Testing for Instrument Independence in the Selection Model

Toru Kitagawa[*][†]
UCL and CeMMAP

February, 2010

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 Richard Blundell, Jean-Pierre Florens, 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, University of Iowa, LSE, MIT, Ohio State, Penn, Rochester, Texas Austin, Toulouse School of Economics, UCL, Warwick, 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

The use of instrumental variables is one of the most important development in econometrics, and many empirical findings in economics rely on the method of instrumental variables. The crucial condition that validates the instrumental variable method is the instrument exclusion restriction; the instrument is assigned irrespective of the unobserved outcome heterogeneity in the population. Recently, Manski (2003) and Pearl (1994b, 2000) clarified refutability of the instrument exclusion restriction in the models where the exclusion restriction is imposed in terms of statistical independence of the instrument and outcome. As of this date, however, no testing procedure based on such testable implication has been developed in the literature.

This paper develops a nonparametric specification test for the instrument exclusion restriction in the selection model where an instrument Z is specified to be *statistically independent* of the underlying outcome Y given observable covariates X. The selection problem that this paper considers is the missing data problem with an instrument (Gronau (1974), Heckman (1979)): 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 $(Y \cdot D, D, X, Z)$. For example, Y could be potential wages that are observed only for those who are employed, X is the worker's observable demographic characteristics, 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. The test procedure developed in this paper can be applied to the model with discrete instruments and discrete covariates. Our object of interest is $f_{Y|X}$, the population conditional distribution of Y given X. Also, the identification of $f_{Y|X}$ 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.[1]

The testable implication of the instrument exclusion restriction in this context is obtained by examining identification of $f_{Y|X}$ without imposing point-identifying restrictions for $f_{Y|X}$. That is, our object of interest is the *identification region* for $f_{Y|X}$: the set of outcome distributions conditional on the covariates that are compatible with the empirical evidence and the model restrictions.[2] Manski (2003) analyzes the identification region for the outcome distribution $f_{Y|X}$ under the independence restriction between Y and Z given X. In the partially identified model, the empty identification region implies a misspecification of the imposed restriction, so our specification test infers from data the emptiness of the identification region constructed under the exclusion restriction. Specification tests based on the emptiness of the identification region for the partially identified parameters have been studied in the

[1] The point-identificaton of $f_{Y|X}$ is achieved if an available instrument satisfies the exclusion restriction and the selection probability $\Pr(D = 1|X = x, Z = z)$ 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.

[2] In a sequence of seminal papers, Manski (1989, 1990, 1994, 2003, 2007) analyzes the selection model where some observations of outcome Y can be missing in a nonrandom way, and stimulated research in partial identification analysis. 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.

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, it is not straightforward to express the identification region for the outcome distribution in terms of the moment inequalities. Interestingly, even though the object of interest is in infinite dimensional, 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 and X. In particular, 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.

This paper also discusses practical implementation of the test procedure. The asymptotic distribution of the integrated envelope estimator is given by a supremum functional of a certain Gaussian process 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 partially identified model (Bugni (2010), Canay (2010), and Chernozhukov, Lee, and Rosen (2009)), we first estimate to which asymptotic distribution the bootstrap approximation should target. Given the targeted asymptotic distribution, we bootstrap the empirical processes so as to approximate the Gaussian process component in the targeted asymptotic distribution. In a different context, Anderson, Linton, and Whang (2009) and Lee and Whang (2009) develop an inferential procedure for parameters that are similar to the integrated envelope considered in this paper. Our development of asymptotic theory differs from theirs since our procedure relies on the empirical process theory and it can be applied regardless of smoothness of the underlying outcome density functions.

Blundell, Gosling, Ichimura, and Meghir (2007) consider testing the instrument independence by inferring whether the bounds for the cumulative distribution function (cdf) of $f_{Y|X}$ intersects or not. Our specification test differs from their method in the following ways. First, their procedure tests the emptiness of potentially non-tight cdf bounds for $f_{Y|X}$ while our procedure always tests the emptiness of the *tightest* cdf bounds, so our procedure can asymptotically screen out more violations of the instrument exclusion. Second, the asymptotic 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

[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 (2009), Andrews and Jia (2008), Andrews and Shi (2008), Andrews and Soares (2010), Bugni (2010), Canay (2010), Chernozhukov, Hong, and Tamer (2007), Guggenberger, Hahn, and Kim (2008), Imbens and Manski (2004), Pakes, Porter, Ho, and Ishii (2006), Romano and Shaikh (2008, 2010), and Rosen (2008).

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 given education level, gender, and age. Our test results provide an evidence that the measure of out-of-work income is not independent of the potential wages conditional on a certain coarsening of these covariates.

The remainder of the paper is organized as follows. Section 2 introduces the basic notation and provides the identification region of $f_{Y|X}$. 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, the test procedure is developed with an asymptotically valid bootstrap algorithm. Section 4 provides simulation results and compares the finite sample performance of the bootstrap with other methods. Section 5 tests whether the out-of-work income constructed in Blundell et al. (2003) is independent of the potential wage given covariates. Section 6 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 with its support denoted by $\mathcal{Y} \subset \mathbb{R}$. Our identification analysis allows discrete covariates X with finite support denoted by $\mathcal{X}$, and the distribution of Y conditional on the covariates X is our main interest. We assume that the distribution of Y given X has a probability density function $f_{Y|X}(y|X = x)$ that is absolutely continuous with respect to a known dominating measure μ on $\mathcal{Y}$. Note that Y need not be continuous and we can interpret $f_{Y|X}(y|X = x)$ to be a probability mass at y when μ is the point mass measure. The reason to focus on the probability density is that it is more convenient to present the identification region for the outcome distribution in terms of the densities rather than the distribution functions.

The model has missing data for the outcome Y. We use D to denote the selection indicator: $D = 1$ indicates Y is observed and $D = 0$ indicates Y is missing, and D is observable for all sampled units. We suppose the researcher has an instrumental variable Z that is observed for all sampled units. In the model with selection on unobservables (self-selection), the instrument associated with the instrument exclusion restriction is used to help identify $f_{Y|X}$. The analysis of this paper focuses on the case where the instrument Z is discrete with $K < \infty$ points of support, $Z \in \mathcal{Z} = \{z_1, z_2, \ldots, z_K\}$. The data in our analysis is given as a random sample of $(Y \cdot D, D, X, Z)$. In the example of potential wages in labor economics, Y is potential wage that is observed only when one is employed $(D = 1)$, X can be worker's characteristic that determines potential wages such as education and job experience, and the examples of Z that have been used in the literature would include the number of children, marital status, policy shock in social benefit scheme, the measure of out of work income, etc.

We denote the density function of $(Y \cdot D, D)$ at $D = 1$ given $Z = z_k$ and $X = x \in \mathcal{X}$ by

$$p_{k,x}(y) \equiv f_{Y,D|Z,X}(y, D = 1|Z = z_k, X = x).$$

Note that $(p_{1,x}(y), \ldots, p_{K,x}(y))$ uniquely characterizes the conditional distribution of data given $X = x$ except for the distribution of Z given X. Since the distribution of Z does not play an important role for the later analysis. We represent the *data generating process* of our model by $P = \{(p_{1,x}(y), \ldots, p_{K,x}(y))\}_{x \in \mathcal{X}} \in \mathcal{P}$ where $\mathcal{P}$ represents the class of data generating processes. On the other hand, f is used to refer to the probability density of the *population* that is characterized by a value of (Y, D, Z, X). It is important to keep in mind that the density functions $p_{k,x}(y)$ integrate to the selection probability $\Pr(D = 1|Z = z_k, X = x)$ that is smaller than one in the presence of missing data.

The main concern of this paper is testing for the instrument exclusion restriction that takes the form of statistical independence between the instrument and the outcome given covariates.

Restriction-ER

Exclusion Restriction (ER): Y is statistically independent of Z given X.

ER is a distributional restriction and cannot be represented by a finite number of moment restrictions if Y is continuous. In case our interest is identifying the marginal distribution of Y and the instrument exclusion restriction is imposed in terms of unconditional independence of Z and Y, we do not need covariate information at least for the purpose of identifying f_Y. The analysis of this paper also covers the no-covariate case.

In the classical sample selection model with the structural outcome equation $Y = g(X) + \epsilon$, the standard exogeneity restriction requires that (Z, X) is independent of the unobserved heterogeneity ϵ. This exogeneity restriction implies ER, so rejecting ER allows us to refute independence of (Z, X) and ϵ.

2.2 The identification region of $f_{Y|X}$ and refutability of the exclusion restriction

ER implies that the conditional distribution of Y given Z and X does not depend on Z, $f_{Y|X} = f_{Y|Z,X}$. By applying the law of total probability to the conditional distribution $f_{Y|Z,X}$, we can decompose $f_{Y|X}$ into the conditional density of the observed outcomes and that of the missing outcomes. Using the notation introduced above, we have, for every $k = 1, \ldots, K$ and $x \in \mathcal{X}$,

$$f_{Y|X}(y|X = x) = f_{Y|Z,X}(y|Z = z_k, X = x) = p_{k,x}(y) + f_{Y,D|Z}(y, D = 0|Z = z_k, X = x), \tag{2.1}$$

ER allows us to interpret that the observed outcome distributions $\{p_{k,x}(y)\}_{k=1}^K$ provide distinct identifying information for the common $f_{Y|X}(y|X = x)$. We aggregate these identifying information for $f_{Y|X}(y|X = x)$ by taking the envelope,

$$\underline{f_{Y|X}}(y|X = x) \equiv \max_k \{p_{k,x}(y)\}.$$

We refer to $\underline{f_{Y|X}}$ as the *density envelope at* $X = x$ and the area below the density envelope at $X = x$ as the *integrated envelope at* $X = x$,[4]

$$\delta_x = \int_{\mathcal{Y}} \underline{f_{Y|X}}(y|X = x) d\mu.$$

ER is not sufficient to point-identify $f_{Y|X}$, so we consider constructing *the identification region of* $f_{Y|X}$ *under ER*, whose definition is stated as follows.

Definition 2.1 (the identification region under ER) *Let a data generating process* P *be given. The identification region for* $f_{Y|X}$ *under ER, denoted by* $IR_{f_{Y|X}}(P|ER)$ *is the collection of* $f_{Y|X}$*, conditional distributions of* Y *given* X*, for each of which we can find a probability distribution of* (Y, D, Z) *given* X *that is compatible with the data generating process and ER.*

The identity (2.1) tells that if the density $f_{Y|X}$ belongs to the identification region under ER, then for every k and $X = x$, $f_{Y|X}(y|X = x) \geq p_{k,x}(y)$ holds because $f_{Y,D|Z}(y, D = 0|Z = z_k, X = x)$ appearing in the left hand side of (2.1) must be nonegative. Along this reasoning, Manski (2003) derives The identification region under ER, that can be given in the following form.

Proposition 2.1 (the identification region under ER) *Assume that the population distribution of* Y *given* X *has the conditional probability density* $f_{Y|X}$ *with respect to a dominating measure* μ *on* $\mathcal{Y}$*. Let* $\underline{f_{Y|X}}(y|X = x)$ *be the density envelope and* δ_x *be the integrated envelope at* $X = x$ *defined above.*
(i) The identification region of $f_{Y|X}$ *under ER is*

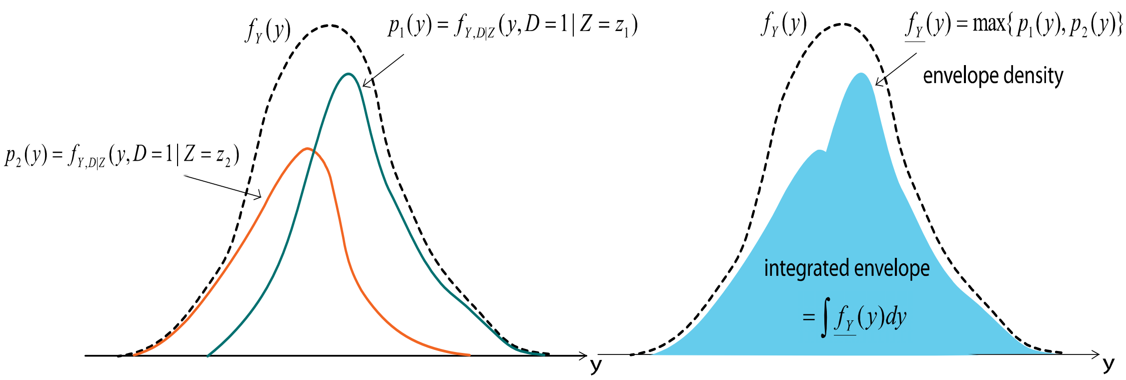
$$\begin{aligned} & IR_{f_{Y|X}}(P|ER) \\ = & \Big\{ \{f_{Y|X}(y|X = x)\}_{x \in \mathcal{X}} : \int_{\mathcal{Y}} f_{Y|X}(y|X = x) d\mu = 1 \\ & \qquad \text{and } f_{Y|X}(y|X = x) \geq \underline{f_{Y|X}}(y|X = x) \ \mu\text{-a.e. at every } x \in \mathcal{X}. \Big\} \end{aligned} \tag{2.2}$$

(ii) $IR_{f_{Y|X}}(P|ER)$ *is empty if and only if* $\delta \equiv \max_{x \in \mathcal{X}} \delta_x > 1$.

Proof. See Appendix A. ■

Figure 1 provides a graphical illustration of the identification region for the binary instrument case without covariates.

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



$f_Y(y)$
$p_1(y) = f_{Y,D|Z}(y, D=1 | Z=z_1)$
$p_2(y) = f_{Y,D|Z}(y, D=1 | Z=z_2)$
y
$f_Y(y)$
$\underline{f_Y}(y) = \max\{p_1(y), p_2(y)\}$
envelope density
integrated envelope
$= \int \underline{f_Y}(y)dy$
y

Figure 1: *Consider the case with a continuous Y and a binary $Z \in \{z_1, z_2\}$ without covariates. 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 $p_1(y)$ and $p_2(y)$ must lie below f_Y. This implies that any f_Y which cover both $p_1(y)$ and $p_2(y)$ are compatible with ER and the empirical evidence. Hence, the identification region of f_Y is obtained as the collection of the probability distributions such that the individual densities each cover both $p_1(y)$ and $p_2(y)$. The right-hand side figure shows the envelope density $\underline{f_Y}(y) = \max\{p_1(y), p_2(y)\}$. The integrated envelope $\delta = \int \underline{f_Y}(y)dy$ is the area below the envelope density (shadow area). If δ exceeds one, then, no probability density function can cover the entire envelope density and we obtain the empty identification region.*

Notice that $IR_{f_{Y|X}}(P|ER)$ becomes empty if and only if the integrated envelope δ_x exceeds one at some $x \in \mathcal{X}$. This is because the probability density function $f_{Y|X}$ must integrate to one at every $X = x$, and, if δ_x exceeds one at $X = x^*$, then there do not exist the probability density $f_{Y|X}(y|X = x^*)$ that can cover the entire density envelope at this value of x^*. Thus, refutability of ER depends only on the estimable parameter $\delta = \sup_{x \in \mathcal{X}} \delta_x$ and testing the emptiness of $IR_{f_{Y|X}}(P|ER)$ is reduced to inferring whether $\delta > 1$ from data. Note that the same refutability condition for instrument independence is found in Manski (2003) in the missing data model and Pearl (1994b) in the context of causal inference with an instrument.

