Proceedings of Machine Learning Research 323:1–38, 2026 5th Conference on Causal Learning and Reasoning

Learning control variables and instruments for causal analysis in observational data

Nicolas Apfel NICOLAS.APFEL@UIBK.AC.AT
Department of Economics Faculty of Economics and Statistics, University of Innsbruck, Universitaetsstrasse 15, 6020

Julia Hatamyar JULIA.HATAMYAR@YORK.AC.UK
Centre for Health Economics, Heslington, York, YO10 5DD, United Kingdom

Martin Huber MARTIN.HUBER@UNIFR.CH
University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland

Jannis Kueck KUECK@DICE.HHU.DE
Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

This study introduces a data-driven, machine learning-based method to detect suitable control variables and instruments for assessing the causal effect of a treatment on an outcome in observational data. Our approach tests the joint existence of instruments, which are associated with the treatment but not directly with the outcome (at least conditional on observables), and suitable control variables, conditional on which the treatment is exogenous, and learns the partition of instruments and control variables from the observed data. The detection of sets of instruments and control variables relies on the condition that proper instruments are conditionally independent of the outcome given the treatment and suitable control variables. We establish the consistency of our method for detecting control variables and instruments under certain regularity conditions, investigate the finite sample performance through a simulation study, and provide an empirical application to health data from the Oregon Health Insurance Experiment.

Keywords: treatment effects, causality, conditional independence, instrument, covariates

1. Introduction

Methods for causal analysis, aimed at quantifying the impact of a treatment on an outcome variable, rely on identifying assumptions considered untestable. For example, the well-known selection-on-observables, unconfoundedness, conditional independence, or ignorability assumption requires the treatment to be exogenous when conditioning on observed control variables, hereafter referred to as covariates. The selection of covariates is typically justified based on theoretical and/or empirical reasoning, intuition, domain expertise, or prior empirical findings. Nonetheless, in most empirical scenarios, this selection is debatable, given that the optimal set of covariates meeting the selection-on-observables assumption remains fundamentally uncertain.

In this paper, we suggest a machine learning (ML)-based procedure to simultaneously test the presence of (i) covariates satisfying the selection-on-observables (SOO) assumption and (ii) relevant and valid instrumental variables (IVs) in observational data, as well as learning which variables in the data belong to either the set of covariates or IVs. When we refer to relevant and valid IVs, we mean variables that are associated with the treatment (relevance) but have no direct association with

© 2026 N. Apfel, J. Hatamyar, M. Huber & J. Kueck.

the outcome other than through the treatment (validity) conditional on covariates. We demonstrate that appropriate sets of covariates satisfying the identification requirements for treatment effects based on the SOO assumption, as well as relevant and valid IVs, can be detected in a data-driven way instead of being assumed by the researcher. For testing and learning covariates and instruments, we exploit a conditional independence condition that must hold when both relevant and valid instruments as well as covariates that satisfy the SOO assumption exist: The IVs must be conditionally independent of the outcome, given the treatment and the covariates, see for instance the discussions in de Luna and Johansson (2014), Black et al. (2015), and Huber and Kueck (HK, 2022).

The contributions of this paper are two-fold. (1) We propose the first data-driven method that is able to assess identification without having to impose SOO or IV validity a priori and which can learn the partition of variables into controls and IVs. We consider a setting with nonlinearities and without assuming homogeneous treatment effects. We also show the theoretical, large-sample properties of our procedure. (2) We propose a new orthogonalized score which can be interesting in more general settings and which is also normally distributed under the alternative hypothesis. A detailed literature review can be found in appendix A.

Our approach consists of the following steps. First, within the combined set of potential covariates and IVs, we sequentially test which variable is strongly associated with the treatment conditional on all remaining variables in that set. Second, we consider each of these strong predictors of the treatment as candidate IVs and sequentially test whether each of them is conditionally independent of the outcome when controlling for the treatment and all remaining variables in the combined set of potential covariates and IVs. If the conditional independence assumption is satisfied by (at least) one candidate IV, then the instrument validity and SOO assumptions hold. This implies that the treatment is as good as random conditional on the remaining variables within the combined set of potential covariates and IVs. Treatment effects can then be estimated using methods that control for observed covariates, such as matching, regression, inverse probability weighting, or doubly robust techniques (Huber, 2023). In other words, the output of the algorithm is a decision on whether the SOO assumption is fulfilled and whether the researcher should continue with their analysis. Indeed, what we require is that the algorithm reliably indicates the presence of at least one valid IV while the SOO assumption holds; we do not require, nor can we guarantee, the correct selection of all valid IVs.

Our test focuses on the conditional mean (rather than full) independence of the IV, which implies the identification of average treatment effects (ATE). When assuming a limited set of observed variables (relative to the sample size), we employ regression for both selecting the candidate IVs in the first step and testing the conditional mean independence of the IV in the second step. More concisely, testing is based on the mean squared difference in outcome prediction when regressing the outcome (1) on the treatment, the control variables, and the candidate IV and (2) on the treatment and the control variables (but not the candidate IV). This approach builds on the mean squared difference test based on a quadratic score function in HK, but applies it sequentially across all candidate IVs. Our test flips the original setup: the H_0 states that identification fails; the H_1 states it holds. We demonstrate that our method is consistent for correctly determining identification, which we illustrate in a simulation study with ten covariates. As a word of caution for empirical applications, we find that the test might require a large sample.

We apply our method to health data from the Oregon Health Insurance (OHI) Experiment, previously analysed by Finkelstein et al. (2012), in which low-income adults were randomly assigned the opportunity to apply for Medicaid, a public health insurance program in the US. The random

assignment provides a plausible IV for actual Medicaid enrollment - the treatment of interest - provided that assignment itself has no direct effect on health outcomes such as doctor visits. Our approach indeed selects random assignment as a valid IV, and indicates that Medicaid enrollment is exogenous, conditional on a rich set of more than 200 pre-assignment covariates.

The remainder of this study is organized as follows. Section 2 discusses the identifying assumptions. Based on these, Section 3 proposes ML-based procedures for jointly testing the IV and SOO assumptions. Section 4 suggests an algorithm that detects strong and valid IVs as well as covariate sets satisfying the SOO assumption. Section 5 provides a simulation study analyzing the finite sample performance of our method. Section 6 presents an application to the Oregon Health Insurance Experiment. Section 7 concludes. All proofs, an extension to the multivalued IV case, a literature review, pseudo-code and the full simulation results can be found in the appendix.

2. Identifying assumptions and testable conditional independence

First, we briefly review the implication that we will use for testing. de Luna and Johansson (2014), Black et al. (2015), and HK imply that under IV validity and a SOO assumption concerning the treatment, the IV is conditionally independent of the outcome given the treatment and observed covariates. To formalize the assumptions, let us denote by D a treatment whose causal effect on an outcome variable Y is of interest. Both D and Y might be discretely or continuously distributed. Using the potential outcomes framework (Neyman, 1923; Rubin, 1974), we denote by $Y(d)$ the potential outcome when exogenously setting the treatment D of a subject to value d in the support of the treatment. More generally, we will use capital and lower case letters for referring to random variables and specific values thereof, respectively.[1] Furthermore, we denote by X and Z sets of observed covariates and IVs, whose properties are yet to be defined. Based on this notation, we consider the same identifying assumptions as HK.

The first assumption imposes some causal structure. It rules out the existence of reverse causality[2] and enforces the principle of causal faithfulness. We formalise this causal structure using the previously mentioned potential outcome notation, by applying the latter also to other variables. To this end, let $A(b)$ and $A(b,c)$ correspond to the potential value of variable A when setting variable B to b, or variables B and C to b and c, respectively.

Assumption 1 (Causal structure)

$$D(y) = D, \quad X(d,y) = X, \textit{ and } Z(d,y) = Z \quad \forall d \in \mathcal{D} \textit{ and } y \in \mathcal{Y},$$

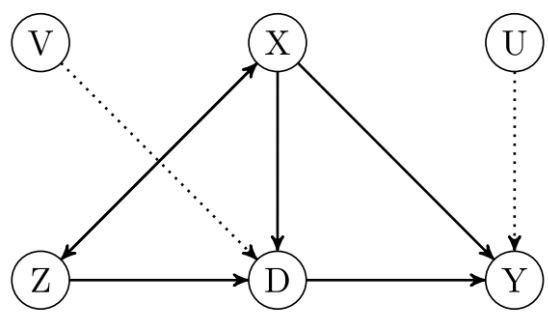
only variables which are d-separated in some causal model are statistically independent.

$\mathcal{D}$ and $\mathcal{Y}$ denote the support of D and Y, respectively. The first line of A 1 rules out a causal effect of outcome Y on D, X, or Z and of treatment D on X or Z. However, it allows for the possibility

1. By representing the potential outcome $Y(d)$ as a function solely dependent on a subject's own treatment status $D = d$, we implicitly adhere to the assumption that the potential outcomes of one subject are not influenced by the treatment status of others. This is known as the 'stable unit treatment value assumption' (SUTVA, see the discussion in Rubin, 1980; Cox, 1958), and is invoked throughout.
2. This means that the outcome cannot causally influence any other variables, and the treatment cannot causally affect any variables other than the outcome. This is in line with the conventional practice of measuring covariates and IVs before treatment assignment, eliminating the potential for reverse causality between D and Y and the pre-treatment variables X and Z.

of both X and Z affecting D, Y, or even each other. This assumption aligns with the directed acyclic graph (DAG, see e.g. Pearl, 2000) presented in Figure 1, where causal relationships between variables are indicated by arrows: Z and X affect D, and D and X affect Y. Additionally, X may influence Z or vice versa, denoted by the bidirectional arrow. The DAG also features unobserved terms U and V that affect Y and D, respectively, with dashed arrows denoting the unobservable nature of these effects. The second line of A 1 enforces causal faithfulness, ensuring that only variables which are d-separated in the sense of Pearl (1988), i.e. not associated with each other via some causal paths (possibly conditional on other variables) are statistically independent (or conditionally independent).[3] While d-separation is generally a sufficient condition for the (conditional) independence of two variables, it is a necessary condition under causal faithfulness.

Figure 1: Causal graph satisfying Assumption 1



Note: Z is an instrument, D is the treatment variable, Y is the outcome, X is an observed control, V and U are unobservables, with dotted lines denoting that the relation between variables is unobservable.

The second assumption is a common support assumption concerning the treatment and the IV:

Assumption 2 (Common support)

$$\mathbb{P}(D = d, Z = z|X) > 0 \quad \forall d \in \mathcal{D} \textit{ and } z \in \mathcal{Z},$$

where $\mathcal{Z}$ denotes the support of Z. Under continuous treatment and/or IV variables, joint probabilities are to be replaced by joint density functions conditional on X. A 2 implies that both $\mathbb{P}(D = d|X)$, the so-called treatment propensity score, and $\mathbb{P}(Z = z|D, X)$, the IV propensity score, are larger than zero. The third assumption imposes a statistical association between the IV and the treatment conditional on the covariates and works as a relevance or first stage assumption.

3. d-separation relies on blocking causal paths between variables. Formally, a path between two (sets of) variables A and B is blocked when conditioning on a (set of) control variable(s) C if

   1. the path between A and B is a causal chain, implying that $A \rightarrow M \rightarrow B$ or $A \leftarrow M \leftarrow B$, or a confounding association, implying that $A \leftarrow M \rightarrow B$, and variable (set) M is among control variables C (i.e. controlled for),
   2. the path between A and B contains a collider, implying that $A \rightarrow S \leftarrow B$, and variable (set) S or any variable (set) causally affected by S is not among control variables C (i.e. not controlled for).

   Based on this definition of blocking, the d-separation criterion states that A and B are d-separated when conditioning on control variable(s) C if and only if C blocks all paths between D and Y.

Assumption 3 (Conditional dependence between the treatment and instrument)

$$D \not\perp\!\!\!\perp Z|X,$$

where $\not\perp\!\!\!\perp$ denotes statistical dependence. This assumption is satisfied in Figure 1, where Z affects D. The fourth assumption invokes SOO, i.e. quasi-random treatment assignment conditional on X as e.g. considered in Imbens (2004):

Assumption 4 (Conditional independence of the treatment)

$$Y(d) \perp\!\!\!\perp D|X \quad \forall d \in \mathcal{D},$$

where $\perp\!\!\!\perp$ denotes statistical independence. A 4 implies that conditional on covariates X, there exist no unobserved confounders jointly affecting outcome Y and treatment D.

Assumption 5 (Conditional independence of the instrument)

$$Y(d) \perp\!\!\!\perp Z|X \quad \forall d \in \mathcal{D}.$$

A 5 rules out unobserved confounders jointly affecting Y and Z when controlling for X. Moreover, by assuming that the potential outcome is solely a function of d (and not z), A 5 also implies that the IV does not have a direct impact on the outcome, except through its impact on the treatment, conditional on X. By this exclusion restriction, it holds that conditional on X, $Y(d,z) = Y(d,z') = Y(d)$ for any IV values z and z'. Otherwise, A5 would be violated, because it would follow that $Y(d) = Y(d,Z)$ and $Y(d,Z) \not\perp\!\!\!\perp Z|X$. A1 is maintained and defines the causal ordering of D, Y and candidate variables. Propensity score trimming ensures that A2 holds. A3 is evaluated in the first-stage screening. A4 and A5 are not imposed, instead, the procedure tests a joint implication of these assumptions via conditional mean-independence. When this restriction is supported for at least one candidate IV, this provides empirical support for identification of the ATE via covariate adjustment, without requiring prior knowledge of which variables are valid IV or controls.

