MAX-PLANCK-INSTITUT FÜR SOZIALRECHT UND SOZIALPOLITIK
MAX PLANCK INSTITUTE FOR SOCIAL LAW AND SOCIAL POLICY

mea Munich Center for the Economics of Aging

Instrument Validity Tests with Causal Forests

Helmut Farbmacher, Raphael Guber, Sven Klaassen

13-2020

MEA DISCUSSION PAPERS

mea – Amalienstr. 33_D-80799 Munich_Phone +49 89 38602-355_Fax +49 89 38602-390_www.mpisoc.mpg.de

Instrument Validity Tests with Causal Forests

Helmut Farbmacher, Raphael Guber, Sven Klaassen

Abstract:

Assumptions that are sufficient to identify local average treatment effects (LATEs) generate necessary conditions that allow instrument validity to be refuted. The degree to which instrument validity is violated, however, probably varies across subpopulations. In this paper, we use causal forests to search and test for such local violations of the LATE assumptions in a data-driven way. Unlike previous instrument validity tests, our procedure is able to detect local violations. We evaluate the performance of our procedure in simulations and apply it in two different settings: parental preferences for mixed-sex composition of children and the Vietnam draft lottery.

Zusammenfassung:

Annahmen zur Identifizierung von Local Average Treatment Effects (LATEs) generieren notwendige Bedingungen in den Daten, die es erlauben, die Validität von Instrumentalvariablen (IV) zu widerlegen. Die Stärke einer möglichen Invalidität variiert sehr wahrscheinlich in Subpopulationen. In diesem Papier verwenden wir einen Causal Forest, um datengetrieben nach solch lokalen Verletzungen der LATE-Annahmen zu suchen und zu testen. Im Gegensatz zu früheren Testverfahren, kann unser Verfahren lokale Verletzungen erkennen. Wir zeigen dies in Simulationen und in zwei Anwendungen.

Keywords:

Specification tests, causal forest, LATE, treatment effects, instrument validity

JEL Classification:

C12, C18, C26

Instrument Validity Tests with Causal Forests

Helmut Farbmacher* Raphael Guber

Munich Center for the Economics of Aging, Max Planck Society

Sven Klaassen

Institute of Statistics, University of Hamburg

May 25, 2020

Abstract

Assumptions that are sufficient to identify local average treatment effects (LATEs) generate necessary conditions that allow instrument validity to be refuted. The degree to which instrument validity is violated, however, probably varies across subpopulations. In this paper, we use causal forests to search and test for such local violations of the LATE assumptions in a data-driven way. Unlike previous instrument validity tests, our procedure is able to detect local violations. We evaluate the performance of our procedure in simulations and apply it in two different settings: parental preferences for mixed-sex composition of children and the Vietnam draft lottery.

Keywords: Specification tests, causal forest, LATE, treatment effects, instrument validity

JEL codes: C12, C18, C26

*farbmacher@mpisoc.mpg.de.

1 Introduction

Producing credible estimates of causal effects in empirical research often entails a heavy reliance on instrumental variables (IVs). IVs, however, have to meet strong assumptions to be valid. Discussions about threats to these identifying assumptions and approaches to checking their robustness constitute a crucial part of many empirical papers. Recently, Kitagawa (2015), Huber and Mellace (2015) and Mourifié and Wan (2017) developed tests that allow the validity of IVs to be refuted based on necessary conditions in the data. These conditions are generated by the joint assumptions sufficient to identify local average treatment effects (LATEs), namely the exclusion restriction, randomization/unconfoundedness and monotonicity (Imbens and Angrist, 1994; Angrist et al., 1996). The unifying idea across these three tests is that, given a treatment status, the estimated density of compliers must be non-negative at any point in the distribution of the outcome variable; compliers comprise the unobserved subpopulation of individuals whose treatment status is causally affected by the instrument.

The concept underlying our paper is that the degree to which the LATE assumptions are violated may vary across subpopulations that can be defined by observed characteristics. For example, a direct effect of the instrument on the outcome variable may be large in a relatively small subpopulation but, in the entire population, watered down to the point at which it can no longer be detected by existing tests. By reformulating the necessary conditions into a form similar to that employed to learn the sign of treatment effects, we are able to leverage recent progress in using machine learning to estimate heterogeneous treatment effects. This heterogeneity is conceptually restricted to non-negative values if the LATE assumptions hold, but may take on negative values otherwise.

In recent years, a growing body of research has used machine learning to estimate heterogeneous treatment effects (among others, Tian et al., 2014; Wager and Athey, 2018; Athey et al., 2019; Künzel et al., 2019; Nie and Wager, 2019). Our test proceeds in three steps. In the spirit of subgroup testing, we use shallow regression trees (Breiman et al., 1984) to split the sample along covariate values. Some of these splits form subgroups, which are promising for finding violations of the LATE assumptions. We apply a simple selection procedure to identify these.

Then, we use the double machine learning framework from Chernozhukov et al. (2018) combined with causal forests developed by Wager and Athey (2018) and Athey et al. (2019) to estimate the magnitude of the violations in these promising subgroups. Lastly, we test for local violations of the LATE assumptions using Bonferroni-corrected critical values. Finding violations in at least one subgroup casts doubt on the instrument, because we cannot rule out further undetected violations. Our proposed test procedure can be easily implemented with existing software packages. Additionally, we provide the R package `LATEtest`.

We consider a setup with a binary (endogenous) treatment and a binary instrumental variable. There are several examples in this framework: for instance, the Vietnam draft lottery (Angrist, 1990), the preference for mixed-sex children (Angrist and Evans, 1998) or the Oregon health insurance experiment (Finkelstein et al., 2012). When we apply our data-driven procedure to detect local violations of IV validity, we have to assume that the instrument is randomized (or at least unconfounded conditional on covariates). Randomization (or unconfoundedness) itself is therefore not testable with our procedure. However, randomization (or unconfoundedness) is often fulfilled by design in applied work (for instance, in the Vietnam draft lottery or in the Oregon health insurance experiment), while the most controversial assumption is the exclusion restriction and to some extent also monotonicity. We thus have applications in mind in which the researcher is interested in testing the exclusion restriction and/or monotonicity.

