Journal of Econometrics 212 (2019) 522–555

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Instrumental variables and the sign of the average treatment effect[☆]

Cecilia Machado [a], Azeem M. Shaikh [b], Edward J. Vytlacil [c,*]

[a] FGV EPGE Escola Brasileira de Economia e Finanças, Brazil
[b] Department of Economics, University of Chicago, United States
[c] Department of Economics, Yale University, United States

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 17 January 2018
Accepted 27 April 2018
Available online 26 June 2019

JEL classification:
C12
C31
C35
C36

Keywords:
Average treatment effect
Endogeneity
Instrumental variables
Union of moment inequalities
Bootstrap
Uniform validity
Multiple testing
Familywise error rate
Gatekeeping

ABSTRACT

This paper considers identification and inference about the sign of the average effect of a binary endogenous regressor (or treatment) on a binary outcome of interest when a binary instrument is available. In this setting, the average effect of the endogenous regressor on the outcome is sometimes referred to as the average treatment effect (ATE). We consider four different sets of assumptions: instrument exogeneity, instrument exogeneity and monotonicity on the outcome equation, instrument exogeneity and monotonicity on the equation for the endogenous regressor, or instrument exogeneity and monotonicity on both the outcome equation and the equation for the endogenous regressor. For each of these sets of conditions, we characterize when (i) the distribution of the observed data is inconsistent with the assumptions and (ii) the distribution of the observed data is consistent with the assumptions and the sign of ATE is identified. A distinguishing feature of our results is that they are stated in terms of a reduced form parameter from the population regression of the outcome on the instrument. In particular, we find that the reduced form parameter being far enough, but not too far, from zero, implies that the distribution of the observed data is consistent with our assumptions and the sign of ATE is identified, while the reduced form parameter being too far from zero implies that the distribution of the observed data is inconsistent with our assumptions. For each set of restrictions, we also develop methods for simultaneous inference about the consistency of the distribution of the observed data with our restrictions and the sign of the ATE when the distribution of the observed data is consistent with our restrictions. We show that our inference procedures are valid uniformly over a large class of possible distributions for the observed data that include distributions where the instrument is arbitrarily "weak." A novel feature of the methodology is that the null hypotheses involve unions of moment inequalities.

1. Introduction

This paper considers identification and inference about the sign of the average effect of an endogenous regressor on an outcome of interest when an instrumental variable is available. In order to obtain simple, closed-form results and for ease

☆ We would like to thank Ivan Canay, Bo Honore, and Joerg Stoye for helpful comments on this paper. The research of the second author was supported by National Science Foundation Grants DMS-0820310 and SES-1227091 and the Alfred P. Sloan Foundation. The research of the third author was supported by National Science Foundation Grant SES-0551089.

* Corresponding author.
E-mail addresses: cecilia.machado@fgv.br (C. Machado), amshaikh@uchicago.edu (A.M. Shaikh), edward.vytlacil@yale.edu (E.J. Vytlacil).

https://doi.org/10.1016/j.jeconom.2018.04.007
0304-4076/

of exposition, we focus on the case where the outcome of interest Y, endogenous regressor D and instrumental variable Z, whose joint distribution we denote by P, are all binary. In this setting, the endogenous regressor is sometimes referred to as the treatment and the average effect of the endogenous regressor on the outcome of interest is sometimes referred to as the average treatment effect (ATE). We consider four different sets of assumptions: instrument exogeneity, instrument exogeneity and monotonicity on the outcome equation, instrument exogeneity and monotonicity on the equation for the endogenous regressor, or instrument exogeneity and monotonicity on both the outcome equation and the equation for the endogenous regressor. Here, monotonicity in the outcome equation requires that different individuals do not have opposite responses to the endogenous regressor, whereas monotonicity in the equation for the endogenous regressor requires that different individuals do not have opposite responses to the instrumental variable. These conditions generally only provide partial identification of ATE.

For each set of assumptions, we show that the sign of the ATE is identified to be positive if and only if the reduced form parameter

$$\begin{aligned}\Delta(P) &= E_P[Y|Z=1] - E_P[Y|Z=0] \\ &= P\{Y=1|Z=1\} - P\{Y=1|Z=0\} \end{aligned} \tag{1}$$

lies in a particular region that depends only on P and that the sign of the ATE is identified to be negative if and only if $\Delta(P)$ lies in another region that, again, depends only on P. When imposing instrument exogeneity and monotonicity in only the equation for the endogenous regressor, we find that when $\Delta(P)$ is sufficiently large in magnitude and positive (negative), one can conclude that the sign of the ATE is positive (negative). When imposing instrument exogeneity and monotonicity in both the outcome equation and the equation for the endogenous regressor, we find that the sign of the ATE simply equals the sign of $\Delta(P)$. Finally, when imposing only instrument exogeneity or when imposing instrument exogeneity and monotonicity in only the outcome equation, we not only find that the sign of the ATE need not equal the sign of $\Delta(P)$, but that it is possible for $\Delta(P)$ to be so large in magnitude and positive (negative) that one concludes the sign of the ATE is in fact negative (positive). For each set of restrictions, we show further that a value for $\Delta(P)$ sufficiently far from zero implies that our assumptions are false. These results may be viewed as formalizing applied researchers' suspicions of empirical results using instrumental variables when the reduced form parameter is "too large" (or, by re-scaling appropriately, when the usual linear instrumental variables estimand is "too large" – see Remark 2.1).

Our analysis reveals that instrument exogeneity alone results in the same ability to determine the sign of the average treatment effect as instrument exogeneity and monotonicity in the equation for the endogenous regressor; instrument exogeneity and monotonicity in the equation for the endogenous regressor has less ability to determine the sign of the average treatment effect than instrument exogeneity and monotonicity in the outcome equation; and instrument exogeneity and monotonicity in the outcome equation has less ability to determine the sign of the average treatment effect than instrument exogeneity and monotonicity in both the outcome equation and the equation for the endogenous regressor. On the other hand, instrument exogeneity alone imposes weaker testable restrictions than instrument exogeneity and monotonicity in the outcome equation; instrument exogeneity and monotonicity in the outcome equation imposes weaker testable restrictions than instrument exogeneity and monotonicity in the equation for the endogenous regressor; and instrument exogeneity and monotonicity in the equation for the endogenous regressor imposes the same testable restrictions as instrument exogeneity and monotonicity in both the outcome equation and the equation for the endogenous regressor.

For each set of restrictions, we develop methods for simultaneous inference about the consistency of the distribution of the observed data with our restrictions and the sign of the ATE when the distribution of the observed data is consistent with our restrictions. For this purpose, we consider a multiple testing problem with three null hypotheses, where rejection of the first null hypothesis means that P is consistent with the assumptions, rejection of the first and second null hypotheses means that P is both consistent with the assumptions and only a positive ATE, and rejection of the first and third null hypotheses means that P is both consistent with the assumptions and only a negative ATE. The multiple testing procedure we develop is an example of a "gatekeeping" multiple testing procedure in that it only considers testing the second and third null hypotheses when the first null hypothesis has been rejected. Another novel feature of the analysis is that some of the null hypotheses involve unions of moment inequalities. We develop a bootstrap-based testing procedure for this family of null hypotheses that controls the familywise error rate – the probability of any false rejection – uniformly over a large class of possible distributions for P that include distributions where the instrument is arbitrarily "weak".

In the context of instrument exogeneity and instrument exogeneity and monotonicity in the equation for the endogenous regressor, our analysis is most closely related to Balke and Pearl (1997), who study partial identification of the ATE and also characterize when P is consistent with their assumptions. A characterization of consistency that does not require Y to be binary can be found in Kitagawa (2015), who builds upon the work of Imbens and Rubin (1997) and Huber and Mellace (2011). Kitagawa (2015) and Bhattacharya et al. (2012) also develop tests for the null hypothesis of instrument exogeneity and monotonicity in the equation for the endogenous regressor. Other related literature includes the local average treatment effect literature (LATE) (Imbens and Angrist, 1994) and the local instrumental variables/non-parametric selection model literature (Heckman and Vytlacil, 2001b), both of which impose instrument exogeneity and monotonicity in the equation for the endogenous regressor. Related results are obtained in Richardson and Robins (2010). In the context of instrument exogeneity and monotonicity in both the outcome equation and the equation for the endogenous regressor, our analysis is most closely related to Bhattacharya et al. (2012) and Shaikh and Vytlacil (2005, 2011), who study partial identification of the ATE, but do not characterize when P is consistent with the assumptions. Related results are obtained by Chiburis (2010), though under a different instrument exogeneity assumption. See also Abrevaya et al. (2010), who focus on inference about the sign of the average treatment effect in a semi-parametric model in a related context while allowing for the treatment to be non-binary and allowing for covariates. In the context of monotonicity in the outcome equation, the most closely related results are found in Chiburis (2010), though, as mentioned previously, under a different instrument exogeneity assumption.

The remainder of the paper proceeds as follows. In Section 2, we define our notation and the assumptions that will be used in the remainder of the paper. For each set of assumptions, we characterize in terms of $\Delta(P)$ in Section 3 when (i) P is inconsistent with the assumptions, (ii) when P is consistent with the assumptions and only a positive ATE, and (iii) when P is consistent with the assumptions and only a negative ATE. We further explore when P is inconsistent with our assumptions in Section 4. Finally, in Section 5, methods for inference are developed. Proofs of all results along with a numerical exploration of some of our results and a simulation study of the behavior of our inference procedure in finite samples can be found in Appendix.

2. Notation and assumptions

Let Y denote a binary outcome of interest, D denote a binary endogenous regressor, and Z denote a binary instrument. For example, Y might denote mortality one year after the start of the experiment, D might denote receipt of the medical treatment, and Z random assignment to the medical treatment, where the randomized experiment suffers from noncompliance so that Z differs from D with positive probability. Further denote by Y_1 the potential outcome if treated, by Y_0 the potential outcome if not treated, by D_1 the potential value of the endogenous regressor if the instrument were to be externally set to 1, and by D_0 the potential value of the endogenous regressor if the instrument were to be externally set to 0. Following Angrist et al. (1996), we will refer to realizations with $D_1 > D_0$ as "compliers", realizations with $D_1 < D_0$ as "defiers", realizations with $D_1 = D_0 = 1$ as "always takers", and realizations with $D_1 = D_0 = 0$ as "never takers". In this notation,

$$Y = DY_1 + (1 - D)Y_0 \tag{2}$$

$$D = ZD_1 + (1 - Z)D_0. \tag{3}$$

Let P denote the distribution of (Y, D, Z) and Q denote the distribution of (Y_0, Y_1, D_0, D_1, Z). Since

$$(Y, D, Z) = T(Y_0, Y_1, D_0, D_1, Z),$$

where T is characterized by (2) and (3), we have that

$$P = QT^{-1}.$$

Below we will restrict $Q \in \mathbf{Q}$, where $\mathbf{Q}$ is a set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying certain restrictions. In particular, we will require Z to be an instrument in the sense that every $Q \in \mathbf{Q}$ satisfies the following exogeneity condition:

Assumption 2.1 (*Instrument Exogeneity*)**.** $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$ under Q.

We will additionally consider the restriction that every $Q \in \mathbf{Q}$ satisfy one or both of the following monotonicity conditions:

Assumption 2.2 (*Monotonicity of D in Z*)**.** $Q\{D_1 \geq D_0\} = 1$ or $Q\{D_1 \leq D_0\} = 1$.

Assumption 2.3 (*Monotonicity of Y in D*)**.** $Q\{Y_1 \geq Y_0\} = 1$ or $Q\{Y_1 \leq Y_0\} = 1$.

We do not impose instrument relevance, i.e., we allow for $P\{D = 1|Z = 1\} = P\{D = 1|Z = 0\}$. Without loss of generality, we will order Z such that $P\{D = 1|Z = 1\} \geq P\{D = 1|Z = 0\}$. Given this ordering and Assumption 2.1, we have that Assumption 2.2 is equivalent to the restriction that $Q\{D_1 \geq D_0\} = 1$.

Our object of interest is the average effect of the endogenous regressor on the outcome, defined to be

$$E_Q[Y_1 - Y_0] = Q\{Y_1 = 1\} - Q\{Y_0 = 1\}. \tag{4}$$

This quantity is typically referred to in the treatment effect literature as the average treatment effect (ATE). It will be useful to partition $\mathbf{Q}$ as $\mathbf{Q} = \mathbf{Q}_+ \cup \mathbf{Q}_0 \cup \mathbf{Q}_-$, where

$$\begin{aligned}
\mathbf{Q}_+ &= \{Q \in \mathbf{Q} : Q\{Y_1 = 1\} - Q\{Y_0 = 1\} > 0\} \\
\mathbf{Q}_0 &= \{Q \in \mathbf{Q} : Q\{Y_1 = 1\} - Q\{Y_0 = 1\} = 0\} \\
\mathbf{Q}_- &= \{Q \in \mathbf{Q} : Q\{Y_1 = 1\} - Q\{Y_0 = 1\} < 0\},
\end{aligned}$$

and define

$$\mathbf{Q}_{0,+} = \mathbf{Q}_+ \cup \mathbf{Q}_0$$

$$\mathbf{Q}_{0,-} = \mathbf{Q}_- \cup \mathbf{Q}_0.$$

In other words, $\mathbf{Q}_-$ ($\mathbf{Q}_{0,-}$) is the set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying our restrictions and having a (weakly) negative ATE, $\mathbf{Q}_0$ is the set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying our restrictions and having a zero ATE, and $\mathbf{Q}_+$ ($\mathbf{Q}_{0,+}$) is the set of distributions for (Y_0, Y_1, D_0, D_1, Z) satisfying our restrictions and having a (weakly) positive ATE. In this notation, the ATE is identified to be positive if

$$P \in \mathbf{Q}_+ T^{-1} \cap (\mathbf{Q}_{0,-} T^{-1})^c, \tag{5}$$

where $\mathbf{Q}_+ T^{-1} = \{QT^{-1} : Q \in \mathbf{Q}_+\}$; $\mathbf{Q}_- T^{-1}$, $\mathbf{Q}_{0,-} T^{-1}$ and $\mathbf{Q}_{0,+} T^{-1}$ are defined similarly. In other words, we identify the ATE to be positive if the distribution of (Y, D, Z) is consistent with our restrictions holding with a positive ATE but not consistent with our restrictions holding with a zero or negative ATE. Symmetrically, the ATE is identified to be negative if

$$P \in \mathbf{Q}_- T^{-1} \cap (\mathbf{Q}_{0,+} T^{-1})^c. \tag{6}$$

Analogously, the distribution of the observed data, P, is consistent with our restrictions if

$$P \in \mathbf{Q} T^{-1}. \tag{7}$$

For completeness, we note that the identified set for the ATE, as a function of P, is given by

$$\{E_Q[Y_1 - Y_0] : Q \in \mathbf{Q} \text{ and } P = QT^{-1}\}.$$

Remark 2.1. Our results below will be stated in terms of the reduced form parameter $\Delta(P)$, defined in (1). In the biostatistics literature, when Z is random assignment to treatment with possible non-compliance, $\Delta(P)$ is sometimes referred to as the "intention-to-treat" parameter. If the instrument is relevant, i.e., $P\{D = 1|Z = 1\} \neq P\{D = 1|Z = 0\}$, then, under mild regularity conditions, the usual linear instrumental variables estimand in this setting is simply $\Delta(P)$ divided by $P\{D = 1|Z = 1\} - P\{D = 1|Z = 0\}$. Under our assumptions, the sign of $\Delta(P)$ and the usual linear instrumental variables estimand are therefore the same. As a result, it will be straightforward to re-scale our results to state them in terms of this quantity. ■

Remark 2.2. Note that Assumption 2.2 is the same monotonicity assumption found in Imbens and Angrist (1994), who also refer to it as an assumption of "no defiers". It follows from results in Vytlacil (2002) that this assumption is equivalent to the selection model of Heckman and Vytlacil (2001b, 2005). In particular, it is equivalent to assuming that there exists a representation of the model as

$$D_z = I\{\delta_0 + \delta_1 z + \eta \geq 0\} \tag{8}$$

with δ_1 being nonrandom. Similarly, Assumption 2.3 is equivalent to assuming that there exists a representation of the model as

$$Y_d = I\{\beta_0 + \beta_1 d + \epsilon \geq 0\} \tag{9}$$

with β_1 nonrandom, and Assumptions 2.2 and 2.3 are equivalent to assuming both (8) and (9) with δ_1 and β_1 nonrandom. In this way, the monotonicity assumptions considered in this paper are implicit in many models with constant coefficients. Note further that Assumption 2.3 is considerably weaker than the "monotone treatment response" assumption considered in Manski and Pepper (2000). ■

Remark 2.3. A stronger version of Assumption 2.3 in which it is assumed further that the direction of the monotonicity is known *a priori* is referred to as the "monotone treatment response" assumption by Manski (1997) and Manski and Pepper (2000). They characterize the identified set for the ATE under this type of restriction. As discussed by Bhattacharya et al. (2008), these results do not hold if only Assumption 2.3 is assumed. In some settings, it may not be reasonable to assume that the direction of the effect is known *a priori*. Our analysis, which focuses on the sign of the ATE, is useful in such settings. ■

3. Identifying the sign of the average treatment effect from IV

In this section, for each of our four possible restrictions on $\mathbf{Q}$, we characterize whether P satisfies (5), (6) or (7) in terms of $\Delta(P)$.

3.1. Instrument exogeneity and monotonicity of D in Z

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.2. In this case, our results essentially follow from Balke and Pearl (1997), who characterize the identified set for the ATE under these assumptions and also when P is consistent with these restrictions. See also Heckman and Vytlacil (2001a) and Kitagawa (2015), who generalize these results.

In order to state our results, we require some additional notation. Define

$$\begin{aligned} A_1(P) &= \max\{A_1^1(P), A_1^2(P)\} \\ A_2(P) &= -P\{Y=0, D=0|Z=1\} - P\{Y=1, D=1|Z=0\} \\ A_3(P) &= P\{Y=1, D=0|Z=1\} + P\{Y=0, D=1|Z=0\} \\ A_4(P) &= \min\{A_4^1(P), A_4^2(P)\}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} A_1^1(P) &= P\{Y=1, D=0|Z=1\} - P\{Y=1, D=0|Z=0\} \\ A_1^2(P) &= P\{Y=0, D=1|Z=0\} - P\{Y=0, D=1|Z=1\} \\ A_4^1(P) &= P\{Y=1, D=1|Z=1\} - P\{Y=1, D=1|Z=0\} \\ A_4^2(P) &= P\{Y=0, D=0|Z=0\} - P\{Y=0, D=0|Z=1\}. \end{aligned}$$

Note that $A_2(P) \leq A_4(P)$, $A_1(P) \leq A_3(P)$, and $A_2(P) \leq 0 \leq A_3(P)$.