These assumptions can be used to test for the identification of causal effects. HK's Theorem 1 demonstrates that conditional on A 1 and 3, $Y \perp\!\!\!\perp Z|D = d, X$, the testable conditional independence, is necessary and sufficient for the joint satisfaction of A 4 and 5 when considering potential outcomes $Y(d)$ which match the factual treatment assignment $D = d$. Formally,

$$Y(d) \perp\!\!\!\perp D|X, \quad Y(d) \perp\!\!\!\perp Z|X \iff Y \perp\!\!\!\perp Z|D = d, X \quad \forall d \in \mathcal{D}. \tag{1}$$

Instead of verifying $Y \perp\!\!\!\perp Z|D, X$, HK test conditional mean independence of the IV:

$$\mathbb{E}[Y|D,X] = \mathbb{E}[Y|D,X,Z]. \tag{2}$$

Condition (2) is sufficient when considering the identification of average causal effects such as the conditional average treatment effect (CATE) given X, $\mathbb{E}[Y(1) - Y(0)|X]$, or the average treatment effect (ATE), $\mathbb{E}[Y(1) - Y(0)]$.Theorem 2 in HK shows that (2) holds when replacing Assumptions 4 and 5 by the weaker conditional mean independence assumptions $\mathbb{E}[Y(d)|D,X] = \mathbb{E}[Y(d)|X]$ and $\mathbb{E}[Y(d)|Z,X] = \mathbb{E}[Y(d)|X]$ $\forall d \in \mathcal{D}$, as well as A 3 by the first stage condition $\mathbb{E}[D|Z,X] \neq \mathbb{E}[D|X]$, implying that the conditional mean of D varies with Z. Formally, conditional on A 1 and $\mathbb{E}[D|X,Z] \neq \mathbb{E}[D|X]$, it holds that

$$\begin{aligned} & \mathbb{E}[Y(d)|D,X] = \mathbb{E}[Y(d)|X], \quad \mathbb{E}[Y(d)|X,Z] = \mathbb{E}[Y(d)|X] \\ \iff \quad & \mathbb{E}[Y|D=d,X,Z] = \mathbb{E}[Y|D=d,X] \quad \forall d \in \mathcal{D}. \end{aligned} \tag{3}$$

The testable implication $\mathbb{E}[Y|D = d, X, Z] = \mathbb{E}[Y|D = d, X]$ is necessary and sufficient for the joint satisfaction of conditional mean independence of the treatment and the IV when considering potential outcomes $Y(d)$ matching the factual treatment assignment $D = d$. This in turn is sufficient for identifying the conditional mean potential outcome $E[Y(d)|X]$ based on observations with $D = d$ when controlling for X and thus, for identifying the CATE and ATE, see for instance the discussion of weak unconfoundedness in Imbens (2000). This matches the single world intervention graphs (SWIG) framework of Richardson and Robins (2013) under the Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) semantics of Robins (1986).

The testing approach of HK requires the prior specification of the IV, Z, and covariates, X. In contrast, our testing approach, as introduced below, does not require the predefinition of Z and X when testing (2). Instead, it learns them from the data by iteratively considering variables as IV Z. This feature appears attractive in many practical contexts where obvious IVs are not available.

3. Testing based on double machine learning

We henceforth suggest a testing approach based on DML based on the following null hypothesis H_0, which is equivalent to the conditional mean independence of the IV provided in condition (2):

$$H_0 : \mathbb{E}[Y|D = d, X = x, Z = z] - E[Y|D = d, X = x] = 0 \quad \forall d \in \mathcal{D}, x \in \mathcal{X}, z \in \mathcal{Z}. \tag{4}$$

Under the null, H_0, the mean conditional outcome is constant across values of Z given any value of D and X, which may be tested for any values of D, X, and Z in their respective support. However, if one or several variables are of rich support, this implies many testable implications. For this reason, one possible testing approach is to follow HK and test violations of (4) based on the mean squared difference between the conditional mean outcome when including versus excluding the IV in the conditioning set. Denoting the conditional means by $\mu(d, x, z) = \mathbb{E}[Y|D = d, X = x, Z = z]$ and $m(d, x) = \mathbb{E}[Y|D = d, X = x]$, one aims at testing the following implication of eq. (4):

$$\mathbb{E}[(\mu(D, X, Z) - m(D, X))^2] = 0, \tag{5}$$

based on a moment condition which uses the following Neyman (1959)-orthogonal score:

$$\phi(W, \theta, \eta) = (\eta_1(W) - \eta_2(W))^2 - \theta + \zeta. \tag{6}$$

$W = (Y, D, X, Z, \zeta)$ are random variables and $\eta = (\eta_1, \eta_2)$ are the so-called nuisance parameters, whose true values correspond to $\eta_{0,1}(W) = \mu(D, X, Z)$ and $\eta_{0,2}(W) = m(D, X)$. We note that the independent mean-zero random variable ζ in (6) is added to avoid a degenerate distribution of the estimator under H_0, a common problem in specification tests, see e.g. Hong and White (1995) and Wooldridge (1992). A disadvantage of testing based on the score function in eq. (6) is the requirement to choose a random term ζ, as the optimal selection of ζ in a given dataset is generally unknown. Further, while the estimator based on eq. (6) is asymptotically normal under the null hypothesis, as demonstrated in HK, this is generally not the case under the alternative hypothesis. For this reason, we subsequently propose a new testing approach that is based on a refined score function that does not require user-selected random terms and entails a test statistic that is normally distributed under both the null and alternative hypotheses.[4]

4. An alternative testing approach for conditional independence satisfying Neyman orthogonality that can be applied in our context is suggested by Lundborg et al. (2024), (see also Kook and Lundborg, 2024).

For the moment, let us assume that the instrument Z is binary. An extension to multivalued IVs is provided in Appendix C. Denote by $p(D, X) = \mathbb{P}(Z = 1|D, X)$ the IV propensity score. The score function considered in this case is given below:

$$\begin{aligned}
&\tilde{\psi}(W, \theta, \eta) \\
&= (\mu(D, X, 1) - \mu(D, X, 0))^2 \\
&+ 2(\mu(D, X, 1) - \mu(D, X, 0)) \left(\frac{(Y - \mu(D, X, 1)) \cdot Z}{p(D, X)} - \frac{(Y - \mu(D, X, 0)) \cdot (1 - Z)}{1 - p(D, X)} \right) \\
&+ \mu(D, X, 1) - \mu(D, X, 0) + \left(\frac{(Y - \mu(D, X, 1)) \cdot Z}{p(D, X)} - \frac{(Y - \mu(D, X, 0)) \cdot (1 - Z)}{1 - p(D, X)} \right) \\
&- \theta,
\end{aligned} \tag{7}$$

with $\theta_0 = \mathbb{E}[(\mu(D, X, 1) - \mu(D, X, 0))^2] + \mathbb{E}[\mu(D, X, 1) - \mu(D, X, 0)]$. Testing based on (7) corresponds to an aggregate L_2-type measure that can be used to test violations across values of D, X, and Z, which is common in specification tests based on nonparametric regression.[5] In addition to the squared difference in conditional mean outcomes, the score function notably contains a term in which the difference in conditional mean outcomes is multiplied with a difference in expressions obtained by inverse probability weighting (IPW) with the IV propensity score. In fact, our new score above combines the orthogonalized squared score in (6) with the popular doubly robust score. Just as the squared difference in conditional mean outcomes, the score function $\tilde{\psi}$ is zero in expectations, $\mathbb{E}[\tilde{\psi}(W, \theta_0, \eta_0)] = 0$, under the null hypothesis that $\theta_0 = 0$, which follows from iterated expectations, and satisfies the Neyman orthogonality property (see Appendix B).

When testing, we assume an i.i.d. sample of size n, in which i is the index of an observation and $W_i = (Y_i, D_i, X_i, Z_i)$ are the variable values of observation i in that sample, with $i \in \{1, 2, .., n\}$. We apply cross-fitting as for instance discussed in Chernozhukov et al. (2018) to avoid over-fitting due to a correlation of the estimation of the nuisance parameters and θ_0. Therefore, we split the data into K subsamples of size $N = n/K$. The cross-fitted estimator is given by

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{N,k}[\psi_k(W_i, 0, \hat{\eta})], \tag{8}$$

with $\hat{\eta} = (\hat{\mu}, \hat{p}_1, \ldots, \hat{p}_L)$. Under the regularity conditions in A 6, $\hat{\theta}$ is asymptotically normal and $\sqrt{n}$-consistent, as stated in Theorem 1. The proof is provided in Appendix E.

Assumption 6 (Asymptotic Normality) *Define $U = Y - \mu(D, X, Z)$. The following assumption needs to hold for all $n \geq 3$, $\mathbb{P} \in \mathcal{P}$ and $q > 2$: (i) $\|Y\|_{\mathbb{P},q} < C$ and $\mathbb{E}[U^2 1(Z \in Z_l)] > c$ (ii) Given a random subset I of $[n]$ of size $N = n/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^c})$ obeys $\|\hat{\eta} - \eta_0\|_{\mathbb{P},2q} \leq C$, $\|\hat{\eta} - \eta_0\|_{\mathbb{P},4} \leq \delta_N$, and $\|\hat{\eta} - \eta_0\|_{\mathbb{P},2} \leq \delta_N^{1/2} N^{-1/4}$ with $\mathbb{P}$-probability not less than $1 - o(1)$.*

Theorem 1 *Conditional on Assumptions 6, the estimator in eq. (8) satisfies*

$$\sqrt{n}\sigma^{-1}\hat{\theta} \xrightarrow{d} N(\theta_0, 1), \tag{9}$$

5. See e.g. Racine (1997), Racine et al. (2006), Hong and White (1995) and Wooldridge (1992).

uniformly over $P \in \mathcal{P}$, where $\sigma^2 = E[\psi(W, \theta_0, \eta_0)^2]$. Moreover, the result continues to hold if σ^2 is replaced by $\hat{\sigma}^2 := \mathbb{E}_n[(\psi(W_i, \hat{\theta}, \hat{\eta}))^2]$. Consequently, a test that rejects the null hypothesis H_0, $\theta_0 = 0$, if $|\sqrt{n}\hat{\sigma}^{-1}\hat{\theta}| > \Phi^{-1}(1 - \alpha/2)$ has asymptotic level α.

Theorem 1 states that the test statistic is normally distributed both under the null ($H_0 : \theta_0 = 0$) and alternative hypothesis ($H_1 : \theta_0 \neq 0$). The following Corollary 2 shows that the proposed test is consistent, i.e., the power converges to one as $n \to \infty$.

Corollary 2 *Let $c_\alpha := \Phi^{-1}(1 - \alpha/2)$ be the critical value of the test proposed above. Under the alternative hypothesis ($\theta_0 \neq 0$), it holds*

$$\lim_{n\to\infty} P(|\sqrt{n}\hat{\theta}/\hat{\sigma}| > c_\alpha) = 1.$$

This holds true since

$$P(|\sqrt{n}\hat{\theta}/\hat{\sigma}| > c_\alpha) = P\left(\left|\widehat{\theta}_j\right| \geq c_\alpha \frac{\widehat{\sigma}_j}{\sqrt{n}}\right) \geq P\left(\left|\widehat{\theta}_j - \theta_j\right| \leq |\theta_j| - c_\alpha \frac{\widehat{\sigma}_j}{\sqrt{n}}\right) = 1,$$

as long as $c_\alpha = o(\sqrt{n})$.

4. Selection Method

The tests outlined in Section 3 were conditional on having already defined the instrument Z and covariates X under which the SOO assumption with respect to the treatment supposedly holds. However, a main contribution of this study is a data-driven approach for learning partitions of observed pre-treatment variables into IVs and covariates. To do so, we suggest applying the testing approach iteratively when sequentially considering one variable from the set of all pre-treatment variables, henceforth denoted by Q, as instrument Z and the remaining variables as covariates X. More specifically, our procedure consists of the following steps (details provided in subsections):

(1) Select candidate variables with a strong first-stage effect on D from the observed variables Q, conditional on remaining variables. $\mathcal{S}$ is the set of strong IVs. The statistical criterion to decide on IV strength will be introduced below. $\hat{S}$ then is the set of candidates selected as strong.

(2) Each candidate in $\hat{S}$ is iteratively defined as the instrument, Z, and all remaining variables in Q are defined as covariates, X, then the test of hypothesis (5) is run for each of the candidates.

(3) If hypothesis (5) is not rejected in (2) for a candidate, then assign that candidate to the set of IVs for which mean independence holds, $\mathcal{V}$. If there are multiple candidates that pass the test, select the test with the maximal p-value as the final IV and the remaining variables in Q as final covariates X. If (2) suggests that the null is violated in all iterations, then implication (2) is rejected.

Step (1) is required to select candidates which satisfy $E[D|X, Z] \neq E[D|X]$, the first stage condition. Let us assume that Q is low-dimensional, meaning that sample size n is larger than the number of variables in Q, denoted by p. In this case, step (1) might be implemented based on a first stage regression of D on Q and selecting all regressors with statistically significant associations after controlling for multiple hypothesis testing issues into the set of candidate instruments $\hat{S}$. In *high*-dimensional settings where $p > n$, regularization can be applied to select strong IV candidates.