If $IR_{f_{Y|X}}(P|ER)$ is nonempty, then for each $x \in \mathcal{X}$, $f_{Y|X}(y|X = x)$ has the representation of a mixture of two probability densities weighted by δ_x,

$$f_{Y|X}(y|X = x) = \delta_x \left(\underline{f_{Y|X}}(y|X = x)/\delta_x \right) + (1 - \delta_x)\gamma_x(y), \tag{2.3}$$

where $\underline{f_{Y|X}}/\delta_x$ is the normalized envelope density depending only on the data generating process and $\gamma_x(y)$ is a probability density function that can be arbitrarily chosen to span the identification region. Thus, another way to view $IR_{f_{Y|X}}(P|ER)$ is the set of probability distributions generated from (2.3) by choosing an arbitrary probability density $\gamma_x(y)$. By this way of representing $IR_{f_{Y|X}}(P|ER)$, $F_{Y|X}$ the conditional cdf of Y given X whose density belongs to $IR_{f_{Y|X}}(P|ER)$ is written as

$$F_{Y|X}(y|X = x) = \int_{(-\infty, y]} \underline{f_{Y|X}}(t|X = x)d\mu + (1 - \delta_x)\Gamma_x(y),$$

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

$$\int_{(-\infty,y]} \underline{f_{Y|X}}(t|X=x)d\mu \leq F_{Y|X}(y|X=x) \leq \int_{(-\infty,y]} \underline{f_{Y|X}}(t|X=x)d\mu + 1 - \delta_x. \tag{2.4}$$

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 $E(Y|X)$ also follow from (2.3). Let Y have a compact support $\mathcal{Y} = [y_l, y_u]$. By specifying $\gamma_x(y)$ as the degenerate distribution at the lower or upper bound of the outcome support, we obtain the tight bounds for $E(Y|X)$ under ER,

$$(1-\delta_x)y_l + \int_{\mathcal{Y}} y\underline{f_{Y|X}}(y|X=x)d\mu \leq E(Y|X=x) \leq \int_{\mathcal{Y}} y\underline{f_{Y|X}}(y|X=x)d\mu + (1-\delta_x)y_u. \tag{2.5}$$

Since statistical independence is a stronger restriction than the moment type restriction, these mean bounds can be strictly narrower than the tight mean bounds under the mean independence restriction, $E(Y|Z,X) = E(Y|X)$, constructed in Manski (1994).

Recall that ER only assumes conditional independence between the outcome Y and instrument Z given X, while it is silent about how the instrument affects one's selection response and how it is related to the unobserved heterogeneity in the selection mechanism. One might wonder whether one can strengthen the refutability condition by introducing the structural selection equation and imposing *instrument joint independence* of Y and the selection heterogeneity conditional on X. It turns out that, if we allow the selection equation to have threshold crossing with *nonadditive* errors, i.e., $D = 1\{u(Z,X,U) \geq 0\}$ where U represents the unobserved heterogeneity in the selection response, the identification region for $f_{Y|X}$ does not change even when we strengthen ER to the instrument joint independence (See Appendix C for a proof of this claim).

Furthermore, if threshold crossing selection with an *additive* error is assumed in addition to instrument joint independence, i.e., $D = 1\{\tilde{u}(Z,X) - U \geq 0\}$ with a scalar unobserved heterogeneity U in the selection response, then we must observe $p_{k,x}(y) \geq p_{k',x}(y)$ μ-a.e. or $p_{k,x}(y) \leq p_{k',x}(y)$ μ-a.e. for every $k \neq k'$ and $x \in \mathcal{X}$. That is, for each $x \in \mathcal{X}$, the K observed densities $\{p_{k,x}(y)\}_{k=1}^{K}$ must show the nesting configurations where a density with higher selection probability $\Pr(D=1|Z=z_k, X=x)$ nests ones with lower selection probabilities, and none of the observed densities intersect with the others. Conversely, it can be shown that if data exhibits such nesting configuration, the identification region under joint independence and the threshold crossing selection with an additive error takes the identical form to $IR_{f_{Y|X}}(P|ER)$ (See Appendix C for details). This implies that the structural selection model with additively separable latent utility *constrains the data generating process without further narrowing the identification region than ER.* The density envelope provides the maximal identifying information for $f_{Y|X}$ based only on the empirical evidence, and optimality of this aggregating scheme is free from the assumptions that only constrain the data generating process.[5]

[5] Note that the condition of $p_{k,x}(y) \geq p_{k',x}(y)$ μ-a.e. or $p_{k,x}(y) \leq p_{k',x}(y)$ μ-a.e. for every $k \neq k'$ and

3 Estimation of the integrated envelope

Our identification analysis clarified that the emptiness of the identification region under ER is summarized by the estimable parameter δ. Hence, the rest of the paper focuses on estimation and inference for δ so as to develop a specification test for the instrument independence assumption. In this section, we first consider the case where there are no covariates and ER is given as the unconditional independence of Z and Y. A slightly more complicated situation where there are covariates and ER takes the form of conditional independence of Z and Y given X is considered in Section 3.4.

Without losing any distributional information of data, we may denote an outcome observation recorded in data by $Y_{data} \equiv DY + (1-D)\{mis\}$ and express data as i.i.d observations of $(Y_{data,i}, Z_i)$, $i = 1, \dots, N$, where $\{mis\}$ indicates that the observation of Y is missing. Note that, except for the marginal distribution of Z, the data generating process in this case is characterized by the conditional distributions of Y_{data} given $Z = z_k$ for $k = 1, \dots, K$, which have the support $\mathcal{Y} \cup \{mis\}$. On the support $\mathcal{Y}$, $p_k(y)$ defined in our identification analysis can be seen as the density function of Y_{data} given $Z = z_k$ that is assumed to be absolutely continuous with respect to a dominating measure μ on $\mathcal{Y}$. So, for a subset $V \subset \mathcal{Y}$, we have

$$P_k(V) \equiv \Pr(Y_{data} \in V | Z = z_k) = \int_V p_k(y) d\mu,$$

and the data generating process is represented by $P = \{P_k\}_{k=1}^K$. We divide the full sample into K subsamples based on the assigned value of $Z \in \{z_1, \dots, z_K\}$. We denote the size of these subsamples by n_k. We assume $\lambda_k \equiv \Pr(Z_i = z) > \epsilon$ for some $\epsilon > 0$ and let $\lambda = (\lambda_1, \dots, \lambda_K)$ and $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_K)$ where $\hat{\lambda}_k \equiv n_k/N$, $k = 1, \dots, K$. We adopt the K-sample problem with nonrandom sample size, i.e., our asymptotic analysis is conditional on the sequence $\{Z_i : i = 1, 2, ...\}$. Since $n_k \to \infty$ and $\hat{\lambda}_k \to \lambda_k$ with probability one for every $k = 1, \dots, K$ as $N \to \infty$, we interpret the stochastic limit with respect to $N \to \infty$ equivalent to the limit with respect to $n_k \to \infty$ and $\hat{\lambda}_k \to \lambda_k$ for all $k = 1, \dots, K$.

The test strategy considered in this paper is as follows. The null hypothesis is that the identification region of f_Y under ER is nonempty, that is, $\delta = \int_{\mathcal{Y}} \max_k\{p_k(y)\} d\mu \leq 1$. 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 δ such that $\sqrt{N}(\hat{\delta} - \delta)$ has an asymptotic distribution,

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

where "$\rightsquigarrow$" denotes weak convergence and $J(\cdot; P, \lambda)$ represents the cdf of the asymptotic distribution that can depend on P and λ. We infer whether or not $\delta \leq 1$ 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 $J(\cdot; P, \lambda)$, $c_{1-\alpha}(P, \lambda)$, and to check whether the one-sided confidence intervals $[\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}}, \infty)$

$x \in \mathcal{X}$ provides a testable implication for the joint restriction of joint independence and additively separable latent utility. We leave a development of testing procedure of the nesting configuration for future research, and, in this paper, we focus on testing ER with the refutability condition $\delta > 1$.

contain 1 or not. We reject the null hypothesis if we observe $\hat{\delta} - \frac{\hat{c}_{1-\alpha}}{\sqrt{N}} > 1$. This procedure yields a *pointwise* asymptotically size correct test[6] since for every P satisfying the null $\delta \leq 1$, we have

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

Our development of the inferential procedure is decomposed into four parts. First, in order to illustrate the main idea on the inferential procedure, we consider a toy example such that the model has a binary Y and a binary Z (Section 3.1). Second, we develop an estimator of δ and derive the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ for a more general case including continuous Y (Section 3.2). Third, we extend the analysis to the case with discrete covariates X (Section 3.3). Last, we analyze how to get an asymptotically valid critical values (Section 3.4).

3.1 An illuminating example: binary Y and binary Z without covariates.

When $Y \in \{1, 0\}$ and $Z \in \{z_1, z_2\}$, the data generating process P can be represented by a pair of three probability masses. To simplify the notations, we write those probability masses by (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 $Z = z_1$ and $Z = z_2$ respectively. Here, the integrated envelope δ becomes

$$\delta \equiv \max\{p^1, q^1\} + \max\{p^0, q^0\}. \tag{3.1}$$

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) that 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

[6] 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.

$$\begin{aligned}
\Sigma_{P,\lambda} &= \lambda_1^{-1}\begin{pmatrix} p^1\left(1-p^1\right) & -p^1p^0 \\ -p^1p^0 & p^0(1-p^0) \end{pmatrix}, \\
\Sigma_{Q,\lambda} &= (1-\lambda_1)^{-1}\begin{pmatrix} q^1\left(1-q^1\right) & -q^1q^0 \\ -q^1q^0 & q^0(1-q^0) \end{pmatrix}, \quad \text{and} \quad \lambda_1 = \Pr(Z = z_1).
\end{aligned}$$

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 $p^0 = q^0$.

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

$$\begin{aligned}
&\delta_1 = p^1 + p^0, \quad G_1 = X^1 + X^0, \\
&\delta_2 = p^1 + q^0, \quad G_2 = X^1 + W^0, \\
&\delta_3 = q^1 + p^0, \quad G_3 = W^1 + X^0, \\
&\delta_4 = q^1 + q^0, \quad G_4 = W^1 + W^0,
\end{aligned}$$

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

$$\sqrt{N}(\hat{\delta} - \delta) = \max\left\{\begin{array}{c} \sqrt{N}(\hat{\delta}_1 - \delta) \\ \sqrt{N}(\hat{\delta}_2 - \delta) \\ \sqrt{N}(\hat{\delta}_3 - \delta) \\ \sqrt{N}(\hat{\delta}_4 - \delta) \end{array}\right\} \rightsquigarrow \max_{\{j:\delta_j=\delta\}}\{G_j\}. \tag{3.2}$$

where the index set of the max operator $\{j : \delta_j = \delta\}$ tells which δ_j achieves δ, and the size of this index set indicates whether or not there are ties between (p^1, p^0) and (q^1, q^0). 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, and it raises difficulties in conducting inference on δ since we do not know which asymptotic distribution gives a better approximation for the sampling distribution of $\sqrt{N}(\hat{\delta} - \delta)$. 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 $\sqrt{N}(\hat{\delta} - \delta)$.

There are several procedures available for asymptotically valid inference on δ. One approach estimates the asymptotic distribution $\max_{\{j:\delta_j=\delta\}}\{G_j\}$ in two steps. In the first step, the index set $\mathcal{V}^{\max}\equiv\{j : \delta_j = \delta\}$ is estimated and, in the second step, the estimated joint distribution of G_j's are plugged into the maximum operator. 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 can estimate $\mathbb{V}^{\max}$ using the sequence of *slackness variables* $\eta_N \geq 0$, $N = 1, 2, ...$ This two-step procedure is analogous to the moment selection approach in the inference on moment equalitiy model (Andrews and Soares (2010), Andrews and Jia (2009), Bugni (2010), Canay (2010)), and has also been considered in the test of superior predictive ability (Hansen (2005)).

One such estimator for $\mathbb{V}^{\max}$ is

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

where

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

Another In this construction of $\hat{\mathcal{V}}^{\max}(\eta_N)$, 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 $\mathcal{V}^{\max}$. As we claim in this estimator for $\mathcal{V}^{\max}$ is asymptotically valid in a certain sense if the slackness sequence $\{\eta_N : N \geq 1\}$ meets the above divergence rate, which is implied by the law of iterated logarithm (see, e.g., Shiryaev (1996)). Other estimators for $\mathcal{V}^{\max}$ are available. For example, given that the null hypothesis becomes least favorable at $\delta = 1$, we may estimate $\mathcal{V}^{\max}$ by $\{j \in \{1, 2, 3, 4\} : \sqrt{N}(\hat{\delta} - \hat{\delta}_j) \leq \eta_N\}$

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

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

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 $\hat{\mathcal{V}}^{\max}(\eta_N)$ and $\hat{\delta}_j$, we resample,

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

where $\hat{\delta}^*_j$ is the bootstrapped $\hat{\delta}_j$. Since the standard argument of the bootstrap consistency ensures $\sqrt{N}(\hat{\delta}^*_j - \hat{\delta}_j) \rightsquigarrow G_j$, we can obtain a valid approximation of the asymptotic distribution. In Section 3.3, we extend this approach to a more general setting.[7]

[7] As Andrews (2000) points out, another asymptotically pointwise valid method is subsampling (Politis and Romano (1994)). This is also the case for the inference of the integrated envelope, while as shown in the Monte Carlo studies, subsampling performs relatively poorer than the bootstrap.

3.2 Estimation of the integrated envelope and its asymptotic property without covariates

We generalize the toy example considered above to a general case where Y can be an arbitrary scalar random variable and Z can be a multi-valued discrete instrument with finite points of support.

We first rewrite the integrated envelope without covariates in the following way,

$$\delta = \int_{\mathcal{Y}} \max_k \{p_k(y)\}\, d\mu = \sum_{k=1}^{K} P_k(E_k) \tag{3.3}$$

where $\mathbf{E} = (E_1, \dots, E_K)$ is a K-partition of the outcome support such that each E_k corresponds to the subset in $\mathcal{Y}$ where $p_k(y)$ is greater than the other densities, i.e., $E_k = \{y \in \mathcal{Y} : p_k(y) \geq p_l(y)\ \forall\ l \neq k\}$ and $\mu(E_k \cap E_l) = 0$ for all $k \neq l$. Let $\mathcal{B}(\mathcal{Y})$ be the Borel σ-algebra on $\mathcal{Y}$, and consider a class of K-partitions of the outcome support generated by $\mathcal{B}(\mathcal{Y})$,

$$\mathcal{V}_{\mathcal{B}}= \{\mathbf{V} =(V_1, \dots, V_K) : V_k \in \mathcal{B}(\mathcal{Y})\ \forall k \text{ and } \mu(V_k \cap V_l) = 0\ \forall k \neq l\}.$$

We introduce an equivalence relationship to the partition class such that two partitions $\mathbf{V} =(V_1, \dots, V_K)$ and $\mathbf{V}' = (V_1', \dots, V_K')$ are equivalent if their difference has measure zero, i.e., $\mathbf{V} = \mathbf{V}'$ if $\sum_{k=1}^{K} \mu(V_k \triangle V_k') = 0$.