Theorem 3.1. *If every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.2, then*

(i) $P \in \mathbf{Q}T^{-1}$ if and only if

$$\Delta(P) \in [A_1(P), A_4(P)]. \tag{11}$$

(ii) $P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c$ if and only if

$$\Delta(P) \in (A_3(P), A_4(P)].$$

(iii) $P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c$ if and only if

$$\Delta(P) \in [A_1(P), A_2(P)).$$

Remark 3.1. Part (i) of Theorem 3.1 implies that P is inconsistent with our restrictions if and only if $\Delta(P) \notin [A_1(P), A_4(P)]$. Hence, P is inconsistent with our restrictions if and only if (a) $A_1(P) > A_4(P)$, (b) $A_1(P) \leq A_4(P)$ and $\Delta(P) < A_1(P)$, or (c) $A_1(P) \leq A_4(P)$ and $\Delta(P) > A_4(P)$. If $A_1(P) \leq A_4(P)$ and $\Delta(P) < A_1(P)$, then it is possible to show that $A_1(P) \leq 0$. Similarly, if $A_1(P) \leq A_4(P)$ and $\Delta(P) > A_4(P)$, then it is possible to show that $A_4(P) \geq 0$. In this sense, part (i) of Theorem 3.1 implies that P is inconsistent with our restrictions whenever $\Delta(P)$ is "too far" from zero. ■

Remark 3.2. Parts (ii) and (iii) of Theorem 3.1 imply that we are both unable to reject our restrictions and unable to determine the sign of the ATE whenever $\Delta(P)$ is "too close" to zero, i.e.,

$$\Delta(P) \in [A_2(P), A_3(P)],$$

where $A_2(P) \leq 0 \leq A_3(P)$. The width of the region of indeterminacy is given by

$$P\{D=0|Z=1\} + P\{D=1|Z=0\} = 1 - Q\{D_1 > D_0\},$$

which decreases with the strength of the instrument, as measured by $P\{D=1|Z=1\} - P\{D=1|Z=0\} = Q\{D_1 > D_0\}$. Using results in Imbens and Angrist (1994), we have that

$$\Delta(P) = E_Q[Y_1 - Y_0|D_1 > D_0]Q\{D_1 > D_0\}$$

under Assumptions 2.1 and 2.2. The reduced form parameter $\Delta(P)$ thus combines the strength of the instrument with the strength of the treatment on "compliers". In this way, the sign of the ATE is easier to determine when the instrument is stronger or the effect of the treatment on the "compliers" is stronger. ■

Remark 3.3. Part (i) of Theorem 3.1 is derived from results in Balke and Pearl (1997). A more general result that does not require Y to be binary can be found in Kitagawa (2015), who builds upon the work of Imbens and Rubin (1997). Kitagawa (2015) also develops a testing procedure. For binary Y, Bhattacharya et al. (2012) develop a test of Assumptions 2.1 and 2.2 by comparing the bounds on the ATE in Manski (1990) with those in Heckman and Vytlacil (2001a). The resulting conditions are in fact equivalent to part (i) of Theorem 3.1. ■

3.2. Instrument exogeneity and monotonicity of Y in D and D in Z

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1–2.3. These restrictions have been previously considered in the literature by Bhattacharya et al. (2008, 2012) and Shaikh and Vytlacil (2005, 2011), who find that the sign of ATE equals the sign of $\Delta(P)$. The following theorem re-states this result and additionally characterizes when $P \in \mathbf{Q}T^{-1}$ in terms of $\Delta(P)$. We emphasize that this additional result is not found in either Bhattacharya et al. (2012) or Shaikh and Vytlacil (2005, 2011).

Theorem 3.2. *If every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1–2.3, then*

(i) $P \in \mathbf{Q}T^{-1}$ if and only if

$$\Delta(P) \in [A_1(P), A_4(P)], \tag{12}$$

(ii) $P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c$ if and only if

$$\Delta(P) \in (0, A_4(P)],$$

(iii) $P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c$ if and only if

$$\Delta(P) \in [A_1(P), 0).$$

Remark 3.4. Note that the conditions on $\Delta(P)$ in (12) that determine whether or not P is consistent with our assumptions are exactly the same as the ones in (11). In other words, P is consistent with Assumptions 2.1 and 2.2 if and only if P is consistent with Assumptions 2.1–2.3. ■

Remark 3.5. In contrast to our earlier results, the only circumstance in which we are both unable to reject our restrictions and unable to determine the sign of the ATE is if $\Delta(P) = 0$. ■

3.3. Instrument exogeneity and monotonicity of Y in D

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.3. Note that Assumption 2.3 has not been considered without Assumption 2.2 previously in the literature. In order to state our results, we require some additional notation. Define

$$\begin{aligned} B_1(P) &= \max\{B_1^1(P), B_1^2(P)\} \\ B_2(P) &= \min\{B_2^1(P), B_2^2(P)\} \\ B_3(P) &= \max\{B_3^1(P), B_3^2(P)\} \\ B_4(P) &= \min\{B_4^1(P), B_4^2(P)\}, \end{aligned} \tag{13}$$

where

$$\begin{aligned} B_1^1(P) &= -P\{Y = 1, D = 1|Z = 0\} \\ B_1^2(P) &= -P\{Y = 0, D = 0|Z = 1\} \\ B_2^1(P) &= P\{Y = 1, D = 1|Z = 1\} \\ B_2^1(P) &= P\{Y = 0, D = 0|Z = 0\} \\ B_3^1(P) &= -P\{Y = 0, D = 1|Z = 1\} \\ B_3^2(P) &= -P\{Y = 1, D = 0|Z = 0\} \\ B_4^1(P) &= P\{Y = 0, D = 1|Z = 0\} \\ B_4^2(P) &= P\{Y = 1, D = 0|Z = 1\}. \end{aligned}$$

Note that $B_1(P) \leq 0$ and $B_3(P) \leq 0$, while $B_2(P) \geq 0$ and $B_4(P) \geq 0$. Using this notation, we have the following theorem:

Theorem 3.3. *If every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.3, then*

(i) $P \in \mathbf{Q}T^{-1}$ if and only if

$$\Delta(P) \in [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}], \tag{14}$$

(ii) $P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c$ if and only if

$$\Delta(P) \in [B_1(P), B_2(P)] \setminus [B_3(P), B_4(P)], \tag{15}$$

(iii) $P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c$ if and only if

$$\Delta(P) \in [B_3(P), B_4(P)] \setminus [B_1(P), B_2(P)]. \tag{16}$$

Remark 3.6. Analogously to our earlier results, part (i) of Theorem 3.3 implies that P is inconsistent with our assumptions if and only if $\Delta(P)$ is "too far" from zero. Here, "too far" means $\Delta(P) < \min\{B_1(P), B_3(P)\} \leq 0$ or $\Delta(P) > \max\{B_2(P), B_4(P)\} \geq 0$. Since $A_1(P) \geq B_3(P)$ and $A_4(P) \leq B_2(P)$,

$$[A_1(P), A_4(P)] \subseteq [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}].$$

Furthermore, the inclusion may be strict, so it is possible to reject Assumptions 2.1 and 2.2 without rejecting Assumptions 2.1 and 2.3, while the reverse is not possible. ■

Remark 3.7. Parts (ii) and (iii) of Theorem 3.3 imply that we are both unable to reject our restrictions and unable to determine the sign of the ATE if $\Delta(P)$ is "too close" to zero, i.e.,

$$\Delta(P) \in [\max\{B_1(P), B_3(P)\}, \min\{B_2(P), B_4(P)\}],$$

where this interval necessarily includes zero. Since $A_2(P) \leq B_1(P)$ and $B_4(P) \leq A_3(P)$,

$$[\max\{B_1(P), B_3(P)\}, \min\{B_2(P), B_4(P)\}] \subseteq [A_2(P), A_3(P)].$$

Furthermore, the inclusion may be strict. Thus, it is possible to identify the sign of ATE under Assumptions 2.1 and 2.3 without being able to identify the sign of ATE under Assumptions 2.1 and 2.2, while the reverse is not possible. ■

Remark 3.8. A possibly counterintuitive implication of Theorem 3.3 is that it is possible for $\Delta(P)$ to be so large that one determines that the sign of the ATE is in fact negative or for $\Delta(P)$ to be so small that one determines that the sign of the ATE is in fact positive. The first case happens when

$$\max\{B_2(P), B_4(P)\} = B_4(P) \text{ and } B_2(P) < \Delta(P) \leq B_4(P), \tag{17}$$

whereas the second case happens when

$$\max\{B_1(P), B_3(P)\} = B_3(P) \text{ and } B_1(P) \leq \Delta(P) < B_3(P). \tag{18}$$

In order to better understand this result, it is instructive to note that

$$\Delta(P) = \begin{cases} Q\{Y_1 > Y_0, D_1 > D_0\} - Q\{Y_1 > Y_0, D_1 < D_0\} & \text{if } Y_1 \geq Y_0 \\ Q\{Y_1 < Y_0, D_1 < D_0\} - Q\{Y_1 < Y_0, D_1 > D_0\} & \text{if } Y_1 \leq Y_0. \end{cases}$$

The first case occurs when $Q\{Y_1 < Y_0, D_1 < D_0\} > Q\{Y_1 < Y_0, D_1 > D_0\}$, so we require enough "defiers" with a negative treatment effect, and the second case occurs when $Q\{Y_1 > Y_0, D_1 > D_0\} < Q\{Y_1 > Y_0, D_1 < D_0\}$, so we require enough "defiers" with a positive treatment effect. Note further that

$$Q\{Y_1 > Y_0, D_1 > D_0\} - Q\{Y_1 > Y_0, D_1 < D_0\} \in [B_1(P), B_2(P)]$$
$$Q\{Y_1 < Y_0, D_1 < D_0\} - Q\{Y_1 < Y_0, D_1 > D_0\} \in [B_3(P), B_4(P)].$$

It follows that it must be the case that $Y_1 \leq Y_0$ whenever $\Delta(P) \in (B_2(P), B_4(P)] \subseteq (0, 1]$ and that $Y_1 \geq Y_0$ whenever $\Delta(P) \in [B_1(P), B_3(P)) \subseteq [-1, 0)$. ■

Remark 3.9. In order to gain further insight into Theorem 3.3, it is instructive to consider what happens when $\Delta(P)$ satisifes (11). Recall from the discussion in Remark 3.6 that $\Delta(P)$ satisfying (11) implies that P is not only consistent with Assumptions 2.1 and 2.2, but also with Assumptions 2.1 and 2.3. In that case, it is possible to show that a sufficient condition for (15) is $\Delta(P) \in [A_3(P)/2, A_4(P)]$ and a sufficient condition for (16) is $\Delta(P) \in [A_1(P), A_2(P)/2]$. By comparing these regions with parts (ii) and (iii) of Theorem 3.1, we therefore see that whenever $\Delta(P)$ satisifes (11), the identifying power of Assumptions 2.1 and 2.3 is at least twice that of Assumptions 2.1 and 2.2. Furthermore, a necessary condition for (15) is that $\Delta(P) > 0$ and a necessary condition for (16) is that $\Delta(P) < 0$. As a result, the counterintuitive possibility discussed in Remark 3.8 of determining that the sign of the ATE is positive from a negative value of $\Delta(P)$ or vice versa is not possible whenever $\Delta(P)$ satisifes (11). ■

3.4. Instrument exogeneity

In this section, we assume that every $Q \in \mathbf{Q}$ satisfies Assumption 2.1. In this case, our results essentially follow from Balke and Pearl (1997), who characterize the identified set for the ATE under these assumptions and also when P is consistent with these restrictions.

In order to state the results, we require some additional notation. Define

$$\begin{aligned} C_1(P) &= \max\{C_1^1(P), C_1^2(P)\} \\ C_2(P) &= \max\{C_2^1(P), \dots, C_2^8(P)\} \\ C_3(P) &= \min\{C_3^1(P), \dots, C_3^8(P)\} \\ C_4(P) &= \min\{C_4^1(P), C_4^2(P)\}, \end{aligned} \tag{19}$$

where

$$
\begin{aligned}
C_1^1(P) &= -P\{Y=0,D=0|Z=1\} - P\{Y=1,D=0|Z=0\}\\
C_1^2(P) &= -P\{Y=1,D=1|Z=0\} - P\{Y=0,D=1|Z=1\}\\
C_2^1(P) &= A_2(P)\\
C_2^2(P) &= -P\{Y=0|Z=1\} - P\{Y=1|Z=0\} + P\{Y=1,D=0|Z=1\} + P\{Y=0,D=1|Z=0\}\\
C_2^3(P) &= -P\{Y=0|Z=1\} - P\{Y=1,D=1|Z=0\} + P\{Y=0,D=1|Z=0\}\\
C_2^4(P) &= -P\{Y=0,D=0|Z=1\} - P\{Y=1|Z=0\} + P\{Y=1,D=0|Z=1\}\\
C_2^5(P) &= -2P\{Y=0|Z=1\} - P\{Y=1,D=1|Z=0\} + 2P\{Y=0,D=1|Z=0\}\\
C_2^6(P) &= -2P\{Y=1,D=1|Z=0\} - P\{Y=0,D=0|Z=0\}\\
C_2^7(P) &= -2P\{Y=0,D=0|Z=1\} - P\{Y=1,D=1|Z=1\}\\
C_2^8(P) &= -P\{Y=0,D=0|Z=1\} - 2P\{Y=1|Z=0\} + 2P\{Y=1,D=0|Z=1\}\\
C_3^1(P) &= A_3(P)\\
C_3^2(P) &= P\{Y=1,D=0|Z=1\} + P\{Y=0|Z=0\} - P\{Y=0,D=0|Z=1\}\\
C_3^3(P) &= P\{Y=1|Z=1\} + P\{Y=0|Z=0\} - P\{Y=0,D=0|Z=1\} - P\{Y=1,D=1|Z=0\}\\
C_3^4(P) &= P\{Y=1|Z=1\} + P\{Y=0,D=1|Z=0\} - P\{Y=1,D=1|Z=0\}\\
C_3^5(P) &= 2P\{Y=0,D=1|Z=0\} + P\{Y=1,D=0|Z=0\}\\
C_3^6(P) &= 2P\{Y=1|Z=1\} + P\{Y=0,D=1|Z=0\} - 2P\{Y=1,D=1|Z=0\}\\
C_3^7(P) &= P\{Y=1,D=0|Z=1\} + 2P\{Y=0|Z=0\} - 2P\{Y=0,D=0|Z=1\}\\
C_3^8(P) &= 2P\{Y=1,D=0|Z=1\} + P\{Y=0,D=1|Z=1\}\\
C_4^1(P) &= P\{Y=1,D=1|Z=1\} + P\{Y=0,D=1|Z=0\}\\
C_4^2(P) &= P\{Y=0,D=0|Z=0\} + P\{Y=1,D=0|Z=1\}.
\end{aligned}
$$

Theorem 3.4. *If every $Q \in \mathbf{Q}$ satisfies Assumption 2.1, then*

(i) $P \in \mathbf{Q}T^{-1}$ if and only if

$$\Delta(P) \in [C_1(P), C_4(P)]. \tag{20}$$

(ii) $P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c$ if and only if

$$\Delta(P) \in [C_1(P), C_4(P)] \setminus [C_1(P), C_3(P)].$$

(iii) $P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c$ if and only if

$$\Delta(P) \in [C_1(P), C_4(P)] \setminus [C_2(P), C_4(P)].$$

Remark 3.10. Part (i) of Theorem 3.4 implies that P is inconsistent with our restrictions if and only if $\Delta(P) \notin [C_1(P), C_4(P)]$. Since $C_1(P) \le 0 \le C_4(P)$, part (i) of Theorem 3.4 implies that P is inconsistent with our restrictions whenever $\Delta(P)$ is "too far" from zero. Note further that

$$[\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}] \subseteq [C_1(P), C_4(P)].$$

Furthermore, the inclusion may be strict, so it is possible to reject Assumptions 2.1 and 2.3 without rejecting Assumption 2.1, while the reverse is not possible. ■

Remark 3.11. Balke and Pearl (1997) show that the identified set for the ATE under Assumptions 2.1 and 2.2 is the same as the identified set for the ATE under Assumption 2.1 alone. By combining this observation with Theorem 3.1, we see that if $\Delta(P)$ satisfies (11), then we do not reject Assumption 2.1 and do identify that the sign of the ATE is positive under Assumption 2.1 whenever $\Delta(P) > A_3(P) \ge 0$, do not reject Assumption 2.1 and do identify that the sign of the ATE is negative under that assumption whenever $\Delta(P) < A_2(P) \le 0$, and neither reject Assumption 2.1 nor identify the sign of the ATE under that assumption if $\Delta(P) \in [A_2(P), A_3(P)]$, an interval that necessarily includes zero. ■

Remark 3.12. It is possible to show by construction that the counter-intuitive possibility under Assumptions 2.1 and 2.3 discussed in Remark 3.8 is also possible under Assumption 2.1 alone: it is possible to identify a positive ATE from a negative $\Delta(P)$, or vice versa, under Assumption 2.1 alone. In light of the discussion in Remark 3.11, this phenomenon is only possible when $\Delta(P) \notin [A_1(P), A_4(P)]$. On the other hand, it may occur regardless of whether $\Delta(P)$ satisfies (14), that is, regardless of whether or not P is consistent with Assumptions 2.1 and 2.3. ■

Remark 3.13. If $\Delta(P) \in [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}] \setminus [A_1(P), A_4(P)]$, so P is consistent with Assumptions 2.1 and 2.3, but not with Assumptions 2.1 and 2.2, then it is possible to show that Assumption 2.1 has less ability to determine the sign of the ATE than Assumptions 2.1 and 2.3 in the sense that the set of distributions for which one can identify the sign of the ATE under Assumption 2.1 is a strict subset of the set of distributions for which one can identify the sign of the ATE under Assumptions 2.1 and 2.3. ■

4. Detecting failure of the restrictions

In the preceding section, we characterized when P was consistent with our restrictions in terms of the reduced form parameter $\Delta(P)$. In particular, we showed that in each case a value of $\Delta(P)$ sufficiently far from zero implied that the restrictions were violated. In this section, we first characterize conditions on Q for a violation of instrument exogeneity to be detectable in the sense that they lead to $\Delta(P)$ being sufficiently far from zero. We then, while maintaining instrument exogeneity, characterize which types of violations of the monotonicity assumptions are detectable. To complement the analytical results in this section, we also provide some numerical results in Appendix A, where we explore which violations of the restrictions are detectable in the context of a parametric model for Y and D.

4.1. Instrument exogeneity

Part (i) of Theorem 3.4 shows that P is consistent with Assumption 2.1 if and only if $\Delta(P)$ satisfies (20). The following proposition states conditions on Q for which $\Delta(P)$ fails to satisfy (20). In order to state our results, we require some additional notation. Define

$$\Delta_0(Q) = E_Q[Y_0 \mid Z = 1] - E_Q[Y_0 \mid Z = 0]$$
$$\Delta_1(Q) = E_Q[Y_1 \mid Z = 1] - E_Q[Y_1 \mid Z = 0].$$

Note that Assumption 2.1 implies in particular that $\Delta_0(Q) = \Delta_1(Q) = 0$. More generally, $\Delta_d(Q)$ measures the dependence between Y_d and Z under Q. Further define

$$G_0^1(Q) = -Q\{Y_0 = 1, D = 1|Z = 0\} - Q\{Y_0 = 0, D = 1|Z = 1\}$$
$$G_0^2(Q) = \quad Q\{Y_0 = 1, D = 1|Z = 1\} + Q\{Y_0 = 0, D = 1|Z = 0\}$$
$$G_1^1(Q) = -Q\{Y_1 = 1, D = 0|Z = 0\} - Q\{Y_1 = 0, D = 0|Z = 1\}$$
$$G_1^2(Q) = \quad Q\{Y_1 = 1, D = 0|Z = 1\} + Q\{Y_1 = 0, D = 0|Z = 0\}.$$

In terms of this notation, we have the following result:

Proposition 4.1. *If $P = QT^{-1}$, then $\Delta(P) \notin [C_1(P), C_4(P)]$ if and only if*

$$\Delta_d(Q) \notin [G_d^1(Q), G_d^2(Q)]$$

for some $d \in \{0, 1\}$. Furthermore,

(i) $\Delta_0(Q) \notin [G_0^1(Q), G_0^2(Q)]$ if $|\Delta_0(Q)| > P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$.
(ii) $\Delta_1(Q) \notin [G_1^1(Q), G_1^2(Q)]$ if $|\Delta_1(Q)| > 2 - P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\}$.
(iii) $\Delta_0(Q) \notin [G_0^1(Q), G_0^2(Q)]$ only if $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} < 1$.
(iv) $\Delta_1(Q) \notin [G_1^1(Q), G_1^2(Q)]$ only if $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} > 1$.