4.1. Strong IV Selection

To select strong IVs in the high-dimensional setting, we use first-stage hard-thresholding (FSHT) proposed by Guo et al. (2018). In this approach, IVs are considered irrelevant if their t-statistic does not exceed a predefined threshold. We iteratively consider each variable in Q as instrument Z and any remaining variables as covariates. That is, when considering the jth variable in Q as instrument and denoting it by Q_j, we have that $Z = Q_j$ and $X = Q_{[j]}$, where $Q_{[j]} = Q \setminus Q_j$ and $Q = X \cup Z$ when estimating the first-stage association $E[D|X, Z]$. Leveraging the DML literature, we consider the following partially linear specification to estimate the effect of a variable Z on treatment D:

$$D = \gamma_j^T Z + g(X) + \varepsilon, \tag{10}$$

$$Z = h(X) + v. \tag{11}$$

Here, $g(X)$ and $h(X)$ are general functions of X, and the first-stage effect is indexed by j as it may vary with each variable considered. DML-based estimation of γ_j is asymptotically normal under regularity conditions.[6] As only one IV is considered in turn, the first-stage F-statistic is equal to the square of the t-statistic, $t_j^2 = F_j$, which is asymptotically χ_1^2-distributed with one degree of freedom, $F_j \xrightarrow{d} \chi_1^2$ (e.g. Proposition 3 of Masten and Poirier, 2021). Under the alternative, $\gamma_j \neq 0$, we have $\frac{F_j}{n} \xrightarrow{P} \kappa_j$, where $\kappa_j > 0$ is a constant. The critical values C_n for the F-statistic need to satisfy $C_n \to \infty$ and $C_n = o(n)$ as $n \to \infty$, for the FSHT procedure to successfully identify strong IVs with probability approaching 1, $\lim_{n\to\infty} P(\hat{S} = \mathcal{S}) = 1$. An IV j is considered as strong if $F_j > C_{\tilde{\alpha}}$ with $\tilde{\alpha} = 0.1/\log(n)$.[7] $C_{\tilde{\alpha}}$ denotes the critical value $q_{\chi_1^2}(1 - \tilde{\alpha})$ and $q_{\chi_1^2}(\cdot)$ denotes the quantile function of the χ_1^2-distribution. As one way to implement DML, which is particularly useful in high dimensions ($p > n$), we estimate the nuisance functions $g(X)$ and $h(X)$ using the LASSO.

4.2. Valid IV Selection and Identification Test

In step (2), we test the null hypothesis in eq. (5) iteratively over all candidate IVs that pass the first-stage threshold. Our aim is to find a partition of variables for which the conditional independence of the respective candidate IV holds. To discuss this more formally, we introduce the partition

$$\mathcal{P}_j = \{Z = Q_j, \;\; X = Q_{[j]}\},$$

such that variable j in set Q is chosen to be the IV, while the remaining variables in Q are used as controls. Moreover, let $\mathcal{V}$ denote the set of candidate IVs which are conditionally mean independent of the outcome, satisfying condition (2) and the null in eq. (5):

$$\mathcal{V} = \{j : E[Y|D, X] = E[Y|D, X, Z]\}.$$

Moreover, we denote the set of partitions for which the IV has a first stage effect on the treatment and the conditional independence of the IV holds by $\mathcal{P}^*$:

$$\mathcal{P}^* = \{\mathcal{P}_j : j \in (\mathcal{S} \cap \mathcal{V})\}. \tag{12}$$

6. In particular, estimators of the models of D and Z should attain a convergence rate of $o(n^{-1/4})$. Then $\sqrt{n}\hat{\sigma}_{\gamma,j}^{-1}(\hat{\gamma}_j - \gamma_{0,j}) \xrightarrow{d} N(0, 1)$, where $\hat{\sigma}_{\gamma,j}^{-1}$ is the standard error of $\hat{\gamma}_j$. (Chernozhukov et al., 2018)
7. Windmeijer et al. (2021) exploit a result in Potscher (1983) and Andrews (1999), which states that a sequence of p-values, p_n, satisfying $\mathrm{p}_n \to 0$ and $\log(\mathrm{p}_n) = o(n)$ can be chosen to meet the just-mentioned conditions on C_n. Based on the recommendation in Belloni et al. (2012), they adopt $\mathrm{p}_n = 0.1/\log(n)$.

The corresponding estimated set of partition(s) $\mathcal{P}^*$, denoted by $\hat{\mathcal{P}}_{pass}$, is given by

$$\hat{\mathcal{P}}_{pass} = \left\{ \mathcal{P}_j : j \in \left(\hat{S} \cap \hat{\mathcal{V}} \right) \right\}, \tag{13}$$

where $\hat{\mathcal{V}} = \{j : |\sqrt{n}\hat{\sigma}_j^{-1}\hat{\theta}_j| < c_\alpha\}$ is the set of instruments for which conditional independence is not rejected based on the test defined in Theorem 1 and Corollary 2 with significance level α and critical value c_α. Our main contribution is the development of a new procedure which tests the identification of a causal effect in a data-driven way. Thus, we consider the following hypotheses:

$$H_0 : \text{no identification} \quad vs. \quad H_1 : \text{identification (conditional mean independence)}. \tag{14}$$

We conclude that H_1 is true (the ATE is identified, see the null hypothesis in (4)) if $\hat{\mathcal{P}}_{pass} \neq \emptyset$. The following Theorem helps us understand the type 1 error of our test.

Theorem 3 *Under the assumptions of Th. 1, assuming* $\lim_{n\to\infty} P(\hat{S} = \mathcal{S}) = 1$, *for a given* α, *it holds*

$$\lim_{n\to\infty} P(\hat{\mathcal{P}}_{pass} \subseteq \mathcal{P}^*) = 1.$$

Theorem 3 states that $\lim_{n\to\infty} P(\mathcal{V} \neq \emptyset) = 1$ if $\hat{\mathcal{V}} \neq \emptyset$. Hence, the type 1 error of our proposed identification test in (14) goes to zero as $n \to \infty$. This means that if our test finds identification ($\hat{\mathcal{P}}_{pass} \neq \emptyset$), there is identification with probability 1 for large n. The next theorem provides insights about the type 2 error of our test, i.e., how likely it is that we can find identification if there is identification. Theorem 4 states that the type 2 error is at least bounded by α.

Theorem 4 *Assume that* $\mathcal{P}^* \neq \emptyset$ *(testable identification). Under the assumptions of Theorem 1, assuming* $\lim_{n\to\infty} P(\hat{S} = \mathcal{S}) = 1$, *for a given* α, *it holds*

$$\lim_{n\to\infty} P(\hat{\mathcal{P}}_{pass} \neq \emptyset) \geq 1 - \alpha.$$

If $|\hat{\mathcal{P}}_{pass}| > 1$ such that there is more than one sufficiently strong candidate IV while conditional independence is not rejected, we select the final partition as the one which maximizes the $\mathtt{p}$-value when testing conditional independence:

$$\hat{\mathcal{P}}_{pmax} = \underset{\mathcal{P}_j \in \hat{\mathcal{P}}_{pass}}{argmax} \ \mathtt{p}(\mathcal{P}_j), \tag{15}$$

where $\mathtt{p}(\mathcal{P}_j)$ is the $\mathtt{p}$-value obtained in the conditional independence test of Z and Y. Finally, Algorithm 1 in the appendix describes the steps of our method by means of pseudo-code and we have added a discussion of the computational cost in appendix I.

5. Simulation study

In this section, we briefly summarize the results of the simulation study. The detailed settings and the full results can be found in Appendix J. We consider two main settings: one with a single valid IV and one with multiple valid IVs. These are then again split into two settings, one with binary and one with continuous instruments. The outcome equation is linear, while we choose a linear

index model to generate D. We encode the violations of Assumptions 3 and 5 via parameters in this two-stage model. Variables in Q are correlated and error terms are standard normal. We then vary the sample sizes, setting them to $n = 1000, 4000$ and $16,000$. We use the LASSO for all ML steps, to illustrate our algorithm.

Our results can be summarized as follows: first, we observe that with no violations of SOO and validity, as n increases, the probability of $\hat{\mathcal{P}}_{pass}$ being non-empty, finding identification, increases. The probability of correctly selecting a valid IV in the final partition increases, while that of incorrectly selecting a confounder decreases. These results are especially clear when multiple valid IVs are available. When there is no violation, we expect the violation parameter, $\hat{\theta}$, to be close to zero over the repetitions, and this is also what we observe in the results. Secondly, we model violations of SOO (A4), through an unobserved confounder, and IV validity (A5), through a violation of the exclusion restriction. When one of these is violated, across all settings the probability of $\hat{\mathcal{P}}_{pass}$ being empty quickly increases with sample size. The latter results are particularly clear when the violation parameter is reasonably large as is the case with violations of IV validity in our simulation.

Overall, our procedure consistently selects the correct IV(s) when Assumptions 4 and 5 are satisfied, and the algorithm correctly concludes there is no identification in the case of violations of SOO or IV validity. The method seems to perform particularly well when n is large, when violations are clearly separated from zero and when there are multiple candidate IVs.

Table 1: Empirical Application

PANEL A: Primary Care Visits			
Method	$\hat{\theta}$	se	p-value
LASSO	0.000	0.000	0.997
Random Forest	-0.000	0.013	0.999
XGBoost	-0.070	0.207	0.735
PANEL B: Number of Prescriptions			
LASSO	0.001	0.000	0.148
Random Forest	-0.007	0.010	0.472
XGBoost	0.073	0.088	0.407

Notes: this table reports the estimate $\hat{\theta}$ from eq. (8) with five folds ($K = 5$) when using the doubly robust score functions in eqs. (7) and (18). 'se' and 'p-value' report the standard error and reported p-value for estimate $\hat{\theta}$ on the random program assignment variable.

6. Empirical application

We apply our method to the Oregon Health Insurance Experiment, in which low-income adults were randomly selected by lottery to be eligible to apply for Medicaid. Oregon launched the lottery in early 2008 for about 90,000 participants, with notifications through October 2009. Previous work uses random assignment as an IV for insurance coverage and finds increases in healthcare utilization, reductions in out-of-pocket expenditures, and improvements in self-reported health.[8]

We apply our testing methodology to the experimental OHI data, using the sample definition adopted by Finkelstein et al. (2012), which yields 23,762 observations. The treatment D indicates whether an individual ever enrolled in Medicaid by October 2009. While assignment is randomized, enrollment may be selective due to non-compliance. We consider two outcomes: the number of primary care visits and the number of prescription medications. Along with random assignment as

8. See Finkelstein et al. (2012); Baicker et al. (2013); Taubman et al. (2014); Finkelstein et al. (2019)

a natural IV candidate, the vector Q contains 218 pre-assignment characteristics, including demographics, socioeconomic variables, and pre-treatment health, utilization, and expenditure measures, as well as indicators for missing values.

Testing is based on the cross-fitted estimator $\hat{\theta}$ of eq. (8) with five folds ($K = 5$) when using the doubly robust score functions in eqs. (7) and (18) in the case of binary and continuous IVs, respectively. In the case of a continuous candidate IV, we make use of the score function outlined in Appendix C for multivalued IVs. We partition the support of the continuous candidate IV using quartiles. As in the simulations, nuisance functions are estimated using LASSO, with random forest and gradient boosting as alternatives. We drop observations with extremely low or high propensity scores, which cause estimation instability and lead to high variance.[9]

Results are reported in Table 1. The algorithm selects random program assignment as the only valid IV, with $\hat{\theta}$ close to zero. Although multiple candidates enter $\hat{S}$ after the first stage, none satisfy the trimming requirement except random assignment.[10] Using the 30% threshold, random assignment passes the test across all learners for the doctor visits outcome. For LASSO, the estimate is zero with a p-value of 99.7%, supporting both IV validity and the SOO assumption given the observed covariates. On the one hand, this indicates validity of the IV, implying not only the randomness of program assignment inherent to the experimental design, but also that the assignment does not directly affect the earnings outcome other than through the treatment (for example, through motivation or disappointment when being or not being eligible for the program). On the other hand, the testing result suggests that the SOO assumption holds for the treatment when controlling for the pre-assignment covariates available in the data. For the prescriptions outcome, results are less conclusive: random forest and XGBoost yield p-values above 30%, while LASSO does not. Consequently, given the covariates, our result suggests that we may evaluate the average treatment effect (ATE) on the total population for the doctor visits outcome. In contrast, employing an IV-based approach utilizing random assignment as the IV for effect estimation would, under specific additional assumptions like treatment monotonicity in the IV, only permit assessing the local average treatment effect (LATE) on the subpopulation of compliers, whose training participation aligns with the random assignment (Imbens and Angrist, 1994).

7. Conclusion

In this paper, we introduced an ML-based algorithm based on a novel doubly robust score function designed to detect and test, in a data-adaptive manner, the presence of control variables sufficient for identifying treatment effects in observational data, as well as variables satisfying IV validity. Treatment effects may be heterogeneous, but identification relies on a conditional mean restriction and therefore pertains to the ATE rather than the full distribution of treatment effects. The method searches over partitions of variables into candidate controls and an IV and tests a conditional mean-independence implication that must hold when the SOO assumption and IV validity jointly hold true. This represents an important advancement over previous work which relies on a

9. For a discussion of such trimming based on the propensity score, see, e.g., Crump et al. (2009) and Lechner and Strittmatter (2019). Specifically, we discard observations for which the estimated propensity score $p_l(D, X) = P(Z \in Z_l \mid D, X)$, defined for a partition Z_l of the candidate IV (whether continuous or discrete), falls outside the interval $0.01 < p_l(D, X) < 0.99$. We require that no more than 5% of observations be trimmed using this rule in order for the candidate IV to be considered.

10. We report every variable selected in the first stage to be included in $\hat{S}$, sorted by decreasing p-value, in Appendix Table A1, as well as the number of observations dropped following the trimming rule.

priori assumptions about whether a variable is an IV or a control. We demonstrated that the method consistently detects controls and IVs (if they exist) both through theory and simulations. Moreover, an application to the OHI Experiment confirms that our algorithm correctly selects random assignment into the program as IV, across various ML algorithms for the nuisance function learners and in line with what a researcher would expect. This shows that our new algorithm can be applied as explorative method, where the researcher wants to learn the partition from the data, or as a confirmatory method.

Acknowledgments

We have benefited from comments by Niels Richard Hansen and Leonard Henckel.