The existing test procedures are, ceteris paribus, more powerful when the overall share of compliers in the *entire population* is low, i.e. the instrument is weak on average. However, applied researchers often rely on strong instruments to avoid issues associated with weak IVs (Bound et al., 1995), and because a larger share of compliers may deliver a causal effect with greater external validity. Our test procedure is more powerful when the share of compliers is low *within a subpopulation*. That is, our test can have more power asymptotically than existing tests even if the instrument turns out to be strong on average. The proposed approach also has advantages over simply using existing IV validity tests within arbitrarily defined covariate subgroups. It automatically chooses the covariates and covariate values to partition on, with honest splitting safeguarding the researcher from an overfitting bias. Researchers can, therefore, credibly demonstrate that they have

extensively searched for violations of key identifying assumptions.

This paper contributes to the recent and fast growing literature on adapting machine learning tools to the needs of applied economists who wish to estimate causal effects and detect and characterize the heterogeneity of these. One strand of the literature uses tree and forest algorithms to estimate the heterogeneity of treatment effects (Asher et al., 2016; Athey and Imbens, 2016; Wager and Athey, 2018; Athey et al., 2019). Belloni et al. (2012, 2014b) and Chernozhukov et al. (2015) present methods based on the least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996) for inference in high-dimensional settings where there may be many possible instrument or control variables relative to the number of observations. See also Belloni et al. (2014a) for an overview. Knaus et al. (2020) use Lasso-type estimators to detect treatment effect heterogeneity in job search programs. In the presence of instruments that violate the exclusion restriction, Kang et al. (2016) and Windmeijer et al. (2019) use the Lasso to select valid IVs in linear models. Wager et al. (2016), Bloniarz et al. (2016) and Athey et al. (2018) improve the efficiency of average treatment effect estimates in randomized experiments with Lasso-based balancing.

The rest of this paper proceeds as follows. In the next section, we briefly revisit the testable implications of IV validity. Section 3 describes our procedure to detect, select and test local violations of IV validity. In Sections 4 and 5, we provide results from a simulation study and apply the test to real data. Section 6 concludes.

2 (Local) Violations of LATE

Let the observed outcome Y have support $\mathcal{Y}$ and the endogenous treatment have status $D \in \{0,1\}$, where $D = 1$ indicates treatment, and the binary instrument $Z \in \{0,1\}$. The potential outcomes and treatments are denoted with Y^{dz} and D^z, where $d, z \in \{0,1\}$. Following Imbens and Angrist (1994) and Abadie (2003), three assumptions are sufficient to identify LATEs in this setup.

Assumption A1. *(Exclusion restriction):* $Y^{d1} = Y^{d0} = Y^d$ *for* $d \in \{0,1\}$ *wp 1.*

Assumption A2.a. *(Randomization):* $(Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0) \perp Z$.

Assumption A2.b. *(Unconfoundedness):* $(Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0) \perp Z|X$.

Assumption A3. *(Monotonicity):* $D^1 \geq D^0$ *or* $D^1 \leq D^0$ *wp 1.*

Assumption A1 rules out a direct effect of the instrument on the potential outcomes. Assumption A2.a assumes that Z is jointly independent of the potential outcomes and treatments. In many applications, Assumption A2.a will hold only when conditioning on a set of predetermined covariates X. In this case, Assumption A2.b replaces Assumption A2.a. Assumption A3 rules out the existence of defiers. Without loss of generality, we assume $D^1 \geq D^0$ in the remainder of this paper, i.e. the instrument is to create an incentive to take up the treatment, and we assume that this is known a priori to the researcher. Our arguments would hold symmetrically for negative monotonicity (i.e., $D^1 \leq D^0$).
Let $\mathcal{B}_\mathcal{Y}$ be a collection of Borel sets generated from $\mathcal{Y}$. Imbens and Rubin (1997) show that, under Assumption A1-A3, it must hold for every $J \in \mathcal{B}_\mathcal{Y}$ that

$$P(Y \in J, D = 1|Z = 1) - P(Y \in J, D = 1|Z = 0) = P(Y^1 \in J, C)\,,$$
$$P(Y \in J, D = 0|Z = 0) - P(Y \in J, D = 0|Z = 1) = P(Y^0 \in J, C)\,.$$

Since the share of compliers (C) has to be non-negative at every point in the distribution of Y, the following inequalities must hold

$$P(Y \in J, D = 1|Z = 1) - P(Y \in J, D = 1|Z = 0) \geq 0\,, \tag{2.1}$$
$$P(Y \in J, D = 0|Z = 0) - P(Y \in J, D = 0|Z = 1) \geq 0\,. \tag{2.2}$$

Balke and Pearl (1997) and Heckman and Vytlacil (2005) also discuss these testable implications. Kitagawa (2015) uses a variance-weighted Kolmogorov-Smirnov-type statistic to test (2.1) and (2.2). Mourifié and Wan (2017) rewrite (2.1) and (2.2) as conditional moment inequalities (conditional on $Y = y$) and then apply the intersection bounds approach of Chernozhukov et al. (2013). Huber and Mellace (2015) relax Assumptions A1 and A2 to hold only in expectation because this is sufficient to identify average effects. Kitagawa (2015), Proposition 1.1 and Mourifié and Wan (2017), Theorem 1, establish that (2.1) and (2.2) are sharp, in the sense that they are the strongest testable implications of Assumption A1-A3 given the available data. Laffers and Mellace (2017) prove that the inequalities proposed by Huber and Mellace (2015) are the strongest testable implications when

Assumptions A1 and A2 hold only in expectation.
The inequalities (2.1) and (2.2) must hold at any point x in the covariate space $\mathcal{X}$ as well

$$P(Y \in J, D = 1|Z = 1, X = x) - P(Y \in J, D = 1|Z = 0, X = x) \geq 0\,, \quad (2.3)$$
$$P(Y \in J, D = 0|Z = 0, X = x) - P(Y \in J, D = 0|Z = 1, X = x) \geq 0\,. \quad (2.4)$$

Conditioning on X can be helpful in several ways. For illustration, if we only impose randomization/unconfoundedness, we can derive from (2.3)

$$\begin{aligned} P(Y \in & J, D = 1|Z = 1, X = x) - P(Y \in J, D = 1|Z = 0, X = x) \\ &= P(Y^{11} \in J|C, X = x)P(C|X = x) - P(Y^{10} \in J|F, X = x)P(F|X = x) \\ &\quad + P(Y^{11} \in J|A, X = x)P(A|X = x) - P(Y^{10} \in J|A, X = x)P(A|X = x) \geq 0\,, \end{aligned}$$

where A, C and F denote always-takers, compliers and defiers, respectively. First, the proportion of compliers within some covariate cell, $P(C|X = x)$, might be lower than in the full sample. Second, the fraction of defiers might be overrepresented or exist only in certain covariate cells. In both cases, local violations of Assumption A3 can be detected more easily when we condition on those cells. Third, the direct effect of the instrument might be stronger for some subpopulations, making it easier to find local violations of Assumption A1 when we condition on X.
Kitagawa (2015), Huber and Mellace (2015) and Mourifié and Wan (2017) also apply their tests within covariate cells. However, how to form these cells so that they make finding violations of the LATE assumptions more likely is an open question, in particular when covariates are continuous. Arbitrarily defining subgroups is inefficient. A further problem is the potentially large dimensionality of X, which makes implementation of the above tests for all x infeasible. Therefore, we propose a data-driven way to find and test local violations of IV validity.