Remark 4.1. Since zero always lies in $[G_d^1(Q), G_d^2(Q)]$ and $\Delta_d(Q)$ equals zero whenever $Y_d \perp\!\!\!\perp Z$, it follows from Proposition 4.1 that P is only inconsistent with Assumption 2.1 if $Y_d \not\perp\!\!\!\perp Z$ for some $d \in \{0, 1\}$. Part (i) of Proposition 4.1 implies that even slight deviations from $Y_0 \perp\!\!\!\perp Z$ will be detectable if $P\{D = 1 \mid Z = 1\}$ and $P\{D = 1 \mid Z = 0\}$ are both sufficiently close to zero, and part (ii) of Proposition 4.1 implies that even slight deviations from $Y_1 \perp\!\!\!\perp Z$ will be detectable if $P\{D = 1 \mid Z = 1\}$ and $P\{D = 1 \mid Z = 0\}$ are both sufficiently close to one. On the other hand, part (iii) of Proposition 4.1 implies that no deviation from $Y_0 \perp\!\!\!\perp Z$ can be detected if $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} \geq 1$, and part (iv) of Proposition 4.1 implies no deviation from $Y_1 \perp\!\!\!\perp Z$ can be detected if $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} \leq 1$. In particular, no violation of Assumption 2.1 can be detected if $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} = 1$, which includes both the case in which $P\{D = 1|Z = 1\} = P\{D = 1|Z = 0\} = \frac{1}{2}$ (i.e., Z is irrelevant) and the case in which $P\{D = 1|Z = 1\} = 1, P\{D = 1|Z = 0\} = 0$ (i.e., an experiment with full compliance). ■

4.2. Monotonicity of D in Z (and Y in D) while maintaining instrument exogeneity

Parts (i) of Theorems 3.1 and 3.2 show that P is consistent with Assumptions 2.1 and 2.2 (and 2.3) if and only if $\Delta(P)$ satisfies (11). The following proposition characterizes distributions Q satisfying Assumption 2.1 for which $\Delta(P)$ fails to satisfy (11).

Proposition 4.2. *If $P = QT^{-1}$ for a distribution Q that satisfies Assumption 2.1, then $\Delta(P) \notin [A_1(P), A_4(P)]$ if and only if*

$$Q\{Y_j = k, D_1 < D_0\} > Q\{Y_j = k, D_1 > D_0\}$$

for some $(j, k) \in \{0, 1\}^2$.

Remark 4.2. Given our normalization that $P\{D = 1|Z = 1\} \geq P\{D = 1|Z = 0\}$ and Assumption 2.1, we have that the fraction of "compliers", $Q\{D_1 > D_0\}$ weakly exceeds the fraction of "defiers", $Q\{D_1 < D_0\}$ and does so by the magnitude of $P\{D = 1|Z = 1\} - P\{D = 1|Z = 0\} = Q\{D_1 > D_0\} - Q\{D_1 < D_0\}$. Proposition 4.2 therefore implies that in order to detect a violation of Assumption 2.2 while satisfying Assumption 2.1 it must be the case that the fraction of "defiers" is sufficiently large (which in turn requires the instrument be sufficiently weak in that $P\{D = 1|Z = 1\} - P\{D = 1|Z = 0\}$ is sufficiently small) and that the distribution of potential outcomes among "defiers" and "compliers" differs, i.e., $Q\{Y_j = 1|D_1 < D_0\} \neq Q\{Y_j = 1|D_1 > D_0\}$ for some $j \in \{0, 1\}$. ■

4.3. Monotonicity of Y in D while maintaining instrument exogeneity

Part (i) of Theorem 3.3 showed that P is consistent with Assumptions 2.1 and 2.3 if and only if $\Delta(P)$ satisfies (14). The following proposition characterizes distributions Q satisfying Assumption 2.1 for which $\Delta(P)$ fails to satisfy (14). In order to state our results, we require some additional notation. Define

$$\begin{aligned}
M_1^1(Q) &= Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 < D_0\}\\
M_1^2(Q) &= Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 0, D_1 < D_0\}\\
M_2^1(Q) &= Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 > D_0\}\\
M_2^2(Q) &= Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 0, D_1 > D_0\}\\
M_3^1(Q) &= Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 > D_0\}\\
M_3^2(Q) &= Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 1, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 1, D_1 > D_0\}\\
M_4^1(Q) &= Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 < D_0\}\\
M_4^2(Q) &= Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 1, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 1, D_1 < D_0\}
\end{aligned}$$

and, for $1 \leq j \leq 4$, let

$$M_j(Q) = \min\{M_j^1(Q), M_j^2(Q)\}.$$

Using this notation, we have the following result:

Proposition 4.3. *If $P = QT^{-1}$ for a distribution Q that satisfies Assumption 2.1, then*

$$\Delta(P) \notin [\min\{B_1(P), B_3(P)\}, \max\{B_2(P), B_4(P)\}]$$

if and only if either

$$\begin{aligned}
&Q\{Y_1 > Y_0, D_1 > D_0\} + Q\{Y_1 < Y_0, D_1 < D_0\}\\
&\qquad < \min\{Q\{Y_1 < Y_0, D_1 > D_0\} - M_1(Q), Q\{Y_1 > Y_0, D_1 < D_0\} - M_3(Q)\}
\end{aligned} \tag{21}$$

or

$$\begin{aligned}
&Q\{Y_1 < Y_0, D_1 > D_0\} + Q\{Y_1 > Y_0, D_1 < D_0\}\\
&\qquad < \min\{Q\{Y_1 < Y_0, D_1 < D_0\} - M_2(Q), Q\{Y_1 > Y_0, D_1 > D_0\} - M_4(Q)\}.
\end{aligned} \tag{22}$$

Remark 4.3. Note that if there are no "defiers", then it is impossible for either (21) or (22) to hold. Hence, while satisfying Assumption 2.1, it is only possible to detect violations of Assumption 2.3 if Assumption 2.2 does not hold. ■

Remark 4.4. In order to satisfy (21), there must be strong negative dependence between $Y_1 - Y_0$ and $D_1 - D_0$. In addition, it seems that the probability of being an "always taker" or "never taker" must be small so that $M_1(Q)$ and $M_3(Q)$ will be small. For instance, (21) is satisfied when $Q\{Y_1 < Y_0|D_1 > D_0\} = 1$, $Q\{Y_1 > Y_0|D_1 < D_0\} = 1$ and $Q\{D_1 = D_0\} = 0$. Analogous comments apply to (22). In this sense, it seems that the requirements on Q in order to satisfy either (21) or (22) are rather extreme. The numerical results in Appendix A further highlight the difficulty of detecting violations of Assumption 2.3 when Assumption 2.1 holds. ■

5. Inference

In this section, we let (Y_i, D_i, Z_i), $i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\{0, 1\}^3$ and, for each of the four sets of restrictions considered in the previous sections, consider the problem of simultaneous inference about the consistency of the distribution of the observed data with our restrictions and the sign of the ATE when the distribution of the observed data is consistent with our restrictions. More precisely, for each set of restrictions on $\mathbf{Q}$, we will consider the problem of testing the family of null hypotheses

$$H_j : P \in \mathbf{P}_j \text{ for } 1 \le j \le 3, \tag{23}$$

where $\mathbf{P}_1 \subseteq \mathbf{P}$, $\mathbf{P}_2 \subseteq \mathbf{P}$ and $\mathbf{P}_3 \subseteq \mathbf{P}$ are such that

$$\begin{aligned} \mathbf{P}_1^c &= \{P \in \mathbf{P} : P \in \mathbf{Q}T^{-1}\} \\ \mathbf{P}_2^c \cap \mathbf{P}_1^c &= \{P \in \mathbf{P} : P \in \mathbf{Q}_+T^{-1} \cap (\mathbf{Q}_{0,-}T^{-1})^c\} \\ \mathbf{P}_3^c \cap \mathbf{P}_1^c &= \{P \in \mathbf{P} : P \in \mathbf{Q}_-T^{-1} \cap (\mathbf{Q}_{0,+}T^{-1})^c\}, \end{aligned}$$

in a way that satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} FWER_P \le \alpha. \tag{24}$$

Here, $\mathbf{P}_j^c$ is understood to be relative to $\mathbf{P}$, i.e., $\mathbf{P}_j^c = \mathbf{P} \setminus \mathbf{P}_j$, and

$$FWER_P = P\{\text{any false rejection}\}.$$

Note that $\mathbf{P}_1$ is defined so that $\mathbf{P}_1^c$ equals the set of distributions $P \in \mathbf{P}$ that are consistent with our restrictions (in particular, with our instrument exogeneity restriction and with the specified monotonicity restrictions), $\mathbf{P}_2$ is defined so that $\mathbf{P}_2^c \cap \mathbf{P}_1^c$ equals the set of distributions $P \in \mathbf{P}$ that are both consistent with our restrictions and the sign of the ATE only being positive, and $\mathbf{P}_3$ is defined so that $\mathbf{P}_3^c \cap \mathbf{P}_1^c$ equals the set of distributions $P \in \mathbf{P}$ that are both consistent with our restrictions and the sign of the ATE only being negative. Our testing procedure below will only consider testing H_2 or H_3 when H_1 is rejected; in that sense, H_1 is a "gatekeeper" for H_2 and H_3. See Dmitrienko et al. (2008) for further examples of "gatekeeping" multiple testing procedures. If H_1 is rejected, then we will conclude that P is consistent with our restrictions; if H_1 and H_2 are rejected, then we will conclude that P is consistent with our restrictions and only a positive ATE; if H_1 and H_3 are rejected, then we will conclude that P is consistent with our restrictions and only a negative ATE. The testing procedure will additionally have the feature that it is not possible to reject H_2 and H_3 at the same time. We explore the finite-sample performance of our inference procedures in a small simulation study in Appendix B.

Below we will assume that $\mathbf{P}$ is such that

$$\inf_{P\in\mathbf{P}} \inf_{(y,d,z)\in\{0,1\}^3} P\{Y = y, D = d, Z = z\} > \epsilon \tag{25}$$

for some $\epsilon > 0$. We will also denote by $\hat{P}_n$ the empirical distribution of (Y_i, D_i, Z_i), $i = 1, \ldots, n$.

5.1. Instrument exogeneity and monotonicity of D in Z

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.2. For this choice of $\mathbf{Q}$, it follows from Theorem 3.1 that

$$\mathbf{P}_1 = \{P \in \mathbf{P} : \Delta(P) < A_1(P) \cup \Delta(P) > A_4(P)\} \tag{26}$$

$$\mathbf{P}_2 = \{P \in \mathbf{P} : \Delta(P) \le A_3(P)\} \tag{27}$$

$$\mathbf{P}_3 = \{P \in \mathbf{P} : \Delta(P) \ge A_2(P)\}. \tag{28}$$

In order to describe our testing procedure, it is useful to introduce some further notation. Define

$$\begin{aligned} a_1(P) = -a_8(P) &= A_1^1(P) - \Delta(P) \\ a_2(P) = -a_9(P) &= A_1^2(P) - \Delta(P) \\ a_3(P) = -a_6(P) &= \Delta(P) - A_4^1(P) \\ a_4(P) = -a_7(P) &= \Delta(P) - A_4^2(P) \\ a_5(P) &= \Delta(P) - A_3(P) \\ a_{10}(P) &= A_2(P) - \Delta(P). \end{aligned}$$

For $1 \le j \le 3$, define

$$T_{j,n}^1 = \min_{K\in\mathcal{K}_j^1} \max_{k\in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^a},$$

where

$$\mathcal{K}_1^1 = \{\{6\}, \{7\}, \{8\}, \{9\}\}$$
$$\mathcal{K}_2^1 = \{\{5\}\}$$
$$\mathcal{K}_3^1 = \{\{10\}\},$$

and $\hat{\sigma}^a_{k,n}$ for $1 \leq k \leq 10$ is the usual (unpooled) estimate of the standard deviation of $a_k(\hat{P}_n)$. Note that at most one of $T^1_{2,n}$ and $T^1_{3,n}$ will be strictly positive. Furthermore, the maximum over $k \in K$ is superfluous in the definition of $T^1_{j,n}$, but we retain it to maintain consistency with the subsequent sections.

For $\emptyset \neq \mathcal{K} \subseteq 2^{\{1,\ldots,10\}} \setminus \{\emptyset\}$, define

$$\hat{c}_{1,n}(\mathcal{K}, 1-\alpha) = \max_{K\in\mathcal{K}} J^{-1}_{1,n}(1-\alpha, K, \hat{P}_n), \tag{29}$$

where

$$J_{1,n}(x, K, P) = P\left\{\max_{k\in K} \frac{a_k(\hat{P}_n) - a_k(P)}{\hat{\sigma}^a_{k,n}} \leq x\right\}.$$

For $\emptyset \neq S \subseteq \{1, 2, 3\}$, further define

$$\mathcal{K}^1(S) = \{\cup_{j\in S} C_j : C_j \in \mathcal{K}^1_j\}.$$

Using this notation, the testing procedure is given by the following algorithm:

Algorithm 5.1.

Step 1: Reject H_1 if

$$T^1_{1,n} > \hat{c}_{1,n}(\mathcal{K}^1(\{1\}), 1-\alpha).$$

Step 2: If H_1 is rejected, then further reject any additional H_j with

$$T^1_{j,n} > \hat{c}_{1,n}(\mathcal{K}^1(\{2, 3\}), 1-\alpha).$$

Theorem 5.1. *Consider testing (23) with $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ given by (26), (27)* and *(28), respectively. If $\mathbf{P}$ satisifes (25), then Algorithm 5.1 satisfies (24).*

Remark 5.1. It may be of interest to test only H_2 and H_3 simultaneously without testing H_1. The argument used to establish Theorem 5.1 implies that the test that rejects any H_j with $T^1_{j,n} > \hat{c}_{1,n}(\mathcal{K}^1(\{2, 3\}), 1-\alpha)$ satisfies (24) for this smaller family of null hypotheses. ■

Remark 5.2. It may be of interest to test the null hypothesis that P is consistent with our restrictions, $P \in \mathbf{P}^c_1$ (as opposed to H_1 above, which specifies that $P \in \mathbf{P}_1$). By arguing as in the proof of Theorem 5.1, it is possible to show that the test

$$\phi^1_n = I\left\{\max_{k\in\{1,2,3,4\}} \frac{a_k(\hat{P}_n)}{\hat{\sigma}^a_{k,n}} > J^{-1}_{1,n}(1-\alpha, \{1, 2, 3, 4\}, \hat{P}_n)\right\}$$

satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}^c_1} E_P[\phi^1_n] \leq \alpha. \quad ■$$

Remark 5.3. The critical value $\hat{c}_{1,n}(\mathcal{K}, 1-\alpha)$ in (29) may be viewed as a "least favorable" critical value in the same way that critical values based on assuming that all moments are binding in the moment inequality literature are "least favorable". To see this, it is useful to note that $\hat{c}_{1,n}(\mathcal{K}, 1-\alpha)$ is the same critical value that would be used to test the null hypothesis that

$$P \in \bigcup_{K\in\mathcal{K}} \bigcap_{k\in K} \{P \in \mathbf{P} : a_k(P) \leq 0\}$$

at level α using the test statistic

$$\min_{K\in\mathcal{K}} \max_{k\in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}^a_{k,n}}.$$

In contrast to the moment inequality literature, where the null hypotheses only involve a single set of inequalities, the null hypothesis involves a union of different sets of inequalities. As a result, there is no longer a single "least favorable" critical value, but rather one for each set of inequalities in the union. It is for this reason that the maximum appears in (29). It is possible to construct critical values that are not "least favorable" by modifying other approaches in the moment inequality literature, such as the "generalized moment selection" approach of Andrews and Soares (2010) or the recent approach by Romano et al. (2012). Indeed, "generalized moment selection" critical values may be constructed simply by replacing K in (29) with

$$\hat{K}(K) = \left\{ k \in K : \frac{a_k(\hat{P}_n)}{\hat{\sigma}^a_{k,n}} > -\epsilon_n \right\}$$

for $\epsilon_n \to \infty$, but satisfying $\epsilon_n/\sqrt{n} \to 0$. Analogous comments apply to each of our subsequent theorems. See Canay and Shaikh (2016) for an overview of these and related methods in the context of inference for partially identified models. ■

5.2. Instrument exogeneity and monotonicity of Y in D and D in Z

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1–2.3. For this choice of $\mathbf{Q}$, it follows from Theorem 3.2 that

$$\mathbf{P}_1 = \{P \in \mathbf{P} : \Delta(P) < A_1(P) \cup \Delta(P) > A_4(P)\} \tag{30}$$
$$\mathbf{P}_2 = \{P \in \mathbf{P} : \Delta(P) \leq 0\} \tag{31}$$
$$\mathbf{P}_3 = \{P \in \mathbf{P} : \Delta(P) \geq 0\}. \tag{32}$$

Recall the definitions of $a_k(P)$ and $\hat{\sigma}^a_{k,n}$ for $1 \leq k \leq 10$ in Section 5.1 and define

$$a_{11}(P) = -a_{12}(P) = \Delta(P).$$

For $1 \leq j \leq 3$, define

$$T^2_{j,n} = \min_{K \in \mathcal{K}^2_j} \max_{k \in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}^a_{k,n}},$$

where

$$\mathcal{K}^2_1 = \{\{6\}, \{7\}, \{8\}, \{9\}\}$$
$$\mathcal{K}^2_2 = \{\{11\}\}$$
$$\mathcal{K}^2_3 = \{\{12\}\},$$

and $\hat{\sigma}^a_{k,n}$ for $11 \leq k \leq 12$ is the usual (unpooled) estimate of the standard deviation of $a_k(\hat{P}_n)$. Note that at most one of $T^2_{2,n}$ and $T^2_{3,n}$ will be strictly positive. For $\emptyset \neq \mathcal{K} \subseteq 2^{\{1,\ldots,12\}} \setminus \{\emptyset\}$, define

$$\hat{c}_{2,n}(\mathcal{K}, 1-\alpha) = \max_{K \in \mathcal{K}} J^{-1}_{2,n}(1-\alpha, K, \hat{P}_n),$$

where

$$J_{2,n}(x, K, P) = P\left\{ \max_{k \in K} \frac{a_k(\hat{P}_n) - a_k(P)}{\hat{\sigma}^a_{k,n}} \leq x \right\}.$$

For $\emptyset \neq S \subseteq \{1, 2, 3\}$, further define

$$\mathcal{K}^2(S) = \{\cup_{j \in S} C_j : C_j \in \mathcal{K}^2_j\}.$$

Using this notation, the testing procedure is given by the following algorithm:

Algorithm 5.2.

