Testing for Instrument Independence in the Selection Model

Toru Kitagawa*†
UCL and CeMMAP

September, 2009

Abstract

We develop a specification test for the independent instrument assumption in the sample selection model. We test the emptiness of the identification region of Manski (2003): the set of outcome distributions that are compatible with data and the restriction of statistical independence between the instrument and outcome. The size of the identification region is characterized by a scalar parameter, the integrated envelope, and in particular the identification region is empty if and only if the integrated envelope exceeds one. Since the empty identification region implies a violation of the exclusion restriction, we obtain a nonparametric specification test for the instrument exclusion restriction by developing a testing procedure for whether the integrated envelope exceeds one. This test procedure has a non-pivotal asymptotic distribution and it is well-known that in this case the standard nonparametric bootstrap is not valid to obtain the critical values. We therefore develop a modified bootstrap procedure and show its validity. Monte Carlo simulations examine the finite sample performance of this bootstrap procedure. We use the procedure to test the independence of the instrument used by Blundell et al. (2003).

Keywords: Partial Identification, instrumental variable, sample selection, missing data, specification test, bootstrap.

JEL Classification: C12, C15, C24.

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

†I am deeply indebted to Frank Kleibergen for thoughtfuI discussions and constant encouragement. I have also benefitted from comments from Stefan Hoderlein, Joe Hogan, Guido Imbens, Sung Jae Jun, Tony Lancaster, Tao Liu, Leandro Magnusson, Sophocles Mavroeidis, and Blaise Melly as well as the seminar participants at the Brown Center for Statistical Science, Cornell, Iowa, Ohio State, Penn, Rochester, Tesas Austin, UCL, and Yale. I also thank the UK Economic and Social Data Service and Costas Meghir for providing me with the data. Financial support from the Brown University Department of Economics Merit Dissertation Fellowship and CeMMAP is gratefully acknowledged.

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 papers, Manski (1989, 1990, 1994, 2003) analyzes the selection model 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 Y. The selection problem that this paper considers is the missing data problem with an instrument: the outcome Y is observed if the selection indicator D is one while it is missing if D is zero, and the researcher has a random sample of $(Y \cdot D, D, Z)$. 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 specifed to be independent of one's potential wage while it can affect one's employment status. For example, a list of instruments that has been used in this potential wage example includes the number of kids, marital status, a measure of out-of-work income, etc. Our object of interest is f_Y, the population distribution of Y, and the identification of f_Y leads to identification of location parameters such as the mean or quantiles of Y. In the potential wage example, this problem arises when the researcher is interested in estimating the wage gaps between male and female, black and white, or skilled and unskilled. Although an instrument Z combined with the exclusion restriction plays a crucial role in identifying f_Y in the (semi)nonparametric sample selection model,[2] no testing procedures have been proposed for the instrument exclusion restriction.

This paper provides 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, this paper first analyzes identification of f_Y without imposing point-identifying restrictions for f_Y. That is, our object of interest is the *identification region* for f_Y: the set of outcome distributions that are compatible with the empirical evidence and the model restrictions. 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. We provide a closed-form representation of the identification region

[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.

[2] The point-identificaton of f_Y is achieved if an available instrument satisfies the exclusion restriction and the selection probability $\Pr(D = 1|Z = z)$ attains one for some z. This is the identification at infinity argument (Andrews and Schafgans (1998), Chamberlain (1986), and Heckman (1990)) based on an extrapolation by the instrument exclusion restriction. See, e.g., Mulligan and Rubinstein (2008) for an application of the identification at infinity strategy to the estimation of wage gaps between genders.

that extend his result to a wider range of settings where Y can be continuous.

The main contribution of this paper is the development of a specification test. An empty identification region implies a misspecification of the exclusion restriction, so 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.[3] 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. 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, as Manski (2003) noticed, the identification region is empty if and only if the integrated envelope exceeds one. Therefore, a nonparametric specification test for the instrument exclusion restriction is obtained by developping 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 Pearl (1994b) and Manski (2003), 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.

The third 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 and hence our procedure can screen out more violations of the instrument exclusion. Second, the asymptotic

[3] 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 analyzes the confidence sets in the moment inequality model contains Andrews, Berry and Jia (2004), Andrews and Guggenberger (2008), Andrews and Jia (2008), Andrews and Soares (2009), Bugni (2009), Canay (2009), Chernozhukov, Hong, and Tamer (2007), Guggenberger, Hahn, and Kim (2008), Imbens and Manski (2004), Pakes, Porter, Ho, and Ishii (2006), Romano and Shaikh (2008, 2009), and Rosen (2008).

validity of their bootstrap procedure is not formally investigated and its asymptotic property is not known. Our bootstrap algorithm 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.

In addition to the inference on the emptiness of the identification region, this paper provides contributions to the identification aspect of the selecton model. We use the expression for the identification region 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. This implies that a further identification gain from the joint independence restriction, which is known to exist in the counterfactual causal model with an instrument (Balke and Pearl (1997)), does *not* exist in the selection model with a single outcome. 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.

The remainder of the paper is organized as follows. Section 2 introduces the basic notation and provides the identification region of f_Y. It also provides a refutability result of instrument independence based on the integrated envelope. Section 3 develops the estimator for the integrated envelope and derives its asymptotic distribution. Based on this asymptotic distribution, Section 4 formalizes the test procedure by developing an asymptotically valid bootstrap algorithm. We also demonstrate the validity of subsampling. Section 5 provides simulation results and compares the finite sample performance of the bootstrap with subsampling. For simplicity of exposition, our analytical framework is limited to the case with a binary instrument up to Section 5. In Section 6, we extend the framework to the case with a multi-valued discrete instrument. Using this extended framework, Section 7 tests whether the out-of-work income constructed in Blundell et al. (2003) is independent of the potential wage. Section 8 concludes. Proofs are provided in Appendix A.

2 The identification region of the outcome distribution

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 Section 6 and Appendix E.

We do not introduce covariates X into our analysis. When the exclusion restriction of the instrument is specifed 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 $Q = (Q(\cdot), Q_{mis})$,

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

$P(\cdot)$ 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

[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.

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 (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 (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, and the bounds for $E(Y)$ under ER can be strictly narrower than the bounds for $E(Y)$ under MI (see Appendix C 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?).

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

$$\begin{aligned} f_Y(y) &= f_{Y|Z}(y|Z=1) = p(y) + f_{Y,D|Z}(y, D=0|Z=1), \\ f_Y(y) &= f_{Y|Z}(y|Z=0) = q(y) + f_{Y,D|Z}(y, D=0|Z=0). \end{aligned} \tag{2.1}$$

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 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 (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

$$IR_{f_Y}(P,Q) = \left\{ f_Y : \int_{\mathcal{Y}} f_Y(y) d\mu = 1,\ f_Y(y) \geq \underline{f}(y)\ \mu\text{-a.e.} \right\}. \tag{2.2}$$

Figure 1 provides a graphical illustration for the identification region.

