{c}^{sub}_{1-\alpha} = \inf\left\{ x : \frac{1}{N_m N_n} \sum_{k=1}^{N_m} \sum_{l=1}^{N_n} I\left\{ \sqrt{B}(\hat{\delta}^*_{k,l} - \hat{\delta}) \leq x \right\} \geq 1 - \alpha \right\}. \tag{2.3.2.15}$$

Using the obtained $\hat{c}^{sub}_{1-\alpha}$, we reject the null hypothesis if $\hat{\delta} - \frac{\hat{c}^{sub}_{1-\alpha}}{\sqrt{N}} > 1$.

The construction of $\hat{c}^{sub}_{1-\alpha}$ is similar to the one in Politis and Romano except it sums over every combination of the two subsamples. This scheme is required since we cannot define the estimator $\hat{\delta}$ if there are no observations from one of the samples.

The next proposition demonstrates the pointwise validity of subsampling.

Proposition 2.3.2 (subsampling validity) *Assume (A1) through (A4). Let* $(b_m, b_n) \to (\infty, \infty)$, $(b_m/m, b_n/n) \to (0,0)$, *and* $b_m/(b_m + b_n) \to \lambda$ *as* $N \to \infty$. *Then, the test procedure using the subsampling critical value* $\hat{c}^{sub}_{1-\alpha}$ *is pointwise asymptotically size correct, that is, for every* P *and* Q *satisfying* $\delta(P,Q) \leq 1$,

$$\lim_{N\to\infty} Prob_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}^{sub}_{1-\alpha}}{\sqrt{N}} > 1 \right) \leq \alpha.$$

When m and n are large, computing the critical values through (2.3.2.15) is difficult because of the large values of N_m and N_n. In this case, $\hat{c}^{sub}_{1-\alpha}$ can be approximated by randomly chosen subsamples (Politis et al. (1999)). Specifically, we construct the subsamples by repeatedly sampling b_m and b_n observations from $\mathbf{Y}^1_{data,m}$ and $\mathbf{Y}^0_{data,n}$ *without* replacement. Note that, analogous to the slackness sequence η_N in the modified bootstrap, subsampling also has a practical difficulty in choosing the blocksizes (b_m, b_n).

2.3.3 Power of the test against fixed alternatives

Due to the restriction of $\mathbb{V}$ to a VC-class, the test procedure is not able to screen out all the data generating processes that have $\delta(P,Q) > 1$. In order for asymptotic power of the test to be one against a fixed alternative, the alternative must meet the following condition.

Definition 2.3.1 (consistent alternatives) *The data generating process P and Q is a consistent alternative with respect to a VC-class $\mathbb{V}$ if*

$$\sup_{V \in \mathbb{V}} \{\delta(V)\} > 1.$$

In the discrete Y case, any data generating processes that have $\delta(P,Q) > 1$ are the consistent alternatives. On the other hand, for a continuous Y, $\delta(P,Q) > 1$ does not imply that the data generating process is a consistent alternative since $\mathbb{V}$ is strictly smaller than $\mathcal{B}(\mathcal{Y})$. This implies that a specification of $\mathbb{V}$ affects the refutability of the test procedure in the sense that as we specify a smaller $\mathbb{V}$, less alternatives can be screened out by the test. This can be seen as another aspect of the trade-off between precision of the estimator $\hat{\delta}$ and the fineness of $\mathbb{V}$.

The next proposition shows that the proposed test procedures are consistent in power against the consistent alternatives.

Proposition 2.3.3 (power against fixed alternatives) *The test procedures based on the proposed bootstrap and subsampling are consistent in power against the consistent alternatives, i.e., for each consistent alternative P and Q,*

$$\begin{aligned} \lim_{N \to \infty} Prob_{P,Q,\lambda_N} \left(\hat{\delta} - \frac{\hat{c}^{boot}_{1-\alpha}}{\sqrt{N}} > 1 \right) &= 1, \\ \lim_{N \to \infty} Prob_{P,Q,\lambda_N} \left(\hat{\delta} - \frac{\hat{c}^{sub}_{1-\alpha}}{\sqrt{N}} > 1 \right) &= 1. \end{aligned}$$

2.4 Monte Carlo simulations

In order to evaluate the finite sample performance of the proposed test procedures, we conduct Monte Carlo studies for various specifications of P and Q. Since the asymptotically valid test procedures attain the nominal size when $\delta(P,Q) = 1$, we set the integrated envelope equal to one for every specification. Throughout our simulation experiments, we consider two samples with equal size, $m = n$.

We specify Y to be continuous on the unit interval $\mathcal{Y} = [0,1]$. As for a specification of $\mathbb{V}$, we employ the half unbounded interval class $\mathbb{V}_{half}$ as defined in (2.2.2.11). Our Monte Carlo specifications all satisfy the optimal partition condition of Assumption (A3).

Let $\phi(\mu,\sigma)$ be the normal density with mean μ and standard deviation σ whose support is restricted on $[0,1]$ (the truncated normal). The following four specifications of P and Q are simulated (see Figure 2.3).

$$
\begin{aligned}
&\text{Design 1: } \textit{No ties,} && p(y) = 0.54 \times \phi(0.65, 0.10), \\
& && q(y) = 0.54 \times \phi(0.35, 0.10), \\
&\text{Design 2: } \textit{No ties,} && p(y) = 0.84 \times \phi(0.60, 0.20), \\
& && q(y) = 0.75 \times \phi(0.46, 0.23), \\
&\text{Design 3: } \textit{Partially tied} && p(y) = \begin{cases} 0.70 \times \phi(0.50, 0.20) & \text{for } y \leq 0.66 \\ 0.58 \times \phi(0.70, 0.25) & \text{for } y > 0.66 \end{cases}, \\
& && q(y) = \begin{cases} 0.70 \times \phi(0.50, 0.20) & \text{for } y > 0.34 \\ 0.58 \times \phi(0.30, 0.25) & \text{for } y \leq 0.34 \end{cases} \\
&\text{Design 4: } \textit{Completely tied,} && p(y) = q(y) = \phi(0.50, 0.23).
\end{aligned}
$$

In Design 1 and Design 2, there are no ties between $p(y)$ and $q(y)$, while $p(y)$ and $q(y)$ differ more significantly in Design 1 than in Design 2. Design 3 represents the case where $p(y)$ and $q(y)$ are tied on a subset of the outcome support. As an extreme case, Design 4 features a $p(y)$ that is identical to $q(y)$.

We estimate the critical values using four different methods. The first method uses the critical values implied from asymptotic normality (Corollary 2.2.1). The second method uses the naive implementation of the nonparametric bootstrap, that is, given $\hat{\delta}$, we resample $\sqrt{N}(\hat{\delta}^* - \hat{\delta})$ where $\hat{\delta}^*$ is the bootstrap analogue of $\hat{\delta}$. The third method is subsampling. We consider three different choices of the blocksizes, $(b_m, b_n) = (m/3, n/3), (m/6, n/6)$, and $(m/10, n/10)$. As the fourth method, we apply our bootstrap procedure with three choices of the slackness variable, $\eta_N = 5.0$, 2.0, and 0.5. The Monte Carlo simulations are replicated 3000 times. Subsampling and bootstrap are iterated 300 times for each Monte Carlo replication.

Table 1 shows the simulated rejection probabilities for nominal test size, $\alpha = 0.25$, 0.10, 0.05, and 0.01. The result shows that, except for Design 1, the normal approximation and the naive bootstrap over-reject the null. In particular, their test size is seriously biased when the two densities have ties, as our asymptotic analysis predicts. It is worth noting that, against the asymptotic normality in Corollary 2.2.1, the normal approximation does not perform well in Design 2. This is because the finite sample distribution of the statistic is approximated better by the distribution with ties than the normal distribution. Although the naive bootstrap is less size-distorted than the normal approximation, we can confirm that it also suffers from ties (Design 3 and 4). Thus, our simulation results indicate that, except for the case where $p(y)$ and $q(y)$ are significantly different as in Design 1, the normal approximation and the naive bootstrap are not useful for inferring δ.

Subsampling shows a good finite sample performance for Design 1 and Design 2 when the blocksizes are specified as $(m/10, n/10)$. However, if the blocksize is large such as $(m/3, n/3)$, the test performance is as bad as the normal approximation. Although Proposition 2.3.2 validates subsampling for any data generating processes, the simulation results suggest that the subsampling is contaminated by the ties.

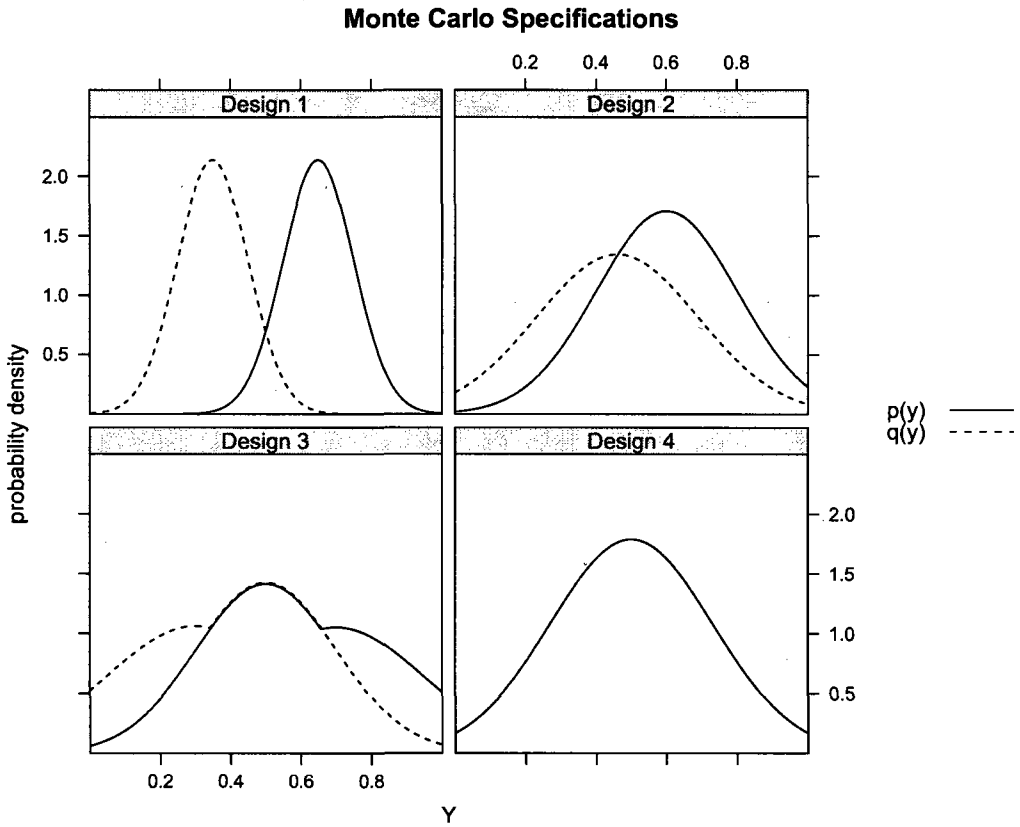

Figure 2.3: *There are no ties in Design 1 and Design 2. In Design 3, the two densities are partially tied. In Design 4, the two densities are identical.*

Among the four methods simulated, the modified bootstrap has the best size performance given an appropriate tuning of η_N, i.e., $\eta_N = 0.5$ for Design 2, $\eta_N = 2$ for Design 3, and $\eta_N = 5$ for Design 4. However, test size is rather sensitive to the choice of η_N. As we set η_N larger than optimal, we obtain a smaller rejection rate and the test becomes conservative. On the other hand, by setting η_N smaller than optimal, the rejection rate tends to be upwardly biased and approaches that of the naive bootstrap.

Table 1-I (Design 1): Simulated Rejection Rates
3000 MC replications. 300 subsampling/bootstrap replications.

Sample size		$m = n = 300$				$m = n = 1000$			
Nominal rejection prob.		25%	10%	5%	1%	25%	10%	5%	1%
Normal Approx.		28.6%	13.2%	6.5%	1.6%	26.9%	12.1%	6.9%	1.3%*
Naive bootstrap		26.0%*	10.8%*	5.8%*	1.7%	25.9%*	10.7%*	6.1%	1.6%
Subsampling	$(m/3, n/3)$	31.6%	16.1%	10.7%	4.4%	29.4%	15.4%	10.6%	4.1%
	$(m/6, n/6)$	27.5%	13.5%	7.6%	2.4%	26.6%*	12.8%	7.6%	2.4%
	$(m/10, n/10)$	25.9%*	12.2%	6.9%	1.9%	24.7%*	11.2%	6.4%	1.8%
Our bootstrap	$\eta_N = 5$	12.9%	4.6%	2.3%	0.6%*	14.7%	5.6%	2.4%	0.6%*
	$\eta_N = 2$	17.1%	6.1%	3.2%	0.9%*	18.1%	7.1%	3.3%	0.7%*
	$\eta_N = 0.5$	21.1%	8.5%	4.4%*	1.1%*	21.8%	9.3%*	4.8%*	1.0%*
Blundell et al.'s bootstrap		0%	0%	0%	0%	0%	0%	0%	0%
s.e.		0.8%	0.5%	0.4%	0.2%	0.8%	0.5%	0.4%	0.2%

*: the estimated rejection rate is not significantly different from the nominal size at the 1% level.

Table 1-II (Design 2)

3000 MC replications. 300 subsampling/bootstrap replications.

Sample size		$m=n=300$				$m=n=1000$			
Nominal rejection prob.		25%	10%	5%	1%	25%	10%	5%	1%
Normal Approx.		41.8%	20.1%	10.4%	2.7%	37.2%	16.9%	9.3%	2.0%
Naive bootstrap		32.4%	14.1%	8.2%	2.4%	29.4%	13.3%	7.0%	1.8%
Subsampling	$(m/3, n/3)$	38.8%	20.0%	13.6%	5.7%	33.9%	18.5%	12.5%	4.9%
	$(m/6, n/6)$	30.3%	14.8%	9.0%	3.1%	28.2%	13.4%	7.6%	2.4%
	$(m/10, n/10)$	26.3%*	12.1%	7.3%	2.4%	24.6%*	11.3%	6.1%	2.0%
Our bootstrap	$\eta_N = 5$	11.8%	5.1%	2.5%	0.5%	12.3%	4.6%	2.3%	0.6%*
	$\eta_N = 2$	15.8%	6.2%	3.3%	0.8%*	15.6%	6.0%	3.0%	0.8%*
	$\eta_N = 0.5$	25.6%*	10.7%*	6.0%*	1.5%	23.6%*	9.9%*	5.1%*	1.3%*
Blundell et al.'s bootstrap		2.7%	0.3%	0.1%	0%	2.0%	0.1%	0%	0%
s.e.		0.8%	0.5%	0.4%	0.2%	0.8%	0.5%	0.4%	0.2%

*: the estimated rejection rate is not significantly different from the nominal size at the 1% level.

Table 1-III (Design 3)

3000 MC replications. 300 subsampling/bootstrap replications.

Sample size		$m=n=300$				$m=n=1000$			
Nominal rejection prob.		25%	10%	5%	1%	25%	10%	5%	1%
Normal Approx.		61.5%	35.0%	21.5%	5.9%	62.2%	35.9%	23.0%	5.9%
Naive bootstrap		45.5%	24.2%	14.1%	4.6%	46.2%	25.8%	15.4%	4.6%
Subsampling	$(m/3, n/3)$	53.0%	32.6%	23.6%	10.5%	52.0%	33.7%	24.5%	10.8%
	$(m/6, n/6)$	42.7%	23.7%	15.2%	5.7%	43.3%	24.8%	15.5%	5.9%
	$(m/10, n/10)$	37.3%	20.3%	11.6%	4.3%	38.5%	20.3%	12.2%	4.0%
Our bootstrap	$\eta_N = 5$	21.5%	8.9%	4.5%*	0.8%*	23.2%*	9.0%*	4.9%*	1.1%*
	$\eta_N = 2$	23.6%*	9.8%*	5.2%*	1.1%*	25.8%*	10.3%*	5.3%*	1.5%
	$\eta_N = 0.5$	37.3%	17.9%	10.2%	3.0%	39.5%	20.2%	10.7%	3.1%
Blundell et al.'s bootstrap		10.5%	2.7%	0.9%	0.1%	10.9%	1.9%	0.7%	0%
s.e.		0.8%	0.5%	0.4%	0.2%	0.8%	0.5%	0.4%	0.2%

*: the estimated rejection rate is not significantly different from the nominal size at the 1% level.

Table 1-IV (Design 4)

3000 MC replications. 300 subsampling/bootstrap replications.

Sample size		$m=n=300$				$m=n=1000$			
Nominal rejection prob.		25%	10%	5%	1%	25%	10%	5%	1%
Normal Approx.		99.8%	82.8%	56.8%	18.8%	99.9%	82.5%	55.8%	17.9%
Naive bootstrap		77.9%	50.7%	32.2%	10.9%	77.9%	48.9%	31.6%	10.4%
Subsampling	$(m/3,n/3)$	82.7%	63.6%	49.3%	23.4%	83.4%	63.6%	45.8%	22.9%
	$(m/6,n/6)$	69.6%	43.3%	31.4%	13.2%	67.7%	41.5%	27.4%	10.9%
	$(m/10,n/10)$	63.7%	36.4%	23.0%	9.3%	56.8%	32.2%	20.3%	7.4%
Our bootstrap	$\eta_N=5$	24.6%*	10.0%*	5.3%*	1.3%*	23.3%*	9.4%*	5.2%*	1.4%*
	$\eta_N=2$	34.7%	19.1%	10.8%	2.5%	33.2%	16.6%	9.9%	2.7%
	$\eta_N=0.5$	68.3%	39.8%	24.7%	7.3%	69.2%	40.0%	23.9%	7.2%
Blundell et al.'s bootstrap		49.6%	22.2%	11.5%	2.9%	50.4%	23.2%	12.1%	2.8%
s.e.		0.8%	0.5%	0.4%	0.2%	0.8%	0.5%	0.4%	0.2%

*: the estimated rejection rate is not significantly different from the nominal size at the 1% level.