Step 1: Reject H_1 if

$$T^2_{1,n} > \hat{c}_{2,n}(\mathcal{K}^2(\{1\}), 1-\alpha).$$

Step 2: If H_1 is rejected, then further reject any additional H_j with

$$T^2_{j,n} > \hat{c}_{2,n}(\mathcal{K}^2(\{2, 3\}), 1-\alpha).$$

Theorem 5.2. *Consider testing (23) with $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ given by (30), (31)* and *(32), respectively. If $\mathbf{P}$ satisifes (25), then Algorithm 5.2 satisfies (24).*

Remark 5.4. As in the previous section, it may be of interest to test only H_2 and H_3 simultaneously. The argument used to establish Theorem 5.2 implies that the test that rejects any H_j with $T^2_{j,n} > \hat{c}_{2,n}(\mathcal{K}^2(\{2,3\}), 1-\alpha)$ satisfies (24) for this smaller family of null hypotheses. ■

Remark 5.5. As in the previous section, it may be of interest to test the null hypothesis that $P \in \mathbf{P}_1^c$ (as opposed to H_1 above, which specifies that $P \in \mathbf{P}_1$). Since $\mathbf{P}_1^c$ under instrument exogeneity and monotonicity of both D in Z and Y in D equals $\mathbf{P}_1^c$ under instrument exogeneity and monotonicity of D in Z alone, the test described in Remark 5.2 may be used for this purpose. ■

5.3. Instrument exogeneity and monotonicity of Y in D

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumptions 2.1 and 2.3. For this choice of $\mathbf{Q}$, it follows from Theorem 3.3 that

$$\mathbf{P}_1 = \{P \in \mathbf{P} : \Delta(P) < \min\{B_1(P), B_3(P)\} \cup \Delta(P) > \max\{B_2(P), B_4(P)\}\} \tag{33}$$

$$\mathbf{P}_2 = \{P \in \mathbf{P} : B_2(P) \leq B_4(P) \cup \Delta(P) \leq B_4(P), B_3(P) \leq B_1(P) \cup \Delta(P) \geq B_3(P)\} \tag{34}$$

$$\mathbf{P}_3 = \{P \in \mathbf{P} : B_2(P) \geq B_4(P) \cup \Delta(P) \leq B_2(P), B_3(P) \geq B_1(P) \cup \Delta(P) \geq B_1(P)\}. \tag{35}$$

In order to describe our testing procedure, it is useful to introduce some further notation. Define

$$\begin{aligned}
b_1(P) &= -b_{19}(P) = B_1^1(P) - \Delta(P)\\
b_2(P) &= -b_{20}(P) = B_1^2(P) - \Delta(P)\\
b_3(P) &= -b_{31}(P) = B_3^1(P) - \Delta(P)\\
b_4(P) &= -b_{32}(P) = B_3^2(P) - \Delta(P)\\
b_5(P) &= -b_{13}(P) = \Delta(P) - B_2^1(P)\\
b_6(P) &= -b_{14}(P) = \Delta(P) - B_2^2(P)\\
b_7(P) &= -b_{25}(P) = \Delta(P) - B_4^1(P)\\
b_8(P) &= -b_{26}(P) = \Delta(P) - B_4^2(P)\\
b_9(P) &= -b_{21}(P) = B_2^1(P) - B_4^1(P)\\
b_{10}(P) &= -b_{23}(P) = B_2^1(P) - B_4^2(P)\\
b_{11}(P) &= -b_{22}(P) = B_2^2(P) - B_4^1(P)\\
b_{12}(P) &= -b_{24}(P) = B_2^2(P) - B_4^2(P)\\
b_{15}(P) &= -b_{27}(P) = B_3^1(P) - B_1^1(P)\\
b_{16}(P) &= -b_{29}(P) = B_3^2(P) - B_1^1(P)\\
b_{17}(P) &= -b_{28}(P) = B_3^1(P) - B_1^2(P)\\
b_{18}(P) &= -b_{30}(P) = B_3^2(P) - B_1^2(P).
\end{aligned}$$

For $1 \leq j \leq 3$, define

$$T^3_{j,n} = \min_{K \in \mathcal{K}^3_j} \max_{k \in K} \frac{b_k(\hat{P}_n)}{\hat{\sigma}^b_{k,n}},$$

where

$$\begin{aligned}
\mathcal{K}^3_1 &= \{A \cup B : A \in \{\{13\},\{14\}\}, B \in \{\{25\},\{26\}\}\} \cup \{A \cup B : A \in \{\{19\},\{20\}\}, B \in \{\{31\},\{32\}\}\}\\
\mathcal{K}^3_2 &= \{A \cup B : A \in \{\{7,8\},\{9,10\},\{11,12\}\}, B \in \{\{3,4\},\{15,16\},\{17,18\}\}\}\\
\mathcal{K}^3_3 &= \{A \cup B : A \in \{\{5,6\},\{21,22\},\{23,24\}\}, B \in \{\{1,2\},\{27,28\},\{29,30\}\}\},
\end{aligned}$$

and $\hat{\sigma}^b_{k,n}$ for $1 \leq k \leq 32$ is the usual (unpooled) estimate of the standard deviation of $b_k(\hat{P}_n)$. Note that at most one of $T^3_{2,n}$ and $T^3_{3,n}$ will be strictly positive. For $\emptyset \neq \mathcal{K} \subseteq 2^{\{1,\ldots,32\}} \setminus \{\emptyset\}$, define

$$\hat{c}_{3,n}(\mathcal{K}, 1-\alpha) = \max_{K \in \mathcal{K}} J^{-1}_{3,n}(1-\alpha, K, \hat{P}_n),$$

where

$$J_{3,n}(x, K, P) = P\left\{\max_{k \in K} \frac{b_k(\hat{P}_n) - b_k(P)}{\hat{\sigma}^b_{k,n}} \leq x\right\}.$$

For $\emptyset \neq S \subseteq \{1, 2, 3\}$, further define

$$\mathcal{K}^3(S) = \{\cup_{j \in S} C_j : C_j \in \mathcal{K}_j^3\}.$$

Using this notation, the testing procedure is given by the following algorithm:

Algorithm 5.3.

Step 1: Reject H_1 if

$$T_{1,n}^3 > \hat{c}_{3,n}(\mathcal{K}^3(\{1\}), 1-\alpha).$$

Step 2: If H_1 is rejected, then further reject any additional H_j with

$$T_{j,n}^3 > \hat{c}_{3,n}(\mathcal{K}^3(\{2,3\}), 1-\alpha).$$

Theorem 5.3. *Consider testing (23) with $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ given by (33), (34)* and *(35), respectively. If $\mathbf{P}$ satisifes (25), then Algorithm 5.3 satisfies (24).*

Remark 5.6. As in the previous section, it may be of interest to test only H_2 and H_3 simultaneously. The argument used to establish Theorem 5.3 implies that the test that rejects any H_j with $T_{j,n}^3 > \hat{c}_{3,n}(\mathcal{K}^3(\{2,3\}), 1-\alpha)$ satisfies (24) for this smaller family of null hypotheses. ■

Remark 5.7. As in the previous section, it may be of interest to test the null hypothesis that $P \in \mathbf{P}_1^c$ (as opposed to H_1 above, which specifies that $P \in \mathbf{P}_1$). By arguing as in the proof of Theorem 5.3, it is possible to show that the test

$$\phi_n^3 = I\left\{\min_{K \in \tilde{\mathcal{K}}_1^3} \max_{k \in K} \frac{b_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^b} > \max_{K \in \tilde{\mathcal{K}}_1^3} J_{3,n}^{-1}(1-\alpha, K, \hat{P}_n)\right\},$$

where

$$\tilde{\mathcal{K}}_1^3 = \{A \cup B : A \in \{\{1,2\},\{3,4\}\}, B \in \{\{5,6\},\{7,8\}\}\},$$

satisfies

$$\limsup_{n\to\infty} \sup_{P \in \tilde{\mathbf{P}}_1^c} E_P[\phi_n^3] \leq \alpha. \quad ■$$

5.4. Instrument exogeneity

In this section, we assume every $Q \in \mathbf{Q}$ satisfies Assumption 2.1. For this choice of $\mathbf{Q}$, it follows from Theorem 3.4 that

$$\mathbf{P}_1 = \{P \in \mathbf{P} : \Delta(P) < C_1(P) \cup \Delta > C_4(P)\} \quad (36)$$
$$\mathbf{P}_2 = \{P \in \mathbf{P} : \Delta(P) \leq C_3(P)\} \quad (37)$$
$$\mathbf{P}_3 = \{P \in \mathbf{P} : \Delta(P) \geq C_2(P)\}. \quad (38)$$

In order to describe our testing procedure, it is useful to introduce some further notation. Define

$$\begin{aligned}
c_1(P) = -c_5(P) &= C_1^1(P) - \Delta(P)\\
c_2(P) = -c_6(P) &= C_1^2(P) - \Delta(P)\\
c_3(P) = -c_7(P) &= \Delta(P) - C_4^1(P)\\
c_4(P) = -c_8(P) &= \Delta(P) - C_4^2(P)\\
c_9(P) &= \Delta(P) - C_3^1(P)\\
c_{10}(P) &= \Delta(P) - C_3^2(P)\\
c_{11}(P) &= \Delta(P) - C_3^3(P)\\
c_{12}(P) &= \Delta(P) - C_3^4(P)\\
c_{13}(P) &= \Delta(P) - C_3^5(P)\\
c_{14}(P) &= \Delta(P) - C_3^6(P)\\
c_{15}(P) &= \Delta(P) - C_3^7(P)\\
c_{16}(P) &= \Delta(P) - C_3^8(P)\\
c_{17}(P) &= C_2^1(P) - \Delta(P)
\end{aligned}$$

$$c_{18}(P) = C_2^2(P) - \Delta(P)$$
$$c_{19}(P) = C_2^3(P) - \Delta(P)$$
$$c_{20}(P) = C_2^4(P) - \Delta(P)$$
$$c_{21}(P) = C_2^5(P) - \Delta(P)$$
$$c_{22}(P) = C_2^6(P) - \Delta(P)$$
$$c_{23}(P) = C_2^7(P) - \Delta(P)$$
$$c_{24}(P) = C_2^8(P) - \Delta(P).$$

For $1 \leq j \leq 3$, define

$$T_{j,n}^4 = \min_{K \in \mathcal{K}_j^1} \max_{k \in K} \frac{c_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^c},$$

where

$$\mathcal{K}_1^4 = \{\{5\}, \{6\}, \{7\}, \{8\}\}$$
$$\mathcal{K}_2^4 = \{\{9, 10, 11, 12, 13, 14, 15, 16\}\}$$
$$\mathcal{K}_3^4 = \{\{17, 18, 19, 20, 21, 22, 23, 24\}\},$$

and $\hat{\sigma}_{k,n}^c$ for $1 \leq k \leq 24$ is the usual (unpooled) estimate of the standard deviation of $c_k(\hat{P}_n)$. Note that at most one of $T_{2,n}^4$ and $T_{3,n}^4$ will be strictly positive. For $\emptyset \neq \mathcal{K} \subseteq 2^{\{1,\ldots,24\}} \setminus \{\emptyset\}$, define

$$\hat{c}_{4,n}(\mathcal{K}, 1-\alpha) = \max_{K \in \mathcal{K}} J_{1,n}^{-1}(1-\alpha, K, \hat{P}_n), \tag{39}$$

where

$$J_{4,n}(x, K, P) = P\left\{\max_{k \in K} \frac{c_k(\hat{P}_n) - c_k(P)}{\hat{\sigma}_{k,n}^c} \leq x\right\}.$$

For $\emptyset \neq S \subseteq \{1, 2, 3\}$, further define

$$\mathcal{K}^4(S) = \{\cup_{j \in S} C_j : C_j \in \mathcal{K}_j^4\}.$$

Using this notation, the testing procedure is given by the following algorithm:

Algorithm 5.4.

Step 1: Reject H_1 if

$$T_{1,n}^4 > \hat{c}_{4,n}(\mathcal{K}^4(\{1\}), 1-\alpha).$$

Step 2: If H_1 is rejected, then further reject any additional H_j with

$$T_{j,n}^4 > \hat{c}_{4,n}(\mathcal{K}^4(\{2, 3\}), 1-\alpha).$$

Theorem 5.4. *Consider testing (23) with $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ given by (36), (37) and (38), respectively. If $\mathbf{P}$ satisifes (25), then Algorithm 5.4 satisfies (24).*

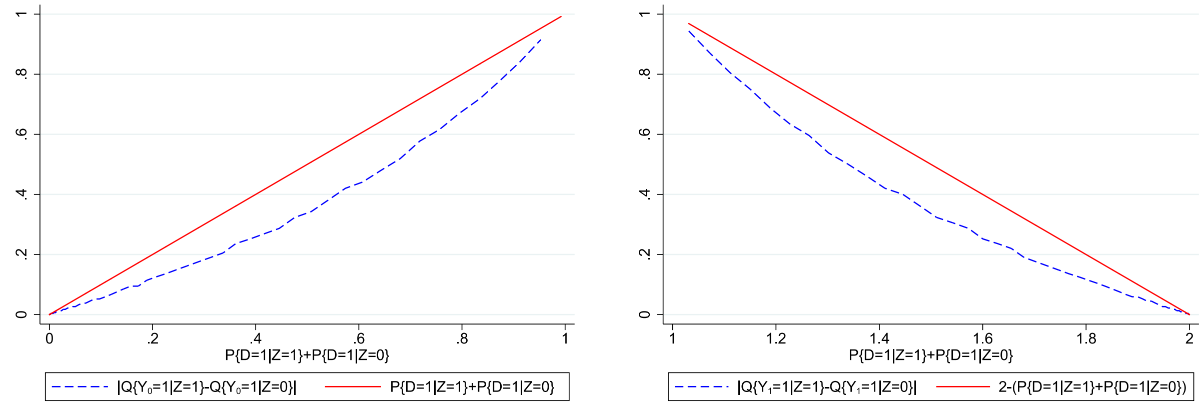
Remark 5.8. As in the previous section, it may be of interest to test only H_2 and H_3 simultaneously. The argument used to establish Theorem 5.4 implies that the test that rejects any H_j with $T_{j,n}^4 > \hat{c}_{4,n}(\mathcal{K}^4(\{2, 3\}), 1-\alpha)$ satisfies (24) for this smaller family of null hypotheses. ■

Remark 5.9. As in the previous section, it may be of interest to test the null hypothesis that $P \in \mathbf{P}_1^c$ (as opposed to H_1 above, which specifies that $P \in \mathbf{P}_1$). By arguing as in the proof of Theorem 5.4, it is possible to show that the test

$$\phi_n^4 = I\left\{\max_{k \in \{1,2,3,4\}} \frac{c_k(\hat{P}_n)}{\hat{\sigma}_{k,n}^c} > J_{4,n}^{-1}(1-\alpha, \{1, 2, 3, 4\}, \hat{P}_n)\right\}$$

satisfies

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}_1^c} E_P[\phi_n^4] \leq \alpha. \quad ■$$



(a) Minimum Detectable Violations of $Y_0 \perp\!\!\!\perp Z$ (b) Minimum Detectable Violations of $Y_1 \perp\!\!\!\perp Z$

Fig. 1. Detecting violations of instrument exogeneity, irrelevant instrument ($\delta = 0$).

Appendix A. Numerical results

Below we provide numerical results to complement the analytical results from Section 4. We consider a parametric, latent variable model for Y and D and examine which parameterizations of the model result in detectable violations of the restrictions. We first examine which parameterizations result in detectable violations of instrument exogeneity – Assumption 2.1. We then examine which parameterizations result in detectable violations of each of the monotonicity restrictions while maintaining instrument exogeneity.

A.1. Instrument exogeneity

Consider the following model for Y and D:

$$\begin{aligned} Y_0 &= I\{\alpha + \nu_0 \geq 0\} \\ Y_1 &= I\{\alpha + \nu_1 \geq 0\} \\ D &= I\{\zeta + \delta Z + \eta \geq 0\} \\ \nu_0 &= \lambda_0 Z + \epsilon \\ \nu_1 &= \lambda_1 Z + \epsilon, \end{aligned}$$

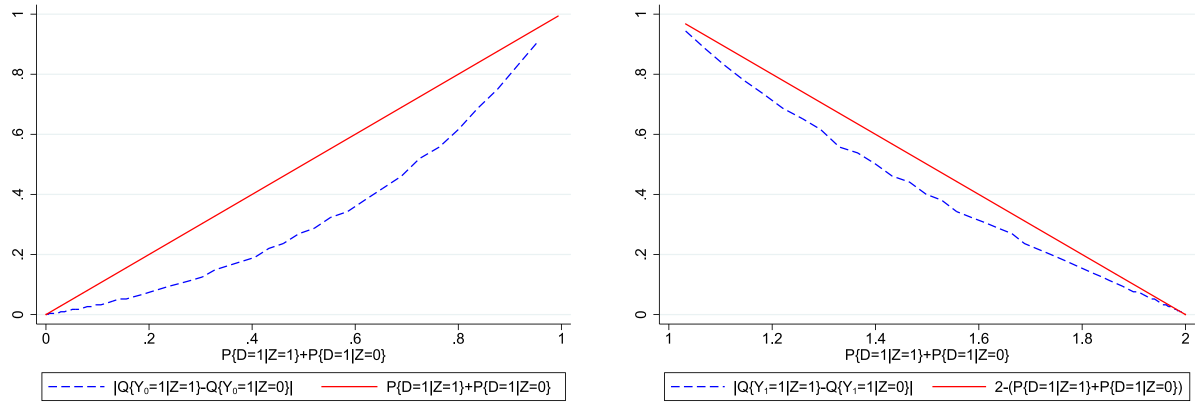
with $\alpha = -2$, $Z \perp\!\!\!\perp (\epsilon, \eta)$, and $(\epsilon, \eta) \sim N(0, I_2)$, where I_2 is the 2-dimensional identity matrix. The outcome Y is determined by (D, Y_0, Y_1) from (2). The parameter λ_d indexes the dependence between Y_d and Z. Using the notation of Section 4.1, we have

$$\Delta_d(Q) = Q\{Y_d = 1 \mid Z = 1\} - Q\{Y_d = 1 \mid Z = 0\} = \Phi(\alpha + \lambda_d) - \Phi(\alpha).$$

Proposition 4.1 relates the ability to detect violations of $Y_d \perp\!\!\!\perp Z$ of given strength of violation $|\Delta_d(Q)|$ to the magnitude of $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$.