Notice that $IR_{f_Y}(P,Q)$ 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 $IR_{f_Y}(P,Q)$ 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 selection model.[7]

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

[7] When Y is discrete, this proposition is reduced to Corollary 2.3 of Manski (2003).

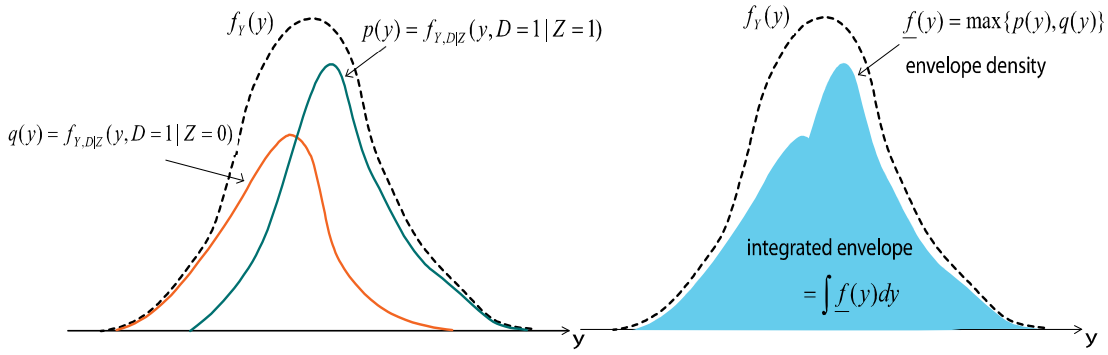

Figure 1: *Consider the case with a continuous Y and a binary Z. The dotted curve represents f_Y the probability density of the outcome Y. The identities (2.1) and the nonnegativity of the missing outcome densities require that the two densities $p(y)$ and $q(y)$ must lie below f_Y. This implies that 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 (shadow 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.*

Proposition 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*

$$\underline{f}(y) \equiv \max\{p(y), q(y)\}, \qquad \delta(P,Q) \equiv \int_{\mathcal{Y}} \underline{f}(y)d\mu. \tag{2.3}$$

(i) The identification region of f_Y under ER is given by

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

(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 $\delta = \delta(P,Q)$,

$$f_Y(y) = \delta\left(\underline{f}(y)/\delta\right) + (1-\delta)\gamma(y), \tag{2.4}$$

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 (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

$$\int_{(-\infty,y]} \underline{f}(t)d\mu \leq F_Y(y) \leq \int_{(-\infty,y]} \underline{f}(t)d\mu + 1 - \delta. \tag{2.5}$$

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

The tight bounds for the mean $E(Y)$ also follow from (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,

$$(1-\delta)y_l + \int_{\mathcal{Y}} y\underline{f}(y)d\mu \leq E(Y) \leq \int_{\mathcal{Y}} y\underline{f}(y)d\mu + (1-\delta)y_u. \tag{2.6}$$

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 C, 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.

2.3 Does selection equation help to identify f_Y?

The structural selection model formulates the selection mechanism as

$$D = I\{v(Z,U) \geq 0\}, \tag{2.7}$$

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?

An importance of this question can be motivated 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 (2.7), it is well known that 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.[8] In constrast to the counterfactual causal model, it has not been clarified whether or not the selection model with a single missing outcome can enjoy an 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[9]

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

and joint independence of Z and (Y, U) is equivalently stated as joint independence of Z and (Y, T) (Pearl (1994a)). Accordingly, the definition of the identification region under joint independence is defined as follows.

Definition 2.2 (the identification region under joint independence) *Given a data generating process P and Q, the identification region for f_Y under the joint independence restriction between Z and (Y, U) 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.*

Appendix D.1 provides a formal analysis on the construction of the identification region under the joint independence restriction between Z and (Y, U). The main result is stated in the next proposition.

Proposition 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 joint independence between Z and (Y, U).*

[8] Balke and Pearl (1997) derives the tight bounds for the average treatment effects $E(Y_1) - E(Y_0)$ under the joint independence restriction $(Y_1, Y_0, U) \perp Z$ for the binary outcome case. Kitagawa (2009) provides a closed-form expression of the identification region as well as the tight bounds of the average treatment effects for the continuous outcome case and shows that the joint independence restriction $(Y_1, Y_0, U) \perp Z$ can narrow the identification region for the distribution of (Y_1, Y_0) relative to the marginal independence restriction $(Y_1, Y_0) \perp Z$.

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

This proposition shows that a further identification gain from the joint independence restriction between Z and (Y,U), which is known to exist in the causal model with an instrument (Balke and Pearl (1997)), does *not* exist in the selection model with a single outcome. This redundancy of the joint independence restriction implies that the marginal independece of Z and Y 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,

$$v(Z,U) = \tilde{v}(Z) - U, \tag{2.8}$$

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 2.3 (the identification region under separable utility) *Given a data generating process P and Q, the identification region for f_Y under joint independence between Z and (Y,U) 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 the joint independence restriction of Z and (Y,T) with either $\Pr(T=c)=0$ or $\Pr(T=d)=0$.*

In Appendix D.2, 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 2.3 (the identification region under separable utility) *The identification region under joint independence between Z and (Y,D_1,D_0) and the specification of threshold crossing selection with an additive error is*

$$\begin{cases} IR_{f_Y}(P,Q) & \text{if } p(y) \geq q(y) \ \mu\text{-a.e. or } q(y) \geq p(y) \ \mu\text{-a.e.} \\ \emptyset & \text{otherwise.} \end{cases} \tag{2.9}$$

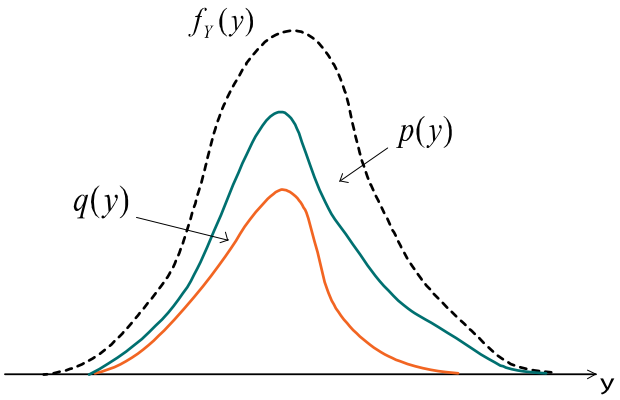

Figure 2: *If the instrument is jointly independent of Y and the unobserved heterogeneities in the latent utility, and threshold crossing selection with an additive error holds in the population, then we must observe that either $p(y)$ or $q(y)$ covers the other on the entire $\mathcal{Y}$, as drawn 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 C).*

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 $IR_{f_Y}(P, Q)$ coincide with the Manski's $E(Y)$ bounds under MI.