A practical difficulty in implementing our bootstrap is that the optimal value of η_N depends on the underlying data generating process. The simulation results indicate that the optimal η_N tends to be larger as the two densities are more similar. To explain this finding, recall the criterion function $\sqrt{N}(\hat{\delta}-\hat{\delta}(V))$, which is used to construct the estimator $\hat{\mathbb{V}}^{\max}(\eta_N)$. For a fixed η_N and $\tilde{V}\in\mathbb{V}^{\max}$, as the distribution of $\sqrt{N}(\hat{\delta}-\hat{\delta}(\tilde{V}))$ shifts toward the positive direction, $\hat{\mathbb{V}}^{\max}(\eta_N)$ becomes less precise in the sense that we are more likely to exclude such $\tilde{V}\in\mathbb{V}^{\max}$ from $\hat{\mathbb{V}}^{\max}(\eta_N)$. In fact, the distribution of $\sqrt{N}(\hat{\delta}-\hat{\delta}(\tilde{V}))$ depends on the underlying $\mathbb{V}^{\max}$. This can be seen from

$$\begin{aligned} E(\sqrt{N}(\hat{\delta}-\hat{\delta}(\tilde{V}))) &= E(\sqrt{N}(\hat{\delta}-\delta(P,Q))) - E(\sqrt{N}(\hat{\delta}(\tilde{V})-\delta(\tilde{V}))) \\ &\approx E\left(\sup_{V\in\mathbb{V}^{\max}}\{G(V)\}\right). \end{aligned}$$

Since the supremum of the Gaussian process tends to be higher as the index set $\mathbb{V}^{\max}$ becomes larger, this approximation implies that the mean of $\sqrt{N}(\hat{\delta}-\hat{\delta}(\tilde{V}))$ at $\tilde{V}\in\mathbb{V}^{\max}$ tends to be higher as the index set $\mathbb{V}^{\max}$ expands. Hence, when the data generating process has more ties, we need to choose a larger value of η_N in order to make the estimator for $\mathbb{V}^{\max}$ more accurate.

The tables also provide simulation results for the bootstrap procedure used in Blundell et al. (2007).[6] Note that the bounds for the cdf of Y constructed in Blundell et al. is not always tight depending on the data generating process. But, for our specifications of the data generating process, the width of their cdf bounds achieves the value of integrated envelope at least one point in the outcome support (see Proposition 1.A.1 in Appendix 1.A.2). Hence, the refuting rule of Blundell et al. such that the upper and lower cdf bounds cross at some y in the outcome support yields an identical conclusion to the one based on the integrated envelope. Nevertheless, our simulation

[6] Blundell et al. (2007) do not provide asymptotic validity of their bootstrap procedure.

results exhibit unstable performance of their bootstrap. For instance, it is very conservative for Design 1 and Design 2, while it over-rejects the null for Design 4.

2.5 Extension to a multi-valued discrete instrument

In this section, we show how the framework of the binary Z can be extended to the case with a multi-valued discrete Z. The analytical framework presented in this section is used in the empirical application of the next section.

Suppose that Z has the support with $K < \infty$ discrete points, $Z \in \{z_1, \ldots, z_K\}$. Denote the probability distribution of Y_{data} conditional on $Z = z_k$ by $P_k = (P_k(\cdot), P_{k,mis})$,

$$\begin{aligned} P_k(A) &\equiv \Pr(Y \in A | D = 1, Z = z_k) \Pr(D = 1 | Z = z_k), \\ P_{k,mis} &\equiv \Pr(D = 0 | Z = z_k). \end{aligned}$$

We use the lowercase letter p_k to denote the density of $P_k(\cdot)$ on $\mathcal{Y}$. The envelope density is defined as

$$\underline{f}(y) = \max_k \{p_k(y)\},$$

and the integrated envelope δ is the integral of $\underline{f}(y)$ over $\mathcal{Y}$.

Now, consider the function $\delta(\cdot)$ as a map from a K-partition of $\mathcal{Y}$ to $\mathbb{R}_+$. That is, given a K-partition of $\mathcal{Y}$, $\mathbf{V} = (V_1, \ldots, V_K)$ such that $\bigcup_{k=1}^{K} V_k = \mathcal{Y}$ and $\mu(V_k \cap V_l) = 0$ for $k \neq l$, we define $\delta(\cdot)$ as

$$\delta(\mathbf{V}) = \sum_{k=1}^{K} P_k(V_k). \tag{2.5.0.16}$$

This can be seen as a generalization of (2.2.2.7) to the case with a multi-valued instrument. Similarly to the binary Z case, $\delta(\cdot)$ is maximized when each subset V_k is given by $\{y : p_k(y) \geq p_l(y) \ \forall l \neq k\}$, $k = 1, \ldots, K$, and the maximum is equal to the integrated envelope. Here, the class of K-partitions as the domain of $\delta(\cdot)$ is written as

$$\mathbb{V} = \left\{ \mathbf{V} = (V_1, \ldots, V_K) : V_1 \in \mathbb{V}_1, \ \ldots, \ V_K \in \mathbb{V}_K, \ \bigcup_{k=1}^{K} V_k = \mathcal{Y}, \ \mu(V_k \cap V_{k'}) = 0 \ \forall \ k \neq k' \right\}, \tag{2.5.0.17}$$

where each $\mathbb{V}_k$, $k = 1, \ldots, K$, is a class of subsets in $\mathcal{Y}$. Then, the integrated envelope has an expression similar to (2.2.2.8),

$$\delta = \sup_{\mathbf{V} \in \mathbb{V}} \{\delta(\mathbf{V})\}, \ \mathbb{V}_1 = \cdots = \mathbb{V}_K = \mathcal{B}(\mathcal{Y}).$$

Let $n_k = \sum_{i=1}^{N} I\{Z_i = z_k\}$ and P_{n_k} the empirical probability distribution of P_k. The estimator $\hat{\delta}$ is obtained by replacing each P_k in (2.5.0.16) with the empirical distribution P_{n_k} and restrict each

$\mathbb{V}_k$ in (2.5.0.17) to a VC-class,

$$\hat{\delta} = \sup_{\mathbf{V}\in\mathbb{V}} \left\{ \hat{\delta}(\mathbf{V}) \right\}, \quad \text{where } \hat{\delta}(\mathbf{V}) = \sum_{k=1}^{K} P_{n_k}(V_k). \tag{2.5.0.18}$$

Under the assumptions analogous to (A1) through (A4) of Section 2.2, $\hat{\delta}$ has the asymptotic distribution given by

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \sup_{\mathbf{V}\in\mathbb{V}^{\max}} \{G(\mathbf{V})\}$$

where $\mathbb{V}^{\max} = \{\mathbf{V} \in \mathbb{V} : \delta(\mathbf{V}) = \delta\}$ and $G(\mathbf{V})$ are tight mean zero Gaussian processes on $\mathbb{V}$.

It is straightforward to accommodate the multi-valued discrete instrument to the bootstrap algorithm given in Section 2.3. The modifications are that the notation for a subset V is replaced with a K-partition $\mathbf{V}$, the class of subsets $\mathbb{V}$ is replaced with the class of partitions (2.5.0.17), and (2.5.0.16) is used for the function $\hat{\delta}(\cdot)$. Note that the rate of divergence of the slackness sequence η_N remained the same. The bootstrap sample is formed by resampling n_k observations with replacement from the subsample $\{Y_{data,i} : Z_i = z_k\}$ for each $k = 1, \ldots, K$.

2.6 An empirical application

We apply our bootstrap procedure to test the exogeneity of an instrument used in the classical problem of self-selection into the labor market. The data set that we use is a subset of the one used in Blundell et al. (2007). The original data source is the U.K. Family Expenditure Survey and our sample consists of the pooled repeated cross sections of individuals of age 23 to 54 for the periods from 1995 to the first quarter of 2000. The main concern of our empirical analysis is whether the out-of-work welfare income is statistically independent of the potential wage or not.

We introduce the conditioning covariates X which include gender, education, and age. As in Blundell et al. (2007), three education groups are defined, "statutory schooling", those who left school by age 16, "high-school graduates", those who left school at age 17 or 18, and "at least some college", those who completed schooling after 18. We form four age groups, 23 -30, 31 - 38, 39 - 46, and 47 - 54. As an instrument, we use the out-of-work income constructed in Blundell et al. (2003), which measures the welfare benefit for which the worker would be eligible when he is out of work (see Blundell et al. (2003) for details). The participation indicator D is one if the worker reported himself being employed or self-employed and earning positive labor income. Wage is measured as the logarithm of the usual weekly earnings divided by the usual weekly working hours and deflated by the quarterly U.K. retail price index.

For each covariate group $X = x$, we discretize the instrument by clustering the percentile ranks of the out-of-work income with every ten percentiles. We denote the instrument category within the group $X = x$ by $z_{k,x}$, $k = 1, \ldots, 10$. The envelope density and the integrated envelope of the

group $X = x$ are written as,

$$\underline{f}(y|x) = \max_{k=1,\ldots,10} \{p_{k,x}(y|x)\}, \qquad \delta_x = \int_{\mathbb{R}} \underline{f}(y|x)dy$$

where $p_{k,x}(y) = f(y|D = 1, Z = z_{k,x}, X = x) \Pr(D = 1|Z = z_{k,x}, X = x)$.

Our specification of the partition class (2.5.0.17) is the histogram class, $\mathbb{V}_1 = \cdots = \mathbb{V}_{10} = \mathbb{V}_{hist}(h, L, \mathcal{Y}_0)$, with binwidth $h = 0.4$, the number of bins $L = 10$, and the possible initial breakpoints $\mathcal{Y}_0$ as the grid points within $[1, 1.4]$ with grid size 0.02. For the multi-valued instrument, the partition class is so large that it is computationally burdensome to construct the estimator of the maximizer subclass $\hat{\mathbb{V}}^{\max}(\eta_N)$ since we need to evaluate $\hat{\delta} - \hat{\delta}(V)$ for all the possible partitions. In order to reduce the computational burden, we develop an algorithm to construct $\hat{\mathbb{V}}^{\max}(\eta_N)$ in Appendix 2.A.3 and use it to obtain the empirical result.

We choose an optimal value of η_N in the following manner. First, we run a Monte Carlo simulation in which the simulated sample size is set to the actual size and the data generating process is specified as the parametric estimate of the observed wage distributions. Specifically, for each x and $k = 1, \ldots, 10$, we specify $p_{k,x}(y)$ as the normal density (multiplied by the sample selection rate) with the mean and variance equal to the sample mean and variance of the observed wage. Accordingly, the population integrated envelope δ_x is obtained by numerically integrating the envelope over the parametric estimates. Second, for each candidate of η_N, we simulate the one-sided confidence intervals $C_{1-\alpha}(\eta_N) = \left[\hat{\delta}_x - \frac{\hat{c}^{boot}_{1-\alpha}(\eta_N)}{\sqrt{N}}, \infty\right]$ 1500 times with the nominal coverage $(1 - \alpha) = 0.75$, 0.90, 0.95, and 0.99 with 300 bootstrap iterations. As for possible values of η_N, we consider the grid points between 0.5 and 12 with grid size 0.5. After simulating the empirical coverage for each η_N, we search the value of η_N that yields the best empirical coverage in terms of minimizing the squared discrepancy from the nominal coverage,

$$\eta_N^* = \arg \min_{\eta_N = 0.5, 1.0, \ldots, 12.0} \left\{ \sum_{\alpha = 0.01,\ 0.05,\ 0.1,\ 0.25} \frac{\left[(1-\alpha) - \hat{Pr}(\delta_x \in C_{1-\alpha}(\eta_N))\right]^2}{\alpha(1-\alpha)} \right\},$$

where $\hat{Pr}(\delta_x \in C_{1-\alpha}(\eta_N))$ is the simulated coverage of the one-sided confidence intervals. As implied by the Monte Carlo study in the previous section, this manner of choosing the slackness variable is reasonable if the estimated normal densities well represent the similarity among the underlying densities $p_{k,x}(y)$. As an illustration for this, Figure ?? draws the kernel density estimates and the estimated normal densities for the group of female workers ages 23 - 31 with some college education. Although some of the kernel density estimates seem multimodal, we can observe that the normal estimates well capture the configuration of the observed wage densities.

Figure ?? shows that the observed wage tends to be higher for the worker with the higher out-of-work income. This is commonly observed in other groups. Two contrasting hypotheses are possible to explain this observation. The first hypothesis is from the perspective of the violation of the exclusion restriction. If the out-of-work income is associated with one's potential wage positively

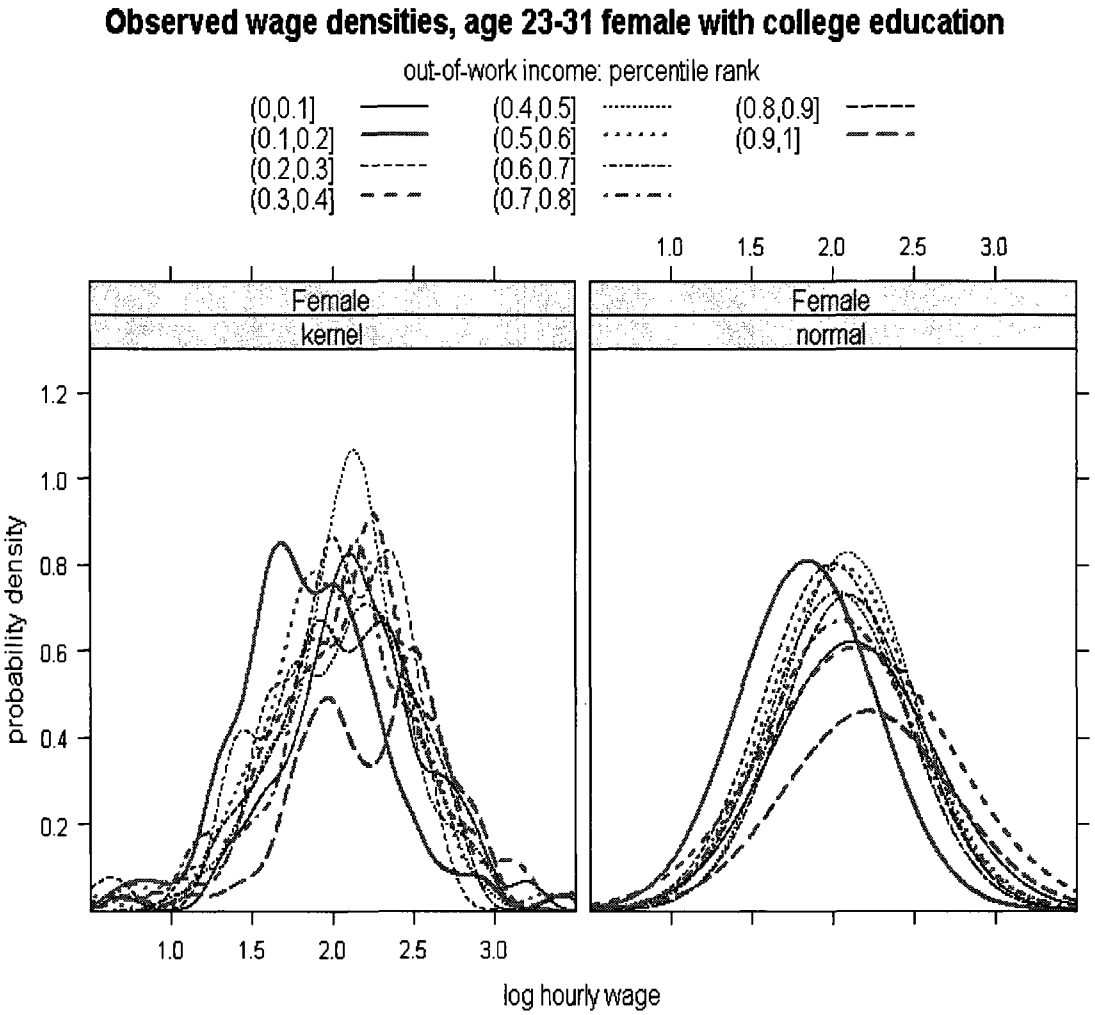
Observed wage densities, age 23-31 female with college education
out-of-work income: percentile rank
(0,0.1]
(0.1,0.2]
(0.2,0.3]
(0.3,0.4]
(0.4,0.5]
(0.5,0.6]
(0.6,0.7]
(0.7,0.8]
(0.8,0.9]
(0.9,1]
Female
kernel
Female
normal
probability density
log hourly wage
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
1.0
1.2

Figure 2.4:

Table 2: The bootstrap specification test of the exogeneity of the out-of-work income
400 Bootstrap iterations

Some college education								
	Male				Female			
	N	$\Pr(D=1\|x)$	p-value	η_N^*	N	$\Pr(D=1\|x)$	p-value	η_N^*
age 23-30	1047	0.84	0.000***	4.0	1196	0.80	0.014**	2.0
31-38	1158	0.81	0.184	7.5	1131	0.69	0.998	6.0
39-46	900	0.77	0.196	7.5	840	0.74	1.000	9.0
47-54	675	0.70	0.886	10.5	594	0.75	0.886	8.0

High-school graduates								
	Male				Female			
	N	$\Pr(D=1\|x)$	p-value	η_N^*	N	$\Pr(D=1\|x)$	p-value	η_N^*
age 23-30	799	0.81	0.016**	5.0	1354	0.72	0.946	3.0
31-38	1014	0.80	0.008***	6.5	1592	0.68	0.998	5.0
39-46	804	0.78	0.968	7.0	990	0.75	0.680	3.5
47-54	561	0.69	0.050**	4.0	698	0.70	0.966	6.5

Note ***: rejection at 1% significance, **: rejection at 5% significance.

and the selection process is nearly random, we can observe that the actual wage is higher as the out-of-work income is higher. Another hypothesis is that a very heterogenous selection process can generate the configuration of the observed densities. That is, the instrument satisfies the exclusion restriction, but the less productive workers tend to exit the labor market as their out-of-work income gets higher. Rejecting the null by our specification test can empirically refute the latter hypothesis.

Table 2 shows the result of the bootstrap specification test.[7] η_N^* indicates the value of the slackness variable obtained from the Monte Carlo procedure described above. We reject the null at a 5% significance level for 5 covariate groups, especially for the workers of younger age. Thus, our test results provide evidence of misspecification of the exclusion restriction for the out-of-work income conditional on the categorized covariates. By the virtue of partial identification analysis, this conclusion is based on the empirical evidence alone and free from any assumptions about the potential wage distribution and the selection mechanism.

2.7 Concluding remarks

This paper develops the specification test for instrument independence in the sample selection model. Our specification test operates by inferring the scalar parameter, integrated envelope, which governs the emptiness of the identification region for the outcome distribution under the instrument exclusion restriction. We propose the estimator for the integrated envelope and derive its asymptotic

[7] For the groups with statutory schooling, the integrated envelope estimates $\hat{\delta}$ do not exceed one due to the low participation rate. Accordingly, we do not reject the null for these groups and the test results for these groups are not presented in Table 2.

distribution. Based on this asymptotic distribution, we develop the asymptotically valid inference for the integrated envelope by inverting the asymptotically valid confidence intervals. Due to ties among the underlying probability densities, the estimator has a non-pivotal asymptotic distribution and therefore, the standard nonparametric bootstrap cannot be used to obtain asymptotically valid critical values. To overcome this, we develop the asymptotically valid bootstrap algorithm for the integrated envelope estimator. Our procedure first selects the target distribution for the bootstrap approximation by estimating whether or not the observable outcome densities have ties. The estimation of the ties uses the slackness variable η_N.