Define a function $\delta(\cdot)$ that maps $\mathcal{V}_{\mathcal{B}}$ to $\mathbb{R}_+$ as

$$\delta(\mathbf{V}) = \sum_{k=1}^{K} P_k(V_k). \tag{3.4}$$

Then, by noting that $\mathbf{E} \in\mathcal{V}_{\mathcal{B}}$ and $\delta(\mathbf{V})$ achieves its maximum at $\mathbf{E}$, the integrated envelope of (3.3) is expressed as

$$\delta = \sup_{\mathbf{V}\in\mathcal{V}_{\mathcal{B}}} \delta(\mathbf{V}). \tag{3.5}$$

Recall, in the binary Y and binary Z example, we could write the true integrated envelope by

$$\begin{aligned}\delta &= \max\left\{\begin{array}{c} p^1+p^0 \\ p^1+q^0 \\ p^0+q^1 \\ q^1+q^0 \end{array}\right\} = \max\left\{\begin{array}{c} P_1(\{1,0\})+P_2(\emptyset) \\ P_1(\{1\})+P_2(\{0\}) \\ P_1(\{0\})+P_2(\{1\}) \\ P_1(\emptyset)+P_2(\{1,0\}) \end{array}\right\} \\ &= \max_{(V_1,V_2)} \{P_1(V_1)+P_2(V_2)\} \end{aligned} \tag{3.6}$$

where (V_1, V_2) is a partition of $\mathcal{Y} =\{1, 0\}$. Here, $P_1(V_1) + P_2(V_2)$ is seen as a function from the partition class on $\{1, 0\}$ to $\mathbb{R}_+$ and the integrated envelope is defined as its maximum over the possible partitions of $\mathcal{Y} =\{1, 0\}$. Thus, the expression (3.5) can be seen as a direct analogue of (3.6) for a more complex $\mathcal{Y}$, and the only complication appears in the class of partitions of $\mathcal{Y}$ on which the supremum operates.

Let $\hat{P}_k$, $k = 1, \ldots, K$, be the empirical probability measures for $\{Y_{data,i} : Z_i = z_k\}$, i.e., for $V \in \mathcal{B}(\mathcal{Y})$, $\hat{P}_k(V) \equiv \frac{1}{n_k} \sum_{i:Z_i=z_k} I\{Y_{data,i} \in V\}$. We define a sample analogue of $\delta(\cdot)$ by replacing the population distribution of $P_k(\cdot)$ in (3.4) with the empirical distributions $\hat{P}_k(\cdot)$,

$$\hat{\delta}(\mathbf{V}) = \sum_{k=1}^{K} \hat{P}_k(V_k) \tag{3.7}$$

Analogous to the construction of the integrated envelope in (3.5), we propose an estimator of δ by maximizing $\hat{\delta}(\cdot)$ over the partition class $\mathcal{V}$ generated by a class of subsets $\mathbb{V} \subset \mathcal{B}(\mathcal{Y})$,[8]

$$\begin{aligned} \hat{\delta} \quad &\equiv \quad \sup_{\mathbf{V} \in \mathcal{V}} \{\hat{\delta}(\mathbf{V})\}, && (3.8) \\ \text{where} \quad \mathcal{V} \quad &= \quad \{\mathbf{V} = (V_1, \ldots, V_K) : V_k \in \mathbb{V} \ \forall k \text{ and } \mu(V_k \cap V_l) = 0 \ \forall k \neq l\}. && (3.9) \end{aligned}$$

When Y is discrete, $\mathbb{V}$ can be specified as the power set of the support points 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 otherwise we can almost surely find the partitions that yield the trivial maximum $\sum_{k=1}^{K} n_k^{-1} \sum_{i:Z_i=z_k} D_i$. Note it provides little information on the true integrated envelope no matter how large the sample size is since it almost surely converges to $\sum_{k=1}^{K} \Pr(D = 1|Z = z_k)$.

A suitable restriction to avoid such overfitting and to guarantee the consistency of the estimator $\hat{\delta}$ to the true integrated envelope δ is that $\mathbb{V}$ is the *Vapnik-Červonenkis class (VC-class)* (see, e.g., Dudley (1999) for the definition of VC-class). For instance, the class of closed intervals in $\mathbb{R} \cup \{-\infty, \infty\}$ including the empty set is an example of the VC-class. Accompanied by such restriction on the class of possible partitions, we shall assume that the partition class $\mathcal{V}$ contains some $\mathbf{V}$ that attain $\delta(\mathbf{V}) = \delta$. This assumption, or, for short, the specification 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 $\mathbb{V}$ as the class of closed intervals, we are imposing a restriction on the configuration of $\{p_k(y)\}_{k=1}^{K}$ such that $\{y \in \mathcal{Y} : p_k(y) \geq p_l(y)$ for every $l \neq k\}$ is convex or empty for each $k = 1, \ldots, K$.[9]

As we saw in the binary Y case, ties among the densities cause a non-Gaussian asymptotic distribution for the estimator of δ. Let us define the *maximizer partition class*

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

[8] 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)).

[9] 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, we will 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.

If $\mathcal{V}^{\max}$ consists of a single element $\mathbf{V}^{\max}$, it implies $\mathbf{V}^{\max}$ is the only one that partitions 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 $p_k(y)$ and $p_l(y)$ for some $k \neq l$ are tied on a set with positive measure, $\mathcal{V}^{\max}$ can contain multiple elements.[10]

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

Condition A

(A1) *Uniform Convergence*: For each $k = 1, \ldots, K$, the set-indexed empirical processes $G_{P_k,n_k}(\cdot) = \sqrt{n_k}\left(\hat{P}_k(\cdot) - P_k(\cdot)\right)$ converge uniformly in law to tight mean zero Gaussian processes in $l^\infty(\mathbb{V})$:

$$G_{P_k,n_k}(V) \rightsquigarrow G_{P_k}(V)$$

where $Cov(G_{P_k}(V), G_{P_k}(V')) = P_k(V \cap V') - P_k(V)P_k(V')$.

(A2) *Optimal partition*: There exists a nonempty *maximizer partition class* $\mathcal{V}^{\max} \subset \mathcal{V}$ *defined by*

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

(A3) *Existence of Maximizers:* There exists random partitions $\hat{\mathbf{V}}_N \in \mathcal{V}$ and $\hat{\mathbf{V}}_N^{\max} \in \mathcal{V}^{\max}$ such that

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

with probability one.

Note condition (A1) generally holds for discrete Y, and, for continuous Y, it is implied by the restriction for $\mathbb{V}$ to a VC-class of measurable subsets in $\mathcal{Y}$. Condition (A2) implies that the partition class $\mathcal{V}$ defined in (3.9) contains at least one optimal subset at which the function $\delta(\cdot)$ achieves the true integrated envelope. Since these partitions maximize $\delta(\cdot)$, we refer to the collection of these partitions as the *maximizer partition class* $\mathcal{V}^{\max}$. We allow $\mathcal{V}^{\max}$ to contain more than one element in order to handle the aforementioned issue of ties among the densities. The condition (A3) says that the supremum of $\hat{\delta}(\mathbf{V})$ on $\mathcal{V}$ and the one on $\mathcal{V}^{\max}$ can be evaluated by $\hat{\delta}(\hat{\mathbf{V}}_N)$ and $\hat{\delta}(\hat{\mathbf{V}}_N^{\max})$ for some sequences $\hat{\mathbf{V}}_N \in \mathcal{V}$ and $\hat{\mathbf{V}}_N^{\max}$ almost surely. It trivially holds if $\mathcal{V}$ is a finite set. This condition is only for mathematical convenience, and it does not seem to have practical restriction.

The derivation of the asymptotic distribution of $\hat{\delta}$ relies on the functional limit theorem for the set index empirical processes as assumed in Condition (A1). By the definition of $\hat{\delta}$, we have

$$\sqrt{N}(\hat{\delta} - \delta) = \sup_{\mathbf{V}\in\mathcal{V}} \left\{\sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V})) + \sqrt{N}(\delta(\mathbf{V}) - \delta)\right\}. \tag{3.10}$$

[10] For instance, suppose that $p_1(y) = p_2(y) > p_k(y)$ for $k = 3, 4, \ldots, K$ on a subset $A \subset \mathcal{B}(\mathcal{Y})$ with positive measure. Let $\mathbf{V} = (V_1, V_2, \ldots, V_K)$ be a maximizer of $\delta(\cdot)$ over $\mathcal{V}$ such that $A \subset V_1$ holds. If $V_1 \setminus A \in \mathbb{V}$ and $V_2 \cup A \in \mathbb{V}$, then $\mathbf{V}' = (V_1 \setminus A, V_2 \cup A, V_3, \ldots, V_K) \in \mathcal{V}$ also maximizes $\delta(\cdot)$, that is, $\delta(\mathbf{V}) = \delta(\mathbf{V}') = \delta$ holds.

The first term in the supremum of (3.10) can be written as the sum of independent empirical processes on $\mathbb{V}$, $\sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V})) = \sum_{k=1}^{K} \hat{\lambda}_k^{-1/2} G_{P_k,n_k}(V_k)$, so that condition (A1) implies the uniform convergence of $\sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V}))$ to tight Gaussian processes on the partition class $\mathcal{V}$ (see the proof of Proposition 3.1 in Appendix A.) On the other hand, the second term in the supremum of (3.10) vanishes for $\mathbf{V} \in \mathcal{V}^{\max}$ and it diverges to negative infinity for $\mathbf{V} \notin \mathcal{V}^{\max}$. Therefore, for large N, the supremum should operate only over $\mathcal{V}^{\max}$. This argument implies that the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is given by the supremum of the partition indexed Gaussian processes over the maximizer partition class $\mathcal{V}^{\max}$.

Proposition 3.1 (asymptotic distribution of $\hat{\delta}$) *Under condition (A1), (A2), and (A3),*

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \sup_{\mathbf{V} \in \mathcal{V}^{\max}} \{G_\delta(\mathbf{V})\}, \tag{3.11}$$

where $G_\delta(\mathbf{V})$ is the mean zero tight Gaussian process in $l^\infty(\mathcal{V})$ with the covariance function, for $\mathbf{V}$, $\mathbf{V}' \in \mathcal{V}$,

$$Cov(G_\delta(\mathbf{V}), G_\delta(\mathbf{V}')) = \sum_{k=1}^{K} \lambda_k^{-1} \left[P_k(V_k \cap V_k') - P_k(V_k) P_k(V_k') \right].$$

In particular, if $\mathcal{V}^{\max}$ is a singleton with the unique element $\mathbf{V}^{\max} = (V_1^{\max}, \ldots, V_k^{\max})$, then $\hat{\delta}$ is $\sqrt{N}$-asymptotically normal,

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

where

$$\sigma^2(P, \lambda) = \sum_{k=1}^{K} \lambda_k^{-1} P_k(V_k^{\max}) \left[1 - P_k(V_k^{\max})\right].$$

Proof. See Appendix A. ■

The asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ depends not only on the data generating process P and λ, but also on the maximizer partition class $\mathcal{V}^{\max}$. If $\mathcal{V}^{\max}$ contains multiple elements, the asymptotic distribution is not normal and it is given by the extremum of the Gaussian processes as we have seen in the toy example of Section 3.1. On the other hand, if $\mathcal{V}^{\max}$ has the unique element $\mathbf{V}^{\max}$, then, the distribution of (3.11) is given by the projection of the Gaussian processes onto $\mathbf{V}^{\max}$ so $\sqrt{N}(\hat{\delta} - \delta(P, Q))$ is asymptotically normal. This 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 the densities $p_k(y)$ do not vary much over k, then it would be reasonable to allow for the situation where $\mathcal{V}^{\max}$ is nonsingleton. In the next section, we consider how to approximate such asymptotic distribution without a priori assuming $\mathcal{V}^{\max}$ to be a singleton.

3.3 Estimation of the integrated envelope with discrete covariates

If we have some covariates X in the model, our interest can be testing for the conditional independence of Y and Z given X. Our approach developed in the previous section can be extended to this case if the covariates are all discrete. If some of the covariates are continuous, discretizing the covariates may reduce the refutability power of the test procedure, while rejection still allows us to refute conditional independence of Y and Z given X.

Let $\mathcal{X}$ be a finite set, and, for a subset $V \subset \mathcal{Y}$, let $P_{k,x}(A) \equiv \Pr(Y_{data} \in A, D = 1|Z = z_k, X = x)$ for each $k = 1, \ldots, K$ and $x \in \mathcal{X}$. We denote its sample counterpart by $\hat{P}_{k,x}(V)$, i.e., with $n_{k,x}$ being the size of subsample with $Z_i = z_k$ and $X_i = x$, We make our asymptotic analysis being conditional on the the sequence of instrument and conditioning covariates $\{(Z_i, X_i)\}_{i=1}^{\infty}$ and, therefore, the subsample size $n_{k,x}$ and $N_x \equiv \sum_{k=1}^{K}$ are nonrandom.

$$\hat{P}_{k,x}(V) = \frac{1}{n_{k,x}} \sum_{i:X_i = x_m \text{ and } Z_i = z_k} I\left\{Y_{data,i} \in V\right\}.$$

With the class of partitions defined in (3.9), the estimator for each δ_x is constructed analogously to (3.8),

$$\hat{\delta}_x = \sup_{\mathbf{V} \in \mathcal{V}} \{\hat{\delta}(\mathbf{V}, x)\} \quad \text{where} \quad \hat{\delta}_x(\mathbf{V}) = \sum_{k=1}^{K} \hat{P}_{k,x}(V_k).$$

By plugging these into $\delta = \max_{x \in \mathcal{X}} \{\delta_x\}$, the estimator for δ is obtained as

$$\hat{\delta} = \max_{x \in \mathcal{X}} \left\{\hat{\delta}_x\right\} = \sup_{(\mathbf{V}, x) \in \mathcal{V} \times \mathcal{X}} \{\hat{\delta}_x(\mathbf{V})\}.$$

Under certain regularity conditions (see Appendix A), the asymptotic property of this estimator is obtained as follows.

Proposition 3.2 *Under the regularity conditions provided in Appendix A,*

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \sup_{(\mathbf{V}, x) \in (\mathcal{V} \times \mathcal{X})^{\max}} \{G_\delta(\mathbf{V}, x)\},$$

where $(\mathcal{V} \times \mathcal{X})^{\max}$ *is the subclass of* $\mathcal{V} \times \mathcal{X}$ *defined as* $\{(\mathbf{V}, x) \in \mathcal{V} \times \mathcal{X} : \delta_x(\mathbf{V}) = \delta\}$ *and* $G_\delta(\mathbf{V}, x)$ *is a mean zero tight Gaussian process in* $l^\infty(\mathcal{V} \times \mathcal{X})$.

Proof. See Appendix A. ■

Note that the expression of the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is similar to the one obtained in Proposition 3.1 except that here the Gaussian processes are indexed by both partitions V and the conditioning covariate value x.

3.4 Obtaining Critical Values by Resampling

In this section, we discuss a resampling method to obtain an asymptotically valid critical values. The resampling methods are particularly useful since the asymptotic distribution of $\sqrt{N}(\hat{\delta}-\delta)$ given in Proposition 3.1 and 3.2 has the form of a supremum functional of the Gaussian processes, and, especially when $\mathcal{V}^{\max}$ or $(\mathcal{V}\times\mathcal{X})^{\max}$ is not a singleton, it is difficult to obtain the critical values analytically (Romano (1988)).