References

Donald WK Andrews. Consistent Moment Selection Procedures for Generalized Method of Moments Estimation. *Econometrica*, 67(3):543–563, 1999.

Joshua D. Angrist. Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114:C52–C83, 2004.

Joshua D Angrist and Miikka Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110: 1331–1344, 2015.

Joshua D Angrist, Peter D Hull, Parag A Pathak, and Christopher R Walters. Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132:871–919, 2017.

Nicolas Apfel and Xiaoran Liang. Agglomerative hierarchical clustering for selecting valid instrumental variables. *Journal of Applied Econometrics*, 39(7):1201–1219, 2024.

Nicolas Apfel, Helmut Farbmacher, Rebecca Groh, Martin Huber, and Henrika Langen. Detecting grouped local average treatment effects and selecting true instruments. *arXiv preprint arXiv:2207.04481*, 2022.

Katherine Baicker, Sarah L Taubman, Heidi L Allen, Mira Bernstein, Jonathan H Gruber, Joseph P Newhouse, Eric C Schneider, Bill J Wright, Alan M Zaslavsky, and Amy N Finkelstein. The oregon experiment—effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6): 2369–2429, 2012.

Marinho Bertanha and Guido Imbens. External validity in fuzzy regression discontinuity designs. *NBER working paper 20773*, 2015.

Dan A. Black, Joonhwi Joo, Robert J. LaLonde, Jeffrey A. Smith, and Evan J. Taylor. Simple tests for selection bias: Learning more from instrumental variables. *IZA Discussion Paper No 9346*, 2015.

Christian N. Brinch, Magne Mogstad, and Matthew Wiswall. Beyond late with a discrete instrument. *Journal of Political Economy*, 125:985 – 1039, 2017.

Tao Chen, Yuanyuan Ji, Yahong Zhou, and Pingfang Zhu. Testing conditional mean independence under symmetry. *Journal of Business & Economic Statistics*, 36:615–627, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Alexander Mangulad Christgau and Niels Richard Hansen. Efficient adjustment for complex covariates: Gaining efficiency with dope. *arXiv preprint 2402.12980*, 2024.

D. Cox. *Planning of Experiments*. Wiley, New York, 1958.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

Xavier de Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2:187–199, 2014.

Xavier De Luna, Ingeborg Waernbaum, and Thomas S Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875, 2011.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial intelligence and statistics*, pages 256–264. PMLR, 2013.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3): 1057–1106, 2012.

Amy Finkelstein, Nathaniel Hendren, and Erzo FP Luttmer. The value of medicaid: Interpreting results from the oregon health insurance experiment. *Journal of Political Economy*, 127(6):2836–2874, 2019.

Jerome Friedman, Robert Tibshirani, and Trevor Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010. doi: 10.18637/jss.v033.i01.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:1–15, 2019.

Xichen Guo, Zheng Li, Biwei Huang, Yan Zeng, Zhi Geng, and Feng Xie. Testability of instrumental variables in additive nonlinear, non-constant effects models. *arXiv preprint arXiv:2411.12184*, 2024.

Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence Intervals for Causal Effects with Invalid Instruments by Using Two-Stage Hard Thresholding with Voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

Yongmiao Hong and Halbert White. Consistent specification testing via nonparametric series regression. *Econometrica: Journal of the Econometric Society*, pages 1133–1159, 1995.

Martin Huber. A simple test for the ignorability of non-compliance in experiments. *Economics Letters*, 120:389–391, 2013.

Martin Huber. *Causal analysis: Impact evaluation and Causal Machine Learning with applications in R*. MIT Press, 2023.

Martin Huber and Jannis Kueck. Testing the identification of causal effects in data. *arXiv preprint 2203.15890*, 2022.

G. W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87:706–710, 2000.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics*, 86:4–29, 2004.

G. W. Imbens and J. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475, 1994.

Markus Kalisch and Peter Bühlmann. Causal structure learning and inference: A selective review. *Quality Technology & Quantitative Management*, 11:3–21, 2014.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144, 2016.

Lucas Kook and Anton Rask Lundborg. Algorithm-agnostic significance testing in supervised learning with multimodal data. *arXiv preprint 2402.14416*, 2024.

Jannis Kueck, Ye Luo, Martin Spindler, and Zigan Wang. Estimation and inference of treatment effects with l2-boosting in high-dimensional settings. *Journal of Econometrics*, 234(2):714–731, 2023. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2022.02.005.

Manabu Kuroki and Zhihong Cai. Instrumental variable tests for directed acyclic graph models. In *AISTATS*, 2005.

Michael Lechner and Anthony Strittmatter. Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews*, 38(2):193–207, 2019.

Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2211.02039*, 2024.

Matthew A Masten and Alexandre Poirier. Salvaging falsified instrumental variable models. *Econometrica*, 89(3):1449–1469, 2021.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, Reprint, 5:463–480, 1923.

J Neyman. *Optimal asymptotic tests of composite statistical hypotheses*, pages 416–444. Wiley, 1959.

Harsh Parikh, Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. A double machine learning approach to combining experimental and observational data. *arXiv preprint 2307.01449*, 2024.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint 1501.01332*, 2015.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, Cambridge, Massachusetts, 2017.

BM Potscher. Order estimation in arma-models by lagrangian multiplier tests. *The Annals of Statistics*, pages 872–885, 1983.

Francesco Quinzan, Ashkan Soleymani, Patrick Jaillet, Cristian R Rojas, and Stefan Bauer. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*, pages 28468–28491. PMLR, 2023.

Jeff Racine. Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics*, 15(3):369–378, 1997.

Jeffery S Racine, Jeffrey Hart, and Qi Li. Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, 25(4):523–544, 2006.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128*, 2013.

J Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

D. Rubin. Comment on 'randomization analysis of experimental data: The fisher randomization test' by d. basu. *Journal of American Statistical Association*, 75:591–593, 1980.

D B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Jaime Sevilla and Alexandra Mayn. A conditional independence test for causality in econometrics. *arXiv preprint 2107.09765*, 2021.

Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.

Ashkan Soleymani, Anant Raj, Stefan Bauer, Bernhard Schölkopf, and Michel Besserve. Causal feature selection via orthogonal search. *Transactions on Machine Learning Research*, 2022.

Sarah L Taubman, Heidi L Allen, Bill J Wright, Katherine Baicker, and Amy N Finkelstein. Medicaid increases emergency-department use: evidence from oregon's health insurance experiment. *Science*, 343(6168):263–268, 2014.

Tyler J VanderWeele and Ilya Shpitser. A new criterion for confounder selection. *Biometrics*, 67: 1406, 2011.

Wolfgang Wiedermann and Dexin Shi. Testing the validity of instrumental variables in just-identified linear non-gaussian models. *British Journal of Mathematical and Statistical Psychology*, 79(1):111–138, 2026.

Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George D Smith. On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

Frank Windmeijer, Xiaoran Liang, Fernando P Hartwig, and Jack Bowden. The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):752–776, 2021.

Jeffrey M Wooldridge. A test for functional form against nonparametric alternatives. *Econometric Theory*, 8(4):452–475, 1992.

Anpeng Wu, Kun Kuang, Junkun Yuan, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representation for counterfactual inference. *arXiv preprint arXiv:2006.07040*, 2021.

Feng Xie, Yangbo He, Zhi Geng, Zhengming Chen, Ru Hou, and Kun Zhang. Testability of instrumental variables in linear non-gaussian acyclic causal models. *Entropy*, 24(4):512, 2022.

Appendix A. Literature review

Our paper contributes to a growing literature on testing identifying assumptions for causal inference. For instance, de Luna and Johansson (2014) and Black et al. (2015) make use of the same conditional independence of the IV as considered here to test the SOO assumption for the treatment when assuming a valid IV, based on matching or regression estimators.[11] In contrast, HK jointly test both IV validity and SOO assumptions with pre-defined instruments and covariates using ML approaches which permit for high-dimensional covariates.[12] This is conceptually closest to our approach, but one important difference is that HK require specifying the sets of supposed IVs and covariates to be used for testing, whereas in this paper, these sets of covariates and IVs are learned from the data and may therefore be a priori unknown.

Moreover, we provide a new orthogonalized quadratic score, which improves upon the quadratic score used in HK. This new score is of broader interest since it can be applied in many other contexts beyond ours. One example is the study by Parikh et al. (2024), who compare experimental and non-experimental treatment effect estimates and use an indicator to distinguish between experimental and non-experimental evaluation designs, which plays a role comparable to the IV in our paper. Imposing external validity of the experiment allows to assess the SOO assumption in the non-experimental design. Conversely, imposing the SOO assumption in the non-experimental design allows to assess the external validity of the experiment, which is equivalent to testing IV validity in our context. Unlike our paper, the authors do not consider testing both assumptions jointly.[13]

Relatedly, one strand of the statistics literature imposes the SOO assumption in observational studies and exploits the conditional independence condition to identify subsets of covariates that are sufficient for identification, implying that the remaining covariates satisfy IV validity; see, e.g, De Luna et al. (2011) and VanderWeele and Shpitser (2011). Considering subsets of covariate information may also allow for more efficient causal effect estimation. For example, Christgau and Hansen (2024) propose a deep learning-based estimator that learns efficient covariate representations from unstructured data and is asymptotically normal. Again, our approach differs from these studies in that it tests the SOO and IV validity assumptions jointly, rather than testing one and assuming the other. Closer to us, Entner et al. (2013) do not pre-impose the SOO assumption when searching for a sufficient set of covariates to control for based on conditional independence and consider a parametric modelling approach for testing, however, without asymptotic guarantees. Here, we suggest a ML-based procedure that allows for nonparametric models and potentially high-dimensional covariates, while having desirable asymptotic properties.

Our study is also related to several contributions in the literature on causal discovery.[14] Peters et al. (2015) make use of pre-defined IVs to learn which of the observed variables are treatments

11. See also the related test by Chen et al. (2018), which (in contrast to other methods) requires symmetrically distributed error terms. Furthermore, Angrist (2004), Brinch et al. (2017), and Huber (2013) assume IV validity to hold unconditionally without controlling for covariates, in order to test the unconditional independence of the treatment and potential outcomes. Bertanha and Imbens (2015) consider the fuzzy regression discontinuity design, where IV validity holds at a cutoff of a running variable which discontinuously affects treatment assignment.
12. Angrist et al. (2017) also propose a joint test for parametric models with low-dimensional covariates.
13. Similarly, Angrist and Rokkanen (2015) employ the same conditional independence of the IV as considered here to test IV validity within the framework of the sharp regression discontinuity design, where SOO for the treatment holds by design, as it is a deterministic function of a cutoff in a running variable. This allows testing whether the running variable is a valid IV, i.e., whether it is not associated with the outcome conditional on the treatment.
14. See, e.g., Kalisch and Bühlmann (2014), Peters et al. (2017), Glymour et al. (2019) for reviews.

in the sense that they directly affect the outcome, assuming that these treatments satisfy the SOO assumption. The approach makes use of IVs in a way that may entail the rejection of treatments which violate the SOO assumptions, thereby providing power to detect identification failures. Our approach is different in that it focuses on a single, pre-defined treatment of interest, tests SOO and IV validity jointly, and learns the sets of covariates and IVs from the data.

Soleymani et al. (2022) and Quinzan et al. (2023) provide algorithms that select treatments, while also controlling for observed covariates based on the double machine learning (DML) framework of Chernozhukov et al. (2018).[15] In contrast to our approach, these algorithms do not exploit IVs to test identifying assumptions, but assume the SOO assumption holds for all treatments.

Another domain of causal discovery related to our study is Y-learning (e.g. Sevilla and Mayn, 2021). Conditional on covariates, Y-learning implies that if two variables are independent of each other when not controlling for the treatment, statistically associated with each other when controlling for the treatment, and both independent of the outcome when controlling for the treatment, then these two variables are relevant and valid IVs. A further implication is that the SOO assumption holds. Our approach differs from Y-learning in that it imposes more causal structure by assuming that potential IVs and covariates are not affected by the treatment. For this reason, our method only requires a single (and a priori unknown) instrument, while Y-learning hinges on the existence of (at least) two IVs.

Another strand of the IV validity and causal discovery literature exploits testable implications of structural causal models. These implications arise either from rank (tetrad/trek) constraints on the covariance matrix in linear settings, or from distributional assumptions such as non-Gaussianity of the errors. Kuroki and Cai (2005) use tetrad (rank) constraints to derive testable implications of IV models. Their approach requires multiple valid IVs (in particular, overidentifying tetrad restrictions arise when at least three valid IVs are available) and the resulting constraints are necessary but not generally sufficient for validity. Silva and Shimizu (2017) propose IV discovery methods based on tetrad constraints and non-Gaussianity in linear non-Gaussian acyclic models (LiNGAM). Their IV-TETRAD procedures require at least two valid IVs to operate and therefore do not apply in settings where only a single IV is valid. Their algorithms return equivalence classes of causal effects. Xie et al. (2022), in contrast, derive testable constraints in linear non-Gaussian acyclic models that can rule out invalid IVs. In contrast to these papers, Guo et al. (2024) consider an additive non-linear non-constant effects model with non-Gaussian errors when exogenous covariates are present. They propose the auxiliary-based independence test condition, which provides a necessary condition for IV validity and becomes sufficient under additional assumptions. Wiedermann and Shi (2026) shows how to test confoundedness of a single IV in a linear model with non-Gaussian errors, but can not test the exclusion restriction. Our method is different from these approaches in that we do not make assumptions on the outcome model or the distribution of the errors and use a test based on mean independence. Moreover, we provide a general, doubly robust, machine learning-based score to implement the test.

Our approach is not intended as a general causal discovery or structure learning method. The causal ordering and structural assumptions are taken as given based on the researcher's substantive knowledge. Within this fixed structure, the procedure tests whether the data support identification of the ATE via the SOO assumption and the existence of a valid instrument.