The Monte-Carlo simulations show that given the appropriate choice of η_N, the proposed bootstrap approximates the finite sample distribution of the statistic accurately. Although the optimal η_N depends on the true data generating process and the test performance is rather sensitive to a choice of η_N, our simulation results indicate that the bootstrap outperforms subsampling over a reasonable range of values of η_N. This paper does not provide a formal analysis on how to choose η_N. In the empirical application, we search the optimal value of η_N through the Monte Carlo simulations where the population data generating process is substituted by its parametric estimate. This way of tuning η_N can be seen as a practical solution for finding its reasonable value.

We apply the proposed test procedure to test whether the measure of out-of-work income constructed in Blundell et al. (2003) is independent of the potential wage. Our test results provide an evidence that the exclusion restriction for the out-of-work income is misspecified. Since our procedure tests the emptiness of the identification region, this conclusion is based on the empirical evidence alone and free from any assumptions on the potential wage distribution and the selection mechanism.

2.A Appendices

2.A.1 Proofs and Lemma

Proof of Proposition 2.2.1 (i).
Since $\delta(P,Q) = \sup_{V\in\mathbb{V}}\{\delta(V)\}$ and $\hat{\delta} = \sup_{V\in\mathbb{V}}\left\{\hat{\delta}(V)\right\}$, $\hat{\delta} - \delta(P,Q)$ is written as

$$\hat{\delta} - \delta(P,Q) = \sup_{V\in\mathbb{V}}\{P_m(V) + Q_n(V^c)\} - \sup_{V\in\mathbb{V}}\{P(V) + Q(V^c)\}.$$

Note that $\hat{\delta} - \delta(P,Q)$ is bounded above by $\sup_{V\in\mathbb{V}}\{(P_m - P)(V) + (Q_n - Q)(V^c)\}$ and bounded below by $\inf_{V\in\mathbb{V}}\{(P_m - P)(V) + (Q_n - Q)(V^c)\}$. Therefore,

$$\begin{aligned}\left|\hat{\delta} - \delta(P,Q)\right| &\leq \sup_{V\in\mathbb{V}}|(P_m - P)(V) + (Q_n - Q)(V^c)| \\ &\leq \sup_{V\in\mathbb{V}}|(P_m - P)(V)| + \sup_{V\in\mathbb{V}}|(Q_n - Q)(V^c)|.\end{aligned}$$

Since $\mathbb{V}$ is the VC-class by Assumption (A2), the Glivenko-Cantelli theorem implies $\sup_{V\in\mathbb{V}}|(P_m - P)(V)| \to$

0 a.s. The class of subsets $\{V^c : V \in \mathbb{V}\}$ is also a VC-class and, therefore, $\sup_{V\in\mathbb{V}} |(Q_n - Q)(V^c)| \to 0$ a.s. as well. Thus, $\hat{\delta}$ is consistent in the strong sense. ■

We use the next lemma in the proof of Proposition 2.2.1 (ii) below.

Lemma 2.A.1 *Assume (A1) through (A4). Let $\hat{V}$ be a maximizer of $\hat{\delta}(\cdot)$ over $\mathbb{V}$ and $\hat{V}^{\max}$ be a maximizer of $\hat{\delta}(\cdot)$ over the maximizer subclass $\mathbb{V}^{\max} = \{V \in \mathbb{V} : \delta(V) = \delta(P,Q)\}$. Then, $d_{P+Q}(\hat{V}, \hat{V}^{\max}) \to 0$ as $N \to \infty$ a.s.*

Proof of Lemma 2.A.1. We first show $\left|\delta(\hat{V}) - \delta(P,Q)\right| \to 0$ a.s. By Assumption (A3), $\mathbb{V}^{\max}$ is nonempty and let us pick an arbitrary element $V^{\max} \in \mathbb{V}^{\max}$. By noting $\delta(V) = \hat{\delta}(V) - (P_m - P)(V) - (Q_n - Q)(V^c)$, we have

$$\begin{aligned}
0 &\leq \delta(P,Q) - \delta(\hat{V}) = \delta(V^{\max}) - \delta(\hat{V}) \\
&= \hat{\delta}(V^{\max}) - \hat{\delta}(\hat{V}) \\
&\quad +(P_m - P)(\hat{V}) + (Q_n - Q)(\hat{V}^c) - (P_m - P)(V^{\max}) - (Q_n - Q)((V^{\max})^c) \\
&\leq (P_m - P)(\hat{V}) + (Q_n - Q)(\hat{V}^c) - (P_m - P)(V^{\max}) - (Q_n - Q)((V^{\max})^c) \\
&\to 0 \text{ a.s.}
\end{aligned}$$

by the Glivenko-Cantelli theorem. Thus, $\delta(\hat{V})$ converges to $\delta(P,Q)$ a.s.
Note that the function $\delta(\cdot)$ is continuous on $\mathbb{V}$ with respect to the semimetric d_{P+Q} since, for V_1, $V_2 \in \mathbb{V}$,

$$\begin{aligned}
|\delta(V_1) - \delta(V_2)| &\leq |P(V_1) - P(V_2)| + |Q(V_1^c) - Q(V_2^c)| \\
&= |P(V_1) - P(V_2)| + |Q(V_1) - Q(V_2)| \\
&\leq P(V_1 \triangle V_2) + Q(V_1 \triangle V_2) \\
&= d_{P+Q}(V_1, V_2).
\end{aligned}$$

Given these results, let us suppose that the conclusion is false, that is, assume that there exist positive ϵ and ζ such that $\mathbb{P}(\{d_{P+Q}(\hat{V}, \hat{V}^{\max}) > \epsilon, \text{ i.o.}\}) > \zeta$. Since the event $\{d_{P+Q}(\hat{V}, \hat{V}^{\max}) > \epsilon\}$ implies $\{\hat{V} \notin \mathbb{V}^{\max}\}$, the continuity of $\delta(\cdot)$ with respect to the semimetric d_{P+Q} and the definition of $\mathbb{V}^{\max}$ imply that we can find $\xi > 0$ such that $\mathbb{P}(\{\delta(P,Q) - \delta(\hat{V}) > \xi, \text{ i.o.}\}) > \zeta$ holds. This contradicts the almost sure convergence of $\delta(\hat{V})$ to $\delta(P,Q)$ shown above. Hence, $d_{P+Q}(\hat{V}, \hat{V}^{\max}) \to 0$ a.s. ■

Proof of Proposition 2.2.1 (ii). Given the VC-class $\mathbb{V}$, the Donsker theorem (theorem 2.5.2 and theorem 2.6.4 in van der Vaart and Wellner (1996)) asserts that the empirical processes $G_{P,m}(V) = \sqrt{m}(P_m - P)(V)$, and $G_{Q,n}(V) = \sqrt{n}(Q_n - Q)(V)$ weakly converge to the tight Brownian bridge processes $G_P(V)$ and $G_Q(V)$ in $l^\infty(\mathbb{V})$. These weakly converging sequences of the empirical processes

$G_{P,m}(V)$ and $G_{Q,n}(V)$ are *asymptotically stochastically equicontinuous* with respect to the seminorm d_P and d_Q respectively (theorem 1.5.7 of van der Vaart and Wellner). That is, for any $\eta > 0$,

$$\lim_{\beta \to 0} \lim_{m \to \infty} \sup \mathbb{P}^* \left(\sup_{d_P(V,V') < \beta} |G_{P,m}(V) - G_{P,m}(V')| > \eta \right) = 0.$$

$$\lim_{\beta \to 0} \lim_{n \to \infty} \sup \mathbb{P}^* \left(\sup_{d_Q(V,V') < \beta} |G_{Q,n}(V) - G_{Q,n}(V')| > \eta \right) = 0.$$

We apply these facts to show that the difference between $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ and $\sup_{V \in \mathbb{V}^{\max}} \{\sqrt{N}(\hat{\delta}(V) - \delta(V))\}$ are asymptotically negligible.
Since $\delta(V) = \delta(P,Q)$ on $\mathbb{V}^{\max} \subset \mathbb{V}$,

$$\begin{aligned} \sup_{V \in \mathbb{V}^{\max}} \{\sqrt{N}(\hat{\delta}(V) - \delta(V))\} &= \sup_{V \in \mathbb{V}^{\max}} \{\sqrt{N}(\hat{\delta}(V) - \delta(P,Q))\} \\ &\leq \sup_{V \in \mathbb{V}} \{\sqrt{N}(\hat{\delta}(V) - \delta(P,Q))\} = \sqrt{N}(\hat{\delta} - \delta(P,Q)) \end{aligned}$$

holds. Let $\hat{V}$ be and $\hat{V}^{\max}$ be the maximizer of $\hat{\delta}(\cdot)$ on $\mathbb{V}$ and $\mathbb{V}^{\max}$ respectively, which are assumed to exist by Assumption (A4). Then,

$$\begin{aligned} 0 &\leq \sqrt{N}(\hat{\delta} - \delta(P,Q)) - \sup_{V \in \mathbb{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}(V) - \delta(P,Q)) \right\} \\ &= \sqrt{N}(\hat{\delta}(\hat{V}) - \hat{\delta}(\hat{V}^{\max})) \\ &= (N/m)^{1/2} \sqrt{m}(P_m(\hat{V}) - P_m(\hat{V}^{\max})) + (N/n)^{1/2} \sqrt{n}(Q_n(\hat{V}^c) - Q_n((\hat{V}^{\max})^c)) \\ &= (N/m)^{1/2}(G_{P,m}(\hat{V}) - G_{P,m}(\hat{V}^{\max})) + (N/n)^{1/2}(G_{Q,n}(\hat{V}^c) - G_{Q,n}((\hat{V}^{\max})^c)). \end{aligned}$$

By Lemma 2.A.1, we have $d_{P+Q}(\hat{V}, \hat{V}^{\max}) \to 0$ a.s. and this implies $d_P(\hat{V}, \hat{V}^{\max}) \to 0$ and $d_Q(\hat{V}^c, (\hat{V}^{\max})^c) \to 0$ a.s. The asymptotic stochastic equicontinuity implies that $G_{P,m}(\hat{V}) - G_{P,m}(\hat{V}^{\max}) \to 0$ and $(G_{Q,n}(\hat{V}^c) - G_{Q,n}((\hat{V}^{\max})^c)) \to 0$ in outer probability. Thus, we conclude $\sqrt{N}(\hat{\delta} - \delta(P,Q)) - \sup_{V \in \mathbb{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}(V) - \delta(V)) \right\} = o_{P^*}(1)$ and the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ is identical to that of $\sup_{V \in \mathbb{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}(V) - \delta(V)) \right\}$. Hence, in the rest of the proof, we focus on deriving the asymptotic distribution of $\sup_{V \in \mathbb{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}(V) - \delta(V)) \right\}$.
The weak convergence of $\sqrt{N}(\hat{\delta}(V) - \delta(V))$ follows from the Donsker theorem,

$$\begin{aligned} \sqrt{N}(\hat{\delta}(V) - \delta(V)) &= (N/m)^{-1/2} G_{P,m}(V) + (N/n)^{-1/2} G_{Q,n}(V^c) \\ &\rightsquigarrow \lambda^{-1/2} G_P(V) + (1-\lambda)^{-1/2} G_Q(V) \equiv G(V), \end{aligned}$$

where G_P are the tight P-brownian bridge processes in $l^\infty(\mathbb{V})$ and G_Q are the tight Gaussian processes in $l^\infty(\mathbb{V})$ with the covariance kernel

$$Cov(G_Q(V_1), G_Q(V_2)) = Q(V_1^c \cap V_2^c) - Q(V_1^c)Q(V_2^c), \quad V_1, V_2 \in \mathbb{V}.$$

Since G_P and G_Q are independent Gaussian processes, the covariance kernel of $G(V) = \lambda^{-1/2}G_P(V) + (1-\lambda)^{-1/2}G_Q(V)$ is given by

$$\begin{aligned} Cov(G(V_1), G(V_2)) &= \lambda^{-1}\left[P(V_1 \cap V_2) - P(V_1)P(V_2)\right] \\ &\quad +(1-\lambda)^{-1}[Q(V_1^c \cap V_2^c) - Q(V_1^c)Q(V_2^c)]. \end{aligned}$$

Lastly, we note that the supremum functional $\sup_{V\in\mathbb{V}^{\max}}\{\cdot\}$ on $l^\infty(\mathbb{V})$ is continuous with respect to the sup metric since for $x_1,\, x_2 \in l^\infty(\mathbb{V})$,

$$\begin{aligned} \left| \sup_{V\in\mathbb{V}^{\max}}\{x_1(V)\} - \sup_{V\in\mathbb{V}^{\max}}\{x_2(V)\} \right| &\leq \sup_{V\in\mathbb{V}^{\max}}\{|x_1(V) - x_2(V)|\} \\ &\leq \sup_{V\in\mathbb{V}}\{|x_1(V) - x_2(V)|\} \\ &= \|x_1 - x_2\|_\infty . \end{aligned}$$

Thus, by applying the continuous mapping theorem of stochastic processes, we obtain the desired result,

$$\sup_{V\in\mathbb{V}^{\max}}\left\{\sqrt{N}(\hat{\delta}(V) - \delta(V))\right\} \rightsquigarrow \sup_{V\in\mathbb{V}^{\max}}\{G(V)\}.$$

■

Proof of Corollary 2.2.1. Given $\mathbb{V}^{\max} = \{V^{\max}\}$, Proposition 2.2.1 (ii) immediately yields the asymptotic normality. Consistency of the plug-in variance estimator follows since

$$\begin{aligned} |P_m(\hat{V}) - P(V^{\max})| &\leq |(P_m - P)(\hat{V})| + |P(\hat{V}) - P(V^{\max})| \\ &\leq |(P_m - P)(\hat{V})| + d_{P+Q}(\hat{V}, V^{\max}) \\ &\to 0 \text{ a.s.} \end{aligned}$$

by the Glivenko Cantelli theorem and Lemma 2.A.1. A similar result holds for $Q_m(\hat{V}^c)$. Hence, $\hat{\sigma}^2 \to \sigma^2(P, Q, \lambda)$ a.s. ■

The next lemma shows that $\hat{\mathbb{V}}^{\max}(\eta_N)$ introduced in the first step of the bootstrap algorithm is consistent to $\mathbb{V}^{\max}$. This lemma is used for the proof of Proposition 2.3.1 below.

Lemma 2.A.2 *Assume (A1) through (A4). Let $\{\eta_N : N \geq 1\}$ be a positive sequence satisfying $\frac{\eta_N}{\sqrt{N}} \to 0$ and $\frac{\eta_N}{\sqrt{\log\log N}} \to \infty$. For the semimetric $d_{P+Q}(V_1, V_2) = P(V_1 \,\triangle\, V_2) + Q(V_1 \,\triangle\, V_2)$, define ϵ-cover of the maximizer subclass $\mathbb{V}^{\max}$ by*

$$\mathbb{V}_\epsilon^{\max} = \left\{V \in \mathbb{V} : \inf_{V'\in\mathbb{V}^{\max}}\{d_{P+Q}(V, V')\} \leq \epsilon\right\}.$$

For the estimator $\hat{\mathbb{V}}^{\max}(\eta_N) = \left\{V \in \mathbb{V} : \sqrt{N}(\hat{\delta} - \hat{\delta}(V)) \leq \eta_N\right\}$ *define a sequence of events*

$$A_N^\epsilon = \left\{\mathbb{V}^{\max} \subseteq \hat{\mathbb{V}}^{\max}(\eta_N) \subseteq \mathbb{V}_\epsilon^{\max}\right\}.$$

Then, for each $\epsilon > 0$,

$$\mathbb{P}\left(\lim_{N\to\infty} \inf A_N^\epsilon\right) = 1,$$

that is, with probability one, A_N^ϵ *occurs for all* N *with the finite number of exceptions.*

Proof of Lemma 2.A.2. We first state the law of the iterated logarithm for empirical processes on VC-classes (LIL, see Alexander and Talagrand (1989)).
For a VC-class $\mathbb{V}$ and set indexed empirical processes, $G_{P,m}(V) = \sqrt{m}(P_m - P)(V)$,

$$\text{(LIL)} \qquad \lim_{m\to\infty} \sup_{V\in\mathbb{V}} \sup \left|\frac{G_{P,m}(V)}{\sqrt{\log\log m}}\right| \leq 1 \qquad \text{a.s.}$$

Let $\tau_{N,m} = \sqrt{N/m}\frac{\sqrt{\log\log m}}{\sqrt{\log\log N}}\frac{\sqrt{\log\log N}}{\eta_N}$ and $\tau_{N,n} = \sqrt{N/n}\frac{\sqrt{\log\log n}}{\sqrt{\log\log N}}\frac{\sqrt{\log\log N}}{\eta_N}$. Consider

$$\sup_{V\in\mathbb{V}}\left|\frac{\sqrt{N}}{\eta_N}(\hat{\delta}(V) - \delta(V))\right| \leq \tau_{N,m} \sup_{V\in\mathbb{V}}\left|\frac{G_{P,m}(V)}{\sqrt{\log\log m}}\right| + \tau_{N,n} \sup_{V\in\mathbb{V}}\left|\frac{G_{Q,n}(V^c)}{\sqrt{\log\log n}}\right|.$$

Since $\tau_{N,m} \to 0$ and $\tau_{N,n} \to 0$ as $N \to \infty$, the right hand side of the above inequality converges to zero a.s. by the LIL. Hence,

$$\lim_{N\to\infty} \sup_{V\in\mathbb{V}}\left|\frac{\sqrt{N}}{\eta_N}(\hat{\delta}(V) - \delta(V))\right| = 0 \qquad \text{a.s.} \tag{A.2}$$

Based on this almost sure result, we next show $\mathbb{P}\left(\liminf\left\{\mathbb{V}^{\max} \subseteq \hat{\mathbb{V}}^{\max}(\eta_N)\right\}\right) = 1$. Note that, by the construction of $\hat{\mathbb{V}}^{\max}(\eta_N)$, $\mathbb{V}^{\max} \subseteq \hat{\mathbb{V}}^{\max}(\eta_N)$ occurs if and only if $\sup_{V\in\mathbb{V}^{\max}}\left\{\frac{\sqrt{N}}{\eta_N}\left(\hat{\delta} - \hat{\delta}(V)\right)\right\} \leq 1$. Therefore, it suffices to show

$$\limsup \sup_{V\in\mathbb{V}^{\max}}\left\{\frac{\sqrt{N}}{\eta_N}\left(\hat{\delta} - \hat{\delta}(V)\right)\right\} \leq 1 \qquad \text{a.s.}$$

Consider

$$\frac{\sqrt{N}}{\eta_N}\left(\hat{\delta} - \hat{\delta}(V)\right) = \frac{\sqrt{N}}{\eta_N}(\hat{\delta} - \delta(P,Q)) - \frac{\sqrt{N}}{\eta_N}\left(\hat{\delta}(V) - \delta(V)\right) + \frac{\sqrt{N}}{\eta_N}(\delta(P,Q) - \delta(V)) \tag{A.3}$$