The expression of the asymptotic distribution (Proposition 3.1 and 3.2) implies that the asymptotic normality fails depending on the size of maximizer subclass of partitions $\mathcal{V}^{\max}$ or $(\mathcal{V}\times\mathcal{X})^{\max}$. In this sense, the asymptotic distribution is not pivotal, and this complicates consistent inference for $\hat{\delta}$. As we illustrated in Section 3.1 with a binary outcome and a binary instrument without covariates, we argued approximating the asymptotic distribution by the distribution of $\sup_{\mathbf{V}\in\hat{\mathcal{V}}^{\max}}\left\{\sqrt{N}(\hat{\delta}^*(\mathbf{V})-\hat{\delta}(\mathbf{V}))\right\}$ where $\hat{\mathcal{V}}^{\max}$ is an estimator for $\mathcal{V}^{\max}$ and $\sqrt{N}(\hat{\delta}^*(\mathbf{V})-\hat{\delta}(\mathbf{V}))$ is the bootstrap analogue of $\sqrt{N}(\hat{\delta}(\mathbf{V})-\delta(\mathbf{V}))$. Below, we generalize this approach to the models covered in Section 3.2 and 3.3.

Let $\mathbf{Y}_{n_{k,x}}$ represent the original sample of $Y_{data,i}$ with $Z_i=z_k$ and $X_i=x$ whose size is $n_{k,x}$. The size of sample with $X_i=x$ is given by $N_x=\sum_{k=1}^K n_{k.x}$. Our bootstrap algorithm is summarized as follows. If there is no covariate in the model, we may drop the subscript x in the description below.

Algorithm: bootstrap for the integrated envelope

1. *GIven the sample, compute* $\hat{\delta}_x(\cdot)$, $\hat{\delta}_x$, *and* $\hat{\delta}$.

2. *Let* $\{\eta_{N_x}: N_x\geq 1\}$ *for each* $x\in X$ *be the slackness sequences that satisfy*

$$\frac{\eta_{N_x}}{\sqrt{N_x}}\to 0,\qquad \frac{\eta_{N_x}}{\sqrt{\log\log N_x}}\to\infty\ \text{ as } N_x\to\infty.$$

3. *Estimate the maximizer partition class by*

$$\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}=\left\{(\mathbf{V},x)\in\mathcal{V}\times\mathcal{X}:\sqrt{N_x}(1-\hat{\delta}_x(\mathbf{V}))\leq\eta_{N_x}\right\}.$$

   If $\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$ *is empty, we do not reject the null. If* $\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$ *is not empty, we proceed to the next step.*[11]

4. *For each* $k=1,\ldots,K$ *and* $x\in X$*, we sample* $n_{k,x}$ *observations from* $\mathbf{Y}_{n_{k,x}}$ *randomly with replacement to construct* $P^*_{k,x}(\cdot)$*, the empirical measure based on the bootstrapped*

[11] Alternatively, it is also possible to estimate $(\mathcal{V}\times\mathcal{X})^{\max}$ by

$$(\mathcal{V}\times\mathcal{X})^{\max}=\left\{(\mathbf{V},x)\in\mathcal{V}\times\mathcal{X}:\sqrt{N_x}\frac{(1-\hat{\delta}_x(\mathbf{V}))}{\sqrt{\widehat{Var}(\hat{\delta}_x(\mathbf{V}))}}\leq\eta_{N_x}\right\}$$

where the criterion function $1-\hat{\delta}_x(\mathbf{V})$ is weighted by the marginal variance of $\hat{\delta}_x(\mathbf{V})$. We may expect that controlling the marginal variance makes it easier to find an appropriate value of η_{N_x}. Nevertheless, we do not so far find such practical gain through our simulation study.

sample. Using the constructed $\left\{P^*_{k,x}(\cdot)\right\}_{k=1}^{K}$*, obtain the bootstrap analogue of* $\hat{\delta}_x(\cdot)$*,*

$$\hat{\delta}^*_x(\mathbf{V}) = \sum_{k=1}^{K} P^*_{k,x}(V_k), \quad \mathbf{V} \in \mathcal{V}.$$

5. *Compute*

$$\sup_{(\mathbf{V},x)\in(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}} \left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V}))\right\}.$$

6. *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.*

7. *Reject the null hypothesis* $\delta \leq 1$ *if* $\hat{\delta} - \frac{\hat{c}^{boot}_{1-\alpha}}{\sqrt{N}} > 1$.

In Step 2, we specify values of the tuning parameters η_{Nx}, $x \in \mathcal{X}$. Given the choice of these tuning parameters, we estmate $(\mathcal{V}\times\mathcal{X})^{\max}$ in Step 3 to be the collection of a pair of partition $\mathbf{V}$ and the covariate value x such that $\sqrt{N_x}(1-\hat{\delta}_x(\mathbf{V}))$ is less than the specified threshold η_{N_x}. The above rate of divergence for η_{N_x} and the construction of $(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}$ guarantees the estimator to be consistent in a certain sense to $(\mathcal{V}\times\mathcal{X})^{\max}$ under the least favorable null $\delta = 1$ (see Lemma A.2 in Appendix A). When the data generating process has $\delta < 1$, then, the probability of obtaining empty $(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}$ will approach one as the sample size gets large. So, for these null hypothesis, the asymptotic rejection probability is zero. The asymptotic argument only governs the speed of divergence of η_{N_x}, and it provides little guidance on how to set their values in practice. We address a practical issue regarding this in the Monte Carlo study of Section 5.

Given $(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}$, in Step 4 and 5, 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 $\sup_{(\mathbf{V},x)\in(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}}\{\cdot\}$. By combining consistency of $(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}$ and bootstrap validity of $\sqrt{N}(\hat{\delta}^*_x(\cdot)-\hat{\delta}_x(\cdot))$ in approximating $G_{\delta,x}(\cdot)$, the statistic $\sup_{(\mathbf{V},x)\in(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}}\left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V})-\hat{\delta}_x(\mathbf{V}))\right\}$ asymptotically replicates the distribution of $\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}}\{G_{\delta,x}(\mathbf{V})\}$.

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

Proposition 3.3 (bootstrap validity) *Under the regularity condition of Proposition 3.2 and the assumption of Lemma A.2 given in Appendix A, the above bootstrap test procedure yields a pointwise asymptotically size correct test for the null* $\delta \leq 1$*, that is, for every data generating process satisfying* $\delta \leq 1$*,*

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

and the equality holds for any null with $\delta = 1$*.*

Proof. See Appendix A. ■

Due to the restriction on the richness of the partition class, the test procedure is not able to screen out all the data generating processes that have $\delta > 1$. In order for asymptotic power of the test to be one against a fixed alternative, the alternative must meet the condition.$\sup_{(\mathbf{V},x)\in\mathcal{V}\times\mathcal{X}}\{\delta_x(\mathbf{V})\} > 1$. This implies that, for continuous Y, a specification of $\mathbb{V}$ from which the partitions are generated affects the asymptotic refutability power of the test procedure. For instance, as we specify a smaller class of partitions, less alternatives can be screened out by the test.

4 Monte Carlo simulations

In order to evaluate the finite sample performance of the proposed test procedures, we conduct Monte Carlo studies for various specifications of data distribution with a binary instrument $Z \in \{1, 2\}$ with no covariates. Since the asymptotically valid test procedure attains the nominal size when $\delta = 1$, we set the integrated envelope equal to one for every specification.

We specify Y to be continuous on the unit interval $\mathcal{Y} = [0, 1]$. As for a specification of the VC-class $\mathbb{V}$, we employ the half unbounded interval class

$$\mathbb{V}_{half} = \{[0, y] : y \in (0, 1]\} \cup \{(y, 1] : y \in (0, 1]\} \cup \emptyset.$$

In particular, the partition class $\mathcal{V}_{half}$ is given by

$$\mathcal{V}_{half} = \{(V, V^c) : V \in \mathbb{V}_{half}\}.$$

Our Monte Carlo specifications all satisfy the optimal partition condition of condition (A2).

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 $p_1(\cdot)$ and $p_2(\cdot)$ are simulated (see Figure 2). We denote size of the sample drawn from $p_1(\cdot)$ and $p_2(y)$ by n_1 and n_2 respectively.

Design 1: *No ties,* $p_1(y) = 0.54 \times \phi(0.65, 0.10)$,
$p_2(y) = 0.54 \times \phi(0.35, 0.10)$,

Design 2: *No ties,* $p_1(y) = 0.84 \times \phi(0.60, 0.20)$,
$p_2(y) = 0.75 \times \phi(0.46, 0.23)$,

Design 3: *Partially tied* $p_1(y) = \begin{cases} 0.70 \times \phi(0.50, 0.20) & \text{for } y \leq 0.66 \\ 0.58 \times \phi(0.70, 0.25) & \text{for } y > 0.66 \end{cases}$,

$p_2(y) = \begin{cases} 0.70 \times \phi(0.50, 0.20) & \text{for } y > 0.34 \\ 0.58 \times \phi(0.30, 0.25) & \text{for } y \leq 0.34 \end{cases}$

Design 4: *Completely tied,* $p_1(y) = p_2(y) = \phi(0.50, 0.23)$.

In Design 1 and Design 2, there are no ties between $p_1(y)$ and $p_2(y)$, while $p_1(y)$ and $p_2(y)$ differ more significantly in Design 1 than in Design 2. Design 3 represents the case where $p_1(y)$ and $p_2(y)$ are tied on a subset of the outcome support. As an extreme case, Design 4 features a $p_1(y)$ that is identical to $p_2(y)$.

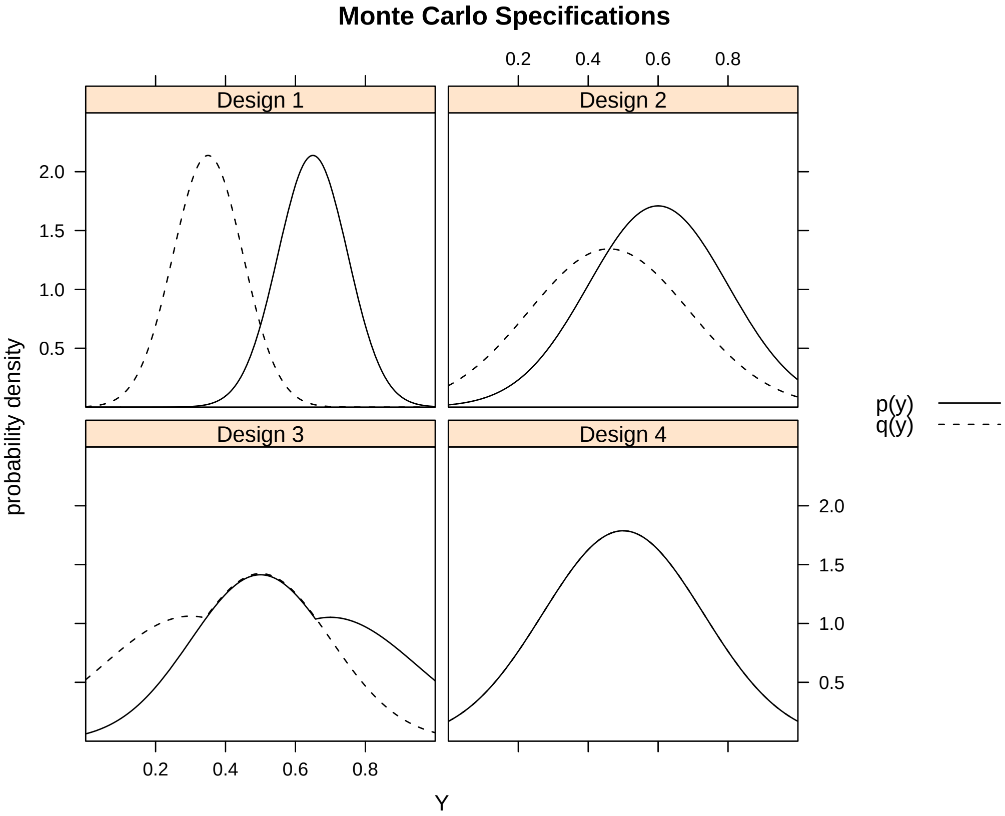


Figure 2: *There are no ties in Design 1 and Design 2. In Design 3, the two densities are partially tied. In Design 4, the two densities are identical.*

We estimate the critical values using four different methods. The first method uses the critical values implied from asymptotic normality of the statement of asymptotic normality in Proposition 3.1. Here, the variance-covariance matrix is estimated by

$$\hat{\sigma}^2 = \frac{1}{2}P_1(\hat{V}_1^{\max})(1 - P_1(\hat{V}_1^{\max})) + \frac{1}{2}P_2((\hat{V}_1^{\max})^c)(1 - P_2((\hat{V}_1^{\max})^c))$$

with $\hat{V}_1^{\max} = \arg\sup_{V_1 \in \mathbb{V}} \left\{ \hat{P}_1(V_1) + \hat{P}_2(V_1^c) \right\}$.

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, which also provides pointwise asymptotically valid critical values. We consider three different choices of the blocksizes, $(b_{n_1}, b_{n_2}) = (n_1/3, n_2/3), (n_1/6, n_2/6)$, and $(n_1/10, n_2/10)$. 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 Proposition 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 $p_1(y)$ and $p_2(y)$ 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 $(n_1/10, n_2/10)$. However, if the blocksize is large such as $(n_1/3, n_2/3)$, the test performance is as bad as the normal approximation. Although subsampling can be shown to be valid for any data generating processes, the simulation results suggest that the subsampling can be 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 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		$n_1= n_2= 300$				$n_1= n_2= 1000$			
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	$(n_1/3, n_2/3)$	31.6%	16.1%	10.7%	4.4%	29.4%	15.4%	10.6%	4.1%
	$(n_1/6, n_2/6)$	27.5%	13.5%	7.6%	2.4%	26.6%*	12.8%	7.6%	2.4%
	$(n_1/10, n_2/10)$	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		$n_1= n_2= 300$				$n_1= n_2= 1000$			
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	$(n_1/3, n_2/3)$	38.8%	20.0%	13.6%	5.7%	33.9%	18.5%	12.5%	4.9%
	$(n_1/6, n_2/6)$	30.3%	14.8%	9.0%	3.1%	28.2%	13.4%	7.6%	2.4%
	$(n_1/10, n_2/10)$	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		$n_1 = n_2 = 300$				$n_1 = n_2 = 1000$			
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	$(n_1/3, n_2/3)$	53.0%	32.6%	23.6%	10.5%	52.0%	33.7%	24.5%	10.8%
	$(n_1/6, n_2/6)$	42.7%	23.7%	15.2%	5.7%	43.3%	24.8%	15.5%	5.9%
	$(n_1/10, n_2/10)$	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		$n_1 = n_2 = 300$				$n_1 = n_2 = 1000$			
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	$(n_1/3, n_2/3)$	82.7%	63.6%	49.3%	23.4%	83.4%	63.6%	45.8%	22.9%
	$(n_1/6, n_2/6)$	69.6%	43.3%	31.4%	13.2%	67.7%	41.5%	27.4%	10.9%
	$(n_1/10, n_2/10)$	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 seem to depend 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, i.e., as the maximizer partition class $\mathcal{V}^{\max}$ becomes larger.

The tables also provide simulation results for the bootstrap procedure used in Blundell et al. (2007).[12] 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

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

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.

5 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. In order to guarantee moderate sample size for each covariate group, we coarsen the covariate values of age by forming four age groups, 23 -30, 31 - 38, 39 - 46, and 47 - 54. Note that conditional independence of Z and Y given such coarsened X is not implied by conditional indepedence of Z and Y given X. Therefore, rejecting the null in the following test procedure does not allow us to refute conditional independence of Z and Y given X. How to incorporate the covariates with many points of support into the test procedure is left for future research.