We fix δ at either 0 or 0.5, vary $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$ by varying ζ, and, for each resulting $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$, compute the minimal value of $|\Delta_d(Q)|$ for which the violation of $Y_d \perp\!\!\!\perp Z$ is detectible. We then compare this minimal value to the upper bound on the minimal detectible violation from Proposition 4.1. On the lefthand-side of Figs. 1 (for $\delta = 0$) and 2 (for $\delta = 0.5$), we consider values of ζ such that $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} < 1$. From Proposition 4.1, for this range of $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$, we have that no violation of $Y_1 \perp\!\!\!\perp Z$ is detectible and an upper bound on the minimal value of $|\Delta_0(Q)|$ such that violation of $Y_0 \perp\!\!\!\perp Z$ is detectible is given by $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$. We graph in that range the actual minimal value of $\Delta_0(Q)$ for which we detect the violation in blue, and graph the upper bound on that value from Proposition 4.1 in red. On the righthand-side of Figs. 1 and 2, we consider values of ζ such that $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\} > 1$. From Proposition 4.1, for this range of $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$, we have that no violation of $Y_0 \perp\!\!\!\perp Z$ is detectible and an upper bound on the minimal value of $|\Delta_1(Q)|$ such that violation of $Y_1 \perp\!\!\!\perp Z$ is detectible is given by $2 - P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\}$. We graph in that range the actual minimal value of $\Delta_1(Q)$ for which we detect the violation in blue, and graph the upper bound on that value from Proposition 4.1 in red.

We find that the minimal value of $|\Delta_d(Q)|$ for which we can detect violation of $Y_d \perp\!\!\!\perp Z$ is, as it must be, below the upper bound on that minimal value, with the gap between the actual minimum and the upper bound on the minimum



(a) Minimum Detectable Violations of $Y_0 \perp\!\!\!\perp Z$ (b) Minimum Detectable Violations of $Y_1 \perp\!\!\!\perp Z$

Fig. 2. Detecting violations of instrument exogeneity, strong instrument ($\delta = 0.5$).

being modest in magnitude and shrinking to zero as $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$ approaches 0 from the right or approaches 1 from the left (for detecting violations of $Y_0 \perp\!\!\!\perp Z$) and as $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$ approaches 1 from the right or 2 from the left (for detecting violations of $Y_1 \perp\!\!\!\perp Z$). The upper bounds on the minimal detectible violation of $Y_0 \perp\!\!\!\perp Z$ shrink monotonically to zero as $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$ approaches 0, and so does the actual minimum detectible violation. The upper bounds on the minimal detectible violation of $Y_1 \perp\!\!\!\perp Z$ shrinks monotonically to zero as $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$ approaches 2, and so does the actual minimum detectible violation.

A.2. Monotonicity of D in Z

Consider the following model for Y and D:

$$\begin{aligned} Y &= I\{\beta D + \epsilon \geq 0\} \\ D &= I\{\delta Z + \eta \geq 0\} \end{aligned} \tag{40}$$

with $Z \perp\!\!\!\perp (\epsilon, \eta, \beta, \delta)$, $(\epsilon, \eta, \beta, \delta) \sim N(\mu, \Sigma)$, and $E[\delta] > 0$. Note that this model satisfies Assumption 2.1 and that Assumption 2.2 is violated whenever $\text{Var}[\delta] > 0$. $\text{Corr}[\beta, \delta]$ measures the dependence between treatment response to the instrument and outcome response to the treatment. $\text{Var}[\delta]$ and $E[\delta]$ measure the strength of the instrument, which is decreasing in $\text{Var}[\delta]$ and increasing in $E[\delta]$. From Proposition 4.2, we have that the ability to detect violations of Assumption 2.2 is increasing in the size of the violation (increasing in fraction of “defiers”, $Q\{D_1 < D_0\}$), and the maximum possible size of the violation is decreasing in the strength of the instrument (decreasing in $P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\}$). In addition, as explained in Remark 4.2, the ability to detect violations requires sufficient difference in the distribution of potential outcomes among “compliers” and “defiers”. The difference between these distributions is increasing in $|\text{Corr}[\beta, \delta]|$. We therefore examine below how the ability to detect violations of Assumption 2.2 varies with $\text{Var}[\delta]$, $E[\delta]$, and $\text{Corr}[\beta, \delta]$. In particular, we consider parameterizations of (40) with

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mu_\delta \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \sigma_{\beta,\delta} & \sigma_\delta^2 \end{pmatrix}$$

and vary μ_δ from 0.1 to 1, σ_δ^2 from 0.2 to 50, and $\sigma_{\beta,\delta}$ so that $\text{Corr}[\beta, \delta]$ varies from -1 to 1.

Fig. 3 displays the minimum value of $\text{Var}[\delta]$ for which it is possible to detect violations of Assumption 2.2 for different values of $E[\delta]$, $\text{Corr}[\beta, \delta]$. For presentation purposes, we have truncated the graph at 50 for the minimum value of $\text{Var}(\delta)$. The minimum value of $\text{Var}[\delta]$ for which it is possible to detect violations of Assumption 2.2 is increasing in $E[\delta]$, though not dramatically so. In contrast, the minimum value of $\text{Var}[\delta]$ for which it is possible to detect violations of Assumption 2.2 asymptotes to infinity as $\text{Corr}[\beta, \delta]$ approaches zero.

Fig. 4 displays the maximum strength of the instrument, as indexed by $E[\delta]$, for which it is possible to detect violations of Assumption 2.2 for different values of $\text{Var}[\delta]$ and $\text{Corr}[\beta, \delta]$. The maximum value of $E[\delta]$ for which it is possible to detect violations is increasing in $\text{Var}[\delta]$: if the violation is more severe, then the instrument can be stronger with the violation still being detectable. As $\text{Corr}[\beta, \delta]$ approaches zero, the maximum value of $E[\delta]$ for which it is possible to detect violations approaches 0. For any $\text{Corr}[\beta, \delta] \neq 0$, there is a strength of instrument sufficiently weak such that the violation of Assumption 2.2 can still be detected. On the other hand, if $\text{Corr}[\beta, \delta] = 0$, then it is not possible to detect violation of Assumption 2.2 for any value of $E[\delta]$ and $\text{Var}[\delta]$.

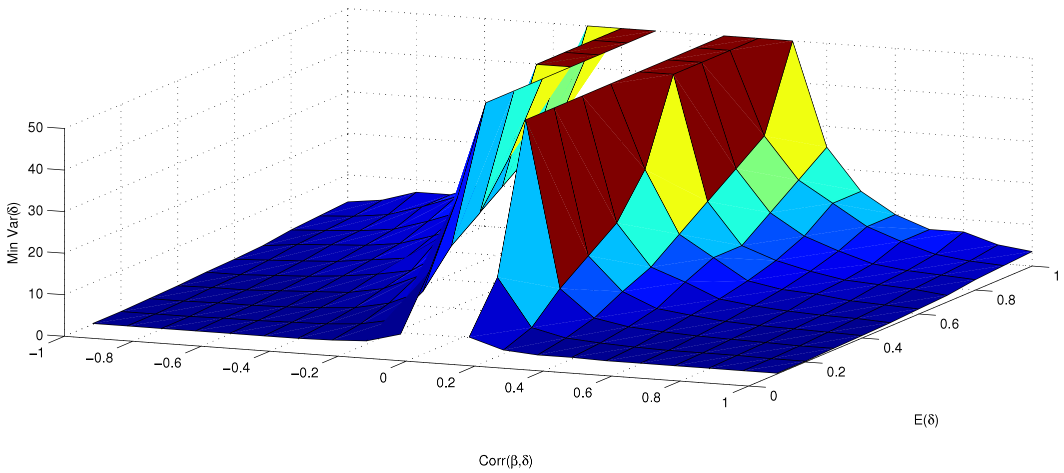


Fig. 3. Detecting violations of D monotonic in Z: Minimum Var[δ].

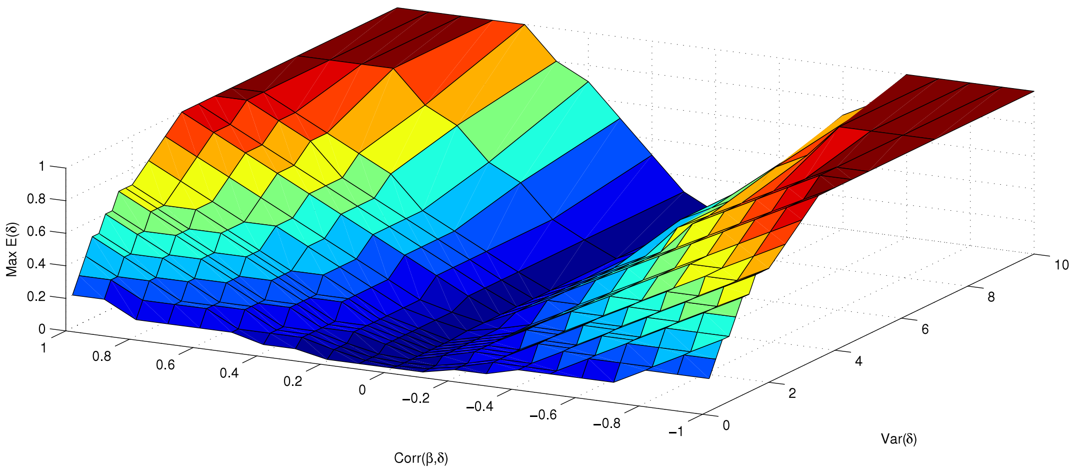


Fig. 4. Detecting violations of D monotonic in Z: Maximum $E[\delta]$.

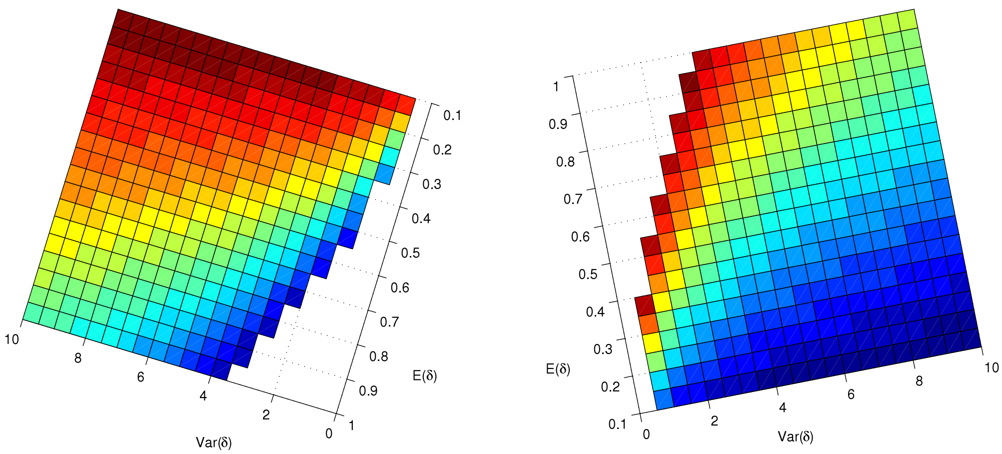


Fig. 5. Detecting violations of D monotonic in Z: Minimum/Maximum Corr[β, δ].

The lefthand-side of Fig. 5 displays the maximum value of Corr[β, δ] < 0 for which we can detect violations of Assumption 2.2 for different values of $E[\delta]$ and Var[δ]; the righthand-side of Fig. 5 displays the minimum value of

$\mathrm{Corr}[\beta, \delta] > 0$ for which we can detect violations of Assumption 2.2 for different values of $E[\delta]$ and $\mathrm{Var}[\delta]$. The figure is plotted from an "overhead" view, with warmer colors indicating higher values for the maximum/minimum value of $\mathrm{Corr}[\beta, \delta]$ for which the violation is detectable and white space for values of $E[\delta]$ and $\mathrm{Var}[\delta]$ for which there is no value of $\mathrm{Corr}[\beta, \delta]$ for which the violation is detectable. The ability to detect the violation of Assumption 2.2 is increasing in $|\mathrm{Corr}(\beta, \delta)|$, but, for a fairly large range of values of $E[\delta]$ and $\mathrm{Var}[\delta]$, there exists no value of $\mathrm{Corr}[\beta, \delta]$ for which the violation is detectable.

A.3. Monotonicity of Y in D

Extensive experimentation revealed that it is difficult to find parameterizations of (40) for which it is possible to detect violations of Assumption 2.3. For example, with $\mathrm{Corr}[\eta, \beta] = 0$, $\mathrm{Corr}[\eta, \delta] = 0$ and $\mathrm{Corr}[\beta, \delta] \approx \pm 1$, we were unable to find any parameterizations for which it is possible to detect violations Assumption 2.3. The only parameterizations we found for which it is possible to detect violations of Assumption 2.3 involved $\mathrm{Corr}[\beta, \delta] \approx 1$, $\mathrm{Corr}[\eta, \beta] \approx -1$, $\mathrm{Corr}[\eta, \delta] \approx -1$, and both $\mathrm{Var}[\beta]$ and $\mathrm{Var}[\delta]$ large. This remained true even for extreme violations of Assumption 2.3, such as $\mathrm{Var}[\beta] = 10{,}000$. The results suggest that in a model of the form of (40), it is difficult to find parameterizations such that the fractions of "always takers" and "never takers" are small enough so that it is possible to detect violations of Assumption 2.3.

Because of the difficulty in finding parameterizations of (40) for which it is possible to detect violations of Assumption 2.3, we consider the following model for Y and D:

$$\begin{aligned} Y &= I\{\beta D + \epsilon \geq 0\} \\ D &= I\{\alpha_t(\delta) + \delta_t(\delta) Z + \eta \geq 0\} \end{aligned} \tag{41}$$

with $Z \perp\!\!\!\perp (\epsilon, \eta, \beta, \delta)$, $(\epsilon, \eta, \beta, \delta) \sim N(\mu, \Sigma)$, and

$$\alpha_t(\delta) = \begin{cases} -t & \text{if } \delta > 0 \\ t & \text{if } \delta \leq 0 \end{cases}$$

$$\delta_t(\delta) = \begin{cases} \delta + 2t & \text{if } \delta > 0 \\ \delta - 2t & \text{if } \delta \leq 0. \end{cases}$$

Here, the parameter $t > 0$ is used as an index to control the fractions of "always takers" and "never takers". In particular, these fractions are decreasing in t. We consider parameterizations of (41) with

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ .1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \sigma_\beta^2 & 0 \\ 0 & 0 & \sigma_{\beta,\delta} & 10 \end{pmatrix}.$$

Extensive experimentation again revealed that it is difficult to find parameterizations of (41) for which it is possible to detect violations of Assumption 2.3 for small values of t, though less difficult as t gets larger (and thus the probability of being an "always taker" or "never taker" approaches zero). For example, when $t \geq 4$, $Q\{D_1 = D_0\} \approx 0$, and, for such t, it is possible to detect violations of Assumption 2.3 if $\mathrm{Corr}[\beta, \delta]$ and $\mathrm{Var}[\beta]$ are sufficiently large, such as $\mathrm{Corr}[\beta, \delta] = .8$ and $\mathrm{Var}[\delta] \geq 3$.

Appendix B. Simulation study

In this section, we investigate the finite-sample performance of our inference procedures developed in Section 5 with a small simulation study. We set:

$$\begin{aligned} Y &= I\{\gamma + \beta D + \nu \geq 0\} \\ D &= I\{\zeta + \delta Z + \eta \geq 0\} \\ \nu &= \lambda Z + \epsilon \end{aligned}$$

with $Z \perp\!\!\!\perp (\epsilon, \eta, \beta, \delta)$, Z binary with $P\{Z = 1\} = P\{Z = 0\} = .5$, $(\epsilon, \eta, \beta, \delta) \sim N(\mu, \Sigma)$, and

$$\mu = \begin{pmatrix} 0 \\ 0 \\ \mu_\beta \\ \mu_\delta \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \sigma_{\eta,\beta} & \sigma_{\eta,\delta} \\ 0 & \sigma_{\eta,\beta} & \sigma_\beta^2 & \sigma_{\beta,\delta} \\ 0 & \sigma_{\eta,\delta} & \sigma_{\beta,\delta} & \sigma_\delta^2 \end{pmatrix}.$$

In Table 1, we list the parameter values for the different designs we consider, and, in Table 2, we report descriptive statistics for each design. Importantly, Assumptions 2.1–2.3 hold in designs (1)–(5), but these designs differ according to the strength of the instrument (indexed by δ) and the strength of the average treatment effect (indexed by β). In this way, these designs allow us to investigate the ability of our inference procedures to correctly determine the sign of ATE

Table 1
Parameterizations of different designs.

Parametrization	Monte Carlo Design								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
μ_β	0	.5	1.5	.5	1.5	0	0	0	0
μ_δ	1.5	.5	1.5	1.5	.5	.5	.5	.5	.5
σ^2_β	0	0	0	0	0	1	5	0	0
σ^2_δ	0	0	0	0	0	5	5	0	0
Corr$[\beta, \delta]$	0	0	0	0	0	.9	.9	0	0
γ	−1	−1	−1	−1	−1	0	0	0	0
ζ	−1	−1	−1	−1	−1	0	0	0	1
Corr$[\eta, \beta]$	0	0	0	0	0	0	−.9	0	0
Corr$[\eta, \delta]$	0	0	0	0	0	0	−.9	0	0
λ	0	0	0	0	0	0	0	1	1

Table 2
Descriptive statistics for different designs.

Statistics	Monte Carlo design								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\Delta(P)$	0.000	0.023	0.284	0.080	0.080	0.097	0.272	0.341	0.341
$A_1(P)$	−0.084	−0.024	−0.084	−0.084	−0.024	0.056	0.203	0.140	0.273
$A_4(P)$	0.084	0.046	0.368	0.164	0.104	0.041	0.069	0.201	0.069
$B_1(P)$	−0.025	−0.049	−0.110	−0.049	−0.110	−0.209	−0.096	−0.049	−0.011
$B_2(P)$	0.110	0.095	0.478	0.213	0.213	0.250	0.250	0.250	0.079
$B_3(P)$	−0.133	−0.133	−0.133	−0.133	−0.095	−0.194	−0.201	−0.110	−0.079
$B_4(P)$	0.049	0.107	0.047	0.049	0.049	0.210	0.181	0.250	0.056
$C_1(P)$	−0.393	−0.262	−0.323	−0.393	−0.205	−0.443	−0.298	−0.299	−0.090
$C_4(P)$	0.243	0.205	0.527	0.323	0.262	0.460	0.431	0.510	0.136
$\Delta(P) - A_3(P)$	−0.182	−0.197	0.186	−0.079	−0.079	−0.363	−0.312	−0.168	−0.136
$\Delta(P) - A_2(P)$	0.285	0.653	0.653	0.389	0.772	0.556	0.549	0.640	0.773
$\Delta(P) - C_3(P)$	−0.182	−0.197	0.186	−0.079	−0.079	−0.363	−0.290	−0.168	0.081
$\Delta(P) - C_2(P)$	0.285	0.653	0.653	0.389	0.772	0.500	0.324	0.409	0.238
$LB, Y \leq D$	−0.182	−0.197	0.186	−0.079	−0.079	−0.363	−0.312	−0.168	−0.136
$UB, Y \geq D$	0.285	0.653	0.653	0.389	0.772	0.500	0.346	0.490	0.500
$E[D_0]$	0.159	0.159	0.159	0.159	0.159	0.500	0.500	0.500	0.841
$E[D_1]$	0.691	0.309	0.691	0.691	0.309	0.581	0.639	0.691	0.933
$E[Y_0]$	0.159	0.159	0.159	0.159	0.159	0.500	0.500	0.670	0.670
$E[Y_1]$	0.159	0.308	0.691	0.308	0.691	0.500	0.500	0.670	0.670
$P\{D_1 > D_0\}$	0.533	0.150	0.533	0.533	0.150	0.225	0.437	0.191	0.092
$P\{D_1 < D_0\}$	0.000	0.000	0.000	0.000	0.000	0.144	0.299	0.000	0.000
$P\{D_1 = D_0\}$	0.467	0.850	0.467	0.467	0.850	0.631	0.264	0.809	0.908
$P\{Y_1 > Y_0\}$	0.000	0.150	0.533	0.150	0.533	0.125	0.183	0.000	0.000
$Pr(Y_1 < Y_0\}$	0.000	0.000	0.000	0.000	0.000	0.125	0.183	0.000	0.000
$P\{Y_1 = Y_0\}$	1.000	0.850	0.467	0.850	0.467	0.750	0.634	1.000	1.000
Corr$[\nu, Z]$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.447	0.447
$P\{Y_1 = 1 \vert Z = 1\} - P\{Y_1 = 1 \vert Z = 0\}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.341	0.341
$P\{Y_0 = 1 \vert Z = 1\} - P\{Y_0 = 1 \vert Z = 0\}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.341	0.341
$P\{D = 1 \vert Z = 1\} + P\{D = 1 \vert Z = 0\}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.192	1.774

Descriptive statistics are computed using 30000000 replications.