Since $\delta(P,Q) - \delta(V) = 0$ on $\mathbb{V}^{\max}$, we have

$$\sup_{V\in\mathbb{V}^{\max}} \frac{\sqrt{N}}{\eta_N}\left(\hat{\delta} - \hat{\delta}(V)\right) \leq \underbrace{\left|\frac{\sqrt{N}}{\eta_N}(\hat{\delta} - \delta(P,Q))\right|}_{(i)} + \underbrace{\sup_{V\in\mathbb{V}^{\max}}\left|\frac{\sqrt{N}}{\eta_N}\left(\hat{\delta}(V) - \delta(V)\right)\right|}_{(ii)}.$$

By the almost sure convergence (A.2), $(ii) \to 0$ a.s. So it suffices to show $(i) \to 0$ a.s. By noting $\hat{\delta} = \hat{\delta}(\hat{V})$, $\hat{\delta}(V) = \delta(V) + (P_m - P)(V) + (Q_n - Q)(V^c)$, and denoting an arbitrary element in $\mathbb{V}^{\max}$ by $V^{\max}$, $(i) \to 0$ a.s. is shown from

$$\begin{aligned}
(i) \quad \leq \quad & \left| \frac{\sqrt{N}}{\eta_N} (\hat{\delta}(\hat{V}) - \delta(\hat{V})) \right| + \frac{\sqrt{N}}{\eta_N} (\delta(V^{\max}) - \delta(\hat{V})) \\
\leq \quad & \left| \frac{\sqrt{N}}{\eta_N} (\hat{\delta}(\hat{V}) - \delta(\hat{V})) \right| + \frac{\sqrt{N}}{\eta_N} (\hat{\delta}(V^{\max}) - \hat{\delta}(\hat{V})) \\
& + \tau_{N,m} \left| \frac{G_{P,m}(\hat{V})}{\sqrt{\log\log m}} \right| + \tau_{N,m} \left| \frac{G_{P,m}(V^{\max})}{\sqrt{\log\log m}} \right| \\
& + \tau_{N,n} \left| \frac{G_{Q,n}(\hat{V}^c)}{\sqrt{\log\log n}} \right| + \tau_{N,n} \left| \frac{G_{Q,n}((V^{\max})^c)}{\sqrt{\log\log n}} \right| \\
\leq \quad & \left| \frac{\sqrt{N}}{\eta_N} (\hat{\delta}(\hat{V}) - \delta(\hat{V})) \right| + \tau_{N,m} \left| \frac{G_{P,m}(\hat{V})}{\sqrt{\log\log m}} \right| + \tau_{N,m} \left| \frac{G_{P,m}(V^{\max})}{\sqrt{\log\log m}} \right| \\
& + \tau_{N,n} \left| \frac{G_{Q,n}(\hat{V}^c)}{\sqrt{\log\log n}} \right| + \tau_{N,n} \left| \frac{G_{Q,n}((V^{\max})^c)}{\sqrt{\log\log n}} \right| \\
\to \quad & 0 \qquad \text{a.s. by LIL.}
\end{aligned}$$

Thus, $\mathbb{P}\left(\liminf \left\{ \mathbb{V}^{\max} \subseteq \hat{\mathbb{V}}^{\max}(\eta_N) \right\}\right) = 1$ is proved.
Next, we show $\mathbb{P}\left(\liminf \left\{ \hat{\mathbb{V}}^{\max}(\eta_N) \subseteq \mathbb{V}_\epsilon^{\max} \right\}\right) = 1$. Since the event $\left\{ \hat{\mathbb{V}}^{\max}(\eta_N) \subseteq \mathbb{V}_\epsilon^{\max} \right\}$ is equivalent to $\inf_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} (\hat{\delta} - \hat{\delta}(V)) \right\} > 1$, it suffices to show

$$\lim_{N \to \infty} \inf \inf_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} (\hat{\delta} - \hat{\delta}(V)) \right\} > 1 \qquad \text{a.s.}$$

We obtain from (A.3)

$$\begin{aligned}
\inf_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} \left(\hat{\delta} - \hat{\delta}(V) \right) \right\} \quad \geq \quad & \frac{\sqrt{N}}{\eta_N} (\hat{\delta} - \delta(P,Q)) - \sup_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} \left(\hat{\delta}(V) - \delta(V) \right) \right\} \\
& + \inf_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} (\delta(P,Q) - \delta(V)) \right\}
\end{aligned}$$

Note that the first two terms have been already proved to converge to zero a.s. For the third term, the continuity of $\delta(\cdot)$ with respect to the semimetric d_{P+Q} (see the proof of Proposition 2.2.1 (ii)) implies that there exists $\zeta(\epsilon) > 0$ such that $\delta(P,Q) - \delta(V) > \zeta(\epsilon)$ for any $V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}$. Since $\frac{\sqrt{N}}{\eta_N} \to \infty$, we obtain $\inf_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} (\delta(P,Q) - \delta(V)) \right\} \geq \frac{\sqrt{N}}{\eta_N} \zeta(\epsilon) \to \infty$. Therefore, $\lim_{N \to \infty} \inf \inf_{V \in \mathbb{V} \setminus \mathbb{V}_\epsilon^{\max}} \left\{ \frac{\sqrt{N}}{\eta_N} (\hat{\delta} - \hat{\delta}(V)) \right\} = \infty$ a.s. and this implies $\mathbb{P}\left(\liminf \left\{ \hat{\mathbb{V}}^{\max}(\eta_N) \subseteq \mathbb{V}_\epsilon^{\max} \right\}\right) = 1$.

Combining these two results completes the proof. ■

Proof of Proposition 2.3.1. We indicate an infinite sequence of $\{(Y_{data,i}, Z_i) : i = 1, 2, \dots\}$ by $\omega \in \Omega$. Denote a random sequence of the probability laws governing the randomness in the bootstrap sample by $\{\mathbb{P}_N : N \geq 1\}$. Once we fix ω, $\{\mathbb{P}_N : N \geq 1\}$ can be seen as a nonrandom sequence of the probability laws. The bootstrap is consistent if, for almost every $\omega \in \Omega$,

$$\sup_{V \in \hat{\mathbb{V}}^{\max}(\eta_N)(\omega)} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} \rightsquigarrow \sup_{V \in \mathbb{V}^{\max}} \{G(V)\}$$

where $G(V)$ is the Gaussian processes obtained in Proposition 2.2.1 (ii). Here, the random objects subject to the probability law of the original sampling sequence are indexed by ω.
By Lemma 2.A.2, for sufficiently large N,

$$\begin{aligned} \sup_{V \in \hat{\mathbb{V}}^*} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} &\leq \sup_{V \in \hat{\mathbb{V}}^{\max}(\eta_N)(\omega)} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} \\ &\leq \sup_{V \in \mathbb{V}_\epsilon^{\max}} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} \end{aligned} \tag{A.4}$$

holds for almost all $\omega \in \Omega$. Let $G^*_{P,m}(\cdot) = \sqrt{m}(P^*_m - P_m)(\cdot)$ and $G^*_{Q,n} = \sqrt{n}(Q^*_n - Q_n)(\cdot)$ be bootstrapped empirical processes where P^*_m and Q^*_n are the empirical probability measures constructed from the bootstrap sample. By the almost sure convergence of the bootstrap empirical processes (Theorem 3.6.3 in van der Vaart and Wellner (1996)),

$$\sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) = \sqrt{\frac{N}{m}} G^*_{P,m}(V) + \sqrt{\frac{N}{n}} G^*_{Q,n}(V^c) \rightsquigarrow G(V),$$

uniformly over $\mathbb{V}$ for almost all ω. Therefore, for the lower bound term and the upper bound term in (A.4), we have

$$\begin{aligned} \sup_{V \in \mathbb{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} &\rightsquigarrow \sup_{V \in \mathbb{V}^{\max}} \{G(V)\}, \\ \sup_{V \in \mathbb{V}_\epsilon^{\max}} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} &\rightsquigarrow \sup_{V \in \mathbb{V}_\epsilon^{\max}} \{G(V)\}. \end{aligned}$$

Since the tight Gaussian processes $G(V)$ are almost surely continuous with respect to d_{P+Q}, the asymptotic stochastic equicontinuity of the Gaussian processes imply

$$\sup_{V \in \mathbb{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} - \sup_{V \in \mathbb{V}_\epsilon^{\max}} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} \to 0$$

in probability with respect to $\{\mathbb{P}_N : N \geq 1\}$ as $\epsilon \to 0$. Hence, from (A.4), we conclude that

$$\sup_{V \in \hat{\mathbb{V}}^{\max}(\eta_N)(\omega)} \left\{ \sqrt{N}(\hat{\delta}^*(V) - \hat{\delta}(V)(\omega)) \right\} \rightsquigarrow \sup_{V \in \mathbb{V}^{\max}} \{G(V)\}.$$

Assumption (A1) and (A2) implies that $G(V)$ are non-degenerate Gaussian processes on $V \in \mathbb{V}^{\max}$ and, therefore, the distribution of $\sup_{V \in \mathbb{V}^{\max}} \{G(V)\}$ is absolutely continuous on $\mathbb{R}$ (see Proposition 11.4 in Davydov, Lifshits, and Smorodina (1998)). Therefore, the $\hat{c}_{1-\alpha}^{boot}$ converges to $c_{1-\alpha}$ in probability with respect to $\{\mathbb{P}_N : N \geq 1\}$ for almost every $\omega \in \Omega$. Hence, for every P and Q with $\delta(P,Q) \leq 1$,

$$\begin{aligned} \text{Prob}_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}^{boot}}{\sqrt{N}} > 1\right) &\leq \text{Prob}_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}^{boot}}{\sqrt{N}} > \delta(P,Q)\right) \\ &= \text{Prob}_{P,Q,\lambda_N}\left(\sqrt{N}(\hat{\delta} - \delta(P,Q)) > \hat{c}_{1-\alpha}^{boot}\right) \\ &\to 1 - J(c_{1-\alpha}; P, Q, \lambda) = \alpha. \end{aligned}$$

■

Proof of Proposition 2.3.2. In order to be explicit about the sample size used to construct the estimator, we notate the estimator by $\hat{\delta}_N$ when the sample with size N is used. Denote the cumulative distribution function of $\sqrt{N}(\hat{\delta}_N - \delta(P,Q))$ by

$$J_N(x, P, Q, \lambda_N) = Prob_{P,Q,\lambda_N}\left\{\sqrt{N}(\hat{\delta}_N - \delta(P,Q)) \leq x\right\}.$$

where $Prob_{P,Q,\lambda_N}(\cdot)$ represents the probability law with respect to the data generating process P and Q with $\lambda_N = m/N$.

Let us define the subsampling estimator for $J_N(x, P, Q, \lambda_N)$ by

$$L_N(x) = \frac{1}{N_m N_n} \sum_{k=1}^{N_m} \sum_{l=1}^{N_n} 1\left\{\sqrt{B}(\hat{\delta}_{k,l}^* - \hat{\delta}_N) \leq x\right\}.$$

Let

$$U_N(x) = \frac{1}{N_m N_n} \sum_{k=1}^{N_m} \sum_{l=1}^{N_n} 1\left\{\sqrt{B}(\hat{\delta}_{k,l}^* - \delta(P,Q)) \leq x\right\},$$

in which $\hat{\delta}_N$ in $L_N(x)$ is replaced with $\delta(P,Q)$. Note that $U_N(x)$ has the representation of the two-sample U-statistic with degree b_m and b_n,

$$U_N(x) = \frac{1}{N_m N_n} \sum_{k=1}^{N_m} \sum_{l=1}^{N_n} h(\mathbf{Y}_{data,b_m,k}^1, \mathbf{Y}_{data,b_n,l}^0),$$

where $\mathbf{Y}_{data,b_m,k}^1$ represents the k-th subsample drawn from $\mathbf{Y}_{data,m}^1$, $\mathbf{Y}_{data,b_n,l}^0$ the l-th subsample drawn from $\mathbf{Y}_{data,}^0$, and $h(\mathbf{Y}_{data,b_m,k}^1, \mathbf{Y}_{data,b_n,l}^0) = 1\left\{\sqrt{B}(\hat{\delta}_{k,l}^* - \delta(P,Q)) \leq x\right\}$. Since for each k and l, $\mathbf{Y}_{data,b_m,k}^1$ and $\mathbf{Y}_{data,b_n,l}^0$ are i.i.d. samples with size b_m and b_n from P and Q, the mean of the kernel of the U-statistic satisfies

$$E(h(\mathbf{Y}_{data,b_m,k}^1, \mathbf{Y}_{data,b_n,l}^0)) = J_B(x, P, Q, \lambda_B),$$

where $J_B(x, P, Q, \lambda_B)$ is the cdf of $\sqrt{B}(\hat{\delta}_B - \delta(P,Q))$ and $\lambda_B = b_m/B$. Then, by the Hoeffding inequality for the two sample U-statistic (p25-p26 of Hoeffding (1963)),

$$Prob_{P,Q,\lambda_B}(|U_N(x) - J_B(x, P, Q, \lambda_B)| \geq \epsilon) \leq 2\exp\left\{-2K\epsilon^2\right\}$$

where

$$K = \min\left\{\frac{m}{b_m}, \frac{n}{b_n}\right\}.$$

By the specification of the blocksizes, $K \to \infty$ holds, so it follows that

$$U_N(x) - J_B(x, P, Q, \lambda_B) \to 0$$

in probability. Since $J_B(\cdot, P, Q, \lambda_B)$ converges weakly to $J(\cdot; P, Q, \lambda)$ the cdf of $\sup_{V\in\mathbb{V}}\{G(V)\}$ and $J(\cdot; P, Q, \lambda)$ is continuous as we addressed in the proof of Proposition 2.3.1, $J_B(x; P, Q, \lambda_B) \to J(x; P, Q, \lambda)$ holds for every x. Therefore, $U_N(x)$ converges to $J(x; P, Q, \lambda)$ in probability. By replicating the argument in Politis and Romano (1994), it follows that $L_{N,B}(x) - U_N(x) \to 0$ in probability. Thus, $L_{N,B}(x) \to J(x; P, Q, \lambda)$ in probability.
Given this result, $\hat{c}^{sub}_{1-\alpha}$ converges to the $(1-\alpha)$-th quantile of $J(\cdot; P, Q, \lambda)$ in probability (see, e.g., lemma 11.2.1 in Lehmann and Romano (2005)). Therefore, for every P and Q with $\delta(P,Q) \leq 1$,

$$\begin{aligned} Prob_{P,Q,\lambda_N}\left(\hat{\delta}_N - \frac{\hat{c}^{sub}_{1-\alpha}}{\sqrt{N}} > 1\right) &\leq Prob_{P,Q,\lambda_N}\left(\hat{\delta}_N - \frac{\hat{c}^{sub}_{1-\alpha}}{\sqrt{N}} > \delta(P,Q)\right) \\ &= Prob_{P,Q,\lambda_N}\left(\sqrt{N}(\hat{\delta}_N - \delta(P,Q)) > \hat{c}^{sub}_{1-\alpha}\right) \\ &\to 1 - J(c_{1-\alpha}; P, Q, \lambda) = \alpha. \end{aligned}$$

■

Proof of Proposition 2.3.3. Fix a consistent alternative P and Q. Let $\tilde{\delta}(P,Q) = \sup_{V\in\mathbb{V}}\{\delta(V)\}$. With a slight abuse of notation, denote by $\mathbb{V}^{\max}$ the class of subsets that attain the supremum of $\delta(V)$. By repeating the same argument as in the proof of Proposition 2.2.1, it is shown that $\sqrt{N}(\hat{\delta} - \tilde{\delta}(P,Q))$ has the asymptotic distribution,

$$\sqrt{N}(\hat{\delta} - \tilde{\delta}(P,Q)) \rightsquigarrow \sup_{V\in\mathbb{V}^{\max}}\{G(V)\} \sim J(\cdot; P, Q, \lambda),$$

where $G(V)$ is the set indexed Gaussian processes obtained in the Proposition 2.2.1 and $J(\cdot; P, Q, \lambda)$ represents its cdf. Let $J_N(\cdot; P, Q, \lambda_N)$ be the cdf of $\sqrt{N}(\hat{\delta} - \tilde{\delta}(P,Q))$.
Note that the bootstrap critical value $\hat{c}^{boot}_{1-\alpha}$ and the subsampling critical value $\hat{c}^{sub}_{1-\alpha}$ are both consistent (in probability) to $c_{1-\alpha}$, the $(1-\alpha)$-th quantile of $J(\cdot; P, Q, \lambda)$. Denote these consistent critical

values by $\hat{c}_{1-\alpha}$. Then, for $\epsilon = \tilde{\delta}(P,Q) - 1 > 0$,

$$\begin{aligned} Prob_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} > 1\right) &= Prob_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} + \epsilon > \tilde{\delta}(P,Q)\right) \\ &= Prob_{P,Q,\lambda_N}\left(\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} + \epsilon > \tilde{\delta}(P,Q)\right) \\ &= Prob_{P,Q,\lambda_N}\left(\sqrt{N}(\hat{\delta} - \tilde{\delta}(P,Q)) > \hat{c}_{1-\alpha} - \sqrt{N}\epsilon\right) \\ &= 1 - J_N(\hat{c}_{1-\alpha} - \sqrt{N}\epsilon; P, Q, \lambda_N) \\ &\to 1 \quad \text{as} \quad N \to \infty. \end{aligned}$$

■

2.A.2 A generalization of Proposition 2.2.1

We use the same notation as in Section 2.5. Here, we provide a generalization of Proposition 2.2.1 to the multi-valued instrument case. The following assumptions that are analogous to (A1) through (A4) of Section 2.2 are imposed.

Assumptions

(A1') *Nondegeneracy*: $P_1, \ldots, P_k$ are nondegenerate distributions on $\mathcal{Y} \cup \{mis\}$ and the integrated envelope is positive $\delta > 0$.

(A2') *VC-class*: $\mathbb{V}_1, \ldots, \mathbb{V}_K$ are VC-classes of measurable subsets in $\mathcal{Y}$.

(A3') *Optimal Partition*: There exists a nonempty *maximizer subclass of partitions* $\mathbb{V}^{\max} \subset \mathbb{V}$,

$$\mathbb{V}^{\max} = \{\mathbf{V} \in \mathbb{V} : \delta(\mathbf{V}) = \delta\}$$

(A4') *Existence of a maximizer:* with probability one, there exists a sequence of random partitions $\hat{\mathbf{V}}_N \in \mathbb{V}$ and $\hat{\mathbf{V}}_N^{\max} \in \mathbb{V}^{\max}$ such that

$$\hat{\delta}(\hat{\mathbf{V}}_N) = \sup_{\mathbf{V} \in \mathbb{V}} \{\hat{\delta}(\mathbf{V})\}, \quad \hat{\delta}(\hat{\mathbf{V}}_N^{\max}) = \sup_{\mathbf{V} \in \mathbb{V}^{\max}} \{\hat{\delta}(\mathbf{V})\}$$

holds for every $N \geq 1$.

A generalization of Proposition 2.2.1 is given as follows. A proof can be given in the same manner as the proof of Proposition 2.2.1, and is therefore omitted for brevity.