3 A Local IV validity test

Let (Y_i, D_i, Z_i, X_i) be i.i.d. observations for $i = 1, \ldots, n$. Define for any $J \in \mathcal{B}_\mathcal{Y}$ the pseudo variables

$$Q_{J,d,i} := \begin{cases} Q_{J,1,i} = -\mathbb{1}\{Y_i \in J\} D_i \\ Q_{J,0,i} = \mathbb{1}\{Y_i \in J\}(1 - D_i) \end{cases}$$

for $d \in \{0, 1\}$. The necessary conditions stated in equations (2.3) and (2.4) can be interpreted as learning the sign of the treatment effect of Z on $Q_{J,d}$ conditional on covariates. Note that the assignment of the instrument is now the "treatment". We first use a causal forest (CF) to estimate conditional average treatment effects (CATEs). Second, we grow shallow trees to search for subgroups in the covariate space, in which we observe heterogeneity in the CATEs. These shallow trees summarize the heterogeneity signals in the CATE, and allow for easy implementation and visualization of our local IV test. Third, we discuss a procedure to select the promising subgroups, in which we may exhibit potential violations. Finally, we test whether the group average treatment effects in these promising subgroups are incompatible with the LATE assumptions. In the following, we describe our test procedure in more detail. Additionally, Appendix A gives further information regarding the implementation of our test in R and collects the main steps of our procedure in pseudo code.

3.1 Estimating heterogeneous treatment effects

Let

$$\tau_{J,d}(x) = \mathbb{E}[Q_{J,d,i} | Z_i = 1, X_i = x] - \mathbb{E}[Q_{J,d,i} | Z_i = 0, X_i = x] \tag{3.1}$$

be the CATE of Z on $Q_{J,d}$ at $X = x$. Under Assumptions A1-A3, $\tau_{J,d}(x) \leq 0$ must hold for every combination of J, d and x. Due to our definition of $Q_{J,d}$, positive signs of $\tau_{J,d}(x)$ now point to violations of the LATE assumptions. For all possible combinations of J and d, we regress out the marginal effects that X has on $Q_{J,d}$ and Z using random forests (Breiman, 2001) to account for the potential

confounding, and use the residuals to estimate $\tau_{J,d}(x)$ with causal forests. Wager and Athey (2018) and Athey et al. (2019) derive pointwise asymptotic normality of the causal forest estimator $\widehat{\tau}_{J,d}(x)$ under certain regularity assumptions. The assumptions to establish causality are

Assumption CF1. $(Q^1_{J,d}, Q^0_{J,d}) \perp Z|X$.

Assumption CF2. *For some* $\epsilon > 0$ *it holds* $\epsilon < \mathrm{P}\left(Z = 1|X = x\right) < 1 - \epsilon$.

Note that Assumption CF1 is implicitly part of Assumption A2 as the pseudo outcomes are functions of Y^{dz} and D^z. Therefore, randomization or unconfoundedness – depending on the application we have in mind – is not testable with our procedure. Neither (A1) nor (A3) are, however, necessary to estimate a CF and can thus be tested with our procedure. CF2 assumes overlap, meaning that the instrument must not be deterministic in X. An empirical example in which this assumption may be violated is the twin birth instrument (Rosenzweig and Wolpin, 1980; Angrist and Evans, 1998; Farbmacher et al., 2018). Twinning strongly depends on maternal age (X) and is a very rare (if not even impossible) event if the expectant mother is very young.
To assess the magnitude of the potential violations of the LATE assumptions, we use the augmented inverse probability weighted scores from Robins et al. (1994)

$$\begin{aligned}\widehat{\Gamma}_{J,d,i} := \; & \widehat{\tau}^{(-i)}_{J,d}(X_i) + \\ & \frac{Z_i - \widehat{e}^{(-i)}(X_i)}{\widehat{e}^{(-i)}(X_i)\big(1 - \widehat{e}^{(-i)}(X_i)\big)} \Big(Q_{J,d,i} - \widehat{\mu}^{(-i)}_{J,d}(X_i) - \big(Z_i - \widehat{e}^{(-i)}(X_i)\big)\widehat{\tau}^{(-i)}_{J,d}(X_i)\Big) \;,\end{aligned} \tag{3.2}$$

where $\widehat{\tau}^{(-i)}_{J,d}(X_i)$, $\widehat{e}^{(-i)}(X_i)$ and $\widehat{\mu}^{(-i)}_{J,d}(X_i)$ denote leave-one-out estimates of $\tau_{J,d}(x)$, $e(x) = \mathrm{P}(Z_i = 1|X_i = x)$ and $\mu_{J,d}(x) = \mathbb{E}\big(Q_{J,d,i}\big|X_i = x\big)$, respectively. Leave-one-out (or out-of-bag) estimates are obtained without using the i-th observation. We average $\widehat{\Gamma}_{J,d,i}$ over all observations i that fall into certain subgroups, which we define in a data-driven way as discussed in the next section.

3.2 Detecting and selecting promising subgroups

We use regression trees as a data-driven approach to partition the data along observable covariates. Trees have already been used to perform subgroup analysis

in the context of heterogeneous effects (for example, Su et al., 2009; Athey and Imbens, 2016). We grow a single tree on each score $\widehat{\Gamma}_{J,d}$ using the classification and regression tree (CART) algorithm (Breiman et al., 1984). The CART algorithm is essentially a data mining tool that recursively adds axis-aligned splits to the tree. It will split the sample at the covariate value that delivers the largest heterogeneity between the newly formed subgroups. We denote the resulting tree structure by $\Pi_{J,d}$, which is a collection of terminal and non-terminal nodes – the terminal nodes are also called leaves. The leaves partition the covariate space into a set of rectangles. The CART algorithm is greedy in the sense that it tries to improve the splitting criterion only at the next split, without considering possible future splits. The splitting ends after certain criteria are met. An important parameter here is the user-defined minimum number of observations ultimately required to be in each leaf.