The empty identification region in (2.9) 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 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.[10]

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.

[10] Kitagawa (2009) 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.

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

Our identification analysis clarified that the emptiness of the identification region under ER is summarized by the integrated envelope $\delta(P,Q)$. We also showed that the joint independence restriction does not tighten $IR_{f_Y}(P,Q)$. These results imply that $\delta(P,Q)$ is the only relevant parameter for the purpose of refuting the instrument exogeneity. Hence, the rest of 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, $\delta(P,Q) \leq 1$. Since this null hypothesis is the necessary but not a sufficient condition of instrument independence, our test is interpreted as a test for a refutable hypothesis (Breusch (1986)). 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[11] since for every (P,Q)

[11] 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.

satisfying the null $\delta(P,Q) \leq 1$, we have

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

We decompose our theoretical development into two parts. First, we develop an estimator of $\delta(P,Q)$ and derive the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P,Q))$ (Section 3). Second, we focus on how to consistently estimate quantiles of the asymptotic distribution $J(\cdot; P, Q, \lambda)$ (Section 4).

3.1 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.

3.1.1 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

$$\delta \equiv \max\{p_1, q_1\} + \max\{p_0, q_0\}. \tag{3.1}$$

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

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

where

$$\begin{aligned} \Sigma_{P,\lambda} &= \lambda^{-1}\begin{pmatrix} p_1(1-p_1) & -p_1p_0 \\ -p_1p_0 & p_0(1-p_0) \end{pmatrix} \text{ and} \\ \Sigma_{Q,\lambda} &= (1-\lambda)^{-1}\begin{pmatrix} q_1(1-q_1) & -q_1q_0 \\ -q_1q_0 & q_0(1-q_0) \end{pmatrix}. \end{aligned}$$

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 (3.1), 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

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

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

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \max_{\{j:\delta_j = \delta\}} \{G_j\}. \tag{3.2}$$

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 $\sqrt{N}(\hat{\delta} - \delta)$.

3.1.2 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

$$\begin{aligned}
\sqrt{N}(\hat{\delta}^* - \hat{\delta}) &= \sqrt{N}(\max\{\hat{p}_1^*, \hat{q}_1^*\} + \max\{\hat{p}_0^*, \hat{q}_0^*\}) - \sqrt{N}(\max\{\hat{p}_1, \hat{q}_1\} + \max\{\hat{p}_0, \hat{q}_0\}) \\
&= \underbrace{\max\left\{\sqrt{N}(\hat{p}_1^* - \hat{q}_1^*), 0\right\} - \max\left\{\sqrt{N}(\hat{p}_1 - \hat{q}_1), 0\right\}}_{(i)} \\
&\quad + \underbrace{\max\left\{\sqrt{N}(\hat{q}_0^* - \hat{p}_0^*), 0\right\} - \max\left\{\sqrt{N}(\hat{q}_0 - \hat{p}_0), 0\right\}}_{(ii)} \\
&\quad + \underbrace{\sqrt{N}(\hat{q}_1^* - \hat{q}_1) + \sqrt{N}(\hat{p}_0^* - \hat{p}_0)}_{(iii)}. \qquad (3.3)
\end{aligned}$$

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 $\omega \in \Omega$,

$$\begin{aligned}
\sqrt{N}\begin{pmatrix} \hat{p}_1^* - \hat{p}_1(\omega) \\ \hat{p}_0^* - \hat{p}_0(\omega) \\ \hat{q}_1^* - \hat{q}_1(\omega) \\ \hat{q}_0^* - \hat{q}_0(\omega) \end{pmatrix} &\rightsquigarrow \begin{pmatrix} X_1 \\ X_0 \\ W_1 \\ W_0 \end{pmatrix}, \qquad (3.4) \\
\hat{q}_0(\omega) - \hat{p}_0(\omega) &\to q_0 - p_0 < 0.
\end{aligned}$$

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 (3.3) is obtained as

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

and the term (iii) weakly converges to $W_1 + X_0$ by (3.4). To sum up, we have for large N

$$\sqrt{N}(\hat{\delta}^* - \hat{\delta}(\omega)) \leq \max\{X_1 - c, W_1\} + X_0 \leq \max\{X_1, W_1\} + X_0, \tag{3.5}$$

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 (3.5) 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.

3.1.3 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[12] 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,

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

[12] 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.

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

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

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

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

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 $\sqrt{N}(\hat{\delta} - \delta)$.

3.2 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 6 and Appendix E).

3.2.1 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

$$\delta(P,Q) = \max_{V\in\mathcal{B}(\{1,0\})}\{P(V)+Q(V^c)\}, \tag{3.6}$$

where $\mathcal{B}(\{1,0\})$ is the power set of $\{1,0\}$, $\mathcal{B}(\{1,0\})$ $=\{\{1,0\},\{1\},\{0\},\emptyset\}$, and $V^c=\{1,0\}\backslash V$, 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 $\delta(\cdot):\mathcal{B}(\mathcal{Y})\to\mathbb{R}_+$,

$$\delta(V) = P(V)+Q(V^c), \tag{3.7}$$

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 $\mathcal{B}(\mathcal{Y})$.[13] Hence, an alternative expression for the integrated envelope $\delta(P,Q)$ is the supremum of $\delta(\cdot)$ over $\mathcal{B}(\mathcal{Y})$,

$$\delta(P,Q) = \sup_{V\in\mathcal{B}(\mathcal{Y})}\{\delta(V)\}. \tag{3.8}$$

[13] 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.

We can see this expression of $\delta(P,Q)$ as a direct analogue of (3.6) 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 (3.7) with the empirical distributions $P_m(\cdot)$ and $Q_n(\cdot)$,

$$\hat{\delta}(V) = P_m(V) + Q_n(V^c). \tag{3.9}$$

Analogous to the construction of the integrated envelope in (3.8), we propose an estimator of $\delta(P,Q)$ by maximizing $\hat{\delta}(\cdot)$ over a class of subsets $\mathbb{V} \subset \mathcal{B}(\mathcal{Y})$,[14]

$$\hat{\delta} \equiv \sup_{V \in \mathbb{V}} \{\hat{\delta}(V)\}. \tag{3.10}$$

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.