Proposition 2.2.1'. *Assume (A1'), (A2'), and (A3')*
(i) $\hat{\delta} \to \delta$ *as* $N \to \infty$ *with probability one.*

(ii) Assume further (A4'). *Let* $\mathbb{V}^{\max}$ *be the maximizer subclass of partitions* $\{\mathbf{V} \in \mathbb{V} : \delta(\mathbf{V}) = \delta\}$. *Then,*

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \sup_{\mathbf{V} \in \mathbb{V}^{\max}} \{G(\mathbf{V})\}. \tag{E.3}$$

Here, $G(\mathbf{V})$ *is the mean zero tight Gaussian processes in* $l^{\infty}(\mathbb{V})$ *with the covariance kernel given by, for* $\mathbf{V}^1 = (V_1^1, \ldots, V_K^1) \in \mathbb{V}$ *and* $\mathbf{V}^2 = (V_1^2, \ldots, V_K^2) \in \mathbb{V}$,

$$Cov(G(\mathbf{V}^1), G(\mathbf{V}^2)) = \sum_{k=1}^{K} \lambda_k^{-1} \left[P_k(V_k^1 \cap V_k^2) - P_k(V_k^1) P(V_k^2) \right],$$

where $\lambda_k = \Pr(Z = z_k)$.

2.A.3 An algorithm to estimate $\mathbb{V}^{\max}$ in the histogram class

This appendix presents an algorithm used in the empirical application (Section 2.6). There, we specify $\mathbb{V}$ as the histogram class, i.e., $\mathbb{V}_1 = \cdots = \mathbb{V}_K = \mathbb{V}_{hist}(h, L, \mathcal{Y}_0)$. The main purpose of the following algorithm is to reduce the computational burden in constructing the estimator of the maximizer subclass of partitions $\hat{\mathbb{V}}^{\max}(\eta_N)$.

Let us fix the number of bins, binwidth, and the initial breakpoint y_0. For each P_{n_k}, let $P_{n_k}(H_0(y_0)), \ldots, P_{n_k}(H_L(y_0))$ be the histogram estimates with respect to the $(L+1)$ bins, $H_0(y_0), \ldots, H_L(y_0)$, as defined in Section 2.2. On each bin $H_l(y_0)$, we infer which P_k achieves $\max_{k'}\{P_{k'}(H_l(y_0))\}$ based on the following criterion: $k = \arg\max_{k'}\{P_{k'}(H_l(y_0))\}$ if

$$\sqrt{N}\left(\max_{k'}\{P_{n_{k'}}(H_l(y_0))\} - P_{n_k}(H_l(y_0))\right) \leq \frac{w_l(y_0)}{\sum_{l=1}^{L} w_l(y_0)}\eta_N, \tag{F.1}$$

where $w_l(y_0) = \sqrt{\lambda_{k^*}^{-1} P_{n_{k^*}}(H_l(y_0))(1 - P_{n_{k^*}}(H_l(y_0)))}$ with $k^* = \arg\max_{k'}\left\{P_{n_{k'}}(H_l(y_0))\right\}$. The weighting term is introduced in order to control the variance of the histogram estimates. That is, for the bin on which $\max_{k'}\{P_{n_{k'}}(H_l(y_0))\}$ is larger, we take a relatively larger margin below $\max_{k'}\{P_{n_{k'}}(H_l(y_0))\}$ to admit other P_k to be tied with P_{k^*} on $H_l(y_0)$. By implementing this procedure for every bin, we obtain a set of indices $\mathcal{I}_k^{\max}(y_0) \subset \{0, 1, \ldots, L\}$ for $k = 1, \ldots, K$ that indicates the bins for which P_{n_k} passes the criterion (F.1). By repeating this procedure for each y_0, we form the estimator of the maximizer subclass by

$$\hat{\mathbb{V}}^{\max}(\eta_N) = \left\{ (V_1, \ldots, V_K) : \bigcup_{k=1}^{K} V_k = \mathcal{Y},\ \mu(V_k \cap V_{k'}) = 0 \text{ for } \forall k \neq k',\ V_1 \in \hat{\mathbb{V}}_1, \ldots, V_K \in \hat{\mathbb{V}}_K \right\} \tag{F.2}$$

where $\hat{\mathbb{V}}_k = \left\{ \bigcup_{l \in \mathcal{I}_k^{\max}(y_0)} H_l(y_0) : y_0 \in \mathcal{Y}_0 \right\}$ for $k = 1, \ldots, K$.

For a fixed y_0, $\mathbb{V}$ contains K^{L+1} partitions and a crude way of constructing $\hat{\mathbb{V}}^{\max}(\eta_N)$ would have the computational complexity $O(K^L)$. The above algorithm reduces the computational complexity

from $O(K^L)$ to $O(KL)$.

Bibliography

[1] Alexander, K. S., and M. Talagrand (1989): "The law of the iterated logarithm for empirical processes on Vapnik-Červonenkis classes," *Journal of Multivariate Analysis*, 30, 155-166.

[2] Andrews, D. W. K. (2000):"Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68, 399-405.

[3] Andrews, D. W. K., S. T. Berry and P. Jia (2004): "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Locations," manuscript, Yale University

[4] Andrews, D. W. K. and P. Guggenberger (2008): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, forthcoming.

[5] Andrews, D. W. K. and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497-517.

[6] Andrews, D. W. K. and G. Soares (2007): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," unpublished manuscript, Cowles Foundation, Yale University.

[7] Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323-363.

[8] Blundell, R., H. Reed, and T. Stoker (2003): "Interpreting Aggregate Wage Growth: The Role of Labor Market Participation," *American Economic Review*, 93, 1114-1131.

[9] Bugni, F. (2008): "Bootstrap Inference in Partially Identified Models," unpublished manuscript, Department of Economics, Northwestern University.

[10] Canay, I. A. (2007): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity" unpublished manuscript, University of Wisconsin - Madison.

[11] Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189-218.

[12] Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica*, 75, 1243-1284.

[13] Davydov, Y. A., M. A. Lifshits, and N. V. Smorodina (1998): *Local Properties of Distributions of Stochastic Functionals.* Providence: American Mathematical Society.

[14] Dudley, R. M. (1999): *Uniform Central Limit Theorem.* Cambridge University Press.

[15] Guggenberger, P., J. Hahn, and K. Kim (2008): "Specification Testing Under Moment Inequalities," *Economics Letters*, 99, 375-378.

[16] Hartigan, J. A. (1987): "Estimation on a convex density contour in two dimensions," *Journal of the American Statistical Association*, Vol. 82, pp. 267-270.

[17] Heckman, J. J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313-318.

[18] Hoeffding, W. (1963): "Probability Inequalities for Sums of Bounded Random Variables" *Journal of the American Statistical Association*, Vol. 58, No. 301, pp. 13-30.

[19] Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845-1857.

[20] Lehmann, E. L. and J. P. Romano (2005): *Testing Statistical Hypotheses, Third ed.* Springer-Verlag, New York.

[21] Manski, C. F. (1989): "Anatomy of the Selection Problem." *Journal of Human Resources*, 24, 343-360.

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

[23] 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.

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

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

[26] Manski, C. F. and J. Pepper (2000): "Monotone Instrument Variables: With Application to the Returns to Schooling," *Econometrica*, 68, 997-1010.

[27] Mulligan, C. B. and Rubinstein, Y. (2008): "Selection, Investment, and Women's Relative Wages," *Quarterly Journal of Economics*, 123, 1061–1110.

[28] Pakes, A., J. Porter, K, Ho, and J. Ishii (2006): "Moment Inequalities and Their Application," manuscript, Harvard University.

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

[30] Pollard, D. (1984): *Convergence of Stochastic Processes*, Springer-Verlag, New York.

[31] Politis, D. N. and J. P. Romano (1994): "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions." *The Annals of Statistics*, 22, 2031-2050.

[32] Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling.* New York: Springer.

[33] Polonik, W. (1995): "Measuring Mass Concentrations and Estimating Density Contour Clusters- An Excess Mass Approach," *The Annals of Statistics*, 23, No. 3, 855-881.

[34] Romano, J. P. (1988): "A Bootstrap Revival of Some Nonparametric Distance Tests." *Journal of American Statistical Association*, 83, 698-708.

[35] Romano, J. P. and A. M. Shaikh (2008a): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference,* forthcoming.

[36] Romano, J. P. and A. M. Shaikh (2008b): "Inference for the Identified Set in Partially Identified Econometric Models", manuscript, University of Chicago.

[37] Rosen, A. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107-117.

[38] Shiryaev, A. N. (1996): *Probability, 2nd ed.* New York: Springer.

[39] van der Vaart, A. W., and J.A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics,* New York: Springer.

Chapter 3

Testing for Instrument Validity in the Heterogeneous Treatment Effect Model

3.1 Introduction

In this paper, we develop a test procedure for the instrumental validity in the heterogeneous treatment effect model. When we suspect that one's participation to a treatment depends on his potential outcomes, a common strategy to extract identifying information for counterfactual causal effects is to employ an instrumental variable Z. As is demonstrated in Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996), and Heckman and Vytlacil (1999, 2001), when the instrument satisfies the two key conditions, we can point-identify the average causal effects for those whose participation decision is strictly randomized by the instrument. (the local average treatment effect) These key conditions consist of i) *random treatment assignment (RTA):* an instrument is assigned independently from individual heterogeneities which affect one's outcome and participation response, and ii) *monotonic participation response to instrument (MPR):* one's participation response to the instrument is uniform in a certain sense over the entire population.[1]

When we analyze (quasi-)experimental data with possible incompliance, we often use the initial treatment assignment as an instrument. In this case, the instrumental validity is reasonably satisfied as far as the initial treatment assignment is completely randomized and incompliance is allowed only for those who are initially assigned to the treatment group (see, for example, Abadie, Angrist, and Imbens (2002) and Kling, Liebman, and Katz (2007)). But, if the incompliance is also allowed for those initially assigned to the control group, we face a risk of violating MPR. Examples of this

[1] MPR considered in this paper stands for the restriction termed as "monotonicity" in Imbens and Angrist (1994). The reason that we call it MPR is to distinguish the monotonicity between one's participation response and instrument from the monotonicity between one's outcome response and instrument considered in Manski and Pepper (2000).

contain the well-known draft lottery of Angrist (1991) and the applications of the fuzzy regression discontinuity design (Campbell (1969), Hahn, Todd, and Van der Klaauw (2001)) where eligibility for a treatment based on one's attribute is used as an instrument. When we conduct an analysis using observational data, the exogeneity of instrument becomes less credible and therefore, not only MPR, but also RTA becomes a threat for the instrument validity. Although validating an instrument is the core of identifying the causal effects, there have been no procedures proposed to empirically test the aforementioned instrumental validity. Because of this, the instrumental validity is simply assumed or justified by indirect evidence outside of data.

The first contribution of this paper is to clarify the testability of the instrument validity in the heterogenous treatment effect model with a binary treatment and a binary instrument. The refutability result of this paper is closely related to the point-identification result of the complier's outcome distributions by Imbens and Rubin (1997). They show that under RTA and MPR, the distribution of complier's treated outcome and that of complier's control outcome are point-identified. But, from the data, the point estimator of the complier's outcome densities can take negative values on some subsets in the outcome support. We focus on this phenomenon as a clue to refute the instrumental validity. That is, if we obtain negative estimates for complier's treated outcome or control outcome density on some regions in the outcome support, we interpret it as a counter-evidence for the joint restriction of RTA and MPR since the probability density function cannot be negative. We derive the condition for the data generating process to yield nonnegative complier's potential outcome densities. We demonstrate that the refuting rule based on that condition is most powerful for screening out the violation of the instrumental validity in the heterogenous treatment effect model.

The second contribution of this paper is to develop a specification test for the instrumental validity based on the aforementioned refutability result. We propose a Kolmogorov-Sminov type test statistic to measure how serious the nonnegativity of the compliers outcome density is violated in data. The asymptotic distribution of the proposed test statistic is not analytically tractable, and therefore the critical values are difficult to obtain. In order to overcome this problem, we develop a bootstrap algorithm to obtain asymptotically valid critical values. As Romano (1988) demonstrated, the bootstrap is widely applicable and easy to implement to obtain the critical values of the general Kolmogorov-Sminov type goodness-of-fit statistic. This is also the case for our test procedure.

The rest of the paper is organized as follows. In Section 3.2, we demonstrate the refutability of the instrumental validity in the heterogeneous treatment effect model. In Section 3.3, we construct a statistic to test the testable implication obtained in Section 3.2 and provide an algorithm of the bootstrap procedure. Monte Carlo simulations and two empirical applications are provided in Section 3.4. Proofs are provided in Appendices.

3.2 Model

Let Y_1 represent the potential outcome with a treatment, and Y_0 represent the potential outcome without the treatment. They are scalar variables and their support is denoted by $\mathcal{Y}$. The observed outcome is denoted by Y_{obs}. Let D indicate the observed participation response such that $D = 1$ when one participates to the treatment while $D = 0$ if one does not. Thus, the observed outcome is written as $Y_{obs} = Y_1 D + Y_0(1 - D)$. We denote a binary instrument by Z. As in Angrist and Imbens (1994), we introduce D_1 as the potential participation decision that one would take if $Z = 1$. Similarly, we define D_0 for $Z = 0$. Associated with the potential selection indicators, we define the individual type T that indicates individual participation response to the instrument Z.

$$
\begin{aligned}
T = c\text{: } &\textit{complier} \quad \text{if } D_1 = 1, D_0 = 0 \\
T = n\text{: } &\textit{never-taker} \quad \text{if } D_1 = 0, D_0 = 0 \\
T = a\text{: } &\textit{always-taker} \quad \text{if } D_1 = 1, D_0 = 1 \\
T = d\text{: } &\textit{defier} \quad \text{if } D_1 = 0, D_0 = 1.
\end{aligned}
$$

The following three assumptions guarantee point-identification of the local average treatment effects for compliers as well as the marginal distributions of the counterfactual outcomes for compliers (see Imbens and Angrist (1994) and Imbens and Rubin (1997)).

Assumption

1. *Random Treatment Assignment (RTA):* Z is jointly independent of (Y_1, Y_0, D_1, D_0).
2. *Monotonic Participation Response to Instrument (MPR): Without loss of generality, assume* $\Pr(D = 1|Z = 1) \geq \Pr(D = 1|Z = 0)$. The potential participation indicators satisfy $D_1 \geq D_0$ *with probability one.*

Note that the above assumptions are defined in terms of the potential variables. RTA is stronger than the conventional instrumental exclusion restriction since it only restricts Z to being independent of the potential outcomes. MPR states that the ordering of the potential participation indicators are identical over the entire population and there are no defiers in the population $\Pr(T = d) = 0$. Since we never observe all the potential variables of the same individual, we cannot directly examine these assumptions from data, and therefore necessary and sufficient testable implications for these assumptions are not available. Hence, we examine the refutability by looking for a testable implication as a necessary condition for the instrumental validity.

To illustrate our analytical framework, we introduce the following notations. Let P and Q be the conditional probability distributions of $(Y_{obs}, D) \in \mathcal{Y} \times \{1, 0\}$ given $Z = 1$ and $Z = 0$ respectively. We interpret the data generating process to have the two-sample structure in terms of the assigned value of Z. For a subset $A \subset \mathcal{Y}$ and $d = 1, 0$, $P(Y_{obs} \in A, D = d)$ and $Q(Y_{obs} \in A, D = d)$ represent

$\Pr(Y_{obs} \in A, D = d|Z = 1)$ and $\Pr(Y_{obs} \in A, D = d|Z = 0)$ respectively. Note that P and Q are the joint distributions of the *observable* variables (Y_{obs}, D), and therefore we can consistently estimate P and Q by data.

We now state the refutability result of the instrumental validity. Provided that the population has a strictly positive fraction of compliers, the conclusion of the next proposition is equivalent to the nonnegativity of the complier's outcome densities pinned down under the instrumental validity (Imbens and Rubin (1997)). A proof is given in Appendix 3.A.1.

Proposition 3.2.1 *If a population distribution of* (Y_1, Y_0, D_1, D_0, Z) *satisfies RTA and MPR, then, the data generating process* P *and* Q *satisfies the following inequalities for arbitrary Borel sets* B *in* $\mathcal{Y}$,

$$\begin{aligned} P(Y_{obs} \in B, D = 1) &\geq Q(Y_{obs} \in B, D = 1), \\ P(Y_{obs} \in B, D = 0) &\leq Q(Y_{obs} \in B, D = 0). \end{aligned} \tag{3.2.0.1}$$

Conversely, if the data generating process P *and* Q *satisfies these inequalities for all Borel sets* B, *then there exists a joint probability law of* (Y_1, Y_0, D_1, D_0, Z) *that is compatible with the data generating process* P *and* Q, *RTA, and MPR.*

Let $p(y, D = d)$ and $q(y, D = d)$ be the probability density function of P and Q on $\mathcal{Y} \times \{d\}$ with respect to a dominating measure μ. . In terms of the density functions, the above two inequalities are equivalent to

$$\begin{aligned} p(y, D = 1) &\geq q(y, D = 1) \quad \mu\text{-a.e.}, \\ p(y, D = 0) &\leq q(y, D = 0) \quad \mu\text{-a.e.} \end{aligned}$$

These inequalities imply that when the instrument is valid, we must observe the configuration of the densities as in Figure 1. The left-hand side figure corresponds to Y_1's distribution and the right figure corresponds to Y_0's distribution. The dotted line in each figure represents the probability density of the potential outcomes, i.e., $f_{Y_1}(y)$ is the marginal density of the treated outcome and $f_{Y_0}(y)$ is the marginal density of the control outcome. The solid lines represent $p(y, D = d)$ and $q(y, D = d)$, which are point-identifiable by data. Note that their integrals are equal to the probability of $D = d$ conditional on Z. Therefore, the scale of $p(y, D = d)$ and $q(y, D = d)$ is smaller than $f_{Y_1}(\cdot)$ and $f_{Y_0}(\cdot)$. Furthermore, $p(y, D = d)$ and $q(y, D = d)$ both lie below the potential outcome density $f_{Y_d}(\cdot)$. This is because RTA implies

$$\begin{aligned} f_{Y_d}(y) &= f_{Y_d|Z}(y|Z = 1) \\ &= f_{Y_d,D|Z}(y, D = d|Z = 1) + f_{Y_d,D|Z}(y, D = 1 - d|Z = 1) \\ &= p(y, D = d) + f_{Y_d,D|Z}(y, D = (1 - d)|Z = 1) \end{aligned}$$

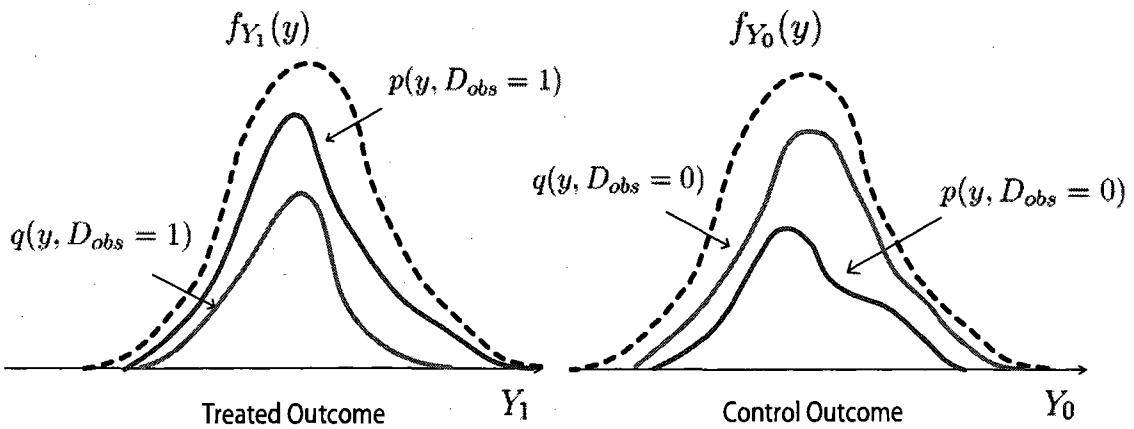

Figure 3.1: *When we observe that the observable densities* $p(y, D = 1)$ and $q(y, D = d)$ *are nested as in this figure, the instrumental validity is not refuted.*

and

$$f_{Y_d}(y) = q(y, D = d) + f_{Y_d, D|Z}(y, D = (1 - d)|Z = 0).$$

The second term in the right hand side of the above equations correspond to the density function for the missing treated or control outcomes, so they must be nonnegative.