15. The idea is to sequentially consider each of the observed variables as treatment variable, while considering all remaining variables as covariates to estimate the direct effect of each candidate treatment on the outcome by DML. The algorithm retains only those variables that exhibit statistically significant effects on the outcome.

The computer science literature closest to us is on representation learning of variable sets. Hassanpour and Greiner (2019) and Wu et al. (2021), for instance, propose deep learning algorithms minimizing global loss functions to simultaneously decompose pre-treatment variables into IVs, confounders, and outcome predictors. Yet, there are several differences between their studies and ours: First, their approaches impose SOO a priori to isolate IVs (whose exclusion from treatment effect estimation can increase efficiency) based on the same conditional independence condition as considered in our paper. However, we exploit the conditional independence condition to test SOO and IV validity jointly. Second, their algorithms, which minimize a global loss function, aim at learning the full set of IVs, which is attractive for maximising efficiency in treatment effect estimation but might be very ambitious to achieve in a finite sample. In contrast, our algorithm pursues the more modest goal of detecting at least one valid IV—a sufficient condition for identification. Third, we demonstrate the consistency of our algorithm in selecting valid IVs under certain regularity conditions, while the asymptotic behaviour of the deep learning methods in Hassanpour and Greiner (2019) and Wu et al. (2021) has not been derived.

We also contribute to a growing literature in statistical learning and econometrics that tries to separate valid from invalid IVs, see for instance Kang et al. (2016), Guo et al. (2018), Windmeijer et al. (2021), Windmeijer et al. (2019), Apfel and Liang (2024), and Apfel et al. (2022). These approaches rely on the assumption that a majority or plurality of (a priori unknown) IVs is valid. In order to detect them Sargan-type tests in combination with IV-based estimators are used. In contrast, our approach does not pre-impose the existence of valid IVs, but tests validity (for a single IV) and SOO assumptions jointly, in order to apply estimation based on the SOO assumption. Moreover, these methods impose a linear model, which we do not require in our method.

Finally, other recent papers also rely on the majority and plurality assumptions from the literature discussed in the preceding paragraph. Kuang et al. (2020) propose Ivy, a method to synthesize information from multiple possibly weak and invalid IVs into a single summary IV, relying on the majority assumption. Hartford et al. (2021) propose modeIV, which uses a plurality assumption in nonlinear models to aggregate IV estimates that cluster around a modal value and therefore relies on homogeneity to interpret deviations as violations of validity. While our procedure also involves sequential testing steps, its objective is fundamentally different from iterative IV screening or aggregation methods such as Ivy and modeIV. Our approach is not designed to construct an IV estimator, but to test whether the identifying assumptions required for ATE estimation via covariate adjustment are supported by the data. Accordingly, we do not rely on majority or plurality assumptions, nor do we assume homogeneity. Instead, we test whether there exists at least one valid instrument that jointly supports instrument validity and the selection-on-observables assumption.

Appendix B. Proof of moment condition and Neyman orthogonality of $\tilde{\psi}$

Equation (7) suggests the following score function for testing when Z is binary:

$$\begin{aligned}
&\tilde{\psi}(W,\theta,\eta) && (16)\\
&= (\mu(D,X,1)-\mu(D,X,0))^2\\
&+ 2(\mu(D,X,1)-\mu(D,X,0))\left(\frac{(Y-\mu(D,X,1))\cdot Z}{p(D,X)} - \frac{(Y-\mu(D,X,0))\cdot(1-Z)}{1-p(D,X)}\right)\\
&+ \mu(D,X,1)-\mu(D,X,0) + \left(\frac{(Y-\mu(D,X,1))\cdot Z}{p(D,X)} - \frac{(Y-\mu(D,X,0))\cdot(1-Z)}{1-p(D,X)}\right)\\
&- \theta && (17)\\
&:= \tilde{\psi}_1(W,\theta,\eta) + \tilde{\psi}_2(W,\theta,\eta) - \theta
\end{aligned}$$

with

$$\begin{aligned}
&\tilde{\psi}_1(W,\theta,\eta)\\
&= (\mu(D,X,1)-\mu(D,X,0))^2\\
&+ 2(\mu(D,X,1)-\mu(D,X,0))\left(\frac{(Y-\mu(D,X,1))\cdot Z}{p(D,X)} - \frac{(Y-\mu(D,X,0))\cdot(1-Z)}{1-p(D,X)}\right)
\end{aligned}$$

and

$$\tilde{\psi}_2(W,\theta,\eta) = \mu(D,X,1)-\mu(D,X,0)+\left(\frac{(Y-\mu(D,X,1))\cdot Z}{p(D,X)} - \frac{(Y-\mu(D,X,0))\cdot(1-Z)}{1-p(D,X)}\right).$$

The moment condition $\mathbb{E}[\tilde{\psi}(W,\theta_0,\eta_0)] = 0$ holds, because

$$\mathbb{E}\left[(\mu_0(D,X,1)-\mu_0(D,X,0))^2 + (\mu_0(D,X,1)-\mu_0(D,X,0))\right] - \theta = 0$$

and

$$\mathbb{E}\left[\left(\frac{(Y-\mu_0(D,X,1))\cdot Z}{p_0(D,X)} - \frac{(Y-\mu_0(D,X,0))\cdot(1-Z)}{1-p_0(D,X)}\right)\right] = 0,$$

as

$$\begin{aligned}
\mathbb{E}\left[\frac{(Y-\mu_0(D,X,1))Z}{p_0(D,X)}\right] &= \mathbb{E}\left[\frac{Z}{p_0(D,X)}\mathbb{E}\left[Y-\mu_0(D,X,1)|D,X,Z\right]\right]\\
&= P(Z=1)\mathbb{E}\left[\frac{1}{p_0(D,X)}\left(\mathbb{E}\left[Y|D,X,Z\right]-\mu_0(D,X,1)\right)\middle| Z=1\right]\\
&= P(Z=1)\mathbb{E}\left[\frac{1}{p_0(D,X)}\left(\mathbb{E}\left[Y|D,X,Z=1\right]-\mu_0(D,X,1)\right)\middle| Z=1\right] = 0
\end{aligned}$$

and analogously $\mathbb{E}\left[\frac{(Y-\mu_0(D,X,0))\cdot(1-Z)}{1-p_0(D,X)}\right] = 0$. Furthermore, by the same argument, we have

$$\mathbb{E}\left[(\mu_0(D,X,1)-\mu_0(D,X,0))\left(\frac{(Y-\mu_0(D,X,1))\cdot Z}{p_0(D,X)} - \frac{(Y-\mu_0(D,X,0))\cdot(1-Z)}{1-p_0(D,X)}\right)\right] = 0.$$

Neyman orthogonality of $\tilde{\psi}$ can be shown by taking the Gateaux derivates w.r.t. the nuisance parameters:

$$
\begin{aligned}
&\partial_r E[\tilde{\psi}_1(W, \theta_0, \eta_0 + r(\eta - \eta_0)]\big|_{r=0} \\
&= E\left[\partial_r \tilde{\psi}_1, (W, \theta_0, \eta_0 + r(\eta - \eta_0))\big|_{r=0}\right] \\
&= 2E\left[((\mu_0(D,X,1) - \mu_0(D,X,0))((\mu(D,X,1) - \mu_0(D,X,1)) - (\mu(D,X,0) - \mu_0(D,X,0)))\}\right] \\
&\quad - 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{Z(\mu(D,X,1) - \mu_0(D,X,1))}{p_0(D,X)}\right] \\
&\quad + 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{(1-Z)(\mu(D,X,0) - \mu_0(D,X,0))}{1 - p_0(D,X)}\right] \\
&\quad - 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{Z(Y - \mu_0(D,X,1))(p(D,X) - p_0(D,X))}{p_0(D,X)^2}\right] \\
&\quad - 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{(1-Z)(Y - \mu_0(D,X,0))(p(D,X) - p_0(D,X))}{(1 - p_0(D,X))^2}\right] \\
&\quad + 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{(1-Z)(Y - \mu_0(D,X,0))(p(D,X) - p_0(D,X))}{(1 - p_0(D,X))^2}\right] \\
&\quad + 2E\Bigg[((\mu(D,X,1) - \mu_0(D,X,1)) - (\mu(D,X,0) - \mu_0(D,X,0))) \\
&\quad \left(\frac{(Y - \mu_0(D,X,1)) \cdot Z}{p_0(D,X)} - \frac{(Y - \mu_0(D,X,0)) \cdot (1-Z)}{1 - p_0(D,X)}\right)\Bigg] \\
&= 2E\left[((\mu_0(D,X,1) - \mu_0(D,X,0))((\mu(D,X,1) - \mu_0(D,X,1)) - (\mu(D,X,0) - \mu_0(D,X,0)))\}\right] \\
&\quad - 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{Z(\mu(D,X,1) - \mu_0(D,X,1))}{p_0(D,X)}\right] \\
&\quad + 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))\frac{(1-Z)(\mu(D,X,0) - \mu_0(D,X,0))}{1 - p_0(D,X)}\right] \\
&= 2E\left[((\mu_0(D,X,1) - \mu_0(D,X,0))((\mu(D,X,1) - \mu_0(D,X,1)) - (\mu(D,X,0) - \mu_0(D,X,0)))\}\right] \\
&\quad - 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))(\mu(D,X,1) - \mu_0(D,X,1))\mathbb{E}\left[\frac{Z}{p_0(D,X)}\middle| D, X\right]\right] \\
&\quad + 2E\left[(\mu_0(D,X,1) - \mu_0(D,X,0))(\mu(D,X,0) - \mu_0(D,X,0))\mathbb{E}\left[\frac{(1-Z)}{1 - p_0(D,X)}\middle| D, X\right]\right] \\
&\quad = 0
\end{aligned}
$$

and

$$
\begin{aligned}
&\partial_r \mathbb{E}[\tilde{\psi}_2(W, \theta_0, \eta_0 + r(\eta - \eta_0)]\big|_{r=0} \\
&= \mathbb{E}\left[\partial_r \tilde{\psi}_2, (W, \theta_0, \eta_0 + r(\eta - \eta_0))\big|_{r=0}\right] \\
&= \mathbb{E}\left[(\mu(D,X,1) - \mu_0(D,X,1))\right] - \mathbb{E}\left[(\mu(D,X,0) - \mu_0(D,X,0))\right] \\
&\quad - \mathbb{E}\left[\frac{Z(\mu(D,X,1) - \mu_0(D,X,1))}{p_0(D,X)}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\mathbb{E}\left[\frac{(1-Z)(\mu(D,X,0)-\mu_0(D,X,0))}{1-p_0(D,X)}\right] \\
&-\mathbb{E}\left[\frac{Z(Y-\mu_0(1,D,X))(p(D,X)-p_0(D,X))}{p_0(D,X)^2}\right] \\
&-\mathbb{E}\left[\frac{(1-Z)(Y-\mu_0(0,D,X))(p(D,X)-p_0(D,X))}{(1-p_0(D,X))^2}\right]=0
\end{aligned}
$$

since $\mathbb{E}[Z|D,X]=p_0(D,X)$, $\mathbb{E}[Z(Y-\mu_0(D,X,1))|D,X]=0$ and $\mathbb{E}[(1-Z)(Y-\mu_0(D,X,0))|D,X]=0$.

Appendix C. Extensions to multivalued IVs

In this appendix, we adapt the score function in eq. (7) to multivalued IVs Z, which is also a key contribution of this paper. In the case of a continuous Z, this requires discretizing its values in some parts of the score function. To this end, let $l=1,\dots,L$ be a partition of its support $\mathcal{Z}$ with $\cup_l Z_l=\mathcal{Z}$. For a discrete IV, $Z_l=z_l$, $l=1,\dots,L$, would be any value Z can take with probability $\mathbb{P}(Z=z_l)>c$, with $c>0$. For a continuous IV, such a partition may be generated based on the quantile function (e.g. percentiles) of Z. Let $1(Z\in Z_l)$ denote the indicator function, which is one if Z falls into the partition Z_l and else is zero, and $p_l(D,X)=\mathbb{P}(Z\in Z_l|D,X)$ denote the corresponding IV propensity score. Then, testing with a multivalued Z can be based on the following score function:

$$
\begin{aligned}
&\psi(W,\theta,\eta) \qquad (18)\\
&=\sum_{l=1}^{L}(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l))^2 \\
&+\sum_{l=1}^{L}2(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l)) \\
&\left(\frac{(Y-\mu(D,X,Z\in Z_l))1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu(D,X,Z\notin Z_l))1(Z\notin Z_l)}{1-p_l(D,X)}\right) \\
&+\sum_{l=1}^{L}(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l)) \\
&+\sum_{l=1}^{L}\left(\frac{(Y-\mu(D,X,Z\in Z_l))1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu(D,X,Z\notin Z_l))1(Z\notin Z_l)}{1-p_l(D,X)}\right).
\end{aligned}
$$

This score has a variance that is bounded away from zero, is zero in expectation under the null hypothesis, $\theta_0=0$, and is Neyman-orthogonal, as formally shown in Appendix D. We may construct cross-fitted estimators of θ_0 based on the score function (18). It is worth noting that the corresponding target parameter in (18) is given by

$$
\mathbb{E}\left[\sum_{l=1}^{L}[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))^2+(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))]\right],
$$

which tests the null hypothesis given in eq. (4) for binary and discrete IVs, and approximates eq. (4) in the case of continuous IVs if the bins defining Z_l become small.