Table 3
Which null hypotheses are false.

Restrictions	Design								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
2.1	H_1	H_1	H_1 & H_2	H_1	H_1	H_1	H_1	H_1	–
2.1, 2.2	H_1	H_1	H_1 & H_2	H_1	H_1	–	–	–	–
2.1, 2.2, 2.3	H_1	H_1 & H_2	H_1 & H_2	H_1 & H_2	H_1 & H_2	–	–	–	–
2.1, 2.3	H_1	H_1	H_1 & H_2	H_1 & H_2	H_1 & H_2	H_1	–	–	–

when these restrictions hold, and how that ability varies with the strength of the treatment effect and the strength of the instrument. In contrast, Assumptions 2.2 and 2.3 are both violated in designs (6)–(7) and Assumption 2.1 is violated in designs (8)–(9). In this way, these designs allow us to investigate the ability of our inference procedure to correctly detect violations of these restrictions. For convenience, in Table 3 we list which null hypotheses are false for each design and set of restrictions.

Table 4
Rejection probabilities for designs (1)–(3).

Restrictions	n	Design					
		1		2		3	
		$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1, H_2\}$	$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1, H_2\}$	$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1, H_2\}$
2.1	200	0.997	0.000	0.889	0.000	0.998	0.528
	500	1.000	0.000	1.000	0.000	1.000	0.953
	1000	1.000	0.000	1.000	0.000	1.000	1.000
	5000	1.000	0.000	1.000	0.000	1.000	1.000
	10000	1.000	0.000	1.000	0.000	1.000	1.000
2.1, 2.2	200	0.471	0.000	0.017	0.000	0.605	0.448
	500	0.929	0.000	0.107	0.000	0.942	0.933
	1000	0.999	0.000	0.240	0.000	0.999	0.999
	5000	1.000	0.000	0.790	0.000	1.000	1.000
	10000	1.000	0.000	0.970	0.000	1.000	1.000
2.1, 2.2, 2.3	200	0.471	0.003	0.017	0.000	0.605	0.599
	500	0.929	0.029	0.107	0.001	0.942	0.942
	1000	0.999	0.051	0.240	0.002	0.999	0.999
	5000	1.000	0.056	0.790	0.443	1.000	1.000
	10000	1.000	0.051	0.970	0.863	1.000	1.000
2.1, 2.3	200	0.588	0.000	0.213	0.000	0.970	0.826
	500	0.982	0.000	0.694	0.000	1.000	0.999
	1000	1.000	0.000	0.939	0.000	1.000	1.000
	5000	1.000	0.000	1.000	0.000	1.000	1.000
	10000	1.000	0.000	1.000	0.000	1.000	1.000

Table 5
Rejection probabilities for designs (4)–(5).

Restrictions	n	Design			
		4		5	
		$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1, H_2\}$	$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1, H_2\}$
2.1	200	0.998	0.000	0.892	0.000
	500	1.000	0.000	1.000	0.000
	1000	1.000	0.000	1.000	0.000
	5000	1.000	0.000	1.000	0.000
	10000	1.000	0.000	1.000	0.000
2.1, 2.2	200	0.605	0.001	0.016	0.000
	500	0.942	0.000	0.101	0.000
	1000	0.999	0.000	0.233	0.000
	5000	1.000	0.000	0.789	0.000
	10000	1.000	0.000	0.970	0.000
2.1, 2.2, 2.3	200	0.605	0.149	0.016	0.002
	500	0.942	0.628	0.101	0.030
	1000	0.999	0.923	0.233	0.166
	5000	1.000	1.000	0.789	0.789
	10000	1.000	1.000	0.970	0.970
2.1, 2.3	200	0.804	0.005	0.548	0.001
	500	0.998	0.043	0.967	0.025
	1000	1.000	0.108	1.000	0.069
	5000	1.000	0.570	1.000	0.458
	10000	1.000	0.910	1.000	0.787

In the simulations, we consider sample sizes of $n = 200$, 500, 1000, 5000 and 10000. For each design, we perform our inference procedures under each of our alternative sets of restrictions: Assumption 2.1 alone, Assumptions 2.1 and 2.2, Assumptions 2.1 and 2.3, and Assumptions 2.1–2.3. For each test, we use a 5% nominal significance level and 500 bootstrap replications when computing the relevant critical values. All results are reported based on 3000 simulations.

Tables 4 and 5 report our results for designs (1)–(5). Designs (1), (3) and (4) have a stronger instrument ($\delta = 1.5$, corresponding to $P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\} = 0.53$), while designs (2) and (5) have a weaker instrument ($\delta = 0.5$, corresponding to $P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\} = 0.15$). Designs (3) and (5) have larger, positive ATEs ($\beta = 1.5$, corresponding to ATE of 0.53), designs (2) and (4) have smaller, positive ATEs ($\beta = 0.5$, corresponding to ATEs of 0.15), and design (1) has an ATE of zero.

The left column for each design reports the probability of rejecting H_1 and thereby correctly concluding that P is consistent with the restrictions. We find that the probability of correctly rejecting H_1 is higher in specifications with a stronger instrument, and also higher under assumptions with weaker testable restrictions. The right column for each

Table 6
Rejection probabilities for designs (6)–(9).

Restrictions	n	Design			
		6	7	8	9
		$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1\}$	$P\{\text{rej. } H_1\}$
2.1	200	0.999	0.636	0.733	0.000
	500	1.000	0.959	0.989	0.000
	1000	1.000	1.000	1.000	0.000
	5000	1.000	1.000	1.000	0.000
	10000	1.000	1.000	1.000	0.000
2.1, 2.2	200	0.000	0.000	0.000	0.000
	500	0.000	0.000	0.000	0.000
	1000	0.000	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000
2.1, 2.2, 2.3	200	0.000	0.000	0.000	0.000
	500	0.000	0.000	0.000	0.000
	1000	0.000	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000
2.1, 2.3	200	0.621	0.012	0.000	0.000
	500	0.952	0.005	0.000	0.000
	1000	1.000	0.004	0.000	0.000
	5000	1.000	0.000	0.000	0.000
	10000	1.000	0.000	0.000	0.000

design reports the probability of rejecting both H_1 and H_2, and thereby concluding that P is consistent with the restrictions and a positive effect. Recall that ATE is zero in design (1) and positive in designs (2)–(5), though not always identified to be positive. From these tables, we see that the procedure is generally conservative, in that, in those cases where H_2 is true, it falsely rejects H_1 and H_2 with probability less than the nominal size. In those cases where both H_1 and H_2 are false, we have very high power to reject H_1 and H_2, and thus to correctly conclude that ATE is positive, when both the instrument and the treatment effect are strong; lower, but still substantial, power when the instrument is strong but the treatment effect is weak; and slightly lower power still when the treatment effect is strong, but the instrument is weak. When both the instrument and the effect are weak (design 2), we identify ATE to be positive only under Assumptions 2.1–2.3, and in that case only have power above nominal size for $n \geq 5000$. As confirmed below by Table 9, this low power is a result of low power to reject H_1 and thus conclude that the P is consistent with the assumptions.

In these tables, we do not report the probability of rejecting H_1 and H_3, as it was estimated to be 0.000 in all cases.

One somewhat paradoxical result from the tables is that in design (3), which features both a strong instrument and a strong treatment effect, the power to conclude correctly that the ATE is positive is slightly weaker under Assumptions 2.1–2.3 as compared to under Assumptions 2.1 and 2.3. This result seems paradoxical as we are imposing more assumptions in the former case than the later, and we have shown that we have great ability to determine the sign of the ATE at the population level in the former case than in the later. The explanation for the paradoxical result is that, because the former case imposes more restrictions on the observable data, it is more difficult to reject H_1 in the former case than in the later case.

Table 6 reports the probability of rejecting H_1 for designs (6)–(9). In designs (6) and (7), Assumption 2.1 holds, but not Assumption 2.2 or 2.3. In designs (8) and (9), the instrument is endogenous, so that Assumption 2.1 is violated, though in a way that is detectable under Assumption 2.1 alone for design (9), but not for design (8). In this table, we only report the probability of rejecting H_1, as the probability of rejecting H_1 and H_2 or of rejecting H_1 and H_3 was estimated to be 0.000 in all cases. We find that the probability of incorrectly rejecting H_1 is very low for all cases for which H_1 is true, while the probability of correctly rejecting H_1 is substantial when H_1 is false. It is worth noting that these statements are for the probability of correctly concluding whether P is consistent with the restrictions, not for the probability of correctly concluding whether the restrictions are valid. For example, in design (8), the procedure with high probability correctly concludes that the data is consistent with Assumption 2.1 even though the instrument is endogenous (in a way that is not detectable).

Tables 7 and 8 report our results for testing only H_2 and H_3 simultaneously as suggested in Remark 5.1. In these tables, we only report the probability of rejecting H_2, as the probability of rejecting H_3 was estimated to be 0.000 in all cases. For designs (2)–(5), where the restrictions hold and the true ATE is positive, we find some increase in power to correctly reject H_2 for the smaller sample sizes compared to our procedure that uses H_1 as a "gatekeeper". In all cases, we have greater power to reject H_2 correctly under Assumptions 2.1–2.3 compared to under Assumptions 2.1 and 2.3. Thus, the paradoxical result described above does not occur when not using H_1 as a "gatekeeper".

The results for designs (2)–(5) show that there is a cost to using H_1 as a "gatekeeper", in that there is some increase in power to reject H_2 correctly for small sample sizes when not using H_1 in this way. In contrast, the results for designs

Table 7
Rejection probabilities for testing only H_2 and H_3 simultaneously for designs (1)–(5).

Restrictions	n	Design				
		1	2	3	4	5
		$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$
2.1	200	0.000	0.000	0.531	0.000	0.000
	500	0.000	0.000	0.953	0.000	0.000
	1000	0.000	0.000	1.000	0.000	0.000
	5000	0.000	0.000	1.000	0.000	0.000
	10000	0.000	0.000	1.000	0.000	0.000
2.1, 2.2	200	0.000	0.000	0.783	0.001	0.001
	500	0.000	0.000	0.991	0.000	0.000
	1000	0.000	0.000	1.000	0.000	0.000
	5000	0.000	0.000	1.000	0.000	0.000
	10000	0.000	0.000	1.000	0.000	0.000
2.1, 2.2, 2.3	200	0.057	0.108	0.994	0.394	0.345
	500	0.052	0.163	1.000	0.685	0.625
	1000	0.051	0.250	1.000	0.924	0.879
	5000	0.056	0.649	1.000	1.000	1.000
	10000	0.051	0.892	1.000	1.000	1.000
2.1, 2.3	200	0.000	0.000	0.856	0.018	0.020
	500	0.000	0.000	0.999	0.044	0.039
	1000	0.000	0.000	1.000	0.108	0.069
	5000	0.000	0.000	1.000	0.570	0.458
	10000	0.000	0.000	1.000	0.910	0.787

Table 8
Rejection probabilities for testing only H_2 and H_3 simultaneously for designs (6)–(9).

Restrictions	n	Design			
		6	7	8	9
		$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$	$P\{\text{rej. } H_2\}$
2.1	200	0.000	0.000	0.000	0.056
	500	0.000	0.000	0.000	0.128
	1000	0.000	0.000	0.000	0.266
	5000	0.000	0.000	0.000	0.946
	10000	0.000	0.000	0.000	1.000
2.1, 2.2	200	0.000	0.000	0.000	0.000
	500	0.000	0.000	0.000	0.000
	1000	0.000	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000
2.1, 2.2, 2.3	200	0.393	0.987	1.000	1.000
	500	0.697	1.000	1.000	1.000
	1000	0.929	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000
2.1, 2.3	200	0.000	0.026	0.007	0.015
	500	0.000	0.145	0.008	0.048
	1000	0.000	0.423	0.007	0.122
	5000	0.000	1.000	0.006	0.744
	10000	0.000	1.000	0.004	0.976

(6)–(9) highlight the advantage of using H_1 as a "gatekeeper". In these designs, the true value of ATE is zero, and H_2 is true in all cases, so that any rejection of H_2 is a false rejection. In these designs, the procedure that only tests H_2 and H_3 often incorrectly rejects H_2 and thus incorrectly concludes that the ATE is positive. For example, consider the results under Assumptions 2.1–2.3. Note that P is incompatible with that set of restrictions under any of the designs (6)–(9). When we use our procedure that includes H_1 a "gatekeeper", the procedure incorrectly rejected H_1 with probability 0.000 in each of those designs for each sample size considered, and thus incorrectly rejected H_2 with probability 0.000 as well. In contrast, when not using H_1 as a "gatekeeper", as reported in Table 8, the procedure incorrectly rejects H_2 and thus incorrectly concludes that the ATE is positive with very high probability. Thus, we find that while using H_1 as a "gatekeeper" does somewhat decrease our power to determine the sign of the ATE correctly when the restrictions are true, it also greatly reduces the probability of incorrectly determining the sign of the ATE when the restrictions are incompatible with the data.

Table 9
Rejection probabilities for testing $H_1^c : P \in \mathbf{P}_1^c$ for designs (1)–(5).

Restrictions	n	Design				
		1	2	3	4	5
		$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$
2.1	200	0.000	0.000	0.000	0.000	0.000
	500	0.000	0.000	0.000	0.000	0.000
	1000	0.000	0.000	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000	0.000
2.1, 2.2	200	0.000	0.003	0.000	0.000	0.004
	500	0.000	0.002	0.000	0.000	0.002
	1000	0.000	0.001	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000	0.000
2.1, 2.2, 2.3	200	0.000	0.003	0.000	0.000	0.004
	500	0.000	0.002	0.000	0.000	0.002
	1000	0.000	0.001	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000	0.000
2.1, 2.3	200	0.000	0.000	0.000	0.000	0.000
	500	0.000	0.000	0.000	0.000	0.000
	1000	0.000	0.000	0.000	0.000	0.000
	5000	0.000	0.000	0.000	0.000	0.000
	10000	0.000	0.000	0.000	0.000	0.000

Table 10
Rejection probabilities for testing $H_1^c : P \in \mathbf{P}_1^c$ for designs (6)–(9).

Restrictions	N	Design			
		6	7	8	9
		$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$	$P\{\text{rej. } H_1^c\}$
2.1	200	0.000	0.000	0.000	0.862
	500	0.000	0.000	0.000	1.000
	1000	0.000	0.000	0.000	1.000
	5000	0.000	0.000	0.000	1.000
	10000	0.000	0.000	0.000	1.000
2.1, 2.2	200	0.114	0.817	0.665	0.973
	500	0.254	0.998	0.976	1.000
	1000	0.485	1.000	1.000	1.000
	5000	0.997	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000
2.1, 2.2, 2.3	200	0.114	0.817	0.665	0.973
	500	0.254	0.998	0.976	1.000
	1000	0.485	1.000	1.000	1.000
	5000	0.997	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000
2.1, 2.3	200	0.000	0.020	0.150	0.967
	500	0.000	0.041	0.473	1.000
	1000	0.000	0.062	0.866	1.000
	5000	0.000	0.254	1.000	1.000
	10000	0.000	0.509	1.000	1.000

Tables 9 and 10 report results that follow Remark 5.2 in testing the null hypothesis that P is consistent with our restrictions, $P \in \mathbf{P}_1^c$, as opposed to H_1 above, which specifies that $P \in \mathbf{P}_1$. In other words, these results are for a model specification test, where the null is correct specification. In designs (1)–(5), P is consistent with each alternative set of restrictions, and we find that in all cases the test rejects $P \in \mathbf{P}_1^c$ incorrectly with probability less than nominal size. In designs (6)–(9), we find generally substantial power to correctly reject that P is consistent with the restrictions. We find higher power to detect violations of those sets of assumptions that impose stronger testable restrictions than those that impose weaker testable restrictions. This finding is in contrast to the results for testing H_1, where we found it more difficult to correctly reject the null for sets of assumptions that implied stronger testable restrictions than for those that implied weaker testable restrictions.

The results for design (8) highlight one possible reason why a researcher who is confident in monotonicity of D in Z and less confident of instrument exogeneity might wish to perform inference maintaining monotonicity of D in

Z: maintaining monotonicity of D in Z makes it far easier to detect violations of instrument exogeneity. In design (8), Assumption 2.1 fails, though the violation of instrument exogeneity is not detectible under Assumption 2.1 alone while it is detectible under Assumptions 2.1 and 2.2. For this design, as reported in Table 10, we see that the violation of instrument exogeneity is correctly detected with very high probability under Assumptions 2.1 and 2.2, while it is not detectible under Assumption 2.1 alone.