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. Therefore, we treat the instrument as a discrete variable with ten points of support $\{z_{k,x}\}_{k=1}^{10}$.

As a VC-class to generate the partitions, we consider in this section the *histogram class* $\mathbb{V}_{hist}$, which is defined as the power set of histogram bins with a fixed width whose breakpoints can float over $\mathbb{R}$. Algebraically, the histogram class is defined as follows. Let $h > 0$ be the bin width and L the number of bins. Pick an initial breakpoint $y_0 \in \mathbb{R}$ and consider equally distanced L points $-\infty < y_0 < y_1 < \cdots < y_{L-1} < \infty$ where $y_l = y_0 + lh$, $l = 1, \ldots, (L-1)$. Denote the $(L+1)$ disjoint intervals formed by these L points by $H_0(y_0, h) = (-\infty, y_0]$, $H_l(y_0, h) = (y_{l-1}, y_l]$, $l = 1, \ldots, (L-1)$, and $H_L(y_0, h) = (y_{L-1}, \infty)$. Let $I_j(L)$, $j = 1, \ldots, 2^{L+1}$ indicate all the possible subsets of the indices $\{0, 1, \ldots, L\}$. Given $\mathcal{Y}_0$ a set of the smallest breakpoint y_0, the histogram class with bin width h and the number of bins L

is expressed as

$$\mathbb{V}_{hist}(h,L,\mathcal{Y}_0)=\left\{\bigcup_{l\in I_j(L)} H_l(y_0,h): y_0\in\mathcal{Y}_0,\ j=1,\ldots,2^{L+1}\right\}. \tag{5.1}$$

where the binwidth can be an additional tuning parameter. In our testing procedure below, we specify 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. Perhaps, the sensitive part in this testing procedure is how to choose a reasonable value of the slackness sequence η_{N_x}. 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 within the group $\{X_i=x, Z_i=z_{k,x}\}$. 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_{Nx}, we consider the grid points between 0.5 and 12 with grid size 0.5. After simulating the empirical coverage for each η_{Nx}, we search the value of η_{Nx} that yields the best empirical coverage in terms of minimizing the squared discrepancy from the nominal coverage,

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

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 $p_{k,x}(y)$.

It is commonly observed in each covariate group that $p_{k,x}(y)$ tends to shift to the right for as the out-of-work income becomes higher. 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, and more productive workers flow to the labor market as the out-of-work income gets higher. Rejecting the null by our specification test can empirically refute the latter hypothesis.

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

Some college education

	Male				Female			
	N_x	$\Pr(D=1\|x)$	p-value	$\eta^*_{N_x}$	N_x	$\Pr(D=1\|x)$	p-value	$\eta^*_{N_x}$
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_x	$\Pr(D=1\|x)$	p-value	$\eta^*_{N_x}$	N_x	$\Pr(D=1\|x)$	p-value	$\eta^*_{N_x}$
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.

Table 2 shows the result of the bootstrap specification test.[13] $\eta^*_{N_x}$ 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.

6 Concluding remarks

From the partial identification point of view, this paper analyzes the identification region under the restriction of instrument independence in the selection model. We focus on the integrated envelope, which is the key parameter for examining the emptiness of the identification region. 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

[13] 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.

densities have ties.

The estimation of the ties uses the slackness variable η_{N_x}. The Monte-Carlo simulations show that given the appropriate choice of η_{Nx}, the proposed bootstrap approximates the finite sample distribution of the statistic accurately. Although the optimal η_{Nx} seems to depend on the true data generating process and the test performance is rather sensitive to a choice of η_{N_x}, our simulation results indicate that the bootstrap outperforms subsampling over a reasonable range of values of η_{N_x}. This paper does not provide a formal analysis on how to choose η_{N_x} nor uniform validity of the test procedure (cf. Andrews and Guggenberger (2009), Andrews and Soares (2010), Romano and Shaikh (2010)), and these issues are left for furture research. In the empirical application, we search the optimal value of η_{Nx} through the Monte Carlo simulations where the population data generating process is substituted by its parametric estimate. This way of tuning η_{Nx} 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 measure of out-of-work income is not independent of the potential wages given the coarsen covariates. 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 $P = \{(p_{1,x}, \ldots, p_{k,x})\}_{x\in\mathcal{X}}$ be given by data and assume $\delta_x \leq 1$ for every $x \in \mathcal{X}$. Let $\mathcal{F}^*$ be the set of conditional distributions of Y given X,

$$\mathcal{F}^* = \left\{ \left\{ f_{Y|X}(y|X=x) \right\}_{x\in\mathcal{X}} : \text{for every } x \in \mathcal{X}, \int_{\mathcal{Y}} f_{Y|X}(y|X=x)d\mu = 1 \right.$$
$$\left. \text{and } f_{Y|X}(y|X=x) \geq \underline{f_{Y|X}}(y|X=x) \ \mu\text{-a.e.} \right\}$$

For an arbitrary $f_{Y|X} \in \mathcal{F}^*$, we shall construct a joint probability law of (Y, D, Z) given X that is compatible with the data generating process P and the identifying restriction ER. Since the distribution of Z given X is irrelevant to the analysis, we focus on the conditional law of (Y, D) given Z and X. Let B be an arbitrary Borel set in $\mathcal{Y}$. In order for the conditional law of (Y, D) given Z and X to be compatible with the data generating process, we need to have, for every $k = 1, \ldots, K$ and $x \in \mathcal{X}$,

$$\Pr(Y \in B, D = 1|Z = z_k, X = x) = \int_B p_{k,x}(y)d\mu,$$

Pin down the probability distribution of $\{Y \in B, D = 0\}$ given $Z = z_k$ and $X = x$ to

$$\Pr(Y \in B, D = 0|Z = z_k, X = x) = \int_B [f_{Y|X}(y|X = x) - p_{k,x}(y)]d\mu.$$

Note that the constructed probabilities are nonnegative by construction and they satisfy ER since $\Pr(Y \in B|Z = z_k, X = x) = \int_B f_{Y|X}(y|X = x)d\mu$. This implies each $f_{Y|X} \in \mathcal{F}^*$ is contained in the identification region under ER.

On the other hand, consider a conditional distribution $f_{Y|X} \notin \mathcal{F}^*$. Then, there exists $x^* \in \mathcal{X}$, $z_{k^*} \in \mathcal{Z}$, and a Borel set A with $\mu(A) > 0$ such that

$$\int_A [f_{Y|X}(y|X = x^*) - p_{k^*,x^*}(y)]d\mu < 0. \tag{A.1}$$

Note that the probabilities of $\{Y \in A, D = 0\}$ given $Z = z_{k^*}$ and $X = x^*$ are written as

$$\begin{aligned}\Pr(Y \in A, D = 0|Z = z_{k^*}, X = x^*) &= \Pr(Y \in A|Z = z_{k^*}, X = x^*) - \Pr(Y \in A, D = 1|Z = z_{k^*}, X = x^*) \\ &= \int_A [f_{Y|Z,X}(y|Z = z_{k^*}, X = x^*) - p_{k^*,x^*}(y)]d\mu\end{aligned}$$

If ER is true, $f_{Y|Z,X} = f_{Y|X}$ must hold. Then, by (A.1) the above probability becomes negative, and therefore we cannot construct a conditional law of (Y, D) given Z and X that is compatible with the data generating process and ER.

Thus, we conclude $\mathcal{F}^*$ is the identification region under ER.

The statement of (ii) is proved as follows. If $\delta_x > 1$ for some $x \in \mathcal{X}$, then no probability density function $f_{Y|X}(y|X = x)$ can cover the entire envelope $\underline{f_{Y|X}}(y|X = x)$ at these x's since the probability density must be integrate to one. On the other hand, if $\delta_x \leq 1$ for all $x \in \mathcal{X}$, there clearly exist some probability densities $f_{Y|X}(y|X = x)$ that can cover the envelope $\underline{f_{Y|X}}(y|X = x)$. Hence, the conclusion follows. ■

Notation: For the rest of this appendix A, 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 K i.i.d infinite sequences of $\{Y_{data,i}(\omega) : Z_i = z_k\}$, $k = 1, \ldots, K$. We abbreviate almost surely with respect to $\mathbb{P}$ by "a.s." and infinitely often by "i.o.". $\mathbb{V}$ always stands for a class of subsets in $\mathcal{Y}$ with respect to which the uniform convergence of the empirical distributions $\{\hat{P}_k(\cdot)\}_{k=1}^K$ holds such as a VC-class. The class of partitions generated by $\mathbb{V}$ is denoted by $\mathcal{V}$ and each partition is denoted by $\mathbf{V} = (V_1, \ldots, V_K)$. We equip $\mathcal{V}$ with the seminorm $d_\rho(\mathbf{V}, \mathbf{V}') = \sum_{k=1}^K \rho(V_k \triangle V_k')$ where ρ denotes a finite nonnegative measure on $\mathcal{B}(\mathcal{Y})$ such that ρ is absolutely continuous with respect to μ and $\rho(B) \geq \max_k \{P_k(B)\}$ holds for any $B \in \mathcal{B}(\mathcal{Y})$. Note that such ρ always exists by the definition of $P_k(\cdot)$. Let $(\hat{P}_k - P_k)(V) \equiv \hat{P}_k(V) - P_k(V)$. $\mathcal{L}_\rho^2$ refers to the space of functions $f : \mathcal{Y} \to \mathbb{R}$ with $\left[\int f^2 d\rho\right]^{1/2} < \infty$. 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 $G_{P_k,n_k}(\cdot) \equiv \sqrt{n_k}(\hat{P}_k - P_k)(\cdot)$. For a nonmeasurable event A, $\mathbb{P}^*(A)$ indicates the outer probability (see van der Vaart and Wellner (1996) for the definition).

We first provide a lemma that will be used in the proof of Proposition 3.1.

Lemma A.1. *Assume condition (A1) through (A3). Let $\hat{\mathbf{V}} = (\hat{V}_1, \ldots, \hat{V}_K)$ and $\hat{\mathbf{V}}^{\max} = (\hat{V}_1^{\max}, \ldots, \hat{V}_K^{\max})$ be sequences of random partitions as defined in condition (A3). Then, $d_\rho(\hat{\mathbf{V}}, \hat{\mathbf{V}}^{\max}) \to 0$ as $N \to \infty$ a.s.*

Proof of Lemma A.1. We first show $\left|\delta(\hat{\mathbf{V}}) - \delta\right| \to 0$ a.s. By condition (A2), $\mathcal{V}^{\max}$ is nonempty, and let us pick an arbitrary element $\mathbf{V}^{\max} = (V_1^{\max}, \ldots, V_K^{\max}) \in \mathcal{V}^{\max}$. By noting $\delta(\mathbf{V}) = \hat{\delta}(\mathbf{V}) - \sum_{k=1}^K (\hat{P}_k - P_k)(V_k)$, we have

$$\begin{aligned} 0 &\leq \delta - \delta(\hat{\mathbf{V}}) = \delta(\mathbf{V}^{\max}) - \delta(\hat{\mathbf{V}}) \\ &= \hat{\delta}(\mathbf{V}^{\max}) - \hat{\delta}(\hat{\mathbf{V}}) + \sum_{k=1}^K (\hat{P}_k - P_k)(\hat{V}_k) - \sum_{k=1}^K (\hat{P}_k - P_k)(V_k^{\max}) \\ &\leq \sum_{k=1}^K (\hat{P}_k - P_k)(\hat{V}_k) - \sum_{k=1}^K (\hat{P}_k - P_k)(V_k^{\max}) \\ &\to 0 \text{ as } N \to \infty \text{ a.s.} \end{aligned}$$

by condition (A1) and the Glivenko-Cantelli theorem. Thus, $\delta(\hat{\mathbf{V}})$ converges to δ a.s.

Note that the function $\delta(\cdot)$ is continuous on $\mathcal{V}$ with respect to the semimetric d_ρ since, for $\mathbf{V}$, $\mathbf{V}' \in \mathcal{V}$,

$$
\begin{aligned}
\left|\delta(\mathbf{V}) - \delta(\mathbf{V}')\right| &\leq \sum_{k=1}^{K} \left|P_k(V_k) - P_k(V_k')\right| \\
&\leq \sum_{k=1}^{K} P_k(V_k \triangle V_k') \\
&\leq \sum_{k=1}^{K} \rho(V_k \triangle V_k') \\
&= d_\rho(\mathbf{V}, \mathbf{V}').
\end{aligned}
$$

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

Proof of Proposition 3.1. Our proof consists of two steps. First, we prove the uniform convergence result for the function $\hat{\delta}(\cdot)$, i.e., $\sqrt{N}(\hat{\delta}(\cdot) - \delta(\cdot)) \rightsquigarrow G_\delta(\cdot)$ where $G_\delta(\cdot)$ is a partition-indexed tight Gaussian process in $l^\infty(\mathcal{V})$. In the second step, we show that the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta)$ is obtained by $\sup_{\mathbf{V} \in \mathcal{V}^{\max}}\{G_\delta(\mathbf{V})\}$.

Let $G_{\delta,N}(\mathbf{V}) = \sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V}))$. By construction, $G_{\delta,N}(\mathbf{V})$ can be written as the sum of the independent empirical processes,

$$
\begin{aligned}
G_{\delta,N}(\mathbf{V}) &= \sum_{k=1}^{K} \hat{\lambda}_k^{-1} \sqrt{n_k}[\hat{P}_k(V_k) - P_k(V_k)] \\
&= \sum_{k=1}^{K} \hat{\lambda}_k^{-1} G_{P_k,n_k}(V_k).
\end{aligned}
$$

Note that, by condition (A1), the asymptotic distribution of $G_{\delta,N}(\mathbf{V})$ for a fixed $\mathbf{V} \in \mathcal{V}$ is Gaussian. So, in order to prove the uniform convergence of $G_{\delta,N}(\mathbf{V})$, it suffices to show that $G_{\delta,N}(\mathbf{V})$ is asymptotically uniformly equicontinuous in probability with respect to the semimetric d_ρ, and that the index set $\mathcal{V}$ is totally bounded with respect to the semimetric d_ρ (see Theorem 1.5.4 and 1.5.7 of van der Vaart and Wellner (1996)).