3.2.2 VC-class

When Y is discrete, $\mathbb{V}$ is specified as the power set of $\mathcal{Y}$ as in the binary Y case (3.6). 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 3, 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.

[14] 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)).

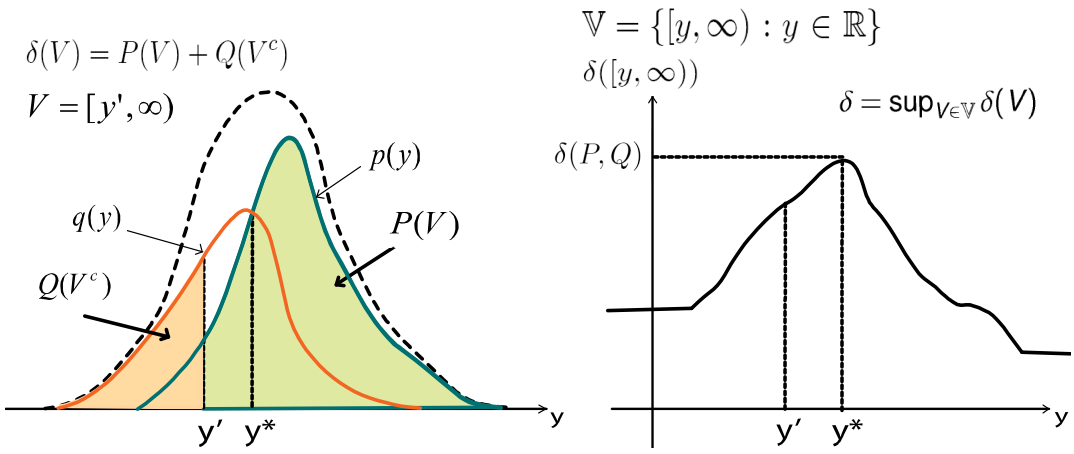

Figure 3: *Let Y be a continuous outcome on $\mathbb{R}$. In order to draw $\delta(\cdot)$ in two dimensions, we plot $\delta(\cdot)$ with respect to the collection of right unbounded intervals $\mathbb{V} = \{[y, \infty) : y \in \mathbb{R}\}$. As the left-hand side figure shows, $P(V)$ corresponds to 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 $\delta([y, \infty))$.*

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

$$\mathbb{V}_{half} = \{\emptyset, \mathbb{R}\} \cup \{(-\infty, y] : y \in \mathbb{R}\} \cup \{[y, \infty) : y \in \mathbb{R}\}. \tag{3.11}$$

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 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 3.

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

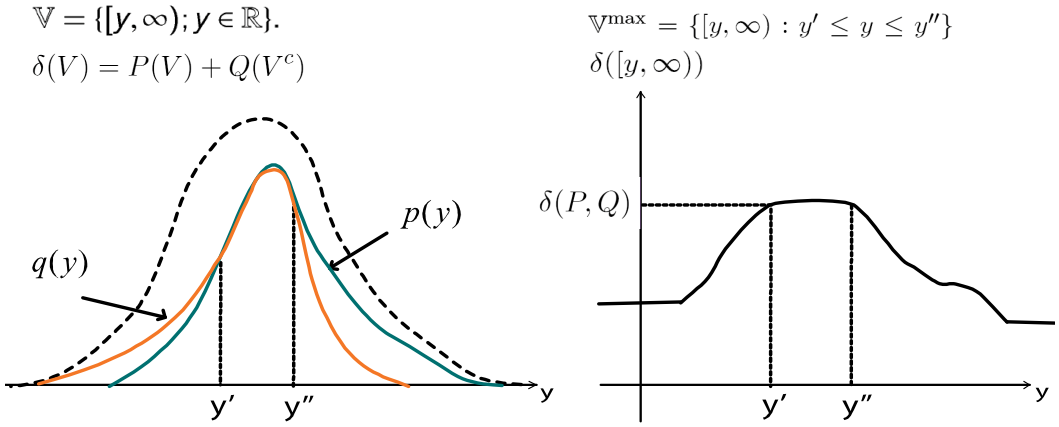

Figure 4: *$p(y)$ and $q(y)$ are tied over $[y', y'']$. Given $\mathbb{V}$ as the collection of right unbounded intervals, $\delta([y, \infty))$ is constant over $[y', y'']$ and there is a continuum of maximizers of $\delta(\cdot)$. Here, the maximizer subclass is given by $\mathbb{V}^{\max} = \{[y, \infty) : y \in [y', y'']\}$.*

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

$$\mathbb{V}_{hist}(h, L, \mathcal{Y}_0) = \left\{ \bigcup_{l \in I_j(L)} H_l(y_0, h) : y_0 \in \mathcal{Y}_0,\ j = 1, \ldots, 2^{L+1} \right\}. \tag{3.12}$$

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 4, 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 $\mu(V_1^{\max} \triangle V_2^{\max}) > 0$, it can be shown that $p(y) = q(y)$ on $V_1^{\max} \triangle V_2^{\max}$, and therefore P and Q are tied on the set with positive measure $V_1^{\max} \triangle V_2^{\max}$.

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.

3.2.3 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

$$\sqrt{N}(\hat{\delta} - \delta(P, Q)) = \sup_{V \in \mathbb{V}} \left\{ \sqrt{N}(\hat{\delta}(V) - \delta(V)) + \sqrt{N}(\delta(V) - \delta(P, Q)) \right\}. \tag{3.13}$$

The first term in the supremum of (3.13) 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 (3.13) 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 3.1 (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,

$$\sqrt{N}(\hat{\delta} - \delta(P,Q)) \rightsquigarrow \sup_{V \in \mathbb{V}^{\max}} \{G(V)\}, \tag{3.14}$$

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 (3.14) 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 3.1 (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.

4 Implementation of resampling methods: bootstrap and subsampling validity

Given the expression of the asymptotic distribution (Proposition 3.1), 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 3.1 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)).

4.1 Resampling method I: a modified bootstrap

The asymptotic distribution given in Proposition 3.1 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*

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

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 estmate $\mathbb{V}^{\max}$ 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 A.2 in Appendix A). 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 5.

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 3.1 (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 4.1 (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* $\delta(P,Q) \leq 1$*,*

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

4.2 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 4.1. 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

$$\hat{c}^{sub}_{1-\alpha} = \inf\left\{ x : \frac{1}{N_m N_n} \sum_{k=1}^{N_m} \sum_{l=1}^{N_n} I\left\{ \sqrt{B}(\hat{\delta}^*_{k,l} - \hat{\delta}) \leq x \right\} \geq 1 - \alpha \right\}. \tag{4.1}$$

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 4.2 (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 (4.1) 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).