When RTA and MPR hold in the population, Proposition 3.2.1 implies that the two identifiable density functions $p(y, D = d)$ and $q(y, D = d)$ must be nested as shown in Figure 3.1. For the treated outcome densities, $p(y, D = 1)$ must lie above $q(y, D = 1)$ and for the control outcome densities, $q(y, D = 0)$ must lie above $p(y, D = 0)$. Under RTA and MSR, we can point-identify the complier's outcome densities by the areas between these two densities rescaled by their area (see the proof of Proposition 3.2.1 in Appendix 3.A.1). Thus, the inequalities of Proposition 3.2.1 constitute necessary conditions for the instrument validity.

The converse statement of Proposition 3.2.1 clarifies that if the data generating process admits the inequalities (3.2.0.1), then we can construct a population distribution of (Y_1, Y_0, D_1, D_0, Z) which does not contradict the data generating process and the instrument validity. This implies that no other refuting rules can screen out violations of the instrument validity more than the refuting rule based on the inequalities (3.2.0.1) does. In this sense, the refuting rule of Proposition 3.2.1 has the most screening power in detecting violation of instrument validity.

Note that Proposition 3.2.1 does not give an if and only if statement for the instrumental validity. That is, an invalid instrument does not necessarily imply a violation of the inequalities. In this sense, testing the inequalities does not guarantee to screen out all the possible violations of the instrumental validity.

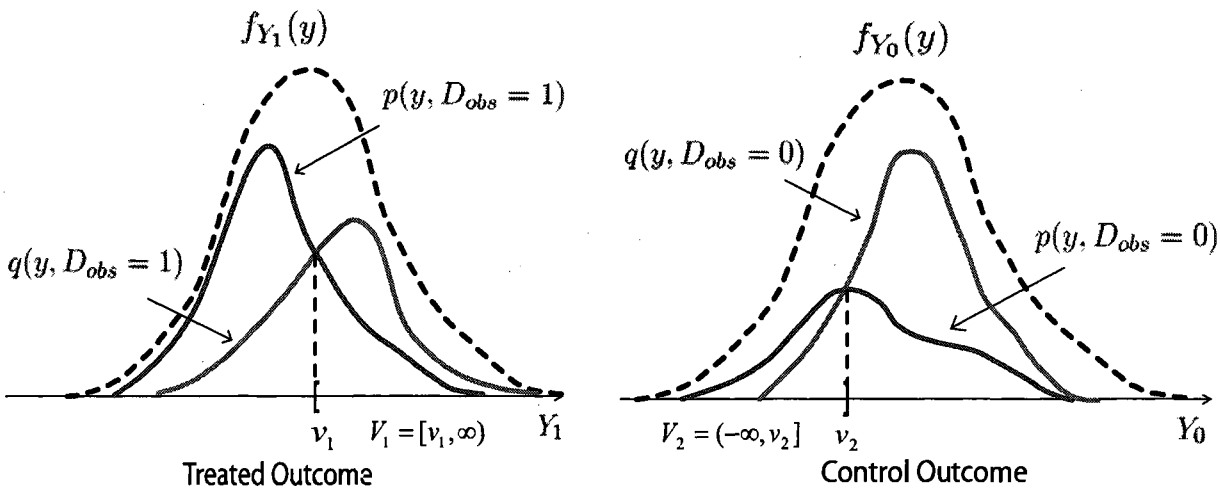

Figure 3.2: *When we observe the above configuration of the densities, we can refute the instrumental validity since at the subset* $V_1 = [v_1, \infty)$, *the first inequality in Proposition 3.2.1 is violated. The right-hand side picture shows that the second inequality in Proposition 1 is violated at* $V_2 = (-\infty, v_2]$.

If we observe the configuration of the densities like Figure 3.2, we can refute at least one of the instrumental validity conditions since some of the inequalities (3.2.0.1) are violated on some subsets of the outcome support. These subsets are labeled as V_1 and V_2 in Figure 3.2. Although observing the configuration of the densities like Figure 3.2 does not tell us which conditions are violated in the population, it allows us to conclude that the chosen instrument is not valid to point-identify the local average treatment effects and, hence, the classical IV-estimator breaks down.

3.3 Test Procedure

P and Q are point-identified by the sampling process, and therefore we can examine the validity of the inequalities (3.2.0.1) by inferring whether estimators for P and Q satisfy them or not.

Let sample consist of N i.i.d observations of (Y_{obs}, D, Z). We divide the sample into two subsamples in terms of the value of Z. Let m be the sample size with $Z_i = 1$ and n the sample size with $Z_i = 0$. Let $(Y^1_{obs,i}, D^1_i)$, $i = 1, \ldots, m$ be the observations with $Z = 1$ and $(Y^0_{obs,j}, D^0_j)$, $j = 1, \ldots, n$ be those with $Z = 0$. We assume $m/N \to \lambda$ as $N \to \infty$ almost surely where $\lambda \in (\epsilon, 1-\epsilon)$ for some $\epsilon > 0$. We estimate P and Q by the empirical distributions,

$$
\begin{aligned}
P_m(V,d) &\equiv \frac{1}{m}\sum_{i=1}^{m} I\{Y^1_{obs,i} \in V \text{ and } D^1_i = d\}, \\
Q_n(V,d) &\equiv \frac{1}{n}\sum_{j=1}^{n} I\{Y^0_{obs,j} \in V \text{ and } D^0_j = d\}.
\end{aligned}
$$

We measure the degree of violation of the inequalities (3.2.0.1) by the next statistic.

$$
T_N = \left(\frac{mn}{N}\right)^{1/2} \max\left\{ \begin{array}{l} \sup_{V\in\mathbb{V}}\{Q_n(V,1) - P_m(V,1)\}, \\ \sup_{V\in\mathbb{V}}\{P_m(V,0) - Q_n(V,0)\} \end{array} \right\}, \tag{3.3.0.2}
$$

where $\mathbb{V}$ is a collection of subsets in $\mathcal{Y}$.

This test statistic is designed to measure the degree of the violations of the inequalities (3.2.0.1) using the empirical distributions. If the sample counterpart of the inequality (3.2.0.1) is violated for a subset V, then, the first supremum in the max operator of the test statistic is positive. Similarly, when the sample counterpart of the inequality (3.2.0.1) is violated for some subset V, then the second term becomes positive. The proposed test statistic returns the maximal deviations of the above inequalities where the maximum is searched over a class of subsets $\mathbb{V}$.

The test statistic can be seen as a variant of the Kolmogorov-Sminov type nonparametric distance test statistic (Romano (1988)). This test statistic is not pivotal due to the discreteness of D and the asymptotic distribution can depend on P and Q. Choice of $\mathbb{V}$ will not affect the size of test while it can affect power of the test.

Although Proposition 1 suggests us to take $\mathbb{V}$ as the Borel σ-algebra of $\mathcal{Y}$, we cannot take it to be as rich as the Borel σ-algebra unless Y is discrete. In order for the above test statistic to have an asymptotic distribution, a specified $\mathbb{V}$ has to guarantee the uniform convergence property of the empirical processes of P_m and Q_n. A class of subsets which meets this requirement is the Vapnik-Červonenkis class (VC-class). For example, a collection of left unbounded intervals $\{(-\infty, y]; y \in \mathbb{R}\}$ and a collection of the finite number of disjoint intervals are the examples of the VC-classes. (See e.g., Dudley (1999) and van der Vaart and Wellner (1996) for the general construction of the VC-classes).

We will employ two specific VC-classes in our Monte Carlo studies and empirical applications given in the next section. They are the *half unbounded interval class* $\mathbb{V}_{half}$ and the *histogram class*

$\mathbb{V}_{hist}$. The half unbounded interval class is simply a collection of right unbounded intervals and left unbounded intervals,

$$\mathbb{V}_{half} = \{(-\infty, y]; y \in \mathbb{R}\} \cup \{[y, \infty); y \in \mathbb{R}\}. \tag{3.3.0.3}$$

The histogram class is the power set of the histogram bins whose breakpoints can float over $\mathbb{R}$. Algebraically, this can be expressed as follows. Let $h > 0$ be a fixed positive number representing the binwidth and L be the number of bins. Pick an initial breakpoint $y_0 \in \mathbb{R}$ and consider equally distanced L points $-\infty < y_0 < y_1 < \cdots < y_{L-1} < \infty$ where $y_l = y_0 + lh$, $l = 1, \ldots, (L-1)$. Denote the $(L+1)$ disjoint intervals formed by these L points by $H_0(y_0, h) = (-\infty, y_0]$, $H_l(y_0, h) = [y_{l-1}, y_l]$, $l = 1, \ldots, (L-1)$, and $H_L(y_0, h) = [y_{L-1}, \infty)$. Let $I_j(L)$, $j = 1, \ldots, 2^{L+1}$ represent all the possible subsets of the indices $\{0, 1, \ldots, L\}$. Given $\mathcal{Y}_0$ a set of the smallest breakpoint y_0, the histogram class with binwidth h and the number of bins L is defined as

$$\mathbb{V}_{hist}(h, L, \mathcal{Y}_0) = \left\{ \bigcup_{l \in I_j(L)} H_l(y_0, h) : y_0 \in \mathcal{Y}_0,\ j = 1, \ldots, 2^{L+1} \right\}. \tag{3.3.0.4}$$

In contrast to a rather complicated expression, the histogram class is flexible and simple to implement.

For the test statistic (3.3.0.2), $P = Q$ is the least favorable null hypothesis among the composite null hypotheses defined by the inequalities (3.2.0.1). Therefore, we will find the critical value with a nominal level α by estimating the $(1-\alpha)$-th quantile of the asymptotic distribution of T_N under the least favorable null $P = Q$. If the estimated critical values are consistent to the $(1-\alpha)$-th quantile of the asymptotic distribution of T_N under the least favorable null, the resulting testing procedure has correct size.

As discussed in Romano (1988), the resampling method is an attractive approach to estimate asymptotically valid critical values for the Kolmogorov-Sminov type test statistic since its asymptotic distribution generally does not have an analytically tractable distribution function. Bootstrap resolves this issue by estimating the null distribution of the statistic by the empirical distribution of the resampled test statistics. Given that the composite null has the least favorable null, bootstrap samples are drawn from $\hat{P}$ and $\hat{Q}$, which is consistent to the least favorable null hypothesis, i.e., $\hat{P} = \hat{Q}$. In the two sample hypothesis testing problem with the null hypothesis given by the equality of the two distributions, one choice of the resampling distribution is the pooled empirical distribution H_N, the empirical distribution of the pooled data $(Y^1_{obs,1}, D^1_1), \ldots, (Y^1_{obs,m}, D^1_m), (Y^0_{obs,1}, D^0_1), \ldots, (Y^0_{obs,n}, D^0_n)$. Abadie (2002) proposes the bootstrap procedure to test hypotheses on distributional features between the complier's treated and control outcomes. Although the null hypothesis and test statistic are different, our bootstrap procedure shown below is analogous to Abadie (2002).

Bootstrap procedure:

1. *Sample* $(Y^*_{obs,i}, D^*_i)$, $i = 1, \ldots, m$ *randomly with replacement from the pooled empirical distribution* H_N *and construct the bootstrap empirical distribution* P^*_m. *Similarly, sample* $(Y^*_{obs,j}, D^*_j)$,

$j = 1, \ldots, n$ randomly with replacement from the pooled empirical distribution H_N and construct the bootstrap empirical distribution Q_n^*.

2. *Compute the test statistic T_N^* defined in (3.3.0.2) by plugging in the bootstrapped empirical distributions P_m^* and Q_n^*.*

3. *Iterate Step 1 and Step 2 and get the empirical distribution of T_N^*. For a chosen nominal level $\alpha \in (0, 1/2)$, we obtain the bootstrapped critical value $\hat{c}_{boot}(1-\alpha)$ from its empirical $(1-\alpha)$-th quantile .*

4. *Reject the null hypothesis if $T_N > \hat{c}_{boot}(1-\alpha)$.*

Note that the bootstrap sample is drawn from the pooled empirical distribution because our interest is in estimating the null distribution of T_N under the least favorable null hypothesis, $P = Q$. This enables us to control the supremum of the asymptotic false rejection probabilities at the chosen nominal level α,

$$\sup_{(P,Q)\in H_0} \lim_{N\to\infty} \Pr(T_N > \hat{c}_{boot}(1-\alpha)) = \alpha. \tag{3.3.0.5}$$

This is the conventional definition of the pointwise consistency of test.

The asymptotic validity of the proposed bootstrap is stated in the next proposition. A proof is given in Appendix 3.A.2.

Proposition 3.3.1 *Let $\mathbb{V}$ be a VC-class and $\alpha \in (0, 1/2)$. (i) For the null hypothesis of P and Q given by the inequalities (3.2.0.1), the proposed bootstrap test procedure provides pointwise correct asymptotic size (3.3.0.5). (ii) If, for a fixed alternative, there exist some $V \in \mathbb{V}$ which violates (3.2.0.1), then the proposed bootstrap testing procedure is consistent, i.e., the rejection probability converges to one as $N \to \infty$.*

3.4 Monte Carlo Studies and Empirical Applications

3.4.1 Small sample performance

To examine the finite sample performance of our bootstrap test, we perform a Monte Carlo simulation. We specify the sampling process as the least favorable null $P = Q$, and therefore the test asymptotically achieves nominal size.

$$p(y, D = 1) = q(y, D = 1) = 0.5 \times \mathcal{N}(1, 1),$$
$$p(y, D = 0) = q(y, D = 0) = 0.5 \times \mathcal{N}(0, 1).$$

Table 1: Test Size in Small Samples
Monte Carlo iterations 2000, Bootstrap iterations 500.

	Specification of $\mathbb{V}$ Nominal test size								
	$\mathbb{V}_{half}$			$\mathbb{V}_{hist}$ binwidth 0.8			$\mathbb{V}_{hist}$ binwidth 0.4		
sample size (m,n)	.10	.05	.01	.10	.05	.01	.10	.05	.01
(50,50)	.085	.042	.008	.098	.049	.009	.106	.053	.010
(50,250)	.124	.073	.022	.098	.046	.008	.118	.058	.014
(100,100)	.108	.054	.015	.113	.052	.015	.104	.054	.001
(500,500)	.092	.046	.011	.104	.057	.017	.112	.062	.014
s.e.	.007	.005	.002	.007	.005	.002	.007	.005	.002

We consider two specifications of $\mathbb{V}$. One is the half unbounded interval class $\mathbb{V}_{half}$ and the other is the histogram class $\mathbb{V}_{hist}$ defined in Section 3.3. The histogram class provides a finer collection of subsets than the half unbounded interval class. This implies that the histogram class has more refutability power in the sense that it can asymptotically reject more alternatives than the half unbounded interval class. In the finite sample situation, however, there will be a trade-off between asymptotic refutability power and finite sample test power. In order to see the effect of a choice of the binwidth of $\mathbb{V}_{hist}$ to test size and power, we consider two different choices of binwidth, 0.8 and 0.4. The number of bins are 12 and 24 respectively. The set of initial breakpoints are $\mathcal{Y}_0 = [-4.4, -3.6)$ for the former histogram class and $\mathcal{Y}_0 = [-4.4, -4.0)$ for the latter.

For each specification of the sample size (m, n), we simulate the test procedure 2000 times with 500 bootstrap iterations. Table 1 shows that for every specification of $\mathbb{V}$, the test has good size performance even for relatively small sample size, $(m, n) = (50, 50)$. The unbalanced sample case, $(m, n) = (50, 250)$, shows a slight size distortion, while size of the test is overall satisfactory. In addition, we can see that size of the test is not affected by the choice of $\mathbb{V}$.