Growing a tree deeply uncovers more heterogeneity and may make it more likely to find violations of the LATE assumptions. A deeper tree, however, also implies smaller sample sizes within the leaves, leading to noisier estimates. A classic solution to solve this bias-variance trade-off is to penalize tree size proportional to a constant, which is determined via K-fold cross-validation. This is called pruning. In the first step, we grow a complete tree without any early stopping criteria other than a minimum leaf size, which can lead to a quite large and complex tree structure. In a second step, we prune this tree using 10-fold cross-validation applying the optimal complexity parameter. Pruning gives us a set of *relevant* subgroups, i.e., groups that potentially exhibit heterogeneity in $\tau_{J,d}(x)$ independent of its sign.

We are, however, particularly interested in finding sign heterogeneity in $\tau_{J,d}(x)$. Under Assumption A1-A3, $\tau_{J,d}(x)$ can vary only between -1 and 0 for any x. However, if the IV is invalid, $\tau_{J,d}(x)$ can vary between -1 and 1. Therefore, local violations of the LATE assumptions may induce observable heterogeneity in the sign of $\tau_{J,d}(x)$. For illustrative purposes, consider testing Assumption A3 separately. This is a special case of our testing procedure, which reduces to finding sign heterogeneity in the first stage effect and testing for the presence of defiers. In this case, we let J cover the whole domain of Y, i.e. $Q_1 = -D$, and Q_0 is redundant. Then, the absolute value of $\tau_1(x) = -E[D^1 - D^0|X = x]$ measures the

conditional average treatment effect of Z on D, which reflects the local fraction of compliers or defiers. Finding positive CATEs would imply that, for some observations, the instrument actually creates a disincentive to take up the treatment, which violates the monotonicity assumption. Note that identifying the LATE may still be possible in such a setting under additional assumptions (de Chaisemartin, 2017).

The following selection procedure aims to find the "most" *promising* subgroups within the relevant ones. We regard a subgroup as promising if we can exhibit potential violations of the LATE assumptions within it, i.e. subgroups in which the CATE is potentially positive. We only use these selected subgroups in the local IV test to increase its power. To ensure the validity of the test, we perform this selection on the training sample. First, we select only the leaves of the pruned trees. If the violations are sufficiently strong, the pruned tree will partition the sample accordingly. Second, we exclude leaves in which the CATE turns out to be clearly negative. Such leaves point to a sizeable fraction of compliers, which makes it hard to detect violations in this subgroup. Finally, we compare each leaf with its left or right pair and use only the leaf side that exhibits a larger average treatment effect. We give additional details about this selection procedure in Appendix A.

3.3 Testing for local violations of IV validity

We use honest estimation to prevent a bias from overfitting (see, for instance, Athey and Imbens, 2016). That is, one randomly chosen part of the sample, called the training sample, is used to build the tree while the remaining sample, called the estimation sample, is used to estimate (group) average treatment effects, which are ultimately used for our local IV test. Instead of training and estimation, we will call the two random halves of the full sample S^A and S^B, and we will swap the roles of the samples to alleviate the inefficiency of the sample splitting (following Chernozhukov et al., 2018). For each combination of J and d, we grow the trees $\Pi_{J,d}^{S^A}$ and $\Pi_{J,d}^{S^B}$, respectively. Due to the sample swapping and as $d \in \{0, 1\}$, we build four times as many pruned trees as we use intervals to discretize Y.

Consider the expectation of $\Gamma_{d,i}$ for a given partition

$$\zeta^A_{J,d,l} = \mathbb{E}\left[\Gamma_{J,d,i} \middle| X_i \in L_l\big(x\,;\,\Pi^{S^B}_{J,d}\big)\right],\ \ \zeta^B_{J,d,l} = \mathbb{E}\left[\Gamma_{J,d,i} \middle| X_i \in L_l\big(x\,;\,\Pi^{S^A}_{J,d}\big)\right],$$

where $L_l\big(x\,;\,\Pi_{J,d}\big)$ denotes the l^{th} element of the set of selected leaves of the tree $\Pi_{J,d}$. The expectations within these partitions are then estimated in sample S^A when the tree has been obtained in sample S^B, and vice versa. In the remainder of the paper, we keep the sample swapping procedure implicit.

We collect the moments of the selected leaves over all combinations of J and d in $\zeta = \big(\zeta^A, \zeta^B\big)$. Positive elements of ζ are local violations of the LATE assumptions. Therefore, we test the following null hypothesis

$$\begin{aligned} H_0 &: \zeta_j \leq 0 \qquad \text{for all} \quad j = 1, \ldots, p \\ H_1 &: \zeta_j > 0 \qquad \text{for some} \quad j = 1, \ldots, p\,, \end{aligned}$$

where $p = |\zeta|$. Rejecting the null means that the LATE assumptions are violated in at least one subpopulation. Finding a violation in any subpopulation casts doubt on the IV validity in the entire population because we cannot rule out further violations in other subpopulations.