Appendix C. Proofs for Section 3

Proof of Theorem 3.1. First consider assertion (i). For $1 \le j \le 2$, $\Delta(P) = A_1^j(P) + A_4^j(P)$, so that $\Delta(P) \ge A_1^j(P) \implies A_4^j(P) \ge 0$ and $\Delta(P) \le A_4^j(P) \implies A_1^j(P) \le 0$. Thus, $\Delta(P) \in [A_1(P), A_4(P)]$ if and only if $A_1(P) \le 0$ and $A_4(P) \ge 0$. The result then follows from Balke and Pearl (1997). Now consider assertions (ii) and (iii). From Balke and Pearl (1997), the identified set for $E_Q[Y_1 - Y_0]$ is given by $[\Delta(P) - A_3(P), \Delta(P) - A_2(P)]$. Combining this result with (i) gives the stated results. ■

Proof of Theorem 3.2. The proof follows the same strategy as in Balke and Pearl (1997), who solve a linear programming problem that maximizes/minimizes the average treatment effect and has as constraints the restrictions between the unobserved latent probability and the distribution of the observed data satisfying Assumptions 2.1 and 2.2. Imposing Assumption 2.3 in addition to Assumptions 2.1 and 2.2 results in the additional constraints in the optimization problem that any candidate Q satisfy either

$$Q\{Y_1 > Y_0, D_1 = j, D_0 = k\} = 0 \text{ for all } (j,k) \in \{0,1\}^2, \tag{42}$$

or

$$Q\{Y_1 < Y_0, D_1 = j, D_0 = k\} = 0 \text{ for all } (j,k) \in \{0,1\}^2. \tag{43}$$

Testable restrictions arise by characterizing admissible values of observed probabilities under which the linear programming problem is feasible, which amounts to checking whether the dual problem is unbounded. We compute the maximum and minimum values for the average treatment effect and specify the conditions that rule out unboundedness of the dual as the testable restrictions. Following this procedure for Q satisfying (43) results in the restriction that $\Delta(P) \in [A_1(P), 0]$, while following this procedure for Q satisfying (42) results in the restriction $\Delta(P) \in [0, A_4(P)]$. The result now follows. ■

Proof of Theorem 3.3. As in the proof of Theorem 3.1, we follow the same linear programming strategy as in Balke and Pearl (1997), but with modifications to the constraint set for the optimization problem. In particular, under Assumptions 2.1 and 2.3, we have the same constraints as Balke and Pearl (1997) except replacing their constraints that

$$Q\{Y_1 = j, Y_0 = k, D_1 < D_0\} = 0 \text{ for all } (j,k) \in \{0,1\}^2,$$

with the constraints that any candidate Q satisfy either

$$Q\{Y_1 < Y_0, D_1 = j, D_0 = k\} = 0 \text{ for all } (j,k) \in \{0,1\}^2, \tag{44}$$

or

$$Q\{Y_1 > Y_0, D_1 = j, D_0 = k\} = 0 \text{ for all } (j,k) \in \{0,1\}^2. \tag{45}$$

Solving the resulting optimization problem for Q satisfying (45) results in the restriction that $\Delta(P) \in [B_1(P), B_2(P)]$ with ATE bounded from below by $|\Delta(P)|$, while solving the resulting optimization problem for Q satisfying (44) results in the restriction that $\Delta(P) \in [B_3(P), B_4(P)]$ with ATE bounded from above by $-|\Delta(P)|$. The result now follows. ■

Proof of Theorem 3.4. Proof follows immediately from the results of Balke and Pearl (1997).

Appendix D. Proofs for Section 4

Proof of Proposition 4.1. $\Delta(P) \in [C_1(P), C_4(P)]$ is equivalent to the following four equalities holding:

$$\begin{aligned}
&P\{Y=1, D=0 \mid Z=1\} + P\{Y=0, D=0 \mid Z=0\} \le 1\\
&P\{Y=1, D=0 \mid Z=0\} + P\{Y=0, D=0 \mid Z=1\} \le 1\\
&P\{Y=1, D=1 \mid Z=1\} + P\{Y=0, D=1 \mid Z=0\} \le 1\\
&P\{Y=1, D=1 \mid Z=0\} + P\{Y=0, D=1 \mid Z=1\} \le 1.
\end{aligned} \tag{46}$$

Plugging $\Delta_d(Q)$, $G_d^0(Q)$, $G_d^1(Q)$, $d \in \{0,1\}$, into (46), we can rewrite the expression as

$$\begin{aligned}
\Delta_0(Q) &\le G_0^2(Q)\\
\Delta_0(Q) &\ge G_0^1(Q)\\
\Delta_1(Q) &\le G_1^2(Q)\\
\Delta_1(Q) &\ge G_1^1(Q).
\end{aligned} \tag{47}$$

Each inequality in (46) holds if and only if the corresponding inequality in (47) holds, and we have thus established the first assertion of the proposition. Part (i) of the proposition now follows from the absolute value of the righthand-side of the first two inequalities in (47) being bounded from above by $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$, and part (ii) follows from the absolute value of the righthand-side of the last two inequalities in (47) being bounded from above by $2 - P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\}$. Part (iii) of the proposition follows from the lefthand-side of the first two terms of (46) being bounded from above by $2 - P\{D = 1 \mid Z = 1\} - P\{D = 1 \mid Z = 0\}$, while part (iv) follows from the lefthand-side of the last two terms of (46) are bounded from above by $P\{D = 1 \mid Z = 1\} + P\{D = 1 \mid Z = 0\}$. ■

Proof of Proposition 4.2. Using (2)–(3) and Assumption 2.1, $\Delta(P)$ may be expressed as

$$\Big(Q\{Y_1 > Y_0, D_1 > D_0\} - Q\{Y_1 < Y_0, D_1 > D_0\}\Big) - \Big(Q\{Y_1 > Y_0, D_1 < D_0\} - Q\{Y_1 < Y_0, D_1 < D_0\}\Big). \tag{48}$$

Furthermore,

$$\begin{aligned}
A_1^1(P) &= Q\{Y_0 = 1, D_1 < D_0\} - Q\{Y_0 = 1, D_1 > D_0\}\\
A_1^2(P) &= Q\{Y_1 = 0, D_1 < D_0\} - Q\{Y_1 = 0, D_1 > D_0\}\\
A_4^1(P) &= Q\{Y_1 = 1, D_1 > D_0\} - Q\{Y_1 = 1, D_1 < D_0\}\\
A_4^2(P) &= Q\{Y_0 = 0, D_1 > D_0\} - Q\{Y_0 = 0, D_1 < D_0\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\Delta(P) - A_1^1(P) &= \quad Q\{Y_1 = 1, D_1 > D_0\} - Q\{Y_1 = 1, D_1 < D_0\}\\
\Delta(P) - A_1^2(P) &= \quad Q\{Y_0 = 0, D_1 > D_0\} - Q\{Y_0 = 0, D_1 < D_0\}\\
\Delta(P) - A_4^1(P) &= -Q(Y_0 = 1, D_1 > D_0\} + Q\{Y_0 = 1, D_1 < D_0\}\\
\Delta(P) - A_4^2(P) &= -Q\{Y_1 = 0, D_1 > D_0\} + Q\{Y_1 = 0, D_1 < D_0\}.
\end{aligned}$$

The desired result now follows immediately. ■

Proof of Proposition 4.3. Using Assumption 2.1, we have that

$$\begin{aligned}
B_1(P) &= -\min\left\{\begin{array}{l} Q\{Y_1 > Y_0, D_1 < D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_0 = 1\},\\ Q\{Y_1 > Y_0, D_1 < D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 0, D_1 = 0\}\end{array}\right\},\\
B_2(P) &= \min\left\{\begin{array}{l} Q\{Y_1 > Y_0, D_1 > D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 1, D_1 = 1\},\\ Q\{Y_1 > Y_0, D_1 > D_0\} + Q\{Y_1 > Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 0, D_0 = 0\}\end{array}\right\},\\
B_3(P) &= -\min\left\{\begin{array}{l} Q\{Y_1 < Y_0, D_1 > D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_1 = 1\}\\ Q\{Y_1 < Y_0, D_1 > D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_0 = Y_1 = 1, D_0 = 0\}\end{array}\right\},\\
B_4(P) &= \min\left\{\begin{array}{l} Q\{Y_1 < Y_0, D_1 < D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 1\} + Q\{Y_1 = Y_0 = 0, D_0 = 1\}\\ Q\{Y_1 < Y_0, D_1 < D_0\} + Q\{Y_1 < Y_0, D_1 = D_0 = 0\} + Q\{Y_1 = Y_0 = 1, D_1 = 0\}\end{array}\right\}.
\end{aligned}$$

so that

$$\begin{aligned}
B_1(P) &= -Q\{Y_1 > Y_0, D_1 < D_0\} - M_1(Q)\\
B_2(P) &= \quad Q\{Y_1 > Y_0, D_1 > D_0\} + M_2(Q)\\
B_3(P) &= -Q\{Y_1 < Y_0, D_1 > D_0\} - M_3(Q)\\
B_4(P) &= \quad Q\{Y_1 < Y_0, D_1 < D_0\} + M_4(Q).
\end{aligned}$$

The desired result now follows immediately. ■

Appendix E. Proofs for Section 5

The proofs of Theorems 5.1 and 5.4 are essentially the same, so we only provide a proof of Theorem 5.1.

Proof of Theorem 5.1. Suppose by way of contradiction that (24) fails. Then there exists a subsequence $\{P_{n_m} \in \mathbf{P} : m \geq 1\}$ and $\alpha' > \alpha$ such that

$$FWER_{P_{n_m}} \to \alpha'. \tag{49}$$

Let

$$I(P) = \{1 \leq j \leq 3 : P \in \mathbf{P}_j\} \subseteq \{1, 2, 3\}.$$

Since there are only finitely many possible values for $I(P)$ and $FWER_P = 0$ when $I(P) = \emptyset$, we may assume further (by considering another subsequence if necessary) that $I(P_{n_m}) = I \neq \emptyset$.

Consider first the case in which $1 \in I$. Note that

$$\{P_{n_m} \in \mathbf{P} : m \geq 1\} \subseteq \bigcup_{K \in \mathcal{K}_1^1} \mathbf{P}_K,$$

where

$$\mathbf{P}_K = \bigcap_{k \in K} \{P \in \mathbf{P} : a_k(P) \leq 0\}.$$

Since there are only finitely many $K \in \mathcal{K}_1^1$, we may assume further (by considering another subsequence if necessary) that there is $K^* \in \mathcal{K}_1^1$ such that

$$\{P_{n_m} \in \mathbf{P} : m \geq 1\} \subseteq \mathbf{P}_{K^*}. \tag{50}$$

Using the fact that $1 \in I$ and the definition of Algorithm 5.1, we have that

$$\begin{aligned} FWER_{P_{n_m}} &= P_{n_m}\left\{T^1_{1,n_m} > \hat{c}_{1,n_m}(\mathcal{K}^1(\{1\}), 1-\alpha)\right\} \\ &\leq P_{n_m}\left\{\max_{k \in K^*} \frac{a_k(\hat{P}_{n_m})}{\hat{\sigma}^a_{k,n_m}} > J^{-1}_{1,n_m}(1-\alpha, K^*, \hat{P}_{n_m})\right\}, \end{aligned} \tag{51}$$

where in (51) we have used the definitions of T^1_{1,n_m} and $\hat{c}_{1,n_m}(\mathcal{K}^1(\{1\}), 1-\alpha)$ as well as the fact that $K^* \in \mathcal{K}_1^1 = \mathcal{K}^1(\{1\})$. Using (50) and Theorem F.1, we see that the righthand-side of (51) tends to α, contradicting (49), and thereby establishing the desired result.

Now consider the case in which $1 \notin I$. Since $I \subseteq \{2, 3\}$, it must be the case that

$$\{P_{n_m} \in \mathbf{P} : m \geq 1\} \subseteq \bigcup_{K \in \mathcal{K}^1(I)} \mathbf{P}_K.$$

Since there are only finitely many $K \in \mathcal{K}^1(I)$, we may assume further (by considering another subsequence if necessary) that there is $K^* \in \mathcal{K}^1(I)$ such that

$$\{P_{n_m} \in \mathbf{P} : m \geq 1\} \subseteq \mathbf{P}_{K^*}. \tag{52}$$

Next, note that

$$\begin{aligned} \max_{j \in I} T^1_{j,n} &= \max_{j \in I} \min_{K \in \mathcal{K}_j^1} \max_{k \in K} \frac{a_k(\hat{P}_n)}{\hat{\sigma}^a_{k,n}} \\ &\leq \max_{k \in K^*} \frac{a_k(\hat{P}_n)}{\hat{\sigma}^a_{k,n}}. \end{aligned} \tag{53}$$

To establish (53), simply note that from the definition of $\mathcal{K}^1(I)$ that for each $j \in I \subseteq \{2, 3\}$ it must be the case that there exists $K \in \mathcal{K}_j^1$ such that $K \subseteq K^*$. The desired inequality thus follows. Since there exists $K \in \mathcal{K}^1(\{2, 3\})$ such that $K^* \subseteq K$, we have further that

$$\begin{aligned} \hat{c}_{1,n}(\mathcal{K}^1(\{2,3\}), 1-\alpha) &= \max_{K \in \mathcal{K}^1(\{2,3\})} J^{-1}_{1,n}(1-\alpha, K, \hat{P}_n) \\ &\geq J^{-1}_{1,n}(1-\alpha, K^*, \hat{P}_n). \end{aligned} \tag{54}$$

Using the fact that $1 \notin I$ and the definition of Algorithm 5.1, we have that

$$\begin{aligned} FWER_{P_{n_m}} &= P_{n_m}\left\{\max_{j \in I} T^1_{j,n_m} > \hat{c}_{1,n_m}(\mathcal{K}^1(\{2,3\}), 1-\alpha)\right\} \\ &\leq P_{n_m}\left\{\max_{k \in K^*} \frac{a_k(\hat{P}_{n_m})}{\hat{\sigma}^a_{k,n_m}} > J^{-1}_{1,n_m}(1-\alpha, K^*, \hat{P}_{n_m})\right\}, \end{aligned} \tag{55}$$

where (55) follows from (53) and (54). Using (52) and Theorem F.1, we see that the righthand-side of (55) tends to α, contradicting (49), and thereby establishing the desired result.

Appendix F. Auxiliary results

In this appendix, we establish the following result:

Theorem F.1. *Let* (X_i, Y_i, Z_i), $i = 1, \ldots, n$ *be an i.i.d. sequence of random variables with distribution* $P \in \mathbf{P}$ *on* $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. *Suppose* $\mathbf{P}$ *is such that*

$$\epsilon < \inf_{P \in \mathbf{P}} P\{Z = 1\} \leq \sup_{P \in \mathbf{P}} P\{Z = 1\} < 1 - \epsilon \tag{56}$$

for some $\epsilon > 0$, *and for each* $1 \leq j \leq k$ *that*

$$\limsup_{\lambda \to \infty} \sup_{P \in \mathbf{P}} E_P \left[\left(\frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)} \right)^2 I \left\{ \left| \frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)} \right| > \lambda \right\} |Z = 1 \right] = 0 \tag{57}$$

and

$$\limsup_{\lambda \to \infty} \sup_{P \in \mathbf{P}} E_P \left[\left(\frac{Y_j - \mu_{Y_j|Z=0}(P)}{\sigma_{Y_j|Z=0}(P)} \right)^2 I \left\{ \left| \frac{Y_j - \mu_{Y_j|Z=0}(P)}{\sigma_{Y_j|Z=0}(P)} \right| > \lambda \right\} |Z = 0 \right] = 0. \tag{58}$$

Let

$$J_n(x, P) = P \left\{ \max_{1 \leq j \leq k} T_{n,j}(P) \leq x \right\}, \tag{59}$$

where

$$T_{n,j}(P) = \frac{\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{j,i} - \mu_{X_j|Z=1}(P) - \frac{1}{n_0} \sum_{1 \leq i \leq n : Z_i = 0} Y_{j,i} - \mu_{Y_j|Z=0}(P)}{\sqrt{\frac{\sigma^2_{X_j|Z=1}(\hat{P}_n)}{n_1} + \frac{\sigma^2_{Y_j|Z=0}(\hat{P}_n)}{n_0}}}.$$

Then

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}} P \left\{ \max_{1 \leq j \leq k} T_{n,j}(P) > J_n^{-1}(1 - \alpha, \hat{P}_n) \right\} \leq \alpha.$$

Before presenting the proof of Theorem F.1, we present a series of useful lemmata.

Lemma F.1. *Let* (X_i, Z_i), $i = 1, \ldots, n$ *be an i.i.d. sequence of random variables with distribution* $P \in \mathbf{P}$ *on* $\mathbf{R} \times \{0, 1\}$. *Suppose* $\mathbf{P}$ *is such that*

$$\inf_{P \in \mathbf{P}} P\{Z = 1\} > \epsilon$$

for some $\epsilon > 0$ *and that*

$$\limsup_{\lambda \to \infty} \sup_{P \in \mathbf{P}} E_P \left[\left| X - \mu_{X|Z=1}(P) \right| I \left\{ \left| X - \mu_{X|Z=1}(P) \right| > \lambda \right\} |Z = 1 \right] = 0. \tag{60}$$

Then, for any $\{P_n \in \mathbf{P} : n \geq 1\}$,

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_i - \mu_{X|Z=1}(P_n) \xrightarrow{P_n} 0,$$

where $n_1 = \sum_{1 \leq i \leq n} Z_i$.

Proof. First assume w.l.o.g. that $\mu_{X|Z=1}(P_n) = 0$. Thus, $E_{P_n}[ZX] = 0$. Next, note that (60) implies that

$$\limsup_{\lambda \to \infty} \frac{1}{P_n\{Z = 1\}} E_{P_n} [|ZX| \, I \{|ZX| > \lambda\} \, |Z = 1] = 0.$$

Since $P_n\{Z = 1\} > \epsilon$, it follows that

$$\limsup_{\lambda \to \infty} E_{P_n} [|ZX| \, I \{|ZX| > \lambda\} \, |Z = 1] = 0.$$

By Lemma 11.4.2 of Romano and Shaikh (2012), we therefore have that

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i Z_i \xrightarrow{P_n} 0.$$

Since $|Z - \mu_Z(P_n)| \leq 1$, we also have that

$$\limsup_{\lambda \to \infty} E_{P_n}\left[|Z - \mu_Z(P_n)| I\left\{|Z - \mu_Z(P_n)| > \lambda\right\} |Z = 1\right] = 0.$$

Thus,

$$\frac{1}{n} \sum_{1 \leq i \leq n} Z_i = P_n\{Z = 1\} + o_{P_n}(1).$$

To complete the argument, note that

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_i = \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i Z_i\right) / \left(\frac{1}{n} \sum_{1 \leq i \leq n} Z_i\right).$$

The desired result now follows since $P_n\{Z = 1\} > \epsilon$. ■

Lemma F.2. *Let (X_i, Y_i, Z_i), $i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Suppose (56) holds for some $\epsilon > 0$ and for all $1 \leq j \leq k$ that (57)* and *(58) hold. Then, for any $\{P_n \in \mathbf{P} : n \geq 1\}$,*

$$\|\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)\| \stackrel{P_n}{\to} 0 \tag{61}$$

$$\|\Omega_{Y|Z=0}(\hat{P}_n) - \Omega_{Y|Z=0}(P_n)\| \stackrel{P_n}{\to} 0, \tag{62}$$

where $\| \cdot \|$ denotes the component-wise maximum of the absolute value of all elements.