Appendix D. Proof of moment condition and Neyman orthogonality of ψ

Equation (18) suggests the following score function for testing when Z is multivalued discrete or continuous:

$$
\begin{aligned}
&\psi(W,\theta,\eta) \qquad (19)\\
&= \sum_{l=1}^{L}(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l))^2\\
&+\sum_{l=1}^{L}2(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l))\\
&\left(\frac{(Y-\mu(D,X,Z\in Z_l))1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu(D,X,Z\notin Z_l))1(Z\notin Z_l)}{1-p_l(D,X)}\right)\\
&+\sum_{l=1}^{L}(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l))\\
&+\sum_{l=1}^{L}\frac{(Y-\mu(D,X,Z\in Z_l))1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu(D,X,Z\notin Z_l))1(Z\notin Z_l)}{1-p_l(D,X)}-\theta\\
&:= \psi_1(W,\theta,\eta)+\psi_2(W,\theta,\eta)-\theta.
\end{aligned}
$$

First, we show that the moment condition holds. By definition, we have

$$
\sum_{l=1}^{L}(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l))^2+\sum_{l=1}^{L}(\mu(D,X,Z\in Z_l)-\mu(D,X,Z\notin Z_l))-\theta=0.
$$

Analogous to the proof in Appendix B, we have

$$
\sum_{l=1}^{L}\mathbb{E}\left[\frac{(Y-\mu_0(D,X,Z\in Z_l))1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu_0(D,X,Z\notin Z_l))1(Z\notin Z_l)}{1-p_l(D,X)}\right]=0.
$$

Hence, we have to show that

$$
\begin{aligned}
&\mathbb{E}\Bigg[\sum_{l=1}^{L}(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\\
&\quad\left(\frac{(Y-\mu_0(D,X,Z\in Z_l))1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu_0(D,X,Z\notin Z_l))1(Z\notin Z_l)}{1-p_l(D,X)}\right)\Bigg]=0,
\end{aligned}
$$

which holds by the same argument. Next, we show that Neyman orthogonality holds. First, note that

$$
\begin{aligned}
&\partial_r\mathbb{E}[\psi_2(W,\theta_0,\eta_0+r(\eta-\eta_0)]\big|_{r=0}\\
&=\mathbb{E}\left[\partial_r\psi_2,(W,\theta_0,\eta_0+r(\eta-\eta_0))\big|_{r=0}\right]\\
&=\sum_{l=1}^{L}\Bigg(\mathbb{E}\left[(\mu(D,X,Z\in Z_l)-\mu_0(D,X,Z\in Z_l))\right]-\mathbb{E}\left[(\mu(D,X,Z\notin Z_l)-\mu_0(D,X,Z\notin Z_l))\right]
\end{aligned}
$$

$$
\begin{aligned}
&- \mathbb{E}\left[\frac{1(Z \in Z_l)(\mu(D,X,Z \in Z_l) - \mu_0(D,X,Z \in Z_l))}{p_l(D,X)}\right] \\
&+ \mathbb{E}\left[\frac{1(Z \notin Z_l)(\mu(D,X,Z \notin Z_l) - \mu_0(D,X,Z \notin Z_l))}{1-p_l(D,X)}\right] \\
&- \mathbb{E}\left[\frac{1(Z \in Z_l)(Y - \mu_0(D,X,Z \in Z_l))(p(D,X) - p_l(D,X))}{p_l(D,X)^2}\right] \\
&- \mathbb{E}\left[\frac{1(Z \notin Z_l)(Y - \mu_0(D,X,Z \notin Z_l))(p(D,X) - p_l(D,X))}{(1-p_l(D,X))^2}\right]\Bigg) = 0
\end{aligned}
$$

since

$$
\begin{aligned}
&\sum_{l=1}^{L} \mathbb{E}\left[\frac{1(Z \in Z_l)(Y - \mu_0(D,X,Z \in Z_l))(p(D,X) - p_l(D,X))}{p_l(D,X)^2}\right] \\
&= \sum_{l=1}^{L} \mathbb{E}\left[\frac{1(Z \notin Z_l)(Y - \mu_0(D,X,Z \notin Z_l))(p(D,X) - p_l(D,X))}{(1-p_l(D,X))^2}\right] = 0,
\end{aligned}
$$

by the same arguments used before,

$$
\begin{aligned}
&\sum_{l=1}^{L} \mathbb{E}\left[\frac{1(Z \notin Z_l)(\mu(D,X,Z \notin Z_l) - \mu_0(D,X,Z \notin Z_l))}{1-p_l(D,X)}\right] \\
&= \sum_{l=1}^{L} \mathbb{E}\left[\mathbb{E}\left[\frac{1(Z \notin Z_l)(\mu(D,X,Z \notin Z_l) - \mu_0(D,X,Z \notin Z_l))}{1-p_l(D,X)}\middle| D, X\right]\right] \\
&= \sum_{l=1}^{L} \mathbb{E}\left[(\mu(D,X,Z \notin Z_l) - \mu_0(D,X,Z \notin Z_l))\mathbb{E}\left[\frac{1(Z \notin Z_l)}{1-p_l(D,X)}\middle| D, X\right]\right] \\
&= \sum_{l=1}^{L} \mathbb{E}\left[(\mu(D,X,Z \notin Z_l) - \mu_0(D,X,Z \notin Z_l))\right]
\end{aligned}
$$

by iterated expectation and

$$
\begin{aligned}
&\sum_{l=1}^{L} \mathbb{E}\left[\frac{1(Z \in Z_l)(\mu(D,X,Z \in Z_l) - \mu_0(D,X,Z \in Z_l))}{p_l(D,X)}\right] \\
&= \sum_{l=1}^{L} \mathbb{E}\left[(\mu(D,X,Z \in Z_l) - \mu_0(D,X,Z \in Z_l))\right].
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
&\partial_r \mathbb{E}[\psi_1(W,\theta_0,\eta_0 + r(\eta-\eta_0))]\big|_{r=0} \\
&= \mathbb{E}\left[\partial_r \psi_1(W,\theta_0,\eta_0 + r(\eta-\eta_0))\big|_{r=0}\right] \\
&= 2\sum_{l=1}^{L} \mathbb{E}\Big[(\mu_0(D,X,Z \in Z_l) - \mu_0(D,X,Z \notin Z_l)) \\
&\quad \cdot ((\mu(D,X,Z \in Z_l) - \mu_0(D,X,Z \in Z_l)) - (\mu(D,X,Z \notin Z_l) - \mu_0(D,X,Z \notin Z_l)))\Big]
\end{aligned}
$$

$$
\begin{aligned}
&- 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot\frac{1(Z\in Z_l)(\mu(D,X,Z\in Z_l)-\mu_0(D,X,Z\in Z_l))}{p_l(D,X)}\right]\\
&+ 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot\frac{1(Z\notin Z_l)(\mu(D,X,Z\notin Z_l)-\mu_0(D,X,Z\notin Z_l))}{1-p_l(D,X)}\right]\\
&- 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot\frac{1(Z\in Z_l)(Y-\mu_0(D,X,Z\in Z_l))(p(D,X)-p_l(D,X))}{p_l(D,X)^2}\right]\\
&- 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot\frac{1(Z\notin Z_l)(Y-\mu_0(D,X,Z\notin Z_l))(p(D,X)-p_l(D,X))}{(1-p_l(D,X))^2}\right]\\
&+ 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot\frac{1(Z\notin Z_l)(Y-\mu_0(D,X,Z\notin Z_l))(p(D,X)-p_l(D,X))}{(1-p_l(D,X))^2}\right]\\
&+ 2\sum_{l=1}^{L}\mathbb{E}\Big[((\mu(D,X,Z\in Z_l)-\mu_0(D,X,Z\in Z_l))-(\mu(D,X,Z\notin Z_l)-\mu_0(D,X,Z\notin Z_l)))\\
&\left(\frac{(Y-\mu_0(D,X,Z\in Z_l))\cdot 1(Z\in Z_l)}{p_l(D,X)}-\frac{(Y-\mu_0(D,X,Z\notin Z_l))\cdot 1(Z\notin Z_l)}{1-p_l(D,X)}\right)\Big]\\
&= 2\sum_{l=1}^{L}\mathbb{E}\Big[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\\
&\cdot((\mu(D,X,Z\in Z_l)-\mu_0(D,X,Z\in Z_l))-(\mu(D,X,Z\notin Z_l)-\mu_0(D,X,Z\notin Z_l)))\Big]\\
&- 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot(\mu(D,X,Z\in Z_l)-\mu_0(D,X,Z\in Z_l))\right]\\
&+ 2\sum_{l=1}^{L}\mathbb{E}\left[(\mu_0(D,X,Z\in Z_l)-\mu_0(D,X,Z\notin Z_l))\cdot(\mu(D,X,Z\notin Z_l)-\mu_0(D,X,Z\notin Z_l))\right]\\
&= 0.
\end{aligned}
$$

Appendix E. Proof of Theorem 1

To prove Theorem 1, we apply Theorem 3.1 in Chernozhukov et al. (2018). Hence, we only need to show Assumption 3.1 and 3.2 in Chernozhukov et al. (2018). All bounds in the proof hold uniformly over $\mathbb{P}\in\mathcal{P}$ but we omit this qualifier for brevity. We use C to denote a strictly positive constant that is independent of n and $\mathbb{P}\in\mathcal{P}$. The value of C may change at each appearance. In Appendix D we have already shown that the moment condition and Neyman orthogonality is satisfied. Next, we note that the score in equation (18) is linear, i.e.

$$\psi(W,\theta,\eta)=\psi^a(W,\eta)\theta+\psi^b(W,\eta),$$

with $\psi^a(W,\eta)=-1$, which is in line with Assumption 3.1 in Chernozhukov et al. (2018). Next, we demonstrate the satisfaction of Assumption 3.2 to complete the proof. We define the following nuisance realization set $\mathcal{T}_n$ as the set of all P-square-integrable functions η such that

$$
\begin{aligned}
\|\eta_0-\eta\|_{P,2q}&\le C,\\
\|\eta_0-\eta\|_{P,4}&\le \delta_N,\\
\|\eta_0-\eta\|_{P,2}&\le \delta_N^{1/2}N^{-1/4},
\end{aligned}
$$

for $\delta_N = o(1)$ and a constant $q > 2$. Note that Assumption 3.2 (a)-(c) hold by construction of the set $\mathcal{T}_N$ and Assumption 6 due to same arguments as in Huber and Kueck (2022). Assumption 3.2 (d) in Chernozhukov et al. (2018) holds since

$$\begin{aligned}
&\mathbb{E}[\psi(W, \theta_0, \eta_0)^2] \\
&\geq \mathbb{E}\left[\left(\sum_{l=1}^{L} \frac{(Y - \mu_0(D, X, Z \in Z_l))1(Z \in Z_l)}{p_l(D, X)} - \frac{(Y - \mu_0(D, X, Z \notin Z_l))1(Z \notin Z_l)}{1 - p_l(D, X)}\right)^2\right] \\
&> 0
\end{aligned}$$

by Assumption 6 and since $p_l(D, X) > c > 0$ by construction for all $l = 1, \ldots, L$. This completes the proof.

Appendix F. Proof of Theorem 3

Proof Since $\lim_{n\to\infty} P(\hat{S} = \mathcal{S}) = 1$ we just have to show that

$$\lim_{n\to\infty} P(\hat{V} \subseteq \mathcal{V}) = 1.$$

If $\hat{\mathcal{P}}_{pass} = \emptyset$, our statement holds trivially. Hence, consider a variable $j : \mathcal{P}_j \in \hat{\mathcal{P}}_{pass}$, i.e, $Z_j \in \hat{S}$ and $Z_j \in \hat{\mathcal{V}}$. Therefore, it holds

$$|\sqrt{n}\hat{\sigma}_j^{-1}\hat{\theta}_j| < c_\alpha$$

by definition of $\hat{\mathcal{V}}$. By Corollary 2, we have that

$$\lim_{n\to\infty} P(|\sqrt{n}\hat{\sigma}_j^{-1}\hat{\theta}_j| > c_\alpha) = 1$$

if $j \notin \mathcal{V}$. This shows that $j \in \hat{\mathcal{V}}$ implies $j \in \mathcal{V}$ which concludes the proof. ■

Appendix G. Proof of Theorem 4

Proof Consider any variable $Z_j \in \mathcal{P}^*$. We know that $E[Y|D, X] = E[Y|D, X, Z_j]$ and hence by Theorem 1,

$$P(\underbrace{|\sqrt{n}\hat{\sigma}_j^{-1}\hat{\theta}_j| > c_\alpha}_{:=A_j}) = \alpha$$

if $n \to \infty$. We can conclude that

$$\begin{aligned}
\lim_{n\to\infty} P(\hat{\mathcal{P}}_{pass} \neq \emptyset) &= 1 - \lim_{n\to\infty} P(\hat{\mathcal{P}}_{pass} = \emptyset) \\
&= 1 - \lim_{n\to\infty} P(\cap_{j=1,\ldots,M} A_j) \geq 1 - \alpha
\end{aligned}$$

with $M := |\mathcal{P}^*| > 0$. Understanding the stochastic dependence structure of the events A_j, $j = 1, \ldots, M$, could obviously lead to sharper bounds. ■

Appendix H. Pseudo-code

Algorithm 1: Testing procedure for selecting partition $\mathcal{P}^*$

Input: Y (Outcome), D (Treatment), Q (Potential controls and instruments)

Select strong instruments $\hat{S}$
for $j \in Q$ **do**
  Predict D by Q_j conditional on $Q \setminus Q_j$ using ML or regression: $\hat{F}_j \leftarrow \hat{t}_j^2$
  if $\hat{F}_j > C_{\tilde{\alpha}}$ *with* $\tilde{\alpha} = 0.1/\log(n)$ **then**
    $j \in \hat{S}$
  end
end