4.3 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 4.1 (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 4.3 (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,*

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

5 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 (3.11). 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 5).

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

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 3.1). 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 3.1, 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 4.2 validates subsampling for any data generating processes, the simulation results suggest that the subsampling is contaminated by the ties.

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

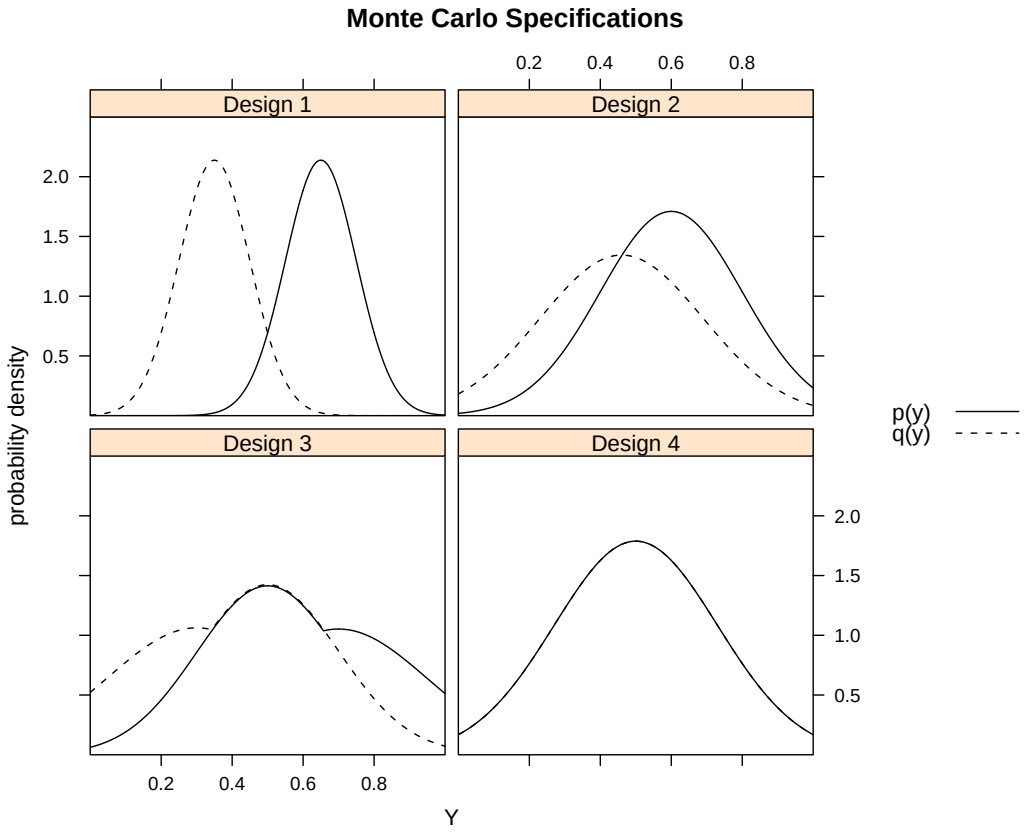

Figure 5: *There are no ties in Design 1 and Design 2. In Design 3, the two densities are partially tied. In Design 4, the two densities are identical.*

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).[15] 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 B.1 in Appendix B). 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 results exhibit unstable performance of their bootstrap. For instance, it is very conservative for Design 1 and Design 2, while it overrejects the null for Design 4.

6 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 appication of the next section. The main focus of this section is a generalization of the estimation and inference procedure for the integrated envelope rather than a generalization of the identification analysis of Section 2 (see E.1 for a generalization of the identification results).

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 $\mathcal{Y}$.

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

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

$$\delta(\mathbf{V}) = \sum_{k=1}^{K} P_k(V_k). \tag{6.1}$$

This can be seen as a generalization of (3.7) 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

$$\mathbb{V} = \left\{ \mathbf{V} = (V_1, \ldots, V_K) : V_1 \in \mathbb{V}_1,\ \ldots,\ V_K \in \mathbb{V}_K,\ \bigcup_{k=1}^{K} V_k = \mathcal{Y},\ \mu(V_k \cap V_{k'}) = 0\ \forall\ k \neq k' \right\}, \tag{6.2}$$

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 (3.8),

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

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 (6.1) with the empirical distribution P_{n_k} and restrict each $\mathbb{V}_k$ in (6.2) to a VC-class,

$$\hat{\delta} = \sup_{\mathbf{V} \in \mathbb{V}} \left\{ \hat{\delta}(\mathbf{V}) \right\},\ \ \text{where } \hat{\delta}(\mathbf{V}) = \sum_{k=1}^{K} P_{n_k}(V_k). \tag{6.3}$$

Under the assumptions analogous to (A1) through (A4) of Section 3.2.3, $\hat{\delta}$ 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 $\mathbb{V}$ (see Appendix E.2 for further details).

It is straightforward to accomodate the multi-valued discrete instrument to the bootstrap algorithm given in Section 4.1. 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 (6.2), and (6.1) 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 $k = 1, \ldots, K$.

7 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 (6.2) 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 F 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 6 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 6 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 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.[16] 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.

8 Concluding remarks

From the partial identification point of view, this paper analyzes the identification power of the restriction of instrument independence in the selection model. By focusing on the

[16] 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.