For ease of notation, let t_j denote the j-th variable that we use to estimate the moment ζ_j from a total of p moments. Furthermore, let $\widehat{\zeta}_j = \frac{1}{n_j}\sum_i t_{ij}$ denote the sample mean of the j-th variable and $\widehat{\sigma}^2_j = \frac{1}{n_j - 1}\sum_i \big(t_{ij} - \widehat{\zeta}_j\big)^2$ its sample variance with n_j the sample size within leaf j. We consider the test statistic

$$T = \max_{1 \leq j \leq p} \frac{\sqrt{n_j}\,\widehat{\zeta}_j}{\widehat{\sigma}_j}\,. \tag{3.3}$$

Under the H_0 it must hold that

$$T \leq \max_{1 \leq j \leq p} \frac{\sqrt{n_j}\,\big(\widehat{\zeta}_j - \zeta_j\big)}{\widehat{\sigma}_j},$$

hence, finding an upper bound for the $(1-\alpha)$ quantile of $\sqrt{n_j}(\widehat{\zeta}_j - \zeta_j)/\widehat{\sigma}_j$ is sufficient to keep the actual size of the test at or below α. Using a Bonferroni correction for

multiple testing, a critical value for T is

$$c(\alpha) = \Phi^{-1}(1 - \alpha/p) \ . \tag{3.4}$$

If one is interested not only in the global null hypothesis, but additionally in obtaining the subgroups that violate the LATE assumptions, the Bonferroni-Holm correction is more powerful.
To obtain our asymptotic result, we will rely on recent findings from the double machine learning framework. Observe that, as described in Athey and Wager (2017), the estimator from (3.2) can be interpreted as

$$\begin{aligned} \widehat{\Gamma}_{J,d,i} := \ & \widehat{m}_{J,d}^{(-k)}(X_i, 1) - \widehat{m}_{J,d}^{(-k)}(X_i, 0) + \\ & \frac{Z_i - \widehat{e}^{(-k)}(X_i)}{\widehat{e}^{(-k)}(X_i)\big(1 - \widehat{e}^{(-k)}(X_i)\big)} \Big(Q_{J,d,i} - \widehat{m}_{J,d}^{(-k)}(X_i, Z_i)\Big) , \\ e(x) = P(Z_i = 1 | X_i = x), \ & m_{J,d}(x, z) = \mathbb{E}\left[Q_{J,d} | Z = z, X_i = x\right] , \end{aligned} \tag{3.5}$$

which aligns with the double machine learning framework introduced by Chernozhukov et al. (2018). Here $m_{J,d}(\cdot)$ and $e(\cdot)$ are unknown nuisance functions. In their theorem 5.1, they establish that the mean of the estimates $\widehat{\Gamma}_{J,d,i}$ is asymptotically Gaussian and efficient, as long as the nuisance estimates $\widehat{m}_{J,d}^{(-k)}(\cdot)$ and $\widehat{e}^{(-k)}(\cdot)$ converge sufficiently fast and are determined by crossfitting, which we indicate by $^{(-k)}$ here. Adapted to our testing problem, the following assumptions are needed to derive asymptotic results.

Assumption T1. *For all J, d and every selected leaf l, it holds*

$$P\left(X \in L_l\big(x\,;\,\Pi_{J,d}\big)\right) \geq c > 0 \quad \text{and}$$

$$\mathbb{E}\left[\left(Q_{J,d} - m_{J,d}(X, Z)\right)^2 | X \in L_l\big(x\,;\,\Pi_{J,d}\big)\right] \geq c > 0.$$

where c is a generic constant independent of n.

Assumption T2. *The nuisance functions are estimated via K-fold crossfitting, and with probability no less than $1 - o(1)$ it holds*

1. $\|\widehat{e}^{(-k)}(X) - \frac{1}{2}\|_{P,\infty} \leq \frac{1}{2} - \epsilon$,

2. $\max\Big(\|\widehat{e}^{(-k)}(X) - e(X)\|_{P,2}, \max_{J,d} \|\widehat{m}_{J,d}^{(-k)}(X,Z) - m_{J,d}(X,Z)\|_{P,2}\Big) \leq \delta_n n^{-1/4}$, *where* $\delta_n = o(\log^{-1}(n))$.

Assumption T1 ensures that the number of observations in the subgroups defined by the selected leaves of the pruned trees increases with the sample size. Moreover, it assumes that the pseudo outcome variable is non-deterministic in each leaf given X and Z. The first part of Assumption T2 assumes that the estimates of the propensity score are bounded away from zero and one, which is a standard assumption in the literature. The second part of Assumption T2 states that, with probability converging to one, all nuisance components are $o(n^{-\frac{1}{4}})$ consistent with respect to the L_2-error. This rate is much weaker than the standard rate of $o(n^{-\frac{1}{2}})$ due to the so-called Neyman orthogonality or double robustness property of the estimator. In principle, other estimators than random forests (e.g., from Kernel regressions) can be used for the nuisance functions, as long as they fulfill Assumption T2. Moreover, Assumption T2 can be weakened at the expense of a more complicated notation. For a detailed discussion of the convergence rates and sharpness of the conditions, see Chernozhukov et al. (2018), Chernozhukov et al. (2016) and Athey and Wager (2017).
The following proposition shows that the probability of rejecting H_0 – although being true – does not exceed α asymptotically when we use $c(\alpha)$ as the critical value.

Proposition 1. *Suppose that Assumptions CF1, CF2, T1 and T2 hold, then we have under the* H_0

$$P\big(T > c(\alpha)\big) \leq \alpha + o(1).$$

Proof. We directly obtain Proposition 1 by relying on theorem 5.1 from Chernozhukov et al. (2018). The corresponding conditions stated in their assumption 5.1 have to be satisfied for each subgroup. Our Assumption CF1 implies their assumption 5.1(a). The first part of our Assumption T1 ensures that the number of observations within all subgroups is $O(n)$, with probability converging to one. Further, the second part of T1 is equivalent to their assumption 5.1(d). Assumptions 5.1(b) and 5.1(e) hold because $|Q_{J,d}| \leq 1$ and 5.1(c) is implied by CF2.

Lastly, (i) of 5.1(f) is directly implied by $|Q_{J,d}| \leq 1$ and our Assumption T2. The proposition then follows by the union bound. □

We derive the subgroups from the leaves of the pruned regression trees after applying our selection procedure described in Section 3.2. If there is no heterogeneity in the CATE, the pruned trees will not split the sample, or splits will occur due to noise only. If the sign of $\tau_d(x)$ is negative everywhere but its magnitude is heterogeneous over covariates, then splits may occur even under H_0 but only a few (or even none) of them will turn out to be "promising". In case the set of hypotheses turns out to be empty (due to a very strong instrument, for example), we report a test based on the root nodes of the pruned trees. Note that the number of subgroups that can be tested is bounded by Assumption T1. Relying on results of Chernozhukov et al. (2019) and Belloni et al. (2018) about multiple testing in a high-dimensional setting, it might be possible to allow for a number of leaves that increases slowly with sample size.