Select instruments passing conditional independence test $\hat{\mathcal{V}}$
for $j \in \hat{S}$ **do**
  $\hat{\mu}(D, X, Z) \leftarrow$ and $\hat{p}(D, X) \leftarrow$ ML estimates
  $\hat{\theta}_j \leftarrow$ constructed via score in eq. (7) (binary case) or (18) (continuous case)
  $\hat{\sigma}_j \leftarrow$ constructed as described in Theorem 1
  $\hat{t}_{ind,j} = \frac{\hat{\theta}_j}{\hat{\sigma}_j} \leftarrow$ test $H_0 : \theta = 0$
  if $\hat{t}_{ind,j} < c_\alpha$ **then**
    $j \in \hat{\mathcal{V}}$
  end
end

Select final partition $\hat{\mathcal{P}}$: $\hat{\mathcal{P}}_{pass} \leftarrow \left\{\mathcal{P}_j : j \in \left(\hat{S} \cap \hat{\mathcal{V}}\right)\right\}$
if $\hat{\mathcal{P}}_{pass} = \emptyset$ **then**
  end and report that H_0 is not rejected (no identification)
end
else if $|\hat{\mathcal{P}}_{pass}| = 1$ **then**
  $\hat{\mathcal{P}} \leftarrow \hat{\mathcal{P}}_{pass}$
end
else if $|\hat{\mathcal{P}}_{pass}| > 1$ **then**
  $\mathtt{p}(\mathcal{P}_j) \leftarrow \mathtt{p\text{-}value}(\hat{t}_{ind,j})$ for $j \in \hat{\mathcal{V}}$
  $\hat{\mathcal{P}}_{pmax} = \underset{\mathcal{P}_j \in \hat{\mathcal{P}}_{pass}}{\arg\max}\ p(\mathcal{P}_j)$
end

If identification is confirmed ($\hat{\mathcal{P}}_{pass} \neq \emptyset$), estimate the treatment effect
Regress: $Y \leftarrow D$ and $X = Q_{[j]} \in \hat{\mathcal{P}}_{pmax}$

Appendix I. Computational cost

In low-dimensional settings, we recommend using the FSHT procedure, which is computationally attractive. We agree that in high-dimensional settings the computational costs can be substantially larger. In the paper (see p. 8), we recommend using double machine learning (DML) to estimate the first-stage effect for each of the p candidates when $p > n$. A (weakly) smaller number of computations is performed for the conditional mean-independence test in the second stage because the number of candidate variables that pass the first-stage screening is typically smaller. Denoting by $s \leq p$ the number of strong IV candidates, the overall computational order can be summarized as

$$O(p \cdot C_{\mathrm{DML}} + s \cdot C_{\mathrm{test}}) \,,$$

where C_{DML} denotes the cost of one DML estimation for a single candidate variable (including K-fold cross-fitting and nuisance estimation), and C_{test} denotes the cost of one second-stage test evaluation. In our simulations and application we use, for instance, random forests as learners for nuisance estimation. The computational cost of a random forest with T trees typically scales (up to constants and implementation details) on the order of $O(T \cdot m \cdot n \log n)$, where m is the number of features considered per split. When p is very large, it is crucial to use efficient methods for running the repeated first-stage estimations across the p candidates. In this case, we recommend L_2-boosting as a computationally attractive alternative for nuisance estimation and screening, as discussed in Section 3.3 of Kueck et al. (2023). Using L_2-boosting primarily reduces the per-fit cost C_{DML} (and hence the overall runtime), while the outer loop over candidates remains. Importantly, the computations are embarrassingly parallel across candidate variables, so runtime can be substantially reduced via parallel computing.

Appendix J. Simulation Details

This section provides a simulation study to investigate the finite sample behavior of our testing approach based on the following data generating process (DGP):

$$\begin{aligned} Y &= D + X'\beta + \gamma Z + W + U, \\ D &= I\{X'\beta + Z + \delta W + V > 0\}, \\ X, Z &\sim Bernoulli(\pi), \\ W &\sim N(0,1), U \sim N(0,1), V \sim N(0,1), \end{aligned}$$

with X, Z, W, U, V being independent of each other. Outcome Y is a linear function of D (whose treatment effect is one), covariates X (for $\beta \neq 0$), the unobservables W and U, and the supposed instrument Z if the coefficient $\gamma \neq 0$. The binary treatment D is a function of X and the unobservable V, as well as W if coefficient $\delta \neq 0$. While the supposed IV Z is binary, the unobserved terms U, V, W are random, standard normally distributed variables that are independent of each other, of Z, and of X. The joint set of covariates and IVs Q is created as follows. We first draw a matrix of multivariate normal variables $\tilde{Q} \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a covariance matrix with $Cov(\tilde{Q}_j, \tilde{Q}_k) = 0.5^{|j-k|}$. We then compute probabilities $\pi_i = \frac{1}{1+exp(-2 \cdot \tilde{Q}_i)}$ and draw $Q_i \sim Bernoulli(\pi_i)$. We generate 10 variables in the set $Q = (X, Z)$, with the last one being the (single) IV Z. Concerning the first nine elements in Q, which are considered as covariates X, β gauges their effects on Y and D, respectively, and thus, the magnitude of confounding due to observables. The jth element in the coefficient vector β is set to $0.8/j$ for $j = 1, ..., 4$, implying a linear decay of covariate importance in terms of confounding for the first 4 covariates. Moreover, for elements $j = 5, ..., 9$, the coefficients are equal to zero, implying that covariates 5 to 9 are random noise variables that are neither confounders, nor IVs.

In a second simulation design, we also consider continuous variables Q, which differ from the preceding DGP in that $X, Z \sim Uniform(-0.5, 0.5)$, i.e. the set of observed covariates X and the IV Z are now drawn from the multivariate uniform distribution, $X, Z \sim Uniform(-0.5, 0.5)$. All other properties of this DGP are identical to the previous scenario with binary Z and X variables.

We investigate the performance of our testing approach in 200 simulations with sample sizes of $n = 1000$, 4000, and 16000. For estimating the (conditional) first stage effect of elements in Q on D to select sufficiently strong IVs, we use DML for partially linear models (Chernozhukov et al., 2018). DML can be applied to both discrete and continuous elements in Q and is implemented in the *DoubleML* package for statistical software R. The so-called nuisance parameters in the first stage are estimated using the LASSO with cross-fitting, setting the number of folds to five (K = 5). Specifically, the classification LASSO is used for discrete elements in Q, modelling probabilities for treatment assignment while the LASSO is used for continuous variables. For testing the conditional independence of any selected candidate IV in the second stage, we consider a cross-fitted estimator $\hat{\theta}$ of eq. (8) with five folds ($K = 5$), which is based on the doubly robust score functions in eqs. (7) and (18) for binary and continuous candidate IVs, respectively. In the case of a continuous IV, its support is partitioned based on the quartiles of its distribution, such that $L = 4$. For nuisance parameter estimation in the second stage, LASSO with default parameters as implemented in the *glmnet* package (Friedman et al., 2010) for the statistical software R is applied. We consider several statistics from our simulations: the estimated violation, $\hat{\theta}$, when using variable Z as the IV, the standard deviation of the estimated violation (std), and the average of the standard error of the

estimated violation across all simulation samples (mean se). These statistics are useful for judging the performance of the test conditional on selecting Z as the IV, which is the correct choice if $\gamma = 0$, while there is no valid IV if $\gamma \neq 0$ and no identification when controlling for covariates if $\delta \neq 0$.

Since the choice of the IV is not fixed a priori but is part of the estimation process, we also consider three further statistics. The first one is the empirical selection rate of Z (sel.Z), which indicates the frequency with which Z is selected as the best candidate IV *and also* has a $\mathtt{p}$-value greater than 30% when testing conditional independence based on the estimate $\hat{\theta}$.[16] In other words, this corresponds to the proportion of simulations in which our method simultaneously selects Z as the IV and at the same time keeps the null hypothesis of conditional independence. In scenarios where the null hypothesis holds ($\delta = \gamma = 0$), this selection rate should approach one as n increases. The second statistic is the analogous selection rate of the noise variables X_5 to X_9 (sel.noconf). It indicates the share of simulations in which one of the non-confounders (that are not IVs either) is selected as the best candidate IV and at the same time yields a $\mathtt{p}$-value greater than 30% when testing conditional independence when using this noise variable as the IV. As n increases, this selection rate should always approach zero, because asymptotically, noise variables do not have a first stage effect on the treatment conditional on other elements in Q. Therefore, noise variables should not be selected as candidate IVs for the conditional independence test. Finally, we report an equivalent selection rate for the confounders X_1 to X_4 (sel.conf). It corresponds to the share of simulations in which one of the confounders is selected as the best candidate IV and at the same time yields a $\mathtt{p}$-value greater than 30% when testing conditional independence when using this confounder as IV. As n increases, this selection rate should always approach zero, because confounders are never valid IVs.

Table 2 reports the simulation results for binary variables in Q in Panel A, and for continuous elements in Q in Panel B, respectively. In both panels, the top rows provide the results for $\delta = \gamma = 0$, implying that both Assumptions 4 and 5 are satisfied and conditional independence holds. As n increases, the test is more likely to simultaneously select the true IV Z and keep the null hypothesis of conditional independence. Specifically, the selection rate of Z (sel.Z) increases from 32% with $n = 1000$ observations to 68% with $n = 16000$ observations for the binary IV, and from 55% with $n = 1000$ to 64% with $n = 16000$ for the continuous IV. This improvement is mirrored by a reduction in the rate at which a confounder variable ($X_1 ... X_4$) is selected as the IV and passes the test (sel.conf). Under $n = 16000$, it only occurs that any confounder is selected as the best IV and passes the conditional independence test 4% of the time for the binary and never for the continuous scenario. However, we observe that the noise variables ($X_5 ... X_9$) sometimes appear to step in, as their selection rate (sel.noconf) transiently increases in response to the decreasing selection rate of confounders. Additionally, we observe that conditional on having selected the true IV, the estimated violation $\hat{\theta}$ is close to zero for any sample size and never statistically significant at conventional significance levels. Figure A1 depicts the distribution of the estimated $\hat{\theta}$ and visually confirms this finding in the first column. The top row of density plots show the binary $\hat{\theta}$ and the bottom row shows the continuous estimates. As sample size increases, the distribution moves closer to zero, as expected.

The intermediate rows of Table 2 and middle column of Figure A1 present the results for a violation of SOO when setting $\delta = 2$, $\gamma = 0$. As n increases, the selection rates of Z (sel.Z)

16. We suggest the threshold of 30% based on examination of the distribution of $\mathtt{p}$-values for both true IVs and confounders, see Figures A2 and A3. This stricter threshold guards against the possible choice of a confounder as an IV, although a more traditional threshold of 10% could also be used.

Table 2: Simulations: Single Instrument

PANEL A: Binary Instrument						
N	sel.Z	sel.noconf	sel.conf	$\hat{\theta}$	std	mean se
	Assumptions 4 and 5 hold ($\delta = 0, \gamma = 0$)					
1000	32%	2%	59%	0.002	0.010	0.008
4000	45%	1%	42%	0.000	0.002	0.002
16000	68%	2%	4%	0.000	0.000	0.000
	A 4 violated, A 5 holds ($\delta = 2, \gamma = 0$)					
1000	50%	3%	32%	0.012	0.012	0.016
4000	15%	4%	44%	0.012	0.006	0.009
16000	0%	4%	20%	0.013	0.003	0.005
	A 4 holds, A 5 violated ($\delta = 0, \gamma = 0.5$)					
1000	30%	2%	57%	0.045	0.040	0.050
4000	3%	2%	68%	0.053	0.019	0.027
16000	0%	3%	8%	0.054	0.010	0.014
PANEL B: Continuous Instrument						
N	sel.Z	sel.noconf	sel.conf	$\hat{\theta}$	std	mean se
	Assumptions 4 and 5 hold ($\delta = 0, \gamma = 0$)					
1000	55%	2%	19%	-0.007	0.029	0.031
4000	65%	4%	8%	-0.001	0.015	0.016
16000	64%	1%	0%	0.000	0.007	0.009
	A 4 violated, A 5 holds ($\delta = 2, \gamma = 0$)					
1000	16%	6%	6%	-0.018	0.031	0.017
4000	8%	2%	1%	-0.015	0.018	0.007
16000	0%	2%	0%	-0.030	0.014	0.003
	A 4 holds, A 5 violated ($\delta = 0, \gamma = 0.5$)					
1000	0%	4%	18%	-0.490	0.192	0.049
4000	0%	6%	8%	-0.552	0.103	0.026
16000	0%	1%	0%	-0.554	0.055	0.013

Notes: columns '$\hat{\theta}$', 'std', and 'mean se' provide the average estimate of θ (the violation of conditional independence) conditional on using Z as IV, its standard deviation, and the average of the standard errors across all samples, respectively. 'sel.Z' gives the share of simulations in which Z is selected as best candidate IV and simultaneously yields a p-value > 0.30 when testing conditional independence based on the estimate of $\hat{\theta}$. 'sel.noconf' is an equivalent measure for covariates X_5 to X_9, corresponding to the share of selecting a non-confounder as best candidate IV and having a p-value $> 30\%$ when testing conditional independence. 'sel.conf' is an equivalent measure for covariates X_1 to X_4, corresponding to the share of selecting an observable confounder the best candidate IV and having a p-value $> 30\%$ when testing conditional independence.

and the confounder variables (sel.conf) both go towards zero, as expected. The rate of finding a nonempty set $\hat{\mathcal{P}}_{pass}$, which is equal to the sum of the selection rates sel.Z, sel.noconf and sel.conf, also goes towards zero, finding no identification as expected. Furthermore, the estimated violation conditional on using Z as an IV ($\hat{\theta}$) is statistically different from zero under the larger sample sizes $n = 4000, 16000$. We note that in settings where A 4 is violated, the selection rate of a confounder (sel.conf) in the binary scenario is still slightly high, around 20% in the largest sample for the binary IV, athough this rate is 0% for the continuous IV.[17] The lower rows of Table 2 and final column of Figure A1 provide the performance measures under a violation of A 5, IV validity, when considering

17. In further experiments (not reported here), strengthening the degree of confounding by increasing the value of γ leads to a substantial reduction in the rate at which a confounder is selected.