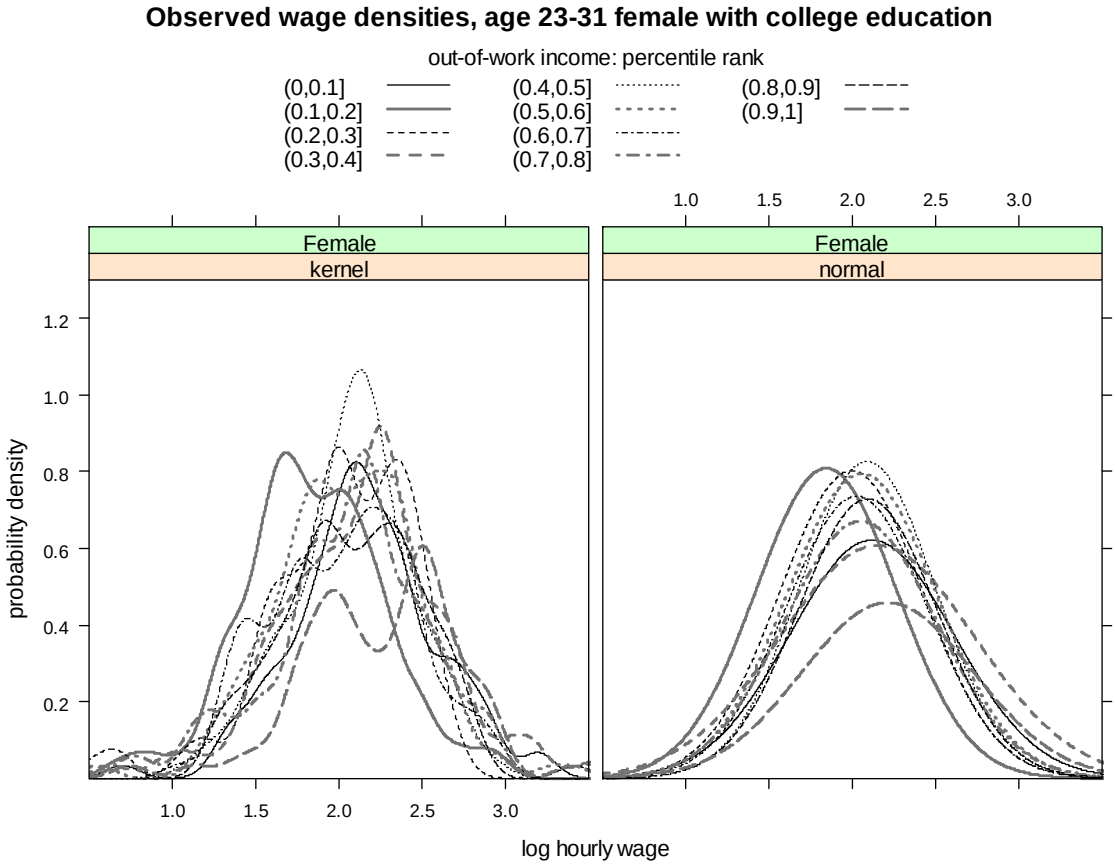

Figure 6: The *left-hand side figure presents the kernel density estimates for the observable densities* $p_{k,x}(y|x)$, *where we use the Gaussian kernel with bandwidth 0.1. The right-hand side figure gives the parametric (normal) estimates for the densities. In the Monte Carlo simulations to look for the optimal* η_N, *the estimated normal densities are specified as the data generating process.*

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.

envelope density, we provide the analytically tractable representation of the identification region for the outcome distribution under the restriction that the instrument is independent of the outcome. We focus on the integrated envelope, which is the key parameter for examining the emptiness of the identification region.

We show that the restriction of the instrument as jointly independent of the outcome and selection heterogeneities does not further tighten the identification region. In addition, we show that threshold crossing selection with an additive error constrains the data generating process but, does not tighten the identification region. These identification results imply that integrating the identifying information for f_Y using the envelope density always provides maximal identification for the outcome distribution under the exclusion restriction.

This paper is the first that analyzes estimation and inference for the integrated envelope. We propose the estimator for the integrated envelope and derive its asymptotic distribution. Using this asymptotic result, we develop the nonparametric specification test for instrument independence. Due to ties among the underlying probability densities, the estimator has a non-pivotal asymptotic distribution and therefore, the standard nonparametric bootstrap is not valid. To overcome this, we consider 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 about the potential wage distribution and the selection mechanism.

Appendix A: Lemmas and Proofs

Proof of Proposition 2.1. (i) Let P and Q be given by data and assume $\delta(P,Q) \leq 1$. Let $\mathcal{F}^*$ be the set of outcome distributions defined as $\mathcal{F}^* = \{f_Y : f_Y(y) \geq p(y) \text{ and } f_Y(y) \geq q(y), \mu\text{-a.e.}\}$. For an arbitrary $f_Y \in \mathcal{F}^*$, 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 $\Pr(Y \in B|Z = 1) = \Pr(Y \in B|Z = 0) = \int_B f_Y(y)d\mu$. This implies each $f_Y \in \mathcal{F}^*$ 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

$$\int_A [f_Y(y) - p(y)]d\mu < 0 \quad \text{or} \quad \int_A [f_Y(y) - q(y)]d\mu < 0. \tag{A.1}$$

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 (A.1) 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 $\mathcal{F}^*$ is the identification region under ER. (ii) is obvious. ■

Proof of Proposition 2.2. See Appendix D.1. ■

Proof of Proposition 2.3. See Appendix D.2. ■

Notation: For the rest of this appendix, we use the following notation. Our analysis is conditional on an infinite sequence of $\{Z_i : i = 1, 2 \ldots, \}$. For the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the sample space Ω consists of pairs of the i.i.d infinite sequences of $\{Y_{data,i}(\omega) : Z_i = 1\}$ and $\{Y_{data,i}(\omega) : Z_i = 0\}$. We abbreviate almost surely with respect to $\mathbb{P}$ by "a.s." and infinitely often by "i.o.". $\mathbb{V}$ always stands for a VC-class of subsets in $\mathcal{Y}$ equipped with the seminorm $d_R(V_1, V_2) = R(V_1 \triangle V_2)$ where R denotes a nonnegative measure on $\mathcal{B}(\mathcal{Y})$. In particular, we define $d_{P+Q}(V_1, V_2) = P(V_1 \triangle V_2) + Q(V_1 \triangle V_2)$. Let $(P_m - P)(V) \equiv P_m(V) - P(V)$ and $(Q_n - Q)(V) \equiv Q_n(V) - Q(V)$. We refer to the space of bounded functions on $\mathbb{V}$ as $l^\infty(\mathbb{V})$ where the metric is the sup metric $\|x\|_\infty = \sup_{V \in \mathbb{V}} |x(V)|$. Set indexed empirical processes which map $\mathbb{V} \to l^\infty(\mathbb{V})$ are denoted by $G_{P,m}(\cdot) \equiv \sqrt{m}(P_m - P)(\cdot)$ and $G_{Q,n}(\cdot) \equiv \sqrt{n}(Q_n - Q)(\cdot)$. For a nonmeasurable event A, $\mathbb{P}^*(A)$ indicates the outer probability (see van der Vaart and Wellner (1996) for the definition).

Proof of Proposition 3.1 (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 3.1 (ii) below.

Lemma 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 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 3.1 (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 $\eta > 0$,

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

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 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 $x_1,\, x_2 \in l^\infty(\mathbb{V})$,

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

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 3.1. Given $\mathbb{V}^{\max} = \{V^{\max}\}$, Proposition 3.1 (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 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 4.1 below.