Proof. We provide only the proof for (61), as the same argument establishes (62). To establish (61), first note that we may assume w.l.o.g. for all $1 \leq j \leq k$ that $\mu_{X_j|Z=1}(P_n) = 0$ and $\sigma_{X_j|Z=1}(P_n) = 1$. The (j, ℓ) element of $\Omega_{X|Z=1}(P_n)$ is thus given by

$$E_{P_n}[X_j X_\ell | Z = 1]$$

and the (j, ℓ) element of $\Omega_{X|Z=1}(\hat{P}_n)$ is given by

$$\frac{\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} X_{i,\ell} - \left(\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j}\right)\left(\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,\ell}\right)}{\sigma_{X_j|Z=1}(\hat{P}_n)\sigma_{X_\ell|Z=1}(\hat{P}_n)},$$

where $n_1 = \sum_{1 \leq i \leq n} Z_i$. From Lemma B.3 in Bhattacharya et al. (2012), we see that

$$\sigma_{X_j|Z=1}(\hat{P}_n) \stackrel{P_n}{\to} 1$$

$$\sigma_{X_\ell|Z=1}(\hat{P}_n) \stackrel{P_n}{\to} 1.$$

From Lemma F.1, we see that

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} \stackrel{P_n}{\to} 0$$

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,\ell} \stackrel{P_n}{\to} 0.$$

Using the inequality

$$|a||b| I\{|a||b| > \lambda\} \leq a^2 I\{|a| > \sqrt{\lambda}\} + b^2 I\{|b| > \sqrt{\lambda}\},$$

we see that

$$\limsup_{\lambda \to \infty} E_{P_n}\left[\left|X_j X_\ell\right| I\left\{\left|X_j X_\ell\right| > \lambda\right\} |Z = 1\right] = 0.$$

Since $|E_{P_n}[X_j X_\ell | Z = 1]| \leq 1$ by the Cauchy-Schwartz inequality, we have further that

$$\limsup_{\lambda \to \infty} E_{P_n}\left[\left|X_j X_\ell - E_{P_n}[X_j X_\ell | Z = 1]\right| I\left\{\left|X_j X_\ell - E_{P_n}[X_j X_\ell | Z = 1]\right| > \lambda\right\} |Z = 1\right] = 0.$$

Thus, Lemma F.1 implies that

$$\frac{1}{n_1} \sum_{1 \leq i \leq n : Z_i = 1} X_{i,j} X_{i,\ell} = E_{P_n}[X_j X_\ell | Z = 1] + o_{P_n}(1).$$

The desired result now follows immediately. ■

Lemma F.3. *Let (X_i, Y_i, Z_i), $i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Suppose (56) holds for some $\epsilon > 0$ and for all $1 \le j \le k$ that (57)* and *(58) hold. Define*

$$D(P) = diag\left(\frac{\frac{\sigma^2_{X_1|Z=1}(P)}{P\{Z=1\}}}{\frac{\sigma^2_{X_1|Z=1}(P)}{P\{Z=1\}} + \frac{\sigma^2_{Y_1|Z=0}(P)}{P\{Z=0\}}}, \ldots, \frac{\frac{\sigma^2_{X_k|Z=1}(P)}{P\{Z=1\}}}{\frac{\sigma^2_{X_k|Z=1}(P)}{P\{Z=1\}} + \frac{\sigma^2_{Y_k|Z=0}(P)}{P\{Z=0\}}} \right).$$

Then,

$$\|D(\hat{P}_n)\Omega_{X|Z=1}(\hat{P}_n) - D(P_n)\Omega_{X|Z=1}(P_n)\| \overset{P_n}{\to} 0 \tag{63}$$

$$\|(I - D(\hat{P}_n))\Omega_{Y|Z=0}(\hat{P}_n) - (I - D(P_n))\Omega_{Y|Z=0}(P_n)\| \overset{P_n}{\to} 0, \tag{64}$$

where I is the k-dimensional identity matrix and $\|\cdot\|$ denotes the component-wise maximum of the absolute value of all elements. Hence,

$$\|V(\hat{P}_n) - V(P_n)\| \overset{P_n}{\to} 0, \tag{65}$$

where

$$V(P) = D(P)\Omega_{X|Z=1}(P) + (I - D(P))\Omega_{Y|Z=0}(P). \tag{66}$$

Proof. We provide only the proof for (63); the same argument establishes (64) and (65) then follows immediately from the triangle inequality. To establish (63), first note that $D(P_n)$ is invertible and that from Lemma B.4 of Bhattacharya et al. (2012)

$$\|D(P_n)^{-1}D(\hat{P}_n) - I\| \overset{P_n}{\to} 0.$$

Next, note for a universal constant C that

$$\begin{aligned} &\|D(\hat{P}_n)\Omega_{X|Z=1}(\hat{P}_n) - D(P_n)\Omega_{X|Z=1}(P_n)\| \\ &\le C\|D(P_n)\|\|D(P_n)^{-1}D(\hat{P}_n)\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)\| \\ &\le C^2\|D(P_n)\|\|\Omega_{X|Z=1}(\hat{P}_n)\| \left(\|D(P_n)^{-1}D(\hat{P}_n) - I\| + \|\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)\|\right) \end{aligned}$$

Since the elements of $D(P_n)$ and $\Omega_{X|Z=1}(\hat{P}_n)$ are all bounded, the norm of these matrices are also bounded. It therefore suffices to show that

$$\|\Omega_{X|Z=1}(\hat{P}_n) - \Omega_{X|Z=1}(P_n)\| \overset{P_n}{\to} 0,$$

which follows from Lemma F.2. ■

Lemma F.4. *Let (X_i, Y_i, Z_i), $i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Suppose (56) holds for some $\epsilon > 0$ and for all $1 \le j \le k$ that (57)* and *(58) hold. Then, for any $\{P_n \in \mathbf{P} : n \ge 1\}$,*

$$\max_{1\le j\le k}\left\{\int_0^\infty |r_j(\lambda, \hat{P}_n) - r_j(\lambda, P)|d\lambda\right\} \overset{P_n}{\to} 0 \tag{67}$$

$$\max_{1\le j\le k}\left\{\int_0^\infty |s_j(\lambda, \hat{P}_n) - s_j(\lambda, P)|d\lambda\right\} \overset{P_n}{\to} 0, \tag{68}$$

where

$$r_j(\lambda, P) = E_P\left[\left(\frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)}\right)^2 I\left\{\frac{X_j - \mu_{X_j|Z=1}(P)}{\sigma_{X_j|Z=1}(P)} > \lambda\right\} |Z = 1\right] \tag{69}$$

$$s_j(\lambda, P) = E_P\left[\left(\frac{Y_j - \mu_{Y_j|Z=0}(P)}{\sigma_{Y_j|Z=0}(P)}\right)^2 I\left\{\frac{Y_j - \mu_{Y_j|Z=0}(P)}{\sigma_{Y_j|Z=0}(P)} > \lambda\right\} |Z = 0\right]. \tag{70}$$

Proof. We provide only the proof for (67); the same argument establishes (68). To establish (67), consider any $1 \le j \le k$. First note that we may assume w.l.o.g. that $\mu_{X_j|Z=1}(P_n) = 0$ and $\sigma_{X_j|Z=1}(P_n) = 1$. Next, note for any $1 \le j \le k$ that $r_j(\lambda, \hat{P}_n) = A_n - 2B_n + B_n$, where

$$A_n = \frac{1}{\sigma_{X_j|Z=1}(\hat{P}_n)}\frac{1}{n_1}\sum_{1\le i\le n: Z_i=1} X_{i,j}^2 I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda\sigma_{X_j|Z=1}(\hat{P}_n)\}$$

$$B_n = \frac{\mu_{X_j|Z=1}(\hat{P}_n)}{\sigma_{X_j|Z=1}(\hat{P}_n)} \frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} X_{i,j} I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda \sigma_{X_j|Z=1}(\hat{P}_n)\}$$

$$C_n = \frac{\mu_{X_j|Z=1}(\hat{P}_n)^2}{\sigma_{X_j|Z=1}(\hat{P}_n)} \frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda \sigma_{X_j|Z=1}(\hat{P}_n)\}.$$

From Lemma F.1, we see that $\mu_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 0$. From Lemma B.3 in Bhattacharya et al. (2012), we see that $\sigma_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 1$. From Lemma F.1, we also see that

$$\frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} |X_{i,j}| = E_{P_n}[|X_j|] + o_{P_n}(1).$$

Since $E_{P_n}[|X_j|] \le 1$ by the Cauchy-Schwartz inequality, it follows that $B_n = o_{P_n}(1)$ uniformly in λ. A similar argument establishes that $B_n = o_{P_n}(1)$ uniformly in λ. In summary,

$$r_j(\lambda, P_n) = \frac{1}{\sigma_{X_j|Z=1}(\hat{P}_n)} \frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} X_{i,j}^2 I\{|X_{i,j} - \mu_{X_j|Z=1}(\hat{P}_n)| > \lambda \sigma_{X_j|Z=1}(\hat{P}_n)\} + \Delta_n$$

uniformly in λ, where $\Delta_n = o_{P_n}(1)$.

For $\delta > 0$, define the events

$$E_n(\delta) = \{|\mu_{X_j|Z=1}(\hat{P}_n)| < \delta \cap 1 - \delta < \sigma_{X_j|Z=1}(\hat{P}_n) < 1 + \delta\}$$

$$E_n'(\delta) = \left\{ \sup_{t \in \mathbf{R}} \left| \frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[X_j^2 I\{|X_j| > t\} | Z = 1] \right| < \delta \right\}$$

$$E_n''(\delta) = \{|\Delta_n| < \delta\}.$$

We now argue that $P_n\{E_n(\delta) \cap E_n'(\delta) \cap E_n''(\delta)\} \to 1$. Since $\mu_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 0$, $\sigma_{X_j|Z=1}(\hat{P}_n) \xrightarrow{P_n} 1$, and $\Delta_n = o_{P_n}(1)$, it suffices to argue that

$$P_n\{E_n'(\delta)\} \to 1. \tag{71}$$

To see this, note that

$$\begin{aligned}
&\frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[X_j^2 I\{|X_j| > t\} | Z = 1] \\
&= \frac{\frac{1}{P_n\{Z=1\}} \frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\}}{\frac{\bar{Z}_n}{P_n\{Z=1\}}} - \frac{1}{P_n\{Z = 1\}} E_{P_n}[Z X_j^2 I\{|X_j| > t\}] \\
&= \left(1 - \frac{P_n\{Z = 1\}}{\bar{Z}_n}\right) \frac{1}{P_n\{Z = 1\}} \frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\} \\
&\qquad \frac{1}{P_n\{Z = 1\}} \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[Z X_j^2 I\{|X_j| > t\}] \right) \\
&= \left(1 - \frac{P_n\{Z = 1\}}{\bar{Z}_n}\right) \frac{\bar{Z}_n}{P_n\{Z = 1\}} \frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} X_{i,j}^2 I\{|X_{i,j}| > t\} && (72) \\
&\qquad \frac{1}{P_n\{Z = 1\}} \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i X_{i,j}^2 I\{|X_{i,j}| > t\} - E_{P_n}[Z X_j^2 I\{|X_j| > t\}] \right). && (73)
\end{aligned}$$

From Lemma F.1, we see that

$$\frac{\bar{Z}_n}{P_n\{Z = 1\}} \xrightarrow{P_n} 1$$

and

$$\frac{1}{n_1} \sum_{1 \le i \le n: Z_i = 1} X_{i,j}^2 I\{|X_{i,j}| > t\} = E_{P_n}[X_j^2 I\{|X_j| > t\} | Z = 1] + o_{P_n}(1).$$

The Cauchy-Schwartz inequality implies that $E_{P_n}[X_j^2 I\{|X_j| > t\}|Z = 1] \leq 1$. Hence, (72) is $o_{P_n}(1)$ uniformly in t. Note further that the class of functions

$$\{zx^2 I\{|x| > t\} : t \in \mathbf{R}\} \tag{74}$$

is a VC class of functions. Therefore, by Theorem 2.6.7 and Theorem 2.8.1 of van der Vaart and Wellner (1996), we see that the class of functions (74) is Glivenko–Cantelli uniformly over $\mathbf{P}$. Since $P_n\{Z = 1\} > \epsilon$, it follows that the supremum over $t \in \mathbf{R}$ of (73) tends in probability to zero under P_n. The desired conclusion (71) follows.

To complete the argument, it now suffices to argue as in the proof of Lemma S.12.2 in Romano and Shaikh (2012). ■

Lemma F.5. *Let $(X_i, Y_i, Z_i), i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Suppose (56) holds for some $\epsilon > 0$ and for all $1 \leq j \leq k$ that (57) and (58) hold. Then, for any $\{P_n \in \mathbf{P} : n \geq 1\}$,*

$$\rho(\hat{P}_n, P_n) \xrightarrow{P_n} 0,$$

where

$$\rho(Q, P) = \max\Bigg\{ \|V(Q) - V(P)\|, |Q\{Z = 1\} - P\{Z = 1\}|,$$
$$\max_{1\leq j\leq k}\left\{\int_0^\infty |r_j(\lambda, Q) - r_j(\lambda, P)|d\lambda\right\}, \max_{1\leq j\leq k}\left\{\int_0^\infty |s_j(\lambda, Q) - s_j(\lambda, P)|d\lambda\right\}\Bigg\}. \tag{75}$$

Here, $V(P)$, $r_j(\lambda, P)$, and $s_j(\lambda, P)$ are defined as in (66), (69), and (70), respectively, and $\|\cdot\|$ denotes the component-wise maximum of the absolute value of all elements.

Proof. By arguing as in the proof of Lemma F.2, we have that

$$\hat{P}_n\{Z = 1\} - P_n\{Z = 1\} = \bar{Z}_n - P_n\{Z = 1\} \xrightarrow{P_n} 0.$$

The desired result now follows from Lemmas F.3 and F.4. ■

Lemma F.6. *Let $\mathbf{P}$ be a set of distributions on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$ such that (56) holds for some $\epsilon > 0$ and for all $1 \leq j \leq k$ that (57) and (58) hold. Let $\mathbf{P}'$ be the set of all distributions on $\mathbf{R}^k \times \mathbf{R}^k \times \{0, 1\}$. Define $\rho(Q, P)$ as in (75) and $J_n(x, P)$ as in (59). Then, for any $\{Q_n \in \mathbf{P}' : n \geq 1\}$ and $\{P_n \in \mathbf{P} : n \geq 1\}$ satisfying $\rho(Q_n, P_n) \to 0$,*

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}} |J_n(x, Q_n) - J_n(x, P_n)| \to 0. \tag{76}$$

Proof. Consider sequences $\{Q_n \in \mathbf{P}' : n \geq 1\}$ and $\{P_n \in \mathbf{P} : n \geq 1\}$ satisfying $\rho(Q_n, P_n) \to 0$. By arguing as in the proof of Lemma S.12.1 in Romano and Shaikh (2012), we see that

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} r_j(\lambda, P_n) = 0$$
$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} r_j(\lambda, Q_n) = 0$$
$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} s_j(\lambda, P_n) = 0$$
$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} s_j(\lambda, Q_n) = 0.$$

We now establish (76). Suppose by way of contradiction that (76) fails. It follows that there exists a subsequence such that n_m such that $V(P_{n_m}) \to V^*$, $V(Q_{n_m}) \to V^*$, and either

$$\sup_{x\in\mathbf{R}} |J_{n_m}(x, P_{n_m}) - \Phi_{V^*}(x)| \not\to 0 \tag{77}$$

or

$$\sup_{x\in\mathbf{R}} |J_{n_m}(x, Q_{n_m}) - \Phi_{V^*}(x)| \not\to 0. \tag{78}$$

Let $W_n(P_n)$ be the vector whose jth element for $1 \leq j \leq k$ is given by

$$\frac{\frac{1}{n_1}\sum_{1\leq i\leq n: Z_i=1} X_{i,j} - \mu_{X_j|Z=1}(P) - \frac{1}{n_0}\sum_{1\leq i\leq n: Z_i=0} Y_{i,j} - \mu_{Y_j|Z=0}(P)}{\sqrt{\frac{\sigma^2_{X_j|Z=1}(\hat{P}_n)}{n_1} + \frac{\sigma^2_{Y_j|Z=0}(\hat{P}_n)}{n_0}}}.$$

From Lemmas B.4 and B.5 in Bhattacharya et al. (2012) and Slutsky's Lemma, we see that

$$W_{n_m}(P_{n_m}) \xrightarrow{d} \Phi_{V^*}(x)$$

under P_{n_m}. It therefore follows from Polya's Theorem that (77) cannot hold. Similarly, we see that (78) cannot hold. The desired conclusion thus follows. ■

Proof of Theorem F.1. The desired result follows immediately from Lemmas F.5 and F.6 and Theorem 2.4 in Romano and Shaikh (2012). ■

References

Abrevaya, J., Hausman, J., Khan, S., 2010. Testing for causal effects in a generalized model with endogenous regressors. Econometrica 6, 2043–2061.
Andrews, D.W., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78, 119–157.
Angrist, J.D., Imbens, G.D., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Amer. Statist. Assoc. 91, 444–454.
Balke, A., Pearl, J., 1997. Bounds on treatment effects from studies with imperfect compliance. J. Amer. Statist. Assoc. 92, 1171–1176.
Bhattacharya, J., Shaikh, A., Vytlacil, E., 2008. Treatment effect bounds under monotonicity conditions: An application to Swan-Ganz catherization. Am. Econ. Rev. 98, 351–356.
Bhattacharya, J., Shaikh, A.M., Vytlacil, E., 2012. Treatment effect bounds: An application to Swan-Ganz catheterization. J. Econometrics 168, 223–243.
Canay, I., Shaikh, A., 2016. Practical and theoretical advances in inference for partially identified models. Tech. rep. Cemmap Working Paper, Centre for Microdata Methods and Practice.
Chiburis, R., 2010. Semiparametric bounds on treatment effects. J. Econometrics 159, 267–275.
Dmitrienko, A., Tamhane, A.C., Wiens, B.L., 2008. General multistage gatekeeping procedures. Biom. J. 50, 667–677.
Heckman, J., Vytlacil, E., 2001a. Instrumental variables, selection models, and tight bounds on the average treatment effect. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluations of Active Labor Market Policies in Europe. Physica-Verlag, Heidelberg; New York.
Heckman, J.J., Vytlacil, E.J., 2001b. Local instrumental variables. In: C. Hsiao, K. Morimune, Powell, J. (Eds.), Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge.
Heckman, J.J., Vytlacil, E.J., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73, 669–738.
Huber, M., Mellace, G., 2011. Testing instrument valitidy for late identification based on inequality moment constraints. Discussion Paper 2011-43, University of Saint Gallen.
Imbens, G.D., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.
Imbens, G.W., Rubin, D.B., 1997. Estimating outcome distributions for compliers in instrumental variable models. Rev. Econom. Stud. 64, 555–574.
Kitagawa, T., 2015. A test for instrument validity. Econometrica 83, 2043–2063.
Manski, C., 1990. Nonparametric bounds on treatment effects. Am. Econ. Rev. 80, 319–323.
Manski, C.F., 1997. Monotone treatment response. Econometrica 131, 1–1334.
Manski, C., Pepper, J., 2000. Monotone instrumental variables: With an application to the returns to schooling. Econometrica 68, 997–1010.
Richardson, T., Robins, J., 2010. Analysis of the binary instrumental variable model. Working Paper 99, University of Washington.
Romano, J.P., Shaikh, A.M., 2012. On the uniform asymptotic validity of subsampling and the bootstrap. Ann. Statist. 40, 2798–2822.
Romano, J., Shaikh, A., Wolf, M., 2012. A simple two-step method for testing moment inequalities with an application to inference in partially identified models. Tech. rep., University of Chicago.
Shaikh, A.M., Vytlacil, E.J., 2005. Threshold crossing models and bounds on treatment effects: A nonparametric analysis. Technical Working Paper 0307, National Bureau of Economic Research.
Shaikh, A.M., Vytlacil, E.J., 2011. Partial identification in triangular simultaneous equations models with binary dependent variables. Econometrica 79, 949–955.
van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications To Statistics. Springer, New York.
Vytlacil, E., 2002. Independence, monotonicity and latent index models: An equivalence result. Econometrica 70, 331–341.