The pointwise normality of the causal forest estimates could also be used to test for violations of the LATE assumptions at prespecified points in the covariate space. Alternatively, one could rely on uniform confidence bands for $\tau_d(x)$. There exists a vast literature on confidence bands for nonparametric functions (among others, Bickel and Rosenblatt, 1973; Konakov and Piterbarg, 1984; Li et al., 1989; Hall et al., 2013) mostly building on kernel or local polynomial methods. Due to the curse of dimensionality, the performance is reliable up to at most $p = 3$. Since the number of covariates used to ensure unconfoundedness is usually not that small, most of these methods are not applicable in economics. Lee et al. (2017) develop uniform confidence bands for the average treatment effect conditional on a small (≤ 3) subset of covariates, implying that one would still need to select the promising covariates beforehand. Additionally, they assume parametric specifications for the propensity score and for the remaining covariates to avoid the curse of dimensionality. By using random forests, we can avoid such strong structural assumptions. As a result, in our procedure, the number of covariates can be relatively large although not high-dimensional. To use random forests in high-dimensions, modifications to the algorithm and an assumption of sparsity are needed (Wager and Walther, 2015).

4 Simulations

To test the finite sample performance of our new procedure, we run several Monte Carlo simulations. Process 1 simulates a randomized experiment similar to that of Huber and Mellace (2015). While in process 1 Assumption A2.a holds, process 2 simulates a setting in which the instrument is unconfounded (i.e. Assumption A2.b holds). Process 2 uses an easy propensity score of the instrument Z and strong confounding of D. In both processes, we use the same function for Y, which also depends on the covariates X:

$$Y = D + \gamma_x Z + \sum_{k=1}^{3} 0.3 X_k + u$$

$$\text{where } X_k \sim N(0,1) \text{ and } (u, v) \sim N\left(0, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right)$$

Process 1 (randomized):	Process 2 (unconfounded):
$D = \mathbb{1}\{\alpha_x Z + v\}$	$D = \mathbb{1}\{b + \alpha_x Z + v\}$
$Z \sim$ Bernoulli(0.5)	$Z = \mathbb{1}\left\{\sum_{k=1}^{3} 0.2 X_k + w\right\}$
	where $b = 0.5 \log(1 + e^{X_1 + X_2 + X_3})$ and $w \sim N(0,1)$

For both processes we use different values of γ_x and α_x :

- DGP0 (*exogenous but uninformative IV*): $\gamma_x = 0$, $\alpha_x = 0$
- DGP1 (*exogenous and relevant IV*): $\gamma_x = 0$, $\alpha_x = 0.20$
- DGP2 (*local violation of monotonicity, defiers exist in subpopulation*):

$$\gamma_x = 0 \ , \ \alpha_x = \begin{cases} -0.75 & \text{if } X_1 < \Phi^{-1}(0.15) \\ 0.40 & \text{otherwise} \end{cases}$$

- DGP3 (*local violation of exclusion restriction*):

$$\gamma_x = \begin{cases} 1.25 & \text{if } X_2 < \Phi^{-1}(0.15) \\ 0 & \text{otherwise} \end{cases} \ , \ \alpha_x = 0.20$$

- DGP4 (*global violation of exclusion restriction*): $\gamma_x = 0.5$, $\alpha_x = 0.20$
- DGP5 (*global violation of exclusion restriction but with sign heterogeneity*):

$$\gamma_x = \begin{cases} 0.50 & \text{if } X_3 < \Phi^{-1}(0.5) \\ -0.50 & \text{otherwise} \end{cases} , \ \alpha_x = 0.20$$

DGP0 allows us to verify the control of the nominal test size. There are no compliers in this setting and, therefore, the moment inequalities are binding (i.e., $\zeta_j = 0$ for all $j = 1, \ldots, p$). DGP1 represents the case in which the instrument is not only exogenous but also relevant. In this case, the values of ζ_j are all supposed to be negative. The stronger the instrument is, the more conservative the test will be. In DGP1, the choice of $\alpha_x = 0.2$ leads to a complier share of roughly 8%. DGP2 models a local violation of monotonicity. The local share of compliers or defiers is given by $\Phi(|\alpha_x|) - \Phi(0)$, which multiplied by the size of the subpopulation gives the average share of compliers or defiers in the population. DGP2 leads to an average share of defiers in the population of roughly 4.1%, which are hidden in the covariate space. The average complier share is more than three times as large (13.2%), which makes it hard to refute the LATE assumptions with tests based on the entire sample. DGP3 corresponds to a *local* violation of the exclusion restriction, while DGP4 and DGP5 *globally* violate it.

Table 1 shows rejection frequencies for both processes with 1,000 replications. The sample size is 3,000. We discretize Y into four intervals using an equidistant grid from the minimum to the maximum value of Y. We compare our procedure with Kitagawa (2015)'s test, Mourifié and Wan (2017)'s test and Huber and Mellace (2015)'s mean and full independence test. We find for both processes that our procedure does not overreject the null hypothesis of no violation of the LATE assumptions at the 5% nominal level (DGP0). Note that the rejection frequency in DGP0 under process 2 (i.e., the instrument is only valid conditional on X) would be 0.296 if we did *not* regress out the effects that X has on $Q_{J,d}$ and Z. This is, therefore, a crucial step if the instrument is confounded. As expected, the test procedure is conservative if compliers exist and there are no violations of the LATE assumptions (DGP1) – this is in line with the results from Kitagawa (2015)'s test, Mourifié and Wan (2017)'s test and Huber and Mellace (2015)'s test.

Under DGP2, the monotonicity assumption is violated only in a certain area of the covariate space. Consequently, we expect that splitting the sample by covariates leads to a strong improvement in test power. Indeed, our procedure has distinctly larger power than the alternative tests in this setting. In case of DGP3, a local violation of the exclusion restriction, our procedure again clearly outperforms the alternatives. When the violation of the exclusion restriction is global (DGP4), the existing tests perform better than our procedure. This is to be expected, because splitting the data by covariates cannot improve the precision of our test but honest splitting of the sample leads to a lower sample size available for estimation. However, if the global violation of the exclusion restriction actually has different signs in local areas of the covariate space, then our test can again perform better than existing ones. In DGP5 we illustrate this by the extreme case of completely opposite direct effects.