Lemma 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*

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

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 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, $G_{P,m}(V) = \sqrt{m}(P_m - P)(V)$,

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

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

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

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)

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

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 3.1 (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 4.1. 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 3.1 (ii). Here, the random objects subject to the probability law of the original sampling sequence are indexed by ω.
By Lemma A.2, for sufficiently large N,

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

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

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

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

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

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 $\hat{c}^{boot}_{1-\alpha}$ 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 $\delta(P,Q)\leq 1$,

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

■

Proof of Proposition 4.2. 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

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

Let

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

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,

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

where $\mathbf{Y}^1_{data,b_m,k}$ represents the k-th subsample drawn from $\mathbf{Y}^1_{data,m}$, $\mathbf{Y}^0_{data,b_n,l}$ the l-th subsample drawn from $\mathbf{Y}^0_{data,}$, and $h(\mathbf{Y}^1_{data,b_m,k},\mathbf{Y}^0_{data,b_n,l})=1\left\{\sqrt{B}(\hat{\delta}^*_{k,l}-\delta(P,Q))\leq x\right\}$. Since for each k and l, $\mathbf{Y}^1_{data,b_m,k}$ and $\mathbf{Y}^0_{data,b_n,l}$ 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

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

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 4.1, $J_B(x; P, Q, \lambda_B) \to J(x; P, Q, \lambda)$ 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 4.3. 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 3.1, 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 3.1 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 $\epsilon = \tilde{\delta}(P,Q) - 1 > 0$,

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

■

Appendix B: 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 (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 $F_Y(y)$,

$$\begin{aligned} & \max\{P((-\infty, y]), Q((-\infty, y])\} \leq F_Y(y) \\ \leq \quad & \min\{P((-\infty, y]) + P_{mis}, Q((-\infty, y]) + Q_{mis}\}. \end{aligned} \tag{B.1}$$

These bounds, which we call the *naive cdf bounds* hereafter, are not necessarily the tightest possible under ER (Proposition B.1 below). 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 B.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 B.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 of Proposition B.1. (i) Fix $y \in \mathcal{Y}$. For the lower bound of the naive cdf bounds,

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

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

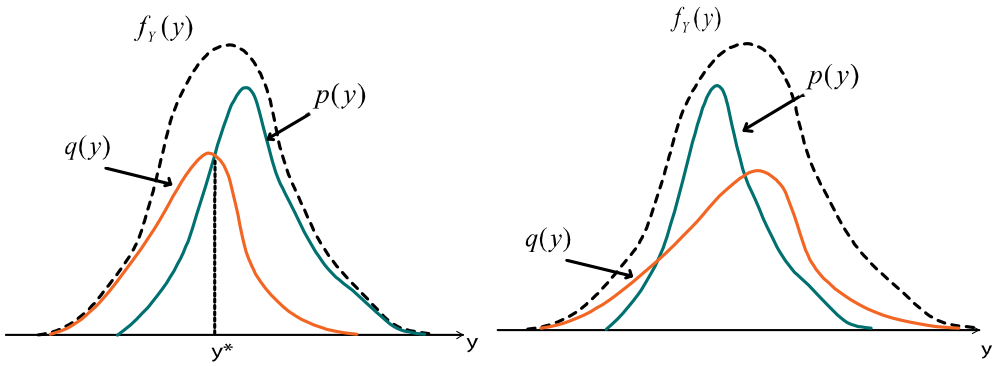

Figure 7: *In the left-hand side figure, the naive cdf bounds at y^* are tight. On the other hand, when $p(y)$ and $q(y)$ are drawn as in the right-hand side figure, the naive cdf bounds are not tight at any $y \in \mathcal{Y}$ (Proposition B.1).*

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 7).

Appendix C: Identification power 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,

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

The next proposition shows the necessary and sufficient condition for the MI mean bounds ($C.1$) to coincide with the ER mean bounds (2.6).

Proposition C.1 (Identification power of ER relative to MI) *The MI mean bounds ($C.1$) coincide with the ER mean bounds (2.6) if and only if the data generating process reveals dominating densities 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 $\min\{P_{mis}, Q_{mis}\} = 1-\delta$ and this holds if the data generating process presents the dominating density on $\mathcal{Y}$. Note that when Y is binary, the above proposition implies the ER mean bounds and the MI mean bounds always coincide. Of course, this must be the case since these two restrictions are equivalent if Y is binary.

Appendix D: No identification gains from the selection equation

In this appendix, we investigate whether the two restrictions on the selection mechanism can further narrow $IR_{f_Y}(P,Q)$. The first restriction is a stronger version of ER, the *random assignment of instrument* (RA, hereafter), which specifies Z to be *jointly* independent of the outcome and the unobserved heterogeneities in the selection equation. The second restriction is the *monotonic selection response to instrument,* which restricts the selection process to the threshold crossing selection with an additive error. Both are common restrictions in the structural selection model.

D.1 Imposing the random assignment restriction

We denote the distribution of types by π_t, $t \in \{c,n,a,d\}$, e.g., $\pi_c \equiv \Pr(T=c) = \Pr(\{U : v(1,U) \geq 0 > v(0,U)\})$.[17] The source of the nonrandom selection process is the dependence between Y and one's selection heterogeneities U. Given the binary instrument Z, this dependence is reduced to the dependence between Y and T, and therefore we can allow distinct outcome distributions conditional on each T (Balke and Pearl (1997) and Imbens and Rubin (1997)). We denote the outcome density conditional on type $T=t$ by $g_t(y) \equiv f_{Y|T}(y|T=t)$, $t=c,n,a,d$.

The random assignment restriction is defined as follows.

Restriction-RA

Random Assignment Restriction (RA): Z is jointly independent of (Y,T).

The implication of imposing RA is summarized by the next lemma.

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

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

where $h_t(y) = \pi_t g_t(y)$.
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 compatible with the data generating process and RA.

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

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

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

$$p(y) = \pi_c g_c(y) + \pi_a g_a(y),$$

[17] It would be most intuitive if we specify an element of the sample space to be an individual in the population, that is, each individual is characterized by the unique value of Y and U.

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

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

We obtain the fourth constraint in a similar manner. This completes the proof of the former statement.
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. ■

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 (*). Proposition 2.2 in the main text shows that this additional requirement does not narrow the identification region $IR_{f_Y}(P, Q)$.