Under condition (A1), for each $k = 1, \ldots, K$, $G_{P_k,n_k}(V_k)$ is asymptotically uniformly equicontinuous in probability, so for arbitrary $\epsilon > 0$ and $\eta > 0$, there exists $\alpha_k > 0$ such that

$$
\lim_{N\to\infty} \inf \mathbb{P}^* \left(\sup_{P_k(V \triangle V') < \alpha_k} \left|G_{P_k,n_k}(V) - G_{P_k,n_k}(V')\right| \leq \hat{\lambda}_k^{1/2} \frac{\epsilon}{K} \right) > 1 - \eta
$$

holds. By the definition of $G_{\delta,N}(\mathbf{V})$,

$$
\left|G_{\delta,N}(\mathbf{V}) - G_{\delta,N}(\mathbf{V}')\right| \leq \sum_{k=1}^{K} \hat{\lambda}_k^{-1/2} \left|G_{P_k,n_k}(V_k) - G_{P_k,n_k}(V_k')\right| \tag{A.2}
$$

holds. Let $\alpha = \min_k \{\alpha_k\}$. The way that we define the semimetric d_ρ implies that, for any $\mathbf{V} = (V_1, \ldots, V_K)$ and $\mathbf{V}' = (V_1', \ldots, V_K')$ with $d_\rho(\mathbf{V}, \mathbf{V}') \leq \alpha$, $P_k(V_k \triangle V_k') \leq \alpha_k$ holds. Hence, the above inequality (A.2) implies

$$
\sup_{d_\rho(\mathbf{V},\mathbf{V}') \leq \alpha} \left|G_{\delta,N}(\mathbf{V}) - G_{\delta,N}(\mathbf{V}')\right| \leq \sum_{k=1}^{K} \hat{\lambda}_k^{-1/2} \sup_{P_k(V \triangle V') < \alpha_k} \left|G_{P_k,n_k}(V_k) - G_{P_k,n_k}(V_k')\right|. \tag{A.3}
$$

Consider the event,

$$\left\{ \sup_{P_k(V \triangle V') < \alpha_k} \left| G_{P_k,n_k}(V) - G_{P_k,n_k}(V') \right| \leq \hat{\lambda}_k^{1/2} \frac{\epsilon}{K} \quad \text{for all } k = 1, \ldots, K \right\}.$$

Since the inequality (A.3) shows that the above event implies $\left\{ \sup_{d_\rho(\mathbf{V},\mathbf{V}') \leq \alpha} \left| G_{\delta,N}(\mathbf{V}) - G_{\delta,N}(\mathbf{V}') \right| \leq \epsilon \right\}$, we have

$$\begin{aligned}
& \mathbb{P}^* \left(\sup_{d_\rho(\mathbf{V},\mathbf{V}') \leq \alpha} \left| G_{\delta,N}(\mathbf{V}) - G_{\delta,N}(\mathbf{V}') \right| \leq \epsilon \right) \\
\geq \; & \mathbb{P}^* \left(\sup_{P_k(V \triangle V') < \alpha_k} \left| G_{P_k,n_k}(V) - G_{P_k,n_k}(V') \right| \leq \hat{\lambda}_k^{1/2} \frac{\epsilon}{K} \quad \text{for all } k = 1, \ldots, K \right) \\
= \; & \prod_{k=1}^{K} \mathbb{P}^* \left(\sup_{P_k(V \triangle V') < \alpha_k} \left| G_{P_k,n_k}(V) - G_{P_k,n_k}(V') \right| \leq \hat{\lambda}_k^{1/2} \frac{\epsilon}{K} \right),
\end{aligned}$$

where the last line follows since the K-empirical processes $\left\{ G_{P_k,n_k}(\cdot) \right\}_{k=1}^{K}$ are mutually independent given the sequence of Z_i. Note for any nonnegative bounded sequences a_N and b_N, $\liminf a_N b_N \geq \liminf a_N \liminf b_N$, so we obtain

$$\begin{aligned}
& \liminf \mathbb{P}^* \left(\sup_{d_\rho(\mathbf{V},\mathbf{V}') \leq \alpha} \left| G_{\delta,N}(\mathbf{V}) - G_{\delta,N}(\mathbf{V}') \right| \leq \epsilon \right) \\
\geq \; & \prod_{k=1}^{K} \liminf \mathbb{P}^* \left(\sup_{P_k(V \triangle V') < \alpha_k} \left| G_{P_k,n_k}(V) - G_{P_k,n_k}(V') \right| \leq \hat{\lambda}_k \frac{\epsilon}{K} \right) \\
> \; & (1 - \eta)^K.
\end{aligned}$$

Since η is arbitrary, $G_{\delta,N}(\mathbf{V})$ is asymptotically uniformly equicontinuous in probability.

Next, we shall show that the semimetric space $(\mathcal{V}, d_\rho)$ is totally bounded. When Y is discrete, the assertion is trivial so that we consider the case with continuous Y and μ being the Lebesgue measure. Since the partition class $\mathcal{V}$ can be seen as a subspace of $\mathcal{B}(\mathcal{Y})^K$, it suffices to show that the semimetric space $(\mathcal{B}(\mathcal{Y})^K, d_\rho)$ is totally bounded. Since $\mathcal{B}(\mathcal{Y})^K$ equiped with the norm d_ρ can be seen as a subspace of K-Cartesian product of $\mathcal{L}^1_\rho = \left\{ f : \int_{\mathcal{Y}} |f| \, d\rho < \infty \right\}$. SInce $\mathcal{L}^1_\rho$ is a Banach space, $(\mathcal{L}^1_\rho)^K$ is totally bounded, and consequently $\mathcal{B}(\mathcal{Y})^K$ equiped with the norm d_ρ is also totally bounded. Hence, by Theorem 1.5.4 and 1.5.7 of van der Vaart and Wellner (1996), $G_{\delta,N}(\mathbf{V})$ converges weakly to a tight Gaussian processes $G_\delta(\mathbf{V})$ in $l^\infty(\mathcal{V})$. The covariance function of $G_\delta(\mathbf{V})$ is obtained by noting that $G_\delta(\mathbf{V})$ is expressed as the weighted sum of the independent Brownian bridges in $l^\infty(\mathbb{V})$.

In the second step, using $G_{\delta,N}(\mathbf{V}) \rightsquigarrow G_\delta(\mathbf{V})$, we shall show that the difference between $\sqrt{N}(\hat{\delta} - \delta)$ and $\sup_{\mathbf{V} \in \mathcal{V}^{\max}} \{ \sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V})) \}$ is asymptotically negligible. Since $\delta(\mathbf{V}) = \delta$ on $\mathcal{V}^{\max} \subset \mathcal{V}$,

$$\begin{aligned}
\sup_{\mathbf{V} \in \mathcal{V}^{\max}} \{ \sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V})) \} &= \sup_{\mathbf{V} \in \mathcal{V}^{\max}} \{ \sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta) \} \\
&\leq \sup_{\mathbf{V} \in \mathcal{V}} \{ \sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta) \} = \sqrt{N}(\hat{\delta} - \delta)
\end{aligned}$$

holds. Let $\hat{\mathbf{V}}$ be and $\hat{\mathbf{V}}^{\max}$ be the maximizer of $\hat{\delta}(\cdot)$ on $\mathcal{V}$ and $\mathcal{V}^{\max}$ respectively, which are assumed to exist by condition (A3). Then,

$$\begin{aligned}
0 &\leq \sqrt{N}(\hat{\delta} - \delta) - \sup_{\mathbf{V} \in \mathcal{V}^{\max}} \left\{ \sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta) \right\} \\
&= \sqrt{N}(\hat{\delta}(\hat{\mathbf{V}}) - \hat{\delta}(\hat{\mathbf{V}}^{\max})) \\
&= \sum_{k=1}^{K} \hat{\lambda}_k^{-1/2} \left[G_{P_k,n_k}(\hat{V}_k) - G_{P_k,n_k}(\hat{V}_k^{\max}) \right]
\end{aligned}$$

By Lemma A.1, we have $d_\rho(\hat{\mathbf{V}}, \hat{\mathbf{V}}^{\max}) \to 0$ a.s. and this implies $P_k(\hat{V}_k, \hat{V}_k^{\max}) \to 0$ a.s. for all $k = 1, \ldots, K$. Then, the asymptotic stochastic equicontinuity for $G_{P_k,n_k}(\cdot)$ implies that $G_{P_k,n_k}(\hat{V}_k) - G_{P_k,n_k}(\hat{V}_k^{\max}) \to 0$ in outer probability. Thus, we conclude $\sqrt{N}(\hat{\delta} - \delta) - \sup_{\mathbf{V} \in \mathcal{V}^{\max}}\{\sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V}))\} = o_{P^*}(1)$ and the asymptotic distribution of $\sqrt{N}(\hat{\delta} - \delta(P, Q))$ is identical to that of $\sup_{\mathbf{V} \in \mathcal{V}^{\max}}\{\sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V}))\} = \sup_{\mathbf{V} \in \mathcal{V}^{\max}}\{G_{\delta,N}(\mathbf{V})\}$. Since the supremum functional $\sup_{\mathbf{V} \in \mathcal{V}^{\max}}\{\cdot\}$ on $l^\infty(\mathcal{V})$ is continuous (with respect to the sup metric), the continuous mapping theorem and $G_{\delta,N}(\mathbf{V}) \rightsquigarrow G_\delta(\mathbf{V})$ yields the desired result,

$$\sup_{\mathbf{V} \in \mathcal{V}^{\max}} \{\sqrt{N}(\hat{\delta}(\mathbf{V}) - \delta(\mathbf{V}))\} \rightsquigarrow \sup_{\mathbf{V} \in \mathcal{V}^{\max}} \{G_\delta(\mathbf{V})\}.$$

■

Regularity conditions for Proposition 3.2.

We impose the following regularity conditions for Proposition 3.2.

Condition B

(B1) *Uniform Convergence*: For each $k = 1, \ldots, K$ and $x \in \mathcal{X}$, the set indexed empirical processes, $G_{P_{k,x},n_{k,x}}(V) \equiv \sqrt{n_{k,x}}(\hat{P}_{k,x}(V) - P_{k,x}(V))$ converge uniformly in law to tight mean zero Gaussian processes in $l^\infty(\mathcal{V})$:

$$G_{P_{k,x},n_{k,x}}(V) \rightsquigarrow G_{P_{k,x}}(\mathcal{V}).$$

(B2) *Optimal partition*: For each $x \in \mathcal{X}$, there exists a nonempty *maximizer partition class* $\mathcal{V}_x^{\max} \subset \mathcal{V}$ *defined by*

$$\mathcal{V}_x^{\max} = \{\mathbf{V} \in \mathcal{V} : \delta_x(\mathbf{V}) = \delta_x\}.$$

In addition, there exists a nonempty subclass in $(\mathcal{V} \times \mathcal{X})^{\max}$ such that

$$(\mathcal{V} \times \mathcal{X})^{\max} = \{(\mathbf{V}, x) \in \mathcal{V} \times \mathcal{X} : \delta_x(\mathbf{V}) = \delta\}.$$

(B3) *Existence of Maximizer*: With probability one, there exists for each $x \in \mathcal{X}$ a sequence of random partitions $\hat{\mathbf{V}}_{N_x} \in \mathcal{V}$ and $\hat{\mathbf{V}}_{N_x}^{\max} \in \mathcal{V}^{\max}$ such that for all $N_x \geq 1$,

$$\hat{\delta}_x(\hat{\mathbf{V}}_{N_x}) = \sup_{\mathbf{V} \in \mathcal{V}} \left\{\hat{\delta}_x(\mathbf{V})\right\}, \quad \hat{\delta}(\hat{\mathbf{V}}_{N_x}^{\max}) = \sup_{\mathbf{V} \in \mathcal{V}^{\max}} \left\{\hat{\delta}_x(\mathbf{V})\right\}$$

holds.

Proof of Proposition 3.2. We will first show the stochastic process indexed by $(\mathbf{V}, x) \in \mathcal{V} \times \mathcal{X}$, $\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))$, converges uniformly to a tight Gaussian process $G_\delta(\mathbf{V}, x)$ in $l^\infty(\mathcal{V} \times \mathcal{X})$. Since $\mathcal{X}$ is a finite set, what matters for the sample path continuity is only in the coordinate of partition $\mathbf{V}$. Furthermore, since we make the analysis conditional on the sequence of subsample size $n_{k,x}$, $\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))$ and $\sqrt{N}(\hat{\delta}_{x'}(\mathbf{V}) - \delta_{x'}(\mathbf{V}))$ are independent for every $x, x' \in \mathcal{X}$. So, in order to prove the claim of uniform convergence, it suffices to show that for each $x \in \mathcal{X}$, $\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))$ converges to a tight Gaussian process in $l^\infty(\mathcal{V})$. Given regularity condition B, this can be shown by replicating the proof of Proposition 3.1, and we obtain

$$\begin{aligned}
\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V})) &= \sum_{k=1}^{K} \frac{\sqrt{N}}{\sqrt{n_{k,x}}} \sqrt{n_{k,x}}(\hat{P}_{k,x}(V_k) - P_{k,x}(V_k)) \\
&\rightsquigarrow \sum_{k=1}^{K} \lambda_{k,x}^{-1/2} G_{P_{k,x}}(V_k) \equiv G_\delta(\mathbf{V}, x).
\end{aligned}$$

where the covariance structure of $G_\delta(\mathbf{V}, x)$ is given by

$$Cov(G_\delta(\mathbf{V}, x), G_\delta(\mathbf{V}', x')) = 1\left\{x = x'\right\} \sum_{k=1}^{K} \lambda_{k,x}^{-1} \left[P_{k,x}(V_k \cap V_k') - P_{k,x}(V_k)P_k(V_k')\right].$$

where $\lambda_{k,x} = \Pr(Z = z_k, X = x)$.

In the next step, we shall show that $\sqrt{N}(\hat{\delta} - \delta)$ weakly converges to $\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \{G_\delta(\mathbf{V}, x)\}$. Under condition (B2), by repeating the same argument as in the proof of Proposition 3.1,

$$\begin{aligned}
\sqrt{N}(\hat{\delta} - \delta) &= \sup_{(\mathbf{V},x)\in\mathcal{V}\times\mathcal{X}} \left\{\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V})) + \sqrt{N}(\delta_x(\mathbf{V}) - \delta)\right\} \\
&= \max_{x\in\mathcal{X}} \left\{\sup_{\mathbf{V}\in\mathcal{V}} \left\{\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V})) + \sqrt{N}(\delta_x(\mathbf{V}) - \delta_x)\right\} + \sqrt{N}(\delta_x - \delta)\right\} \\
&= \max_{x\in\mathcal{X}} \left\{\sup_{\mathbf{V}\in\mathcal{V}_x^{\max}} \left\{\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))\right\} + o_p(1) + \sqrt{N}(\delta_x - \delta)\right\} \\
&= \max_{x\in\mathcal{X}} \left\{\sup_{\mathbf{V}\in\mathcal{V}_x^{\max}} \{G_\delta(\mathbf{V}, x)\} + o_p(1) + \sqrt{N}(\delta_x - \delta)\right\}.
\end{aligned}$$

Since $\sup_{\mathbf{V}\in\mathcal{V}_x^{\max}} \{G_\delta(\mathbf{V}, x)\}$ almost surely bounded and $\sqrt{N}(\delta_x - \delta)$ diverges to negative infinity for $x \in \mathcal{X}$ with $\sqrt{N}(\delta_x - \delta) < 0$, by letting $\mathcal{X}^{\max} = \{x \in \mathcal{X} : \delta_x = \delta\}$, we obtain ■

$$\sqrt{N}(\hat{\delta} - \delta) = \max_{x\in\mathcal{X}^{\max}} \sup_{\mathbf{V}\in\mathcal{V}_x^{\max}} \{G_\delta(\mathbf{V}, x)\} + o_p(1).$$

By combining the maximum and supremum operators into $\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \{\cdot\}$, we obtain the conclusion,

$$\sqrt{N}(\hat{\delta} - \delta) \rightsquigarrow \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{\sqrt{N}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))\right\}.$$

The next lemma is used to prove Proposition 3.3. It shows that $\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$ constructed in Step 3 of the bootstrap algorithm is consistent to $(\mathcal{V}\times\mathcal{X})^{\max}$.