– Table 1 about here –

In Appendix B, we show further simulation results. In Table B.1 we briefly examine the size and power of the tests for three numbers of intervals: 2, 4 and 6. It shows rejection frequencies for all DGPs except for DGP1. In all three specifications, we see that our test controls the size under the null hypothesis. The power tends to decrease with the number of intervals, which may reflect the fact that the effective sample size within an interval decreases when we use more intervals. Moreover, we shed some light on the effectiveness of the selection step in Table B.2. The average number of promising subgroups we obtain after the selection is reduced by up to 45% compared to using all leaves. This reduction is reflected in a rise of the rejection frequencies (particulary, for DGP2, 3 and 5, in which we have a local component).

5 Applications

In this section, we apply our procedure to two widely used instrumental variables, namely parental preferences for mixed-sex composition of children (Angrist and Evans, 1998) and the Vietnam draft lottery (Angrist, 1990). The Vietnam draft lottery is an IV that is randomized by design and, therefore, Assumption A2.a

holds. There are several studies that use the draft lottery to estimate the causal effect of Vietnam-era military service on civilian earnings, schooling, disability status, or health later in life (see, for example, Angrist, 1990; Angrist and Chen, 2011; Angrist et al., 2011, 2010). We are interested in the literature investigating the effect of military service on schooling. Angrist and Chen (2011) argue that schooling gains can be attributed to the use of the GI Bill, which made generous schooling benefits available to veterans. This is an important channel of a causal effect of military service on schooling. A potential direct effect of the lottery on schooling, however, arises due to deferments, which, among other reasons, were issued to men who were still attending school. Card and Lemieux (2001) show that draft avoidance led to a rise in the college enrollment rates of young men. This additional schooling might lead to a violation of the exclusion restriction.
The second application builds on the observation that some parents prefer a mixed-sex composition of their children. Angrist and Evans (1998) propose using the occurrence of same-sex siblings as an IV for the number of children. Rosenzweig and Wolpin (2000) discuss several reasons why this IV may be invalid. Huber (2015) was the first to test the validity of the LATE assumptions in this setting. He finds no violation in the full sample and very few violations across 22 arbitrarily chosen subgroups and concludes that the IV's validity in these data cannot be refuted. Our test procedure can be seen as a flexible extension of table 1 in Huber (2015) or of section VII in Bisbee et al. (2017), in which we derive the promising subgroups in a data-driven way.

5.1 Results with original outcomes

In the first application, we use data from an extract of the 1979 and 1981-85 March Current Population Survey (CPS) (see Angrist and Krueger, 1992, for more details about the data). We consider the 1970 draft lottery, held on December 1969, which affected men born in the period 1944-1950. Since the rate of conscriptions dropped considerably after June 1971, most men who obtained deferments in 1970 (for instance, due to additional schooling) could permanently avoid military service (Card and Lemieux, 2001). The Vietnam draft lottery randomly assigned a number from 1 to 366 to men born in this cohort based on their date of birth. The numbers

determined the order of the call for conscription, starting with the smallest number. At some point during the year, the Selective Service announced a maximum lottery number that would be called. For instance, the ceiling was 195 in the 1970 lottery. In this application, Y measures education (no college, some college, college), D is veteran status and Z indicates whether the individual's date of birth led to a lottery number lower than 200, which considerably increased the risk of conscription. We use age, indicators for ethnicity, region dummies, and the year of the survey as covariates. Table 2 shows the test results for this application. All tests (incl. our test) do not reject the null hypothesis.

The sample in the second application consists of married and unmarried mothers aged 21 to 35 years with at least two children as recorded in the 1980 US census. In this application, Y is the logarithm of annual labor income, Z indicates a same-sex composition of the first two children and D is an indicator of having more than two children at the time of the census. We use mother's age in 1980, age at first birth, educational attainment (three levels) and ethnicity as covariates. The upper panel of Table 3 shows the test results for this application. All tests (incl. our test) do not reject the null hypothesis.

– Tables 2 and 3 about here –

5.2 Results with synthetic outcomes

As reported in the previous section, we do not find local violations of the LATE assumptions using the original outcome Y. To further illustrate the performance of our test procedure in a real application, we add a local direct effect of Z to Y. The direct effect is rather small (1/4 of the SD of Y) and applies only to a small subgroup of mothers in the 1980 US census data (3.4% of the sample, i.e. 7,513 observations), which makes it hard to find in the full sample,

$$Y_{\text{syn}} = Y + 1/4 \; \widehat{\sigma}_Y \, Z \, \mathbb{1}\{\text{age} < 25\} \mathbb{1}\{\text{educ} = \text{HS}\} \, .$$

The mean of the outcome variable is essentially unchanged ($\bar{Y} = 8.939$ and $\bar{Y}_{\text{syn}} = 8.944$) since the manipulated subgroup is rather small. The mean in the manipulated subgroup rises from 8.53 to 8.69. Since we choose where the direct

effect is located in the covariate space, we can check whether different tests in fact succeed in recovering it.

The lower panel of Table 3 shows the test results for the synthetic outcome. Our test clearly rejects the null hypothesis while the p-values of the existing tests are unchanged compared to the results in the upper panel. A traditional subgroup analysis would clearly help them to reject the null hypothesis but cannot be performed since the location of the violation is oracle knowledge. Our data-driven subgroup testing, however, successfully finds the synthetic violation as illustrated in Figure 1, which shows the pruned regression tree leading to the maximum t-statistic.

– Figure 1 about here –

6 Conclusion

Using instrumental variables to identify local average treatment effects is common in empirical research. In studies that do so, however, the validity of instruments is a point of debate. Fortunately, the LATE framework generates empirically testable implications for instrument validity. In this paper, we propose a machine learning based approach to perform IV validity tests in a data-driven way. In the spirit of subgroup testing, our procedure uses the CART algorithm to split the sample along covariate values. Some of these splits form promising subgroups, which can be used to test for local violations of IV validity. We use causal forests to estimate the magnitude of the violations in these subgroups. Our approach can be easily implemented using existing software packages. We provide an R package (`LATEtest`) and apply our test to two widely used instrumental variables, namely parental preferences for mixed-sex composition of children and the Vietnam draft lottery. In line with the previous literature, we do not find violations of the LATE assumptions in either application.