In order to see finite sample power of our test procedure, we simulate the empirical rejection rate of the bootstrap test against a fixed alternative. The data generating process is specified as

$$p(y, D = 1) = 0.55 \times N(1, 1.44), \qquad q(y, D = 1) = 0.45 \times N(0.2, 1)$$
$$p(y, D = 0) = 0.45 \times N(0, 1), \qquad q(y, D = 0) = 0.55 \times N(0, 1).$$

Figure 3.3 presents the densities of the specified data generating process. From this figure, we can observe that the instrumental validity is refuted by the configuration of the treated outcome densities since $p(y, D = 1)$ intersects with $q(y, D = 1)$. Table 2 presents the simulated rejection probabilities. We specify $\mathbb{V}$ as the histogram classes with the binwidth 0.8 or 0.4, the number of bins 12 or 24, and the set of initial breakpoints $\mathcal{Y}_0 = [-6.2, -5.4)$ or $\mathcal{Y}_0 = [-6.2, -5.8)$. For the specified alternative, we find that the simulated power is very poor in the small sample case. It is even lower than nominal size when $(m, n) = (50, 50)$. The test procedure gains power for relatively

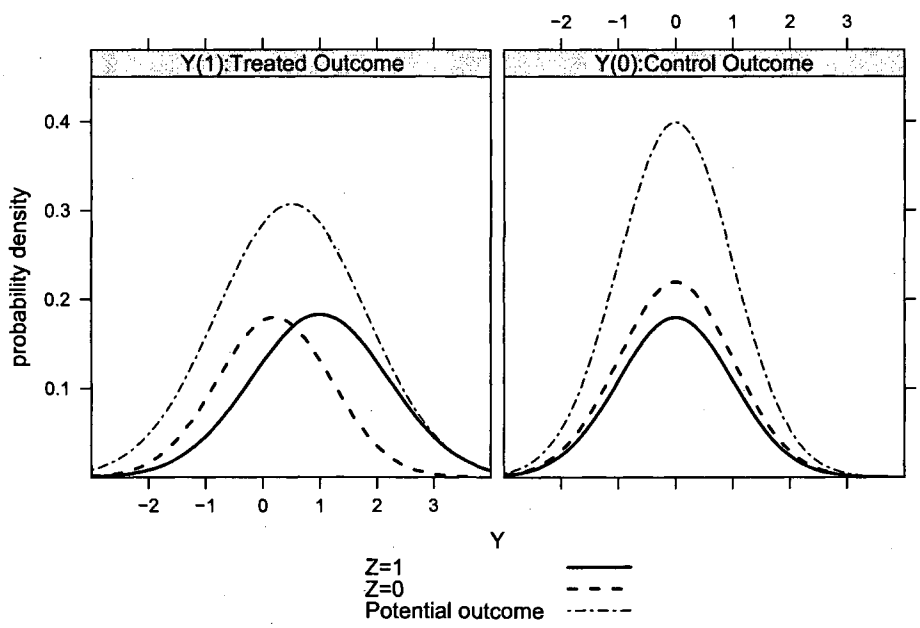

Figure 3.3: **Simulation of Test Power: Specification of Densities.** *The instrumental validity is refuted since for the treated outcomes the two observable densities intersect. Note that in each panel the density drawn to cover the other two represents the probability density of the potential outcomes.*

Table 2: Power against the Fixed Alternative
Monte Carlo iterations 2000, Bootstrap Iterations 500

	specification of $\mathbb{V}$ significance level					
	$\mathbb{V}_{hist}$ with binwidth 0.8			$\mathbb{V}_{hist}$ with binwidth 0.4		
	.10	.05	.01	.10	.05	.01
sample size	rejection probability			rejection probability		
(50,50)	.067	.033	.007	.062	.028	.006
(100,100)	.118	.068	.017	.071	.037	.009
(250,250)	.343	.227	.090	.234	.141	.045
(500,500)	.710	.595	.356	.521	.396	.189

large sample size $(m, n) = (500, 500)$. We can also observe that $\mathbb{V}_{hist}$ with the shorter binwidth is less powerful than that with the wider binwidth. This can be explained that as the binwidth gets finer, the distribution of the test statistic under the least favorable null $P = Q$ has more variance and it raises the bootstrap critical values. This makes our test procedure less powerful. This suggests that given the finite sample there is a trade-off between the richness of $\mathbb{V}$, or equivalently, asymptotic refuting power and the finite sample power. Regardless of its practical importance in choosing $\mathbb{V}$, we make the choice of $\mathbb{V}$ out of scope of this paper and leave that as a part of future research.

3.4.2 Empirical Applications

We illustrate a use of the test procedure with using the following two data sets. The first one is the draft lottery data during Vietnam era used in Angrist (1991). The second one is from Card (1993) on returns to schooling using geographical proximity to college as an instrument.

Draft Lottery Data

The draft lottery data consist of a sample of 10,101 white men, born in 1950-1953. The data source is March Current Population Surveys of 1979 and 1981-1985. The outcome variable is measured in terms of the logarithm of weekly earnings imputed by the annual labor earnings divided by weeks worked. The treatment is whether one has a Vietnam veteran status or not. Since the enrollment for the military service possibly involves self-selection based on one's future earning, the veteran status is not considered to be randomly assigned. In order to solve this endogeneity issue, Angrist (1991) constructs the binary indicator of the draft eligibility, which is randomly assigned based on one's birthdate through the draft lotteries. A justification of the instrumental validity here is that the instrument is generated being independent of any individual characteristics. Hence, it is reasonable to argue that the instrument satisfies RTA. On the other hand, the validity of MPR is less credible since the existence of defiers are not eliminated by the sampling design, i.e., in the sample there are observations who participate to the military service even though they are not initially drafted.

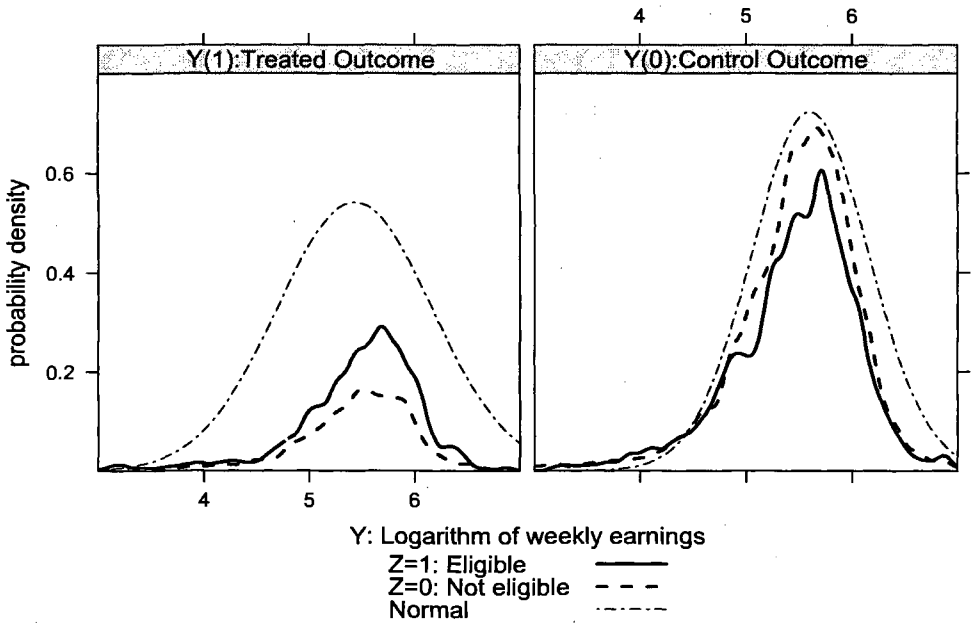

Figure 3.4:

The proposed testing procedure gives a solution to validate these assumptions from data. Figure 4 plots the kernel density estimates for the observed outcome distribution multiplied by the selection probability. We observe that the configuration of the densities in Figure 3.4 is similar to Figure 3.1. Therefore, we do not expect that the instrumental validity is refuted by the testing procedure. As Table 3 shows, p-value of the bootstrap test is almost one, and we do not refute the instrumental validity from the data.

Returns to Education: Proximity to College Data

The Card data is based on National Longitudinal Survey of Young Men (NLSYM) began in 1966 with age 14-24 men and continued with follow-up surveys through 1981. Based on the respondents' county of residence at 1966, the Card data provides the presence of a 4-year college in the local labor market. Observations of years of education and wage level are based on the follow-ups' educational attainment and wage level responded in the interview in 1976.

Table 3: Test Results of the Empirical Applications
Bootstrap iterations 500

	Draft lottery data		Proximity to college data			
	Full sample		Full sample		Restricted sample	
sample size (m,n)	(2780,7321)		(2053,957)		(1047,144)	
$\Pr(D=1\|Z=1)$, $\Pr(D=1\|Z=0)$	0.31, 0.19		0.29, 0.22		0.35,0.24	
$\mathbb{V}_{hist}$ binwidth	0.8	0.4	1.0	0.5	1.0	0.5
Bootstrap test, p-value	0.988	1.00	0.00	0.00	0.997	0.997

The idea of using the proximity to college as an instrument is stated as follows. Presence of a nearby college reduces a cost of college education by allowing students to live at home, while one's inherited ability is presumably independent of his birthplace. Compliers in this context can be considered to be those who grew up in relatively low-income families and who were not able to go to college without living with their parents. We make the educational level as a binary treatment which indicates one's education years to be greater or equal to 16 years. Roughly speaking, the treatment is considered as a four year college degree.

We specify the measure of outcome to be the logarithm of weekly earnings. In the first specification, we do not control any demographic covariates. This simplification raises a concern for the violation of RTA. For instance, one's region of residence, or whether they were born in the standard metropolitan area or rural area may affect one's wage levels and the proximity to colleges if the urban areas are more likely to have colleges and has higher wage level compared with the rural areas. This kind of confounder may contaminate the validity of RTA. In fact, Card (1993) emphasizes an importance of controlling for regions, residence in the urban area, race, job experience, and parent's education in order to make use of the college proximity as an instrument.

Figure 3.5 presents the kernel density estimates for observed outcome densities. In contrast to Figure 3.4, the kernel density estimates in Figure 3.5 intersect especially for those of the control outcomes. That is, the configuration of the densities are similar to Figure 3.1, and this indicates the violation of the instrument validity. Our test procedure yields zero p-value and this provides an empirical evidence that, without any covariates, college proximity is not a valid instrument.

We next look at how the test result changes once we control for some covariates. Controlling discrete covariates can be done by simply making the whole analysis conditional on the specified value of the covariates. We consider restricting the sample to be white workers (black dummy is zero), not living in south states in 1966 (south66 dummy is zero), and living in a metropolitan area in 1966 (SMSA66 dummy is one). That is, we are controlling for race, whether or not one grew up in southern states, and whether or not one grew up in urban area. The size of the restricted sample is 1191 ($m = 1047$, $n = 144$). Figure 3.6 indicates that the kernel density estimates do not reveal a clear evidence for a violation of the instrumental validity. This observation is also supported by the

high p-value of the proposed test. Thus, we conclude that the instrumental validity is not refuted once we control for these covariates.

3.5 Concluding Remarks

In this paper, we develop the bootstrap test procedure to empirically check the conditions of the instrumental validity of Imbens and Angrist (1994). Our testing strategy focuses on the nonnegativity of the complier's outcome densities that are point-identified when the instrument is valid. The nonnegativity of the complier's outcome density is equivalently expressed as the inequalities between the joint probability distributions of Y_{obs} and D conditional on Z. We demonstrate that the inequalities provide the testable implication that has most refuting power. Our test statistic is designed to measure the discrepancy of these inequalities, and it has a form of the supremum statistic on the difference between the two empirical distributions over a specified VC-class of subsets. We develop the bootstrap algorithm to derive the critical values since the asymptotic distribution of the proposed statistic is not analytically tractable.

There are some issues left for future work. First of all, we do not formally investigate how to choose a VC-class $\mathbb{V}$ and how it affects the test performance in the finite sample case. We propose the two different choices of $\mathbb{V}$ in our simulation studies, the half unbounded interval class and the histogram class. We observe that test size is not affected by a choice of $\mathbb{V}$ while power of the test is sensitive to a specification of $\mathbb{V}$.

Second, this paper exclusively considers the binary instrument case. When an instrument is multi-valued, but as long as its support is discrete, it is possible to test the instrument validity for every pair of two instrumental values. However, it is not clear what is a suitable test statistic when we want to test the instrument validity jointly over multiple instrument values. We leave a further discussion of the multi-valued instrument case for future work.

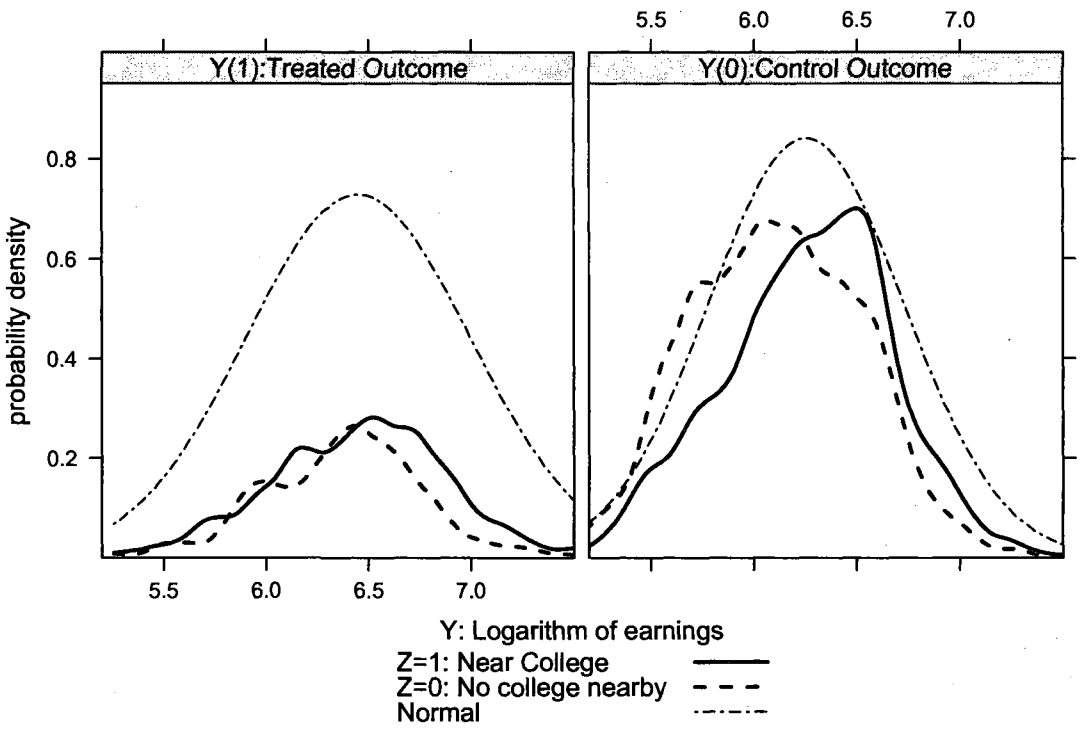

Figure 3.5: **Kernel Density Estimates for the Proximity to College Data (white workers, not living in south states, and living in a metropolitan area).** *The Gaussian kernel with bandwidth 0.1 is used. In each panel, we draw a normal density to illustrate the scale of the estimated densities.*

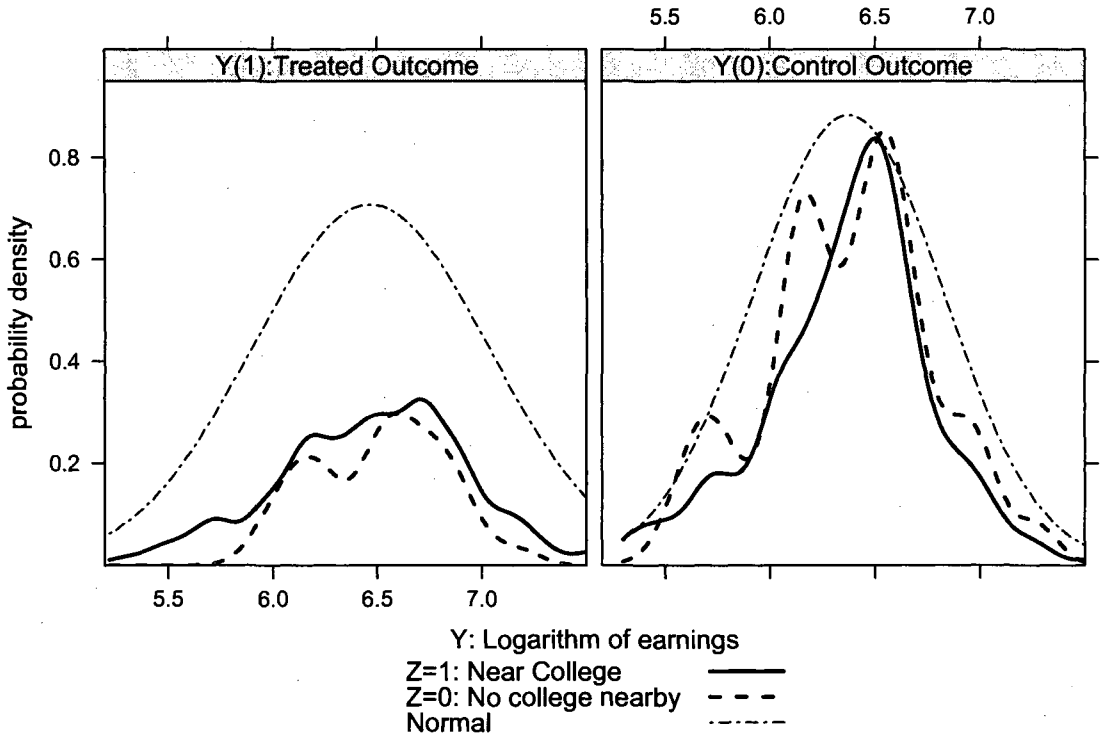

Figure 3.6: **Kernel Density Estimates for the Proximity to College Data (white workers, not living in south states, and living in a metropolitan area).** *The Gaussian kernel with bandwidth 0.1 is used. In each panel, we draw a normal density to illustrate the scale of the estimated densities.*

3.A Appendices

3.A.1 Proof of Proposition 3.2.1

Denote the population distribution of the types by $\pi_t \equiv \Pr(T = t)$, $t \in \{c, n, a, d\}$. Under RTA, $P(B, 1)$, for any Borel set $B \subset \mathcal{Y}$, is expressed as the following.

$$\begin{aligned}
P(B,1) &= \Pr(Y_{obs} \in B | D = 1, Z = 1) \Pr(D = 1 | Z = 1) \\
&= \left[\sum_{t \in \{c,n,a,d\}} \Pr(Y_1 \in V | D_1 = 1, Z = 1, T = t) \Pr(T = t | D_1 = 1, Z = 1) \right] \\
&\qquad \times \Pr(D_1 = 1 | Z = 1) \\
&= \left[\sum_{t \in \{c,n,a,d\}} \Pr(Y_1 \in B | D_1 = 1, T = t) \Pr(T = t | T \in \{c, a\}) \right] \Pr(T \in \{c, a\}) \\
&= \left[\Pr(Y_1 \in B | T = a) \frac{\pi_a}{\pi_a + \pi_c} + \Pr(Y_1 \in V | T = c) \frac{\pi_c}{\pi_a + \pi_c} \right] \\
&\qquad \times (\pi_a + \pi_c) \\
&= \Pr(Y_1 \in V | T = a) \pi_a + \Pr(Y_1 \in V | T = c) \pi_c. \qquad (3.1.1.6)
\end{aligned}$$

The second line follows by the law of total probability and the fact that the conditioning event $\{D = 1, Z = 1\}$ is identical to $\{D_1 = 1, Z = 1\}$. To obtain the third line, we apply RTA to $\Pr(T = t | D_1 = 1, Z = 1)$, $\Pr(D_1 = 1 | Z = 1)$, and $\Pr(Y_1 \in B | D_1 = 1, Z = 1, T = t)$. Note that the type indicator T gives a finer partition of the sample space than D_1, so we obtain $\Pr(Y_1 \in B | D_1 = 1, T = t) = \Pr(Y_1 \in B | T = t)$ and $\Pr(T = t | D_1 = 1, Z = 1) = \Pr(T = t | T \in \{c, a\})$.

The similar operation to $Q(B, 1)$ yields

$$Q(B,1) = \Pr(Y_1 \in B | T = a) \pi_a + \Pr(Y_1 \in B | T = d) \pi_d. \qquad (3.1.1.7)$$

Under MPR, there do not exist defiers in the population, i.e., $\pi_d = 0$. If we take the difference between (3.1.1.6) and (3.1.1.7), we obtain

$$P(B,1) - Q(B,1) = \Pr(Y_1 \in B | T = c) \pi_c \geq 0.$$

This proves the first inequality of the proposition. The second inequality of the proposition is obtained in an analogous way and we omit its derivation for brevity.