Lemma A.2. *Assume condition B. Let $\{\eta_N : N \geq 1\}$ and $\{\eta_{N_x} : N_x \geq 1\}$ be positive sequences defined in the description of bootstrap algorithm. Let $\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$ be the estimator of $(V \times X)^{\max}$ constructed in Step 3 of the bootstrap algorithm,*

$$\widehat{(\mathcal{V}\times\mathcal{X})}^{\max} = \left\{(\mathbf{V}, x) \in \mathcal{V} \times \mathcal{X} : \sqrt{N_x}(1 - \hat{\delta}_x(\mathbf{V})) \leq \eta_{N_x}\right\}.$$

Define a semimetric on $(V \times X)$ as $d_{\mathcal{V}\times\mathcal{X}}((V, x), (V', x')) = d_\rho(V, V') + 1\left\{x = x'\right\}$ and define ϵ-cover of the maximizer partition class by

$$(\mathcal{V}\times\mathcal{X})_\epsilon^{\max} = \left\{(\mathbf{V}, x) \in (\mathcal{V}\times\mathcal{X}) : \inf_{(\mathbf{V}',x')\in(\mathcal{V}\times\mathcal{X})^{\max}} \{d_{\mathcal{V}\times\mathcal{X}}((\mathbf{V}, x), (\mathbf{V}', x'))\} \leq \epsilon\right\}.$$

We further assume that for each $\epsilon > 0$, there exists $\zeta(\epsilon) > 0$ such that

$$\inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})_\epsilon^{\max}} \{\delta - \delta_x(\mathbf{V})\} > \zeta(\epsilon)$$

For the estimator $\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$ define a sequence of events

$$A_N^\epsilon = \left\{(\mathcal{V}\times\mathcal{X})^{\max} \subseteq \widehat{(\mathcal{V}\times\mathcal{X})}^{\max} \subseteq (\mathcal{V}\times\mathcal{X})_\epsilon^{\max}\right\}.$$

If the true integrated envelope δ is one, 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.

If the true integrated envelope is less than one, then for a sequence of events $B_N = \left\{\widehat{(\mathcal{V}\times\mathcal{X})}^{\max} = \emptyset\right\}$, $\mathbb{P}\left(\lim_{N\to\infty} \inf B_N\right) = 1$.

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 $G_{P_{k,x},n_{k,x}}(V) = \sqrt{n_{k,x}}(\hat{P}_{k,x}(V) - P_{k,x}(V))$ indexed by $V \in \mathbb{V}$,

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

First, we consider the case of $\delta = 1$. Let $\tau_{N_x,n_{k,x}} = \sqrt{N_x/n_{k,x}}\frac{\sqrt{\log\log n_{k,x}}}{\sqrt{\log\log N_x}}\frac{\sqrt{\log\log N_x}}{\eta_{Nx}}$ where $\tau_{N_x,n_{k,x}} \to 0$ by the specification of η_{Nx}. Then,

$$\sup_{\mathbf{V}\in\mathcal{V}} \left|\frac{\sqrt{N_x}}{\eta_{Nx}}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))\right| \leq \sum_{k=1}^{K} \tau_{N_x,n_{k,x}} \sup_{V\in\mathbb{V}} \left|\frac{G_{P_{k,x},n_{k,x}}(V)}{\sqrt{\log\log n_{k,x}}}\right|.$$

Since $\tau_{N_x,n_{k,x}} \to 0$ as $N \to \infty$, the right hand side of the above inequality converges to zero a.s. by the LIL. Hence, by the finiteness of $\mathcal{X}$, we obtain

$$\lim_{N\to\infty} \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left|\frac{\sqrt{N_x}}{\eta_{Nx}}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))\right| = 0 \qquad \text{a.s.} \tag{A.4}$$

Based on this almost sure result, we will show $\mathbb{P}\left(\liminf\left\{(\mathcal{V}\times\mathcal{X})^{\max} \subseteq \widehat{(\mathcal{V}\times\mathcal{X})}^{\max}\right\}\right) = 1$. Note that, by the construction of $\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$, $(\mathcal{V}\times\mathcal{X})^{\max} \subseteq \widehat{(\mathcal{V}\times\mathcal{X})}^{\max}$ occurs if and only if $\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}}\left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(1 - \hat{\delta}_x(\mathbf{V}))\right\} \leq 1$. Therefore, it suffices to show

$$\limsup \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(1 - \hat{\delta}_x(\mathbf{V}))\right\} \leq 1 \qquad \text{a.s.}$$

When $\delta = 1$, $\delta_x(\mathbf{V}) = 1$ for $(\mathbf{V}, x) \in (\mathcal{V}\times\mathcal{X})^{\max}$, so

$$\begin{aligned}
\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(1 - \hat{\delta}_x(\mathbf{V}))\right\} &= \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(\delta_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V}))\right\} \\
&\leq \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left(\frac{\sqrt{N_x}}{\eta_{Nx}}(\hat{\delta}_x(\mathbf{V}) - \delta_x(\mathbf{V}))\right) \\
&\to 0 \qquad \text{a.s. by (A.4).}
\end{aligned}$$

Hence, $\limsup \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}}\left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(1 - \hat{\delta}_x(\mathbf{V}))\right\} \leq 1$ a.s. holds and $\mathbb{P}\left(\liminf\left\{(\mathcal{V}\times\mathcal{X})^{\max} \subseteq \widehat{(\mathcal{V}\times\mathcal{X})}^{\max}\right\}\right) = 1$ is proved.

Next, we show $\mathbb{P}\left(\liminf\left\{\widehat{(\mathcal{V}\times\mathcal{X})}^{\max} \subseteq (\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}\right\}\right) = 1$. Since the event $\left\{\widehat{(\mathcal{V}\times\mathcal{X})}^{\max} \subseteq (\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}\right\}$ is equivalent to $\inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\setminus(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}}\left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(1 - \hat{\delta}_x(\mathbf{V}))\right\} > 1$, it suffices to show

$$\lim_{N\to\infty} \inf \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\setminus(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{\frac{\sqrt{N_x}}{\eta_{Nx}}(1 - \hat{\delta}_x(\mathbf{V}))\right\} > 1 \qquad \text{a.s.}$$

We obtain from (**??**)

$$
\begin{aligned}
& \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}(1-\hat{\delta}_x(\mathbf{V})) \right\} \\
= & \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}} \left[(1-\delta_x(\mathbf{V})) - (\hat{\delta}_x(\mathbf{V})-\delta_x(\mathbf{V})\right]) \right\} \\
\geq & \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}((1-\delta_x(\mathbf{V})) \right\} \\
& - \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}(\hat{\delta}_x(\mathbf{V})-\delta_x(\mathbf{V}) \right\}
\end{aligned}
$$

Note that the second term in the right hand side of the above inequality has been already proved to converge to zero a.s. For the first term, the assumption implies that there exists $\zeta(\epsilon)>0$ such that $1-\delta_x(V)>\zeta(\epsilon)$ for any $(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}$. Since $\frac{\sqrt{N_x}}{\eta_{N_x}}\to\infty$ for every x, we obtain

$$
\inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})\backslash(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}(1-\hat{\delta}_x(\mathbf{V})) \right\} \to \infty a.s.
$$

Therefore, $\liminf \inf_{V\in\mathbb{V}\backslash\mathbb{V}^{\max}_{\epsilon}} \left\{ \frac{\sqrt{N}}{\eta_N}(\delta-\hat{\delta}(V)) \right\} = \infty$ a.s. and this implies $\mathbb{P}\left(\liminf \left\{ (\widehat{\mathcal{V}\times\mathcal{X}})^{\max} \subseteq (\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon} \right\}\right) = 1$. Combining these two results completes the proof for $\delta=1$.

Next, we consider the case of $\delta<1$. Note that the event $B_N = \left\{ (\widehat{\mathcal{V}\times\mathcal{X}})^{\max} = \emptyset \right\}$ is equivalent to $\left\{ \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}((1-\hat{\delta}_x(\mathbf{V})) \right\} > 1 \right\}$. Since

$$
\begin{aligned}
\inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}((1-\hat{\delta}_x(\mathbf{V})) \right\} \geq & \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}(1-\delta_x(\mathbf{V})) \right\} \\
& - \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}((\hat{\delta}_x(\mathbf{V})-\delta_x(\mathbf{V})) \right\},
\end{aligned}
$$

and $(1-\delta_x(\mathbf{V}))>0$ for every $(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})$. Thus, the first term in the right hand side diverges to positive infinity, and by (A.4) the second term converges to zero a.s. Hence, $\liminf \inf_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})} \left\{ \frac{\sqrt{N_x}}{\eta_{N_x}}((1-\hat{\delta}_x(\mathbf{V})) \right\} > 1$ a.s. holds and it means $\mathbb{P}(\liminf B_N)=1$. ■

Proof of Proposition 3.3. We first consider the data generating process with $\delta<1$. By Lemma A.2, $(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}=\emptyset$ occurs with probability one for large N. Since we do not reject the null if $(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}=\emptyset$, the rejection probability converges to 0 for the data generating process with $\delta<1$.

From now on, we consider the the data generating process with $\delta=1$. For ease of exposition, we indicate an infinite sampling sequence 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 $\omega\in\Omega$,

$$
\sup_{(\mathbf{V},x)\in(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}(\omega)} \left\{ \sqrt{N}(\hat{\delta}^*_x(\mathbf{V})-\hat{\delta}_x(\mathbf{V})(\omega)) \right\} \rightsquigarrow \sup_{V\in\mathbb{V}^{\max}} \{G_\delta(\mathbf{V},x)\}
$$

where $G_\delta(\mathbf{V},x)$ is the Gaussian processes obtained in Proposition 3.2. 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,

$$
\begin{aligned}
\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{ \sqrt{N}(\hat{\delta}^*_x(\mathbf{V})-\hat{\delta}_x(\mathbf{V})(\omega)) \right\} \leq & \sup_{(\mathbf{V},x)\in(\widehat{\mathcal{V}\times\mathcal{X}})^{\max}(\omega)} \left\{ \sqrt{N}(\hat{\delta}^*_x(\mathbf{V})-\hat{\delta}_x(\mathbf{V})(\omega)) \right\} \\
\leq & \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}_{\epsilon}} \left\{ \sqrt{N}(\hat{\delta}^*_x(\mathbf{V})-\hat{\delta}_x(\mathbf{V})(\omega)) \right\} \qquad \text{(A.5)}
\end{aligned}
$$

holds for almost all $\omega \in \Omega$. Let $G^*_{P_{k,x},n_{k,x}}(\cdot) = \sqrt{n_{k,x}}(\hat{P}^*_{k,x} - P_{k,x})(\cdot)$ be bootstrapped empirical processes where $\hat{P}^*_{k,x}$ is 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)),

$$\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V})(\omega)) = \sum_{k=1}^{K} \lambda_{k,x}^{-1/2} G^*_{P_{k,x},n_{k,x}}(V_k) \rightsquigarrow G_\delta(\mathbf{V},x),$$

for almost all ω. Therefore, for the lower bound term and the upper bound term in (A.5), we have

$$\begin{aligned} \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V})(\omega))\right\} &\rightsquigarrow \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \{G_\delta(\mathbf{V},x)\}, \\ \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}_\epsilon} \left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V})(\omega))\right\} &\rightsquigarrow \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}_\epsilon} \{G_\delta(\mathbf{V},x)\}. \end{aligned}$$

Since the tight Gaussian processes $G_\delta(\mathbf{V},x)$ are almost surely continuous with respect to the semimetric $d_{\mathcal{V}\times\mathcal{X}}$, the asymptotic stochastic equicontinuity of the Gaussian processes imply

$$\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V})(\omega))\right\} - \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}_\epsilon} \left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V})(\omega))\right\} \to 0$$

in probability with respect to $\{\mathbb{P}_N : N \geq 1\}$ as $\epsilon \to 0$. Hence, from (A.5), we conclude that

$$\sup_{(\mathbf{V},x)\in\widehat{(\mathcal{V}\times\mathcal{X})}^{\max}(\omega)} \left\{\sqrt{N}(\hat{\delta}^*_x(\mathbf{V}) - \hat{\delta}_x(\mathbf{V})(\omega))\right\} \rightsquigarrow \sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \{G_\delta(\mathbf{V},x)\}.$$

Since $G_\delta(\mathbf{V},x)$ are non-degenerate Gaussian processes on $(\mathcal{V}\times\mathcal{X})^{\max}$ and, therefore, the distribution of $\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \{G_\delta(\mathbf{V},x)\}$ 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 the $(1-\alpha)$-th quantile of $\sup_{(\mathbf{V},x)\in(\mathcal{V}\times\mathcal{X})^{\max}} \{G_\delta(\mathbf{V},x)\}$ in probability with respect to $\{\mathbb{P}_N : N \geq 1\}$ for almost every $\omega \in \Omega$. Hence, for every data generating process with $\delta = 1$,

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

By combining the these two results, we obtain the conclusion. ■

Appendix B: A Comparison with the cdf bounds in Blundell et al. (2007)

In this appendix, we compare the tightest cdf bounds based on the envelope density (2.4) 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 given X, $F_{Y|Z,X}(y|z,x) = E(I\{Y \in (-\infty,y]\}|Z = z, X = x) = E(I\{Y \in (-\infty,y]\}|X = x) = F_{Y|X}(y|x)$, Blundell et al. (2007) use the mean independence bounds of Manski (1994) to construct the bounds for $F_{Y|X}(y|x)$. Using the notation of the main text of this paper, they are expressed as

$$\begin{aligned} & \max_k \left\{P_{k,x}((-\infty,y])\right\} \leq F_{Y|X}(y|x) \\ \leq\ & \min_k \left\{P_{k,x}((-\infty,y]) + P_{k,x}(\{mis\})\right\}. \end{aligned} \tag{B.1}$$

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 $\{Y \leq y\}$ conditional on X does not depend on Z. This restriction is certainly weaker than the conditional statistical independence restriction since the full statistical independence requires that $\Pr(Y \in A|Z = z, X = x)$ for *any* subsets $A \subset \mathcal{Y}$ do not depend on z.

For stating the main result of this section, we define the dominating density among $\{p_{k,x}\}_{k=1}^K$.

Definition B.1 (dominating density) *(i) $p_{k^*,x}(y)$ $\{p_{k,x}\}_{k=1}^K$* has *a dominating density on $A \subset \mathcal{Y}$ if there exists an instrumental value z_{k^*} such that $p_{k^*,x}(y) \geq p_{l,x}(y)$ for all $k^* \neq l$ holds on μ-a.e. $y \in A$.*

If data reveals the dominating density $p_{k^*,x}(y)$, then the rest of $\{p_{k,x}\}_{k=1}^K$ do not provide identifying information for $f_{Y|X}$ further than $p_{k^*,x}(y)$ because the maximal area under $f_{Y|X}$ is occupied by $p_{k^*,x}(y)$ alone. The existence of the dominating density guarantees the interchangeability between max operation and integration, that is,

$$\int_A \max_k \{p_{k,x}(y)\} d\mu = \max_k \left\{ \int_A p_{k,x}(y) d\mu \right\}.$$

if and only if $\{p_{k,x}\}_{k=1}^K$ has a dominating density 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 $\{p_{k,x}\}_{k=1}^K$ has a dominating density on* $(-\infty, y]$ and (y, ∞).
(ii) The naive cdf bounds are tight under ER for all $y \in \mathcal{Y}$ if and only if $\{p_{k,x}\}_{k=1}^K$ has a dominating density on $\mathcal{Y}$.