$\delta = 0$, $\gamma = 0.5$. Again and as expected, both the selection rates of Z (sel.Z) and the confounder variables (sel.conf) tend to zero as n increases. At the same time, the estimated violation conditional on using Z as an IV ($\hat{\theta}$) is large in absolute terms and highly statistically significant across all sample sizes, and $\hat{\mathcal{P}}_{pass}$ is empty 89% of the time for binary and 99% of the time for continuous IVs for $n = 16000$.

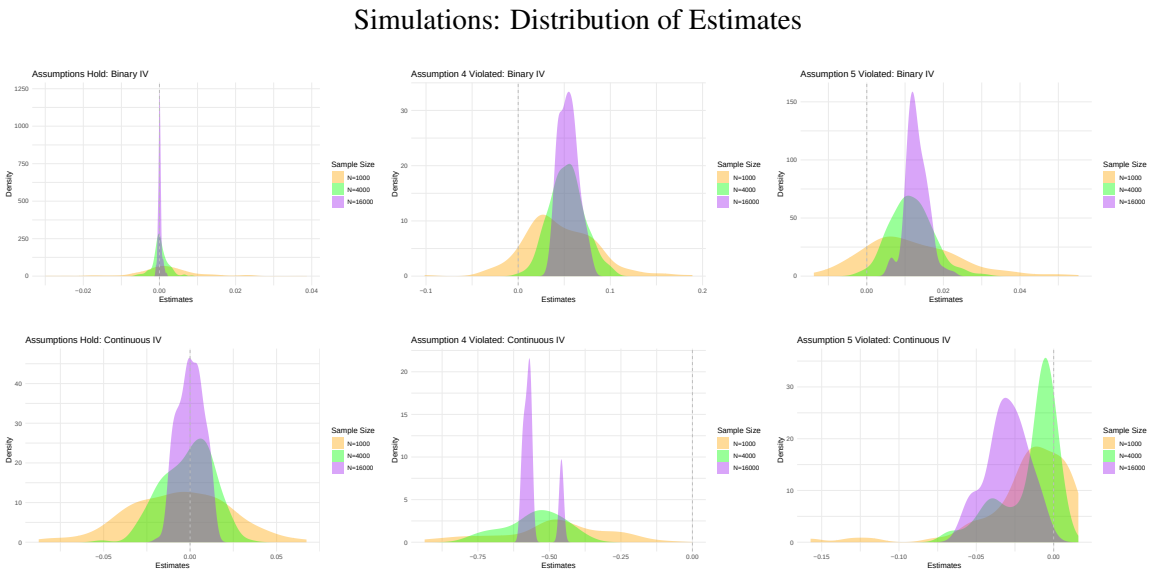
We now consider a modification of the previous DGP that entails the existence of multiple IVs. Specifically, the true set of IVs $(Z_1...Z_3)$ now comprises the last three elements in Q, i.e. variables 8 to 10, which are constructed equivalently to Z in the previous setup. We investigate the performance when A 5 is violated for all of the IVs. This implies that the set of confounders remains the same as before $(X_1...X_4)$, while the set of noise variables now only consists of three elements $(X_5...X_7)$. Results for the multiple IV simulations are shown in Table 3, and are similar to those shown in the single IV results. In the case of multiple binary IVs and no violations of the assumptions, the selection rate of Z (sel.Z) is up to 96% when $n = 16000$. As in the single IV scenario, under a violation of A 4, SOO, the selection rates of the true IVs (sel.Z) and confounding variables (sel.conf) go to zero. For multiple continuous IVs (Panel B), results are consistent with the single IV setting, with the exception of the slightly lower increase of the IV selection rate in the larger sample size compared to the binary version (67% in the case of multiple IVs, compared to 64% under the single IV scenario). However, in all settings of the multiple continuous IV case, the rate of selecting a confounder (sel.conf) goes to zero and the estimate magnitudes and $\hat{\mathcal{P}}_{pass}$ behave as expected, moving towards zero and the empty set under no violation of the assumptions and away from zero otherwise.

Table 3: Simulations: Multiple Instruments

PANEL A: Binary Instruments						
N	sel.Z	sel.noconf	sel.conf	$\hat{\theta}$	std	mean se
	Assumptions 4 and 5 hold ($\delta = 0$, $\gamma = 0$)					
1000	80%	1%	18%	0.002	0.008	0.007
4000	84%	2%	12%	0.000	0.002	0.002
16000	96%	0%	0%	0.000	0.000	0.000
	A 4 violated, A 5 holds ($\delta = 2$, $\gamma = 0$)					
1000	86%	1%	11%	0.010	0.012	0.014
4000	40%	2%	40%	0.010	0.005	0.007
16000	0%	3%	16%	0.011	0.003	0.004
	A 4 holds, A 5 violated ($\delta = 0$, $\gamma = 0.5\ \forall Z$)					
1000	70%	0%	21%	0.053	0.035	0.050
4000	15%	2%	48%	0.058	0.018	0.026
16000	0%	1%	8%	0.058	0.009	0.013
PANEL B: Continuous Instruments						
N	sel.Z	sel.noconf	sel.conf	$\hat{\theta}$	std	mean se
	Assumptions 4 and 5 hold ($\delta = 0$, $\gamma = 0$)					
1000	62%	1%	6%	-0.010	0.035	0.031
4000	64%	1%	4%	-0.003	0.017	0.016
16000	67%	2%	0%	-0.001	0.010	0.008
	A 4 violated, A 5 holds ($\delta = 2$, $\gamma = 0$)					
1000	38%	2%	2%	-0.027	0.036	0.017
4000	14%	0%	1%	-0.025	0.021	0.007
16000	0%	0%	0%	-0.040	0.007	0.003
	A 4 holds, A 5 violated ($\delta = 0$, $\gamma = 0.5\ \forall Z$)					
1000	20%	2%	10%	-0.116	0.090	0.037
4000	0%	2%	6%	-0.142	0.064	0.019
16000	0%	2%	0%	-0.140	0.017	0.010

Notes: columns '$\hat{\theta}$', 'std', and 'mean se' provide the average estimate of θ (the violation of conditional independence) conditional on using Z_3 (the last element in Q) as IV, its standard deviation, and the average of the standard errors across all samples, respectively. 'sel.Z' gives the share of simulations in which one of the IVs Z_1 to Z_3 is selected as best candidate IV and simultaneously yields a p-value > 0.30 when testing conditional independence based on the estimate of $\hat{\theta}$. 'sel.noconf' is an equivalent measure for covariates X_5 to X_7, corresponding to the share of selecting a non-confounder as best candidate IV and having a p-value $> 0.3\%$ when testing conditional independence. 'sel.conf' is an equivalent measure for covariates X_1 to X_4, corresponding to the share of selecting an observable confounder the best candidate IV and having a p-value > 0.3 when testing conditional independence.

Appendix K. Figures



This figure depicts density plots of the estimated violation $\hat{\theta}$ for the true IV (covariate X_{10}), across the simulation settings. The top and bottom rows are the binary and continuous IV cases, respectively. The first plots on the left provide results for $\delta = \gamma = 0$, i.e. Assumptions 4 and 5 are satisfied. The middle plots depict the scenario of A 4 being violated, and the right-most plots A 5 is violated. Orange is the smallest sample size (N = 1000), green the medium sample size (N=4000), and purple is the largest (N=16000).

Figure A2: Distribution of p-values: Oracle IV, Binary Simulations

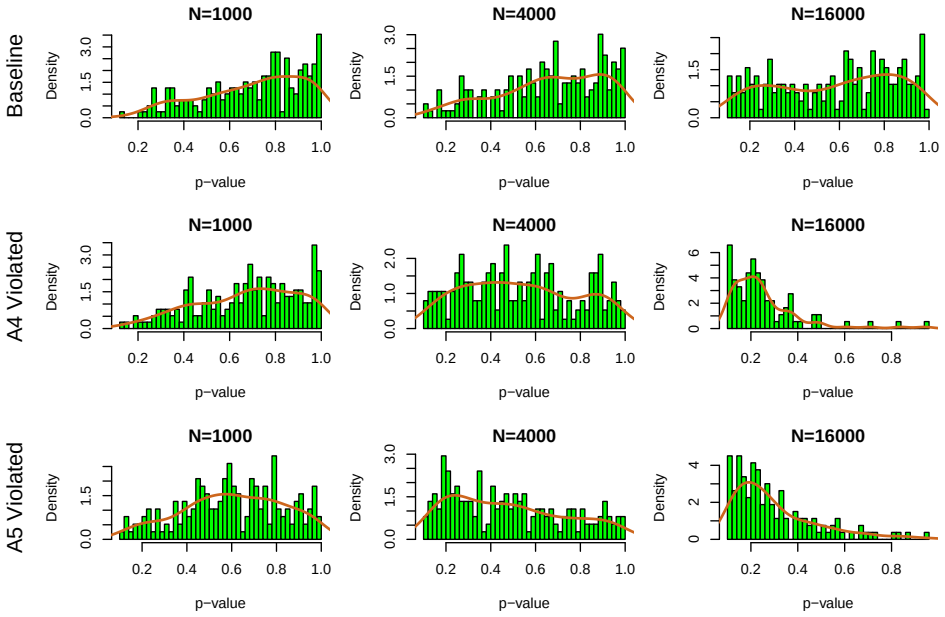
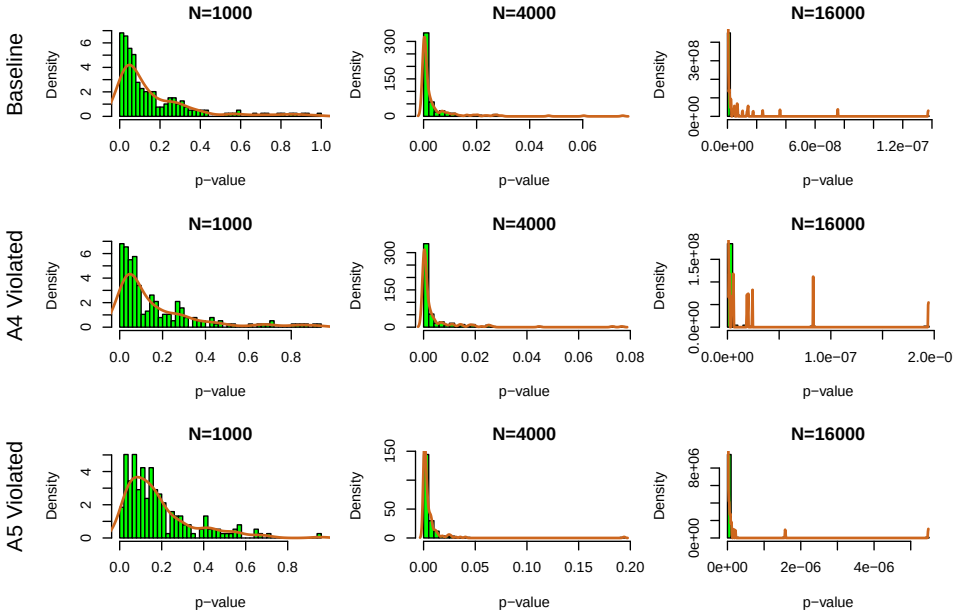


Figure A3: Distribution of p-values: Confounder X_1, Binary Simulations



Appendix L. Tables

Table A1: Application: All $\hat{S}$

Panel A: Number of Prescriptions				
Candidate IV	est	se	pval	pct trimmed
HH Income	0.003	0.038	0.929	0.386
Random Assignment Instrument	-0.007	0.010	0.472	0.001
Living with a partner	0.013	0.009	0.163	0.335
Ever Surveyed	0.013	0.009	0.143	0.294
Any RX	-0.019	0.011	0.078	0.391
Need Dental Assistance Missing	0.060	0.032	0.063	0.905
Application Received	0.024	0.013	0.061	0.381
Employer Insurance	0.025	0.013	0.058	0.551
Application Approved	0.040	0.015	0.007	0.681
Surveyed All Months	-0.540	0.176	0.002	0.771
Retired	0.109	0.031	0.000	0.572
Application Pending	0.057	0.016	0.000	0.703
All Survey End	0.027	0.007	0.000	0.325
OHP Insurance Missing	0.037	0.009	0.000	0.410
Any Hospitalisation	0.057	0.014	0.000	0.656
Panel B: Number of Doctor Visits				
Random Assignment Instrument	-0.000	0.013	0.999	0.002
Any Doctor Visit	-0.002	0.011	0.871	0.344
Application Approved	0.004	0.021	0.867	0.694
Retired	-0.027	0.073	0.705	0.601
Application Received	0.010	0.015	0.496	0.398
Need Dental Work Missing	0.039	0.052	0.447	0.917
Ever Surveyed	0.011	0.011	0.287	0.315
Surveyed in All Months	-0.238	0.217	0.274	0.785
Under 19 All Insured	0.026	0.016	0.111	0.475
Ending Survey	0.023	0.012	0.064	0.343
Application Pending	0.051	0.024	0.036	0.720
HH Income Category	0.112	0.053	0.036	0.402
Any Hospitalisation	0.064	0.025	0.012	0.669
More than 30 hours employed	0.044	0.011	0.000	0.350
Insurance Missing Variable	0.053	0.011	0.000	0.424
OOP Prescription Costs	-0.071	0.002	0.000	0.436
OOP Medical Costs	-0.060	0.001	0.000	0.429