For a proof of converse statement, let a data generating process P and Q satisfying the inequalities (3.2.0.1) be given. Let $p(y, d)$ and $q(y, d)$ be the densities (with respect to a dominating measure μ) of P and Q on $\mathcal{Y} \times \{d\}$. It suffices to show that we can construct a joint distribution of (Y_1, Y_0, T, Z) that is compatible with P and Q and satisfies RTA and MPR. Since the marginal distribution of Z is not important for the analysis, we focus on constructing the conditional distribution of (Y_1, Y_0, T)

given Z. Let us consider the nonnegative functions $h_{Y_d,t}(y)$, $d = 1, 0$, $t \in \{c, n, a, d\}$,

$$\begin{aligned}
h_{Y_1,c}(y) &= p(y,1) - q(y,1), \\
h_{Y_1,n}(y) &= \gamma_{Y_1}(y), \\
h_{Y_1,a}(y) &= q(y,1), \\
h_{Y_1,d}(y) &= 0, \\
h_{Y_0,c}(y) &= q(y,0) - p(y,0), \\
h_{Y_0,n}(y) &= p(y,0), \\
h_{Y_0,a}(y) &= \gamma_{Y_0}(y), \\
h_{Y_0,d}(y) &= 0.
\end{aligned}$$

where $\gamma_{Y_1}(y)$ and $\gamma_{Y_0}(y)$ are arbitrary nonnegative functions satisfying $\int_{\mathcal{Y}} \gamma_{Y_1}(y)d\mu = P(\mathcal{Y}, 0)$ and $\int_{\mathcal{Y}} \gamma_{Y_0}(y)d\mu = Q(\mathcal{Y}, 1)$. We construct a conditional probability law of (Y_1, Y_0, T) given Z as, for an arbitrary Borel sets B_1 and B_0 in $\mathcal{Y}$,

$$\begin{aligned}
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = c|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = c|Z = 0) \\
\equiv\ &\begin{cases} \frac{\int_{B_1} h_{Y_1,c}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_1,c}(y)d\mu} \times \frac{\int_{B_0} h_{Y_0,c}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_0,c}(y)d\mu} \times [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] & \text{if } [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] > 0 \\ 0 & \text{if } [P(\mathcal{Y},1) - Q(\mathcal{Y},1)] = 0 \end{cases} \\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = n|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = n|Z = 0) \\
\equiv\ &\begin{cases} \frac{\int_{B_1} h_{Y_1,n}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_1,n}(y)d\mu} \times \frac{\int_{B_0} h_{Y_0,n}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_0,n}(y)d\mu} \times P(\mathcal{Y},0) & \text{if } P(\mathcal{Y},0) > 0 \\ 0 & \text{if } P(\mathcal{Y},0) = 0 \end{cases} \\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = a|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = a|Z = 0) \\
\equiv\ &\begin{cases} \frac{\int_{B_1} h_{Y_1,a}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_1,a}(y)d\mu} \times \frac{\int_{B_0} h_{Y_0,a}(y)d\mu}{\int_{\mathcal{Y}} h_{Y_0,a}(y)d\mu} \times Q(\mathcal{Y},1) & \text{if } Q(\mathcal{Y},1) > 0 \\ 0 & \text{if } Q(\mathcal{Y},1) = 0 \end{cases} \\
&\Pr(Y_1 \in B_1, Y_0 \in B_0, T = d|Z = 1) = \Pr(Y_1 \in B_1, Y_0 \in B_0, T = d|Z = 0) \\
\equiv\ &0
\end{aligned}$$

Note that this is a valid probability measure since it is nonnegative and satisfies

$$\sum_{t \in \{c,n,a,d\}} \Pr(Y_1 \in \mathcal{Y}, Y_0 \in \mathcal{Y}, T = t|Z = z) = 1, \quad z = 1, 0.$$

Furthermore, the proposed probability distribution satisfies RTA and MPR by construction and it

is consistent with the given data generating process, i.e.,

$$\begin{aligned}
\Pr(Y_{obs} \in B, D = 1 | Z = 1) &= \Pr(Y_1 \in B, T = c | Z = 1) + \Pr(Y_1 \in B, T = a | Z = 1) \\
&= \int_B [h_{Y_1,c}(y) + h_{Y_1,a}(y)] d\mu = P(B, 1), \\
\Pr(Y_{obs} \in B, D = 0 | Z = 1) &= \Pr(Y_0 \in B, T = n | Z = 1) + \Pr(Y_0 \in B, T = d | Z = 1) \\
&= P(B, 0) \\
\Pr(Y_{obs} \in B, D = 1 | Z = 0) &= \Pr(Y_1 \in B, T = a | Z = 0) + \Pr(Y_1 \in B, T = d | Z = 0) \\
&= Q(B, 1) \\
\Pr(Y_{obs} \in B, D = 0 | Z = 0) &= \Pr(Y_0 \in B, T = n | Z = 0) + \Pr(Y_0 \in B, T = c | Z = 0) \\
&= Q(B, 0)
\end{aligned}$$

This completes the proof. ■

3.A.2 Proof of Proposition 3.3.1

Throughout the proof, it is assumed that the probability law of a binary instrument Z is i.i.d Bernoulli with parameter $\lambda \in (\epsilon, 1 - \epsilon)$ for some $\epsilon > 0$.

i)

Step 1: Derive the asymptotic distribution of the test statistic T_N under the null $P = Q$.

Define P_m and Q_n as the empirical probability measure of (Y, D) conditional on $Z = 1$ and $Z = 0$ respectively,

$$P_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{(Y^1_{obs,i}, D^1_i)}, \qquad Q_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{(Y^0_{obs,j}, D^0_j)},$$

where $\delta_{(y,d)}$ represents a unit mass measure on $(Y_{obs}, D) = (y, d)$.

Given $\mathbb{V}$ a VC-class of subsets in $\mathbb{R}$, we define the class of indicator functions on $\mathbb{R} \times \{1, 0\}$, $\mathcal{F}_1$ and $\mathcal{F}_0$,

$$\mathcal{F}_1 = \{1\{(V, 1)\}; V \in \mathbb{V}\}, \qquad \mathcal{F}_0 = \{1\{(V, 0)\}; V \in \mathbb{V}\}$$

where the first coordinate of the indicator function corresponds to a subset $V \subset \mathbb{R}$ and the second coordinate corresponds to the participation indicator D. Following to the notation in van der Vaart and Wellner (1996), for a function $f : \mathbb{R} \times \{1, 0\} \to \mathbb{R}$, Pf stands for the expectation of f with respect to P, $Pf = \int f dP$. Note that $\mathcal{F}_1$ and $\mathcal{F}_0$ are VC-class of functions on $\mathbb{R} \times \{1, 0\}$ since the collection of subsets $\mathbb{V}$ are assumed to be a VC-class.

Consider stochastic processes $G_{1,N} : \mathcal{F} \to \mathbb{R}$ where $\mathcal{F}$ is a class of functions on $\mathbb{R} \times \{1, 0\}$,

$$\begin{aligned} G_{1,N}(\cdot) &= \left(\frac{mn}{N}\right)^{1/2}(Q_n - P_m) \\ &= \left(\frac{m}{N}\right)^{1/2}\sqrt{n}(Q_n - Q) - \left(\frac{n}{N}\right)^{1/2}\sqrt{m}(P_m - P) \\ &\quad + \left(\frac{mn}{N}\right)^{1/2}(Q - P). \end{aligned} \tag{3.1.2.8}$$

Given the above Donsker class of functions $\mathcal{F}_1$, we apply the Donsker theorem (theorem 3.5.1 in van der Vaart and Wellner (1996)) to get the weak convergence of $\sqrt{n}(Q_n - Q)(\cdot)$ and $\sqrt{m}(P_m - P)(\cdot)$ to the brownian bridges on $\mathcal{F}_1$,

$$\begin{aligned} \sqrt{n}(Q_n - Q) &\rightsquigarrow G_Q \quad \text{in} \quad l^\infty(\mathcal{F}_1) \\ \sqrt{m}(P_m - P) &\rightsquigarrow G_P \quad \text{in} \quad l^\infty(\mathcal{F}_1) \end{aligned}$$

where "$\rightsquigarrow$" notates weak convergence, G_P represents the P-brownian bridge, G_Q represents the Q-brownian bridge, and $l^\infty(\mathcal{F})$ denotes the space of l^∞ functions which map from $\mathcal{F}$ into $\mathbb{R}$. Under the null $P = Q$, since $m/N \to \lambda$ almost surely, $G_{1,N}$ converges weakly to a sum of two independent P-brownian bridges G_P and G'_P.

$$G_{1,N} \rightsquigarrow \lambda^{1/2}G_P - (1-\lambda)^{1/2}G'_P.$$

Note that the probability law of the process $\lambda^{1/2}G_P - (1-\lambda)^{1/2}G'_P$ is identical to the P-brownian bridge G_P. Hence, we have $G_{1,N} \rightsquigarrow G_P$ in $l^\infty(\mathcal{F}_1)$. Analogously, for stochastic processes $G_{0,N} : \mathcal{F} \to \mathbb{R}$

$$\begin{aligned} G_{0,N} &= \left(\frac{mn}{N}\right)^{1/2}(P_m - Q_n) \\ &= \left(\frac{n}{N}\right)^{1/2}\sqrt{m}(P_m - P) - \left(\frac{m}{N}\right)^{1/2}\sqrt{n}(Q_n - Q) \\ &\quad + \left(\frac{mn}{N}\right)^{1/2}(P - Q). \end{aligned} \tag{3.1.2.9}$$

we obtain $G_{0,N} \rightsquigarrow G_P$ in $l^\infty(\mathcal{F}_0)$.

Notice that the test statistic is written as

$$T_N = \max\left\{\sup_{f\in\mathcal{F}_1} G_{1,N}f, \sup_{f\in\mathcal{F}_0} G_{0,N}f\right\}.$$

Let $\mathcal{F}^* = \mathcal{F}_1 \cup \mathcal{F}_0$. Note that $\mathcal{F}^*$ is also a Donsker class. For $X \in l^\infty(\mathcal{F}^*)$ with $l^\infty(\mathcal{F}^*)$ equipped with the sup metric, the functional $\sup_{f\in\mathcal{F}_1} Xf$ is continuous with respect to X, since for $X_1, X_2 \in l^\infty(\mathcal{F}^*)$, $|\sup_{f\in\mathcal{F}_1}(X_1 - X_2)f| \le \sup_{f\in\mathcal{F}^*}|(X_1 - X_2)f| \lesssim \|X_1 - X_2\|$ holds. Since the max operator is clearly continuous, the continuous mapping theorem for stochastic processes (see, e.g., Pollard

(1984)) implies

$$T_N \rightsquigarrow T = \max\left\{\sup_{f\in\mathcal{F}_1} G_P f, \sup_{f\in\mathcal{F}_0} G_P f\right\} = \sup_{f\in\mathcal{F}^*} G_P f. \tag{3.1.2.10}$$

This is the limiting probability law of T_N under the null $P = Q$.

Step 2: Prove the asymptotic consistency of the distribution of the bootstrap statistic.

Let us define the bootstrap empirical measure

$$P_m^* = \frac{1}{m}\sum_{i=1}^{m} \delta_{(Y^*_{obs,i}, D_i^*)}, \qquad Q_n^* = \frac{1}{n}\sum_{j=1}^{n} \delta_{(Y^*_{obs,j}, D_j^*)}.$$

where $(Y^*_{obs,i}, D_i^*)$, $i = 1, \ldots, m$, and $(Y^*_{obs,j}, D_j^*)$, $j = 1, \ldots, n$, are drawn randomly from the pooled empirical measure

$$H_N = \frac{m}{N} P_m + \frac{n}{N} Q_n.$$

The bootstrap test statistic is expressed as

$$T_N^* = \max\left\{\sup_{f\in\mathcal{F}_1} G^*_{1,N} f, \sup_{f\in\mathcal{F}_0} G^*_{0,N} f\right\}$$

where $G^*_{1,N} = \left(\frac{mn}{N}\right)^{1/2}(Q_n^* - P_m^*)$ and $G^*_{0,N} = \left(\frac{mn}{N}\right)^{1/2}(P_m^* - Q_n^*)$. The bootstrap consistency is proved if the distribution of T_N^* converges weakly to the one obtained in (3.1.2.10) under the null $P = Q$ for almost every sampling sequences of $\{(Y^1_{obs,i}, D_i^1)\}$ and $\{(Y^0_{obs,j}, D_j^0)\}$.

Let $H = \lambda P + (1-\lambda)Q$. By theorem 3.7.7 in van der Vaart and Wellner (1996), $\sqrt{m}(P_m^* - H_N) \rightsquigarrow G_H$ and $\sqrt{n}(Q_n^* - H_N) \rightsquigarrow G_H$ hold with probability one in terms of the randomness of the sequences, $\{(Y^1_{obs,i}, D_i^1)\}$ and $\{(Y^0_{obs,j}, D_j^0)\}$.

Thus, by the similar argument to Step 1, $G^*_{1,N}$ and $G^*_{0,N}$ weakly converge to the H-brownian bridge, i.e.,

$$\begin{aligned} G^*_{1,N} &= \left(\frac{mn}{N}\right)^{1/2}(Q_n^* - P_m^*) \\ &= \left(\frac{m}{N}\right)^{1/2}\sqrt{n}(Q_n^* - H_N) - \left(\frac{n}{N}\right)^{1/2}\sqrt{m}(P_m^* - H_N) \\ &\rightsquigarrow \lambda^{1/2} G_H - (1-\lambda)^{1/2} G'_H = G_H \end{aligned}$$

and $G^*_{0,N} \rightsquigarrow G_H$ for almost every sequence of $\{(Y^1_{obs,i}, D_i^1)\}$ and $\{(Y^0_{obs,j}, D_j^0)\}$. Therefore, by the continuous mapping theorem,

$$T_N^* \rightsquigarrow \sup_{f\in\mathcal{F}^*} G_H f. \tag{3.1.2.11}$$

Note that, under the null, $H = P$ holds, and therefore the obtained H-brownian bridge is in fact P-brownian bridge. Hence, $T_N^* \rightsquigarrow T$ holds. This implies that the asymptotic distribution of T_N^* coincides with that of T_N under the null for almost every sequence of $\{(Y^1_{obs,i}, D_i^1)\}$ and $\{(Y^0_{obs,j}, D_j^0)\}$.

Step 3: Prove the asymptotic consistency of the rejection probability based on the bootstrap critical value $\hat{c}_{boot}(1-\alpha)$.

Let $J_N(\cdot, H_N)$ be the cdf of the bootstrap statistic T_N^* (conditional on H_N). The bootstrap estimates of the critical value is the $(1-\alpha)$-th quantile of $J_N(\cdot, H_N)$, that is,

$$\hat{c}_{boot}(1-\alpha) \equiv \inf\left\{c : \mathrm{Prob}_{H_N}(T_N^* > c) \leq \alpha\right\}.$$

Let $J(\cdot, H)$ be the cdf of T under the null $P = Q(= H)$ and denote its $(1-\alpha)$-th quantile by $c(1-\alpha)$. Since $J_N(\cdot, H_N)$ converges weakly to $J(\cdot, H)$, $\hat{c}_{boot}(1-\alpha)$ converges to the $c(1-\alpha)$ if $J(\cdot, H)$ is continuous and strictly increasing at its $(1-\alpha)$-th quantile (see, e.g., Lemma 1.2.1. in Politis, Romano, and Wolf (1999)).

The absolute continuity of $J(\cdot, H)$ follows by the absolute continuity theorem for the convex functional of the Gaussian processes (Theorem 11.1 of Davydov, Lifshits, and Smorodina (1998)). Note that the test statistic is a convex functional of $l^\infty(\mathcal{F}^*)$, and for some $f \in \mathcal{F}^*$ with nondegenerate $G_P f$, it holds $\Pr(T \leq 0) \leq \Pr(Gf \leq 0) = 1/2$. Therefore, $J(t, H)$ is absolutely continuous for every $t > 0$. Then, the absolute continuity theorem guarantees that, for $\alpha \in (0, 1/2)$, $J(t, H)$ is absolutely continuous at $c(1-\alpha)$. Thus, $\hat{c}_{boot}(1-\alpha) \to c(1-\alpha)$ almost surely in terms of the randomness of H_N.

Finally, by the Slutsky's Theorem, it follows

$$Prob_{P=Q=H}(T_N > \hat{c}_{boot}(1-\alpha)) \to 1 - J(c(1-\alpha), H) = \alpha.$$

ii)

To examine power of the test against a fixed alternative, consider P and Q such that $(Q-P)f > 0$ for some $f \in \mathcal{F}_1$. Then, the last term in (3.1.2.8) diverges to positive infinity at these f. Since the Brownian bridge processes as the limiting process of $\sqrt{m}(P_m - P)$ and $\sqrt{n}(Q_n - Q)$ are bounded with probability one, $\sup_{f \in \mathcal{F}_1} G_{1,N} f \to \infty$ with probability one. This implies $T_N \to \infty$ with probability one.

On the other hand, the bootstrap critical value are bounded almost surely (with respect to the original sampling sequence) because T_N^* weakly converges to $\sup_{f \in \mathcal{F}^*} G_H f$ with $H = \lambda P + (1-\lambda)Q$. Therefore,

$$Prob_{P=Q=H}(T_N > \hat{c}_{boot}(1-\alpha)) \to 1$$

as $N \to \infty$.

■

Bibliography

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284-292.

[2] 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, 91-117.

[3] Angrist, J. D. (1991): "The Draft Lottery and Voluntary Enlistment in the Vietnam Era," *Journal of the American Statistical Association,* 86, 584-595

[4] 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, 444-455.

[5] Campbell, D. T. (1969): "Reforms as Experiments," *American Psychologist*, 24 (4), 409-429.

[6] Card, D. (1993): "Using Geographical Variation in College Proximity to Estimate the Returns to Schooling", National Bureau of Economic Research Working Paper No. 4, 483.

[7] Davydov, Y.A., Lifshits, M.A., and Smorodina, N.V. (1998): *Local Properties of Distributions of Stochastic Functionals.* Providence: American Mathematical Society.

[8] Dudley, R. M. (1999): *Uniform Central Limit Theorem.* Cambridge University Press.

[9] Hahn, J., Todd, P. E., and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica* 69 (1), 201-209.

[10] Heckman, J. J. and E. Vytlacil (1999): "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

[11] Heckman, J. J. and E. Vytlacil (2001): "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,* 1-46. Cambridge University Press, Cambridge UK.

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

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

[14] Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica,* 75, 83-119.

[15] Romano, J. P. (1988): "A Bootstrap Revival of Some Nonparametric Distance Tests." *Journal of American Statistical Association*, 83, 698-708.

[16] Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling.* New York: Springer.

[17] van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics,* New York: Springer.