Proof of Proposition B.1. (i) Fix $y \in \mathcal{Y}$. For the lower bound of the naive cdf bounds,

$$\begin{aligned} \max_k \left\{ \int_A p_{k,x}(y) d\mu \right\} &\leq \int_A \max_k \{p_{k,x}(y)\} d\mu \\ &= \int_{(-\infty,y]} \underline{f_{Y|X}}(y|X = x) d\mu \\ &= \text{the lower bound of the tight cdf bounds.} \end{aligned}$$

Note that the inequality holds in equality if and only if $\{p_{k,x}\}_{k=1}^K$ has a dominating density on $(-\infty, y]$. For the upper bound of the naive cdf bounds,

$$\begin{aligned} \min_k \left\{ P_{k,x}((-\infty, y]) + P_{k,x}(\{mis\}) \right\} &= \min_k \left\{ 1 - \int_{(y,\infty)} p_{k,x}(y) d\mu \right\} \\ &= 1 - \max_k \left\{ \int_{(y,\infty)} p_{k,x}(y) d\mu \right\} \\ &\geq 1 - \int_{(y,\infty)} \underline{f_{Y|X}}(y|X = x) d\mu \\ &= \int_{(-\infty,y]} \underline{f_{Y|X}}(y|X = x) d\mu + 1 - \delta_x \\ &= \text{the upper bound of the tight cdf bounds,} \end{aligned}$$

where the inequality holds in equality if and only if $\{p_{k,x}\}_{k=1}^K$ has a dominating density 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 condition holds in a rather limited situation (see Figure 3).

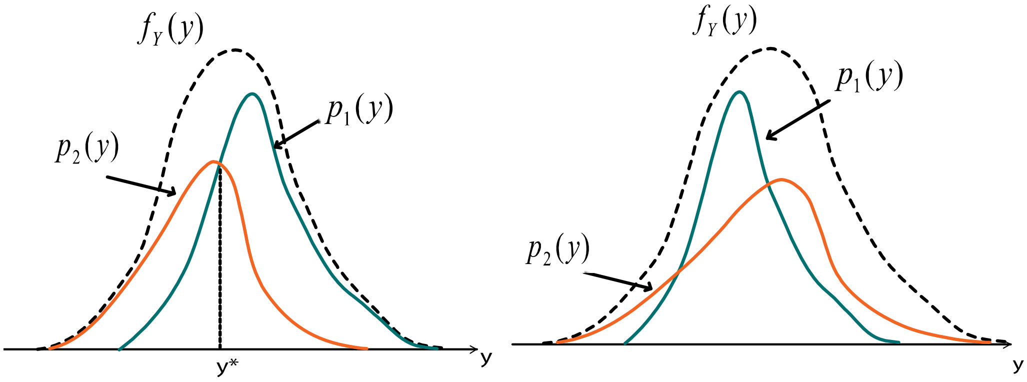


Figure 3: *Consider a model with continuous Y and binary $Z \in \{z_1, z_2\}$, and no covariates. In the left-hand side figure, the naive cdf bounds at y^* are tight. On the other hand, when $p_1(y)$ and $p_2(y)$ are drawn as in the right-hand side figure, the naive cdf bounds are not tight at any $y \in Y$ (Proposition B.1).*

Appendix C: Does selection equation help identify f_Y?

C.1 Identification Region under Joint Independence

The structural selection model formulates the selection mechanism as

$$D = I\{v(Z, X, U) \geq 0\}, \tag{C.1}$$

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. In the case where we believe Z to be independent of any individual unobserved heterogeneities conditional on the observables X, we might want to explicitly impose joint independence between Z and (Y, U) conditional on X. 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 (C.1), it is well known that the joint independence restriction between Z and (Y_1, Y_0, U) yields a narrower identification region for the distribution of the potential outcomes than a pair of marginal independence of Z and Y_1 and Z and Y_0 does.[14] 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 a similar identification gain from the joint independence restriction.

For an ease of exposition, we do not introduce the covariates X into our analysis. Hence, the joint independence restriction to be considered is interpreted as joint independence between Z and (Y, U). The identification analysis given below can be interpreted as the identification analysis for the outcome distribution

[14] 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 a pair of marginal independence restriction $(Y_1 \perp Z)$ and $(Y_0 \perp Z)$.

conditional on each covariate value. We notate the distribution of data by

$$p_k(A) = f_{Y,D|Z}(y, D = 1|Z = z_k),$$

and represent the data generating process by P= $(p_1, \ldots, p_K)$. The density envelope is defined as

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

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 discrete, 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)). 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 $Z = z_k$. Then, 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 C.1 (identification region under joint independence): *Given a data generating process $P = (p_1(y), \ldots, p_K(y))$, 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.*

We denote the distribution of types by $\pi_t = \Pr(T = t)$, $t \in \mathcal{T}$. The source of the nonrandom selection mechanism is the dependence between Y and one's unobserved selection heterogeneities. This dependence is reduced to the dependence between Y and T, and therefore we can allow distinct outcome distributions conditional on each type $T = t$. 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 main result is stated in the next proposition.

Proposition C.1 *$IR_{f_Y}(P|ER)$, is also the identification region of f_Y under joint independence between Z and (Y, U).*

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.

To prove the above proposition, we first provide the next lemma.

Lemma C.2 *If a joint probability distribution on (Y, T, Z) satisfies the joint independence restriction $Z \perp (Y, U)$, then, the following identities hold μ-a.e.* for all $k = 1, \ldots, K$,

$$\begin{aligned} p_k(y) &= \textstyle\sum_{t \in \mathcal{T}_k} h_t(y), \\ f_Y(y) - p_k(y) &= \textstyle\sum_{t \in \mathcal{T} \setminus \mathcal{T}_k} h_t(y), \end{aligned} \tag{C.2}$$

where $h_t(y) = \pi_t g_t(y)$.
Conversely, given a data generating process P, and a marginal distribution of outcome f_Y, if there exist

nonnegative functions $h_t(y)$, $t = c, n, a, d$, that satisfy (C.2) 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.

Lemma C.2 is interpreted that, for a given data generating process $\mathcal{P}$, the set of f_Y for each of which we can find the nonnegative functions $\{h_t(y) : t \in \mathcal{T}\}$ that satisfy *(C.2)*, for all $k = 1, \ldots, K$, constitute the identification region of f_Y under the joint independence restriction.

Proof of Lemma C.2. Assume that a population distribution of (Y, T, Z) satisfies RA. Then,

$$\begin{aligned} p_k(y) &= f_{Y,D|Z}(y, D = 1|Z = z_k) \\ &= f_{Y,T|Z}(y, T \in \mathcal{T}_k|Z = z_k) \\ &= \sum_{t \in \mathcal{T}_k} f_{Y,T|Z}(y, T = t|Z = z_k) \\ &= \sum_{t \in \mathcal{T}_k} f_{Y,T}(y, T = t) \\ &= \sum_{t \in \mathcal{T}_k} \pi_t g_t(y), \end{aligned}$$

which corresponds to the first identity of (C.2). Note that the second line follows since the event $\{Y \in B, D = 1|Z = z_k\}$ is equivalent to $\{Y \in B, T \in \mathcal{T}_k|Z = z_k\}$. The fourth line follows by the joint independence restriction. As for the second identity of (C.2),

$$\begin{aligned} f_Y(y) - p_k(y) &= f_{Y|Z}(y|Z = z_k) - f_{Y,D|Z}(y, D = 1|Z = z_k) \\ &= f_{Y,D|Z}(y, D = 0|Z = z_k) \\ &= f_{Y,T|Z}(y, T \in \mathcal{T} \setminus \mathcal{T}_k|Z = z_k) \\ &= \sum_{t \in \mathcal{T} \setminus \mathcal{T}_k} f_{Y,T|Z}(y, T = t|Z = z_k) \\ &= \sum_{t \in \mathcal{T} \setminus \mathcal{T}_k} \pi_t g_t(y). \end{aligned}$$

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 a marginal distribution f_Y, we have nonnegative functions $h_t(\cdot)$, $t \in \mathcal{T}$ satisfying the constraints (C.2). 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 $f_{Y,T|Z}(y, T = t|Z = z_k)$ and $f_{Y,T|Z}(y, T = t|Z = z_k)$ to be equal to $h_t(y) \geq 0$, $t \in \mathcal{T}$. These yield valid probability measure since $\sum_t \int_{\mathcal{Y}} f_{Y,T|Z}(y, T = t|Z = z_k) = \sum_t \int_{\mathcal{Y}} h_t(y) d\mu = \int_{\mathcal{Y}} f_Y(y) d\mu = 1$, and it satisfies RA by construction. Furthermore, the constructed probability distribution is compatible with the data generating process since $\{h_t(y)\}$ is constructed so as to the identities of (C.2). Thus, the proposed f_Y is contained in the identification region under joint independence.

■

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 (C.2).

Proof of Proposition C.1. If $IR_{f_Y}(P|ER)$ is empty, the identification region under joint independence is clearly empty. So, we assume $IR_{f_Y}(\mathcal{P})$ is nonempty ($\delta \leq 1$).

Pick an arbitrary $f_Y \in IR_{f_Y}(P|ER)$. Our goal is to find the set of nonnegative functions $\{h_t(y)\}_{t \in \mathcal{T}}$ that are compatible with the constraints (C.2).

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 $t \in \mathcal{T} \setminus \mathcal{T}_k$,

$$\bigcup_{t\in\mathcal{T}_k} A(t) = \bigcup_{t\in\mathcal{T}_k} \left(\mathcal{S}_k \cap \left(\bigcap_{l\neq k:D_l=1} \mathcal{S}_l \right) \cap \left(\bigcap_{l\neq k:D_l=0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y} \right), \tag{C.3}$$

$$\bigcup_{t\in\mathcal{T}\setminus\mathcal{T}_k} A(t) = \bigcup_{t\in\mathcal{T}\setminus\mathcal{T}_k} \left(\mathcal{S}_k^c \cap \left(\bigcap_{l\neq k:D_l=1} \mathcal{S}_l \right) \cap \left(\bigcap_{l\neq k:D_l=0} \mathcal{S}_l^c \right) \cap \mathcal{S}_{f_Y} \right). \tag{C.4}$$

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 (C.3) is the union of mutually disjoint and exhaustive partitions of $S_k \cap \mathcal{S}_{f_Y}$. Therefore, the identities (C.3) and (C.4) are reduced to

$$\bigcup_{t\in\mathcal{T}_k} A(t) = \mathcal{S}_k \cap \mathcal{S}_{f_Y} = \mathcal{S}_k,$$

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

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

$$\bigcup_{t\in\mathcal{T}_k} \Pi_y(A(t)) = \Pi_y(S_k),$$

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

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

$$\sum_{t\in\mathcal{T}_k} Leb\left[\Pi_y(A(t))\right] = p_k(y),$$

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

These equations suggest us to pin down each $h_t(y)$ to $Leb\left[\Pi_y(A(t))\right]$. Each $h_t(y)$ is by construction nonnegative and we can see they agree with the constraints (C.2). Since k is arbitrary, this completes the proof. ■

C.2 Imposing the monotonic selection response to an instrument

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,

$$v(Z,U) = \tilde{v}(Z) - U, \tag{C.5}$$

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 identification gain of imposing the additively separable threshold crossing formulation is examined by adding Imbens and Angrist's monotonicity to our analysis.[15] In this appendix, we refer to the monotonicity of Imbens and Angrist, or equivalently, threshold crossing selection with an additive error, as the *monotonic selection response to an instrument (MSR).* Throughout the analysis, we without loss of generarilty 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-monotonic selection response to an instrument (MSR).

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.

Note that the types in $\mathcal{T}_k \cap (\mathcal{T} \setminus \mathcal{T}_{k+1})$ have $D_k = 1$ and $D_{k+1} = 0$. Therefore, in terms of the selection types, MSR is equivalent to $\sum_{t \in \mathcal{T}_k \cap (\mathcal{T} \setminus \mathcal{T}_{k+1})} \pi_t = 0$. Accordingly, the identification region under joint independence and MSR is defined as follows.

Definition C.2 (identification region under separable utility) Given a data generating process P, 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* $\sum_{t \in \mathcal{T}_k \cap (\mathcal{T} \setminus \mathcal{T}_{k+1})} \pi_t = 0$.

Proposition C.2. *Suppose that a population distribution of (Y,T,Z) satisfies the joint independence and MSR. Then, the data generating process P satisfies*

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

Conversely, given the data generating process $P = (p_1(y), \ldots, p_K(y))$, the identification region under joint independence and MSR is given by

$$\begin{cases} IR_{f_Y}(P|ER) & \text{if } p_1(y) \leq p_2(y) \leq, \ldots, \leq p_K(y) \quad \mu\text{-a.e.} \\ \emptyset & \text{otherwise.} \end{cases} \tag{C.6}$$

[15] 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.

This result says that if the data generating process reveals $p_1(y) \leq p_2(y) \leq, \ldots, \leq p_K(y)$ μ-a.e., the identification region under ER is also the identification region under the restrictions of joint independence and monotonic selection response to an instrument. 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, given that we observe $p_1(y) \leq p_2(y) \leq, \ldots, \leq p_K(y)$ μ-a.e., it can be shown that the $E(Y)$ bounds constructed based upon $IR_{f_Y}(P|ER)$ coincide with the Manski's $E(Y)$ bounds under mean independence.

The empty identification region in (C.6) implies that the condition of $p_1(y) \leq p_2(y) \leq, \ldots, \leq p_K(y)$ 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 the observable densities are nested in the order of the selection probabilities $\Pr(D = 1|Z = z_k)$.

Proof of Proposition C.2. (i) From (C.2), 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.

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 $f_Y \in IR_{f_Y}(P|ER)$. We construct a joint distribution of (Y, T, Z) that is compatible with joint independence 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

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

This construction provides nonnegative $h_t(y)$'s. The constructed $h_t(y)$'s satisfy (C.2) 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). ■

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] Anderson, G., O. Linton, and Y.J. Whang (2009): "Nonparametric estimation of a polarization measure," cemmap working paper 14/09.

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

[4] 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

[5] Andrews, D. W. K. and P. Guggenberger (2009): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, Vol. 25(03), pp 669-709.

[6] 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.

[7] 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.

[8] Andrews, D.W.K. and X.Shi (2009): "Inference Based on Conditional Moment Inequalities," unpublished manuscript, Yale University.

[9] Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," Forthcoming in *Econometrica.*

[10] 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.

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

[12] 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.

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

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

[15] Bugni, F. (2010): "Bootstrap Inference in Partially Identified Models," Forthcoming in *Econometrica.*

[16] Canay, I. A. (2010): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity" Forthcoming in *Journal of Econometrics.*

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

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

[19] Chernozhukov, V., S. Lee, and A. Rosen (2009): "Intersection Bounds: estimation and inference," cemmap working paper 19/09.

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

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

[22] Gronau, R., (1974): "Wage Comparisons – A Selectivity Bias", *Journal of Political Economy,* Vol. 82, pp. 1119-1143.

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

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

[25] Hansen, L.P. (2005): “A Test for Superior Predictive Ability,” *Journal of Business and Economic Statistics*, Vol. 23, pp. 365-380.

[26] Heckman, J.J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, Vol. 47, pp. 153-161.

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

[28] 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.

[29] 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.

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

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

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

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

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

[35] Lee S. and Y.J. Whang (2009): "Nonparametric tests of conditional treatment effects," cemmap working paper 37/09.

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

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

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

[39] 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.

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

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

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

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

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

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

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

[47] 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.

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

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

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

[51] Romano, J. P. and A. M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference,* Vol 139. Issue 9. 2786-2807.

[52] Romano, J. P. and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models", Econometrica, Vol 78, No.1, 169-211.

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

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

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

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