Proof of Proposition 2.2. By Lemma D.1, It suffices to show that, for a data generating process P and Q and an arbitrary $f_Y \in IR_{f_Y}(P, Q)$, we can find nonnegative density functions $h_t(y)$, $t = c, n, a, d$, satisfying the constraints (*).
Figure 8 illustrates the proof of this redundancy result. Given a data generating process, $p(y)$ and $q(y)$, pick an arbitrary $f_Y \in IR_{f_Y}(P, Q)$. We can find four partitions in the subgraph of $f_Y(y)$, which are labeled as C, N, A, and D in Figure 8. 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 $f_Y \in IR_{f_Y}(P, Q)$. By Lemma D.1, the conclusion follows. ■

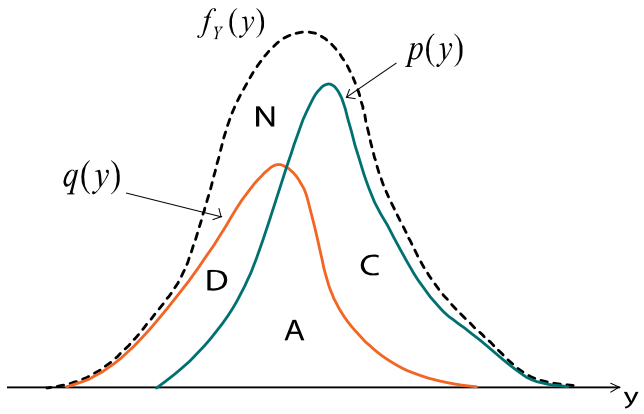

Figure 8: *A graphical illustration of the invariance result of the identification region under RA (Proposition 2.2).*

D.2 Imposing the monotonic selection response to an instrument

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.[18] 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 $\Pr(D_1 = 1) \geq \Pr(D_0 = 1)$. This is equivalent to assuming that the 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 $\Pr(D_1 = 1) \geq \Pr(D_0 = 1)$. The selection process satisfies MSR if $D_1 \geq D_0$ holds for the entire population, that is, no defiers exist in the population $\pi_d = 0$.

From the partial identification point of view, the implication of MSR is summarized in the next proposition, which covers Proposition 2.3 in the main text.

Proposition D.2 (Existence of the dominating density under MSR) *Suppose that a population distribution of (Y, T, Z) satisfies RA and MSR.*
(i) Then, $p(y)$ is the dominating density.
(ii) The MI mean bounds (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

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

Proof of Proposition 2.3 and D.2. (i) From the first two constraints in (*), $\pi_d = 0$ implies that $p(y) - q(y) = \pi_c g_c(y) \geq 0$. (ii) This follows from Proposition C.1. (iii) Suppose that $p(y)$ is the dominating

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

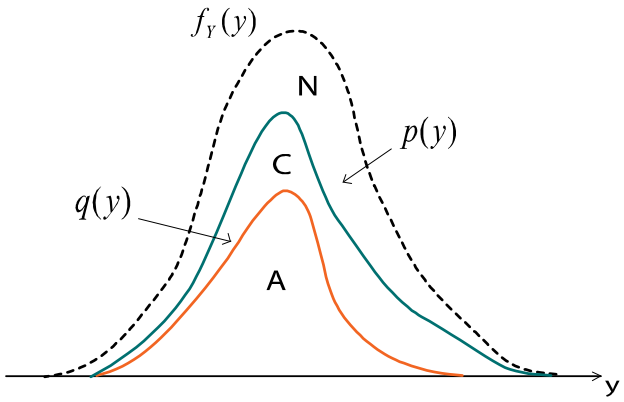

Figure 9: *If RA and MSR are satisfied, we must observe the above configuration of the densities (Proposition 2.3 and D.2).* A *indicates the subgraph of* $q(y)$*. The subgraph of* $p(y)$ *minus that of* $q(y)$ *and the subgraph of* $f_Y(y)$ *minus that of* $p(y)$ *are labeled as* C *and* N*, respectively.*

density. For an arbitrary $f_Y \in IR_{f_Y}(P,Q)$, we want to show that there exists a way to impute the 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,

$$\begin{array}{l} h_c(y) = p(y) - q(y), \\ h_n(y) = f_Y(y) - p(y), \\ h_a(y) = q(y), \\ h_d(y) = 0. \end{array} \tag{D.1}$$

These densities satisfy the constraints (*) and as in the proof of the converse statement of Lemma D.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 $f_Y \in IR_{f_Y}(P,Q)$, we conclude that $IR_{f_Y}(P,Q)$ is the identification region 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 C.1).

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 9. In (D.1), 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 9. This implies that each $f_Y \in IR_{f_Y}(P,Q)$ 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 $IR_{f_Y}(P,Q)$.

Appendix E: Extension to a multi-valued discrete instrument

This appendix provides a framework that covers the case 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 $P_k = (P_k(\cdot), P_{k,mis})$,

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

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.

E.1 A generalization of the identification results

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 non-negative 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 (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

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

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). ■

E.2 A generalization of Proposition 3.1

We use the same notation as in Section 6. Here, we provide a generalization of Proposition 3.1 to the multi-valued instrument case. The following assumptions that are analogous to (A1) through (A4) of Section 3.2.3 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 3.1 is given as follows. A proof can be given in the same manner as the proof of Proposition 3.1, and is therefore omitted for brevity.

Proposition 3.1'. *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 $\lambda_k = \Pr(Z = z_k)$.

Appendix F: An algorithm to estimate $\mathbb{V}^{\max}$ in the histogram class

This appendix presents an algorithm used in the empirical application (Section 7). 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 $P_{n_k}(H_0(y_0))$,..., $P_{n_k}(H_L(y_0))$ be the histogram estimates with respect to the $(L+1)$ bins, $H_0(y_0), \ldots, H_L(y_0)$, as defined in

Section 3.2.2. 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 $O(KL)$.

References

[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 P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," Cowles Foundation Discussion Papers 1676, Cowles Foundation, Yale University.

[6] 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.

[7] Andrews, D. W. K. and G. Soares (2009): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometric*a, forthcoming.

[8] 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.

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

[10] 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.

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

[12] Breusch, T.S. (1986): "Hypothesis Testing in Unidentified Models," *Review of Economic Studies*, 53, 4, 635-651.

[13] Bugni, F. (2009): "Bootstrap Inference in Partially Identified Models," unpublished manuscript, Department of Economics, Duke University.

[14] Canay, I. A. (2009): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity" unpublished manuscript, Northwestern University.

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

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

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

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

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

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

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

[22] 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.

[23] 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.

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

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

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

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

[28] Kitagawa, T. (2009): "Three Essyas in Instrumental Variables" Ph.D dissertation, Brown University Graduate School.

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

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

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

[32] 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.

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

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

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

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

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

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

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

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

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

[42] 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.

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

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

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

[46] Romano, J. P. and A. M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference,* Vol 138, 9, 2786-2807.

[47] Romano, J. P. and A. M. Shaikh (2009): "Inference for the Identified Set in Partially Identified Econometric Models", *Econometrica*, forthcoming.

[48] Rosen, A. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107-117.

[49] Shiryaev, A. N. (1996): *Probability, 2nd ed.* New York: Springer.

[50] van der Vaart, A. W., and J.A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics,* New York: Springer.

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