Our procedure is subject to some restrictions, which offer promising avenues for future research. First, it requires the presence of covariates that are unaffected by the treatment and the instrumental variable. The test results are sensitive to the set of covariates we have available. Second, we follow the literature and

use an equidistant grid to discretize the outcome variable into a finite number of arbitrarily chosen sets. This could lead to some estimated negative densities not being detected because they can average out with nearby positive densities of compliers in the same set. The literature about scan statistics may make our and the existing tests more powerful by providing a way to find these intervals in a data-driven manner as well (see, for example, Walther, 2010). Third, we need to conduct honest splitting to avoid bias from adaptive searching for violations: one half of the sample is used to build pruned trees while the other half is used to estimate the magnitude of the violations. Although we switch the roles of the samples, doing so reduces the number of observations we can use for testing. Fourth, while detecting promising subgroups with pruned trees allows us to interpret and visualize the source of violations easily, it may not be necessary. A particularly interesting topic of future research will be to test whether the estimates of a causal forest are positive at any point of its support.

Acknowledgments

Helpful comments were provided by Martin Huber, Heinrich Kögel, Romuald Méango, Giovanni Mellace, Martin Spindler, Frank Windmeijer, and seminar participants at Konstanz, London, Odense, the IAAE annual meeting Montréal, the Bank of England conference on modelling with big data and machine learning, and the workshop on causal machine learning in St. Gallen. We thank Christina Nießl for excellent research assistance. Raphael Guber acknowledges funding through the International Doctoral Program "Evidence-Based Economics" of the Elite network of Bavaria.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–263.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review 80*(3), 313–336.

Angrist, J. D. and S. H. Chen (2011). Schooling and the Vietnam-era GI Bill: Evidence from the draft lottery. *American Economic Journal: Applied Economics 3*(2), 96–118.

Angrist, J. D., S. H. Chen, and B. R. Frandsen (2010). Did Vietnam veterans get sicker in the 1990s? The complicated effects of military service on self-reported health. *Journal of Public Economics 94*(11-12), 824–837.

Angrist, J. D., S. H. Chen, and J. Song (2011). Long-term consequences of Vietnam-era conscription: New estimates using social security data. *American Economic Review 101*(3), 334–38.

Angrist, J. D. and W. N. Evans (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450–477.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and A. B. Krueger (1992). Estimating the payoff to schooling using the Vietnam-era draft lottery. NBER Working Paper No. 4067.

Asher, S., D. Nekipelov, P. Novosad, and S. P. Ryan (2016). Classification trees for heterogeneous moment-based models. NBER Working Paper No. 22976.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society (Series B) 80*(4), 597–623.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Annals of Statistics 47*(2), 1148–1178.

Athey, S. and S. Wager (2017). Efficient policy learning. *arXiv preprint arXiv:1702.02896*.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized GMM. *arXiv preprint arXiv:1806.01888*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Bickel, P. and M. Rosenblatt (1973). Two-dimensional random fields. In *Multivariate Analysis–III*, pp. 3–15. Elsevier.

Bisbee, J., R. Dehejia, C. Pop-Eleches, and C. Samii (2017). Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect. *Journal of Labor Economics 35*(S1), S99–S147.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences 113*(27), 7383–7390.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*(430), 443–450.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and regression trees*. CRC press.

Card, D. and T. Lemieux (2001). Going to college to avoid the draft: The unintended legacy of the Vietnam war. *American Economic Review 91*(2), 97–102.

Chen, L.-Y. and J. Szroeter (2014). Testing multiple inequality hypotheses: A smoothed indicator approach. *Journal of Econometrics 178*, 678–693.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., D. Chetverikov, and K. Kato (2019). Inference on causal and structural parameters using many moment inequalities. *Review of Economic Studies 86*(5), 1867–1900.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. NBER Working Paper No. 24678.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review 105*(5), 486–490.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: Estimation and inference. *Econometrica 81*(2), 667–737.

de Chaisemartin, C. (2017). Tolerating defiance? Local average treatment effects without monotonicity. *Quantitative Economics 8*(2), 367–396.

Farbmacher, H., R. Guber, and J. Vikström (2018). Increasing the credibility of the twin birth instrument. *Journal of Applied Econometrics 33*(3), 457–472.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012). The Oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics 127*(3), 1057–1106.

Hall, P., J. Horowitz, et al. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *Annals of Statistics 41*(4), 1892–1921.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Huber, M. (2015). Testing the validity of the sibling sex ratio instrument. *Labour 29*(1), 1–14.

Huber, M. and G. Mellace (2015). Testing instrument validity for LATE identification based on inequality moment constraints. *Review of Economics and Statistics 97*(2), 398–411.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (1997). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies 64*(4), 555–574.

Kang, H., A. Zhang, T. T. Cai, and D. S. Small (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association 111*(513), 132–144.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica 83*(5), 2043–2063.

Knaus, M. C., M. Lechner, and A. Strittmatter (2020). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, forthcoming.

Konakov, V. and V. Piterbarg (1984). On the convergence rate of maximal deviation distribution for kernel regression estimates. *Journal of Multivariate Analysis 15*(3), 279–294.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*(10), 4156–4165.

Laffers, L. and G. Mellace (2017). A note on testing instrument validity for the identification of LATE. *Empirical Economics 53*(3), 1281–1286.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*(7), 1207–1225.

Li, K.-C. et al. (1989). Honest confidence regions for nonparametric regression. *Annals of Statistics 17*(3), 1001–1008.

Mourifié, I. and Y. Wan (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics 99*(2), 305–313.

Nie, X. and S. Wager (2019). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Rosenzweig, M. R. and K. I. Wolpin (1980). Life-cycle labor supply and fertility: Causal inferences from household models. *Journal of Political Economy 88*(2), 328–348.

Rosenzweig, M. R. and K. I. Wolpin (2000). Natural "natural experiments" in economics. *Journal of Economic Literature 38*(4), 827–874.

Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research 10*(Feb), 141–158.

Tian, L., A. A. Alizadeh, A. J. Gentles, and R. Tibshirani (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association 109*(508), 1517–1532.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B) 58*(1), 267–288.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences 113*(45), 12673–12678.

Wager, S. and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Annals of Statistics 38*(2), 1010–1033.

Windmeijer, F., H. Farbmacher, N. Davies, and G. D. Smith (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association 114*(527), 1339–1350.

Tables and Figures

Table 1: Simulation results

	FGK	K	MW	HM
Process 1				
DGP0 (exog but uninformative)	0.036	0.060	0.060	–
DGP1 (exog and relevant)	0.009	0.005	0.007	0.000
DGP2 (local defiers)	0.749	0.005	0.006	0.000
DGP3 (local direct effects)	0.743	0.057	0.066	0.002
DGP4 (global direct effects)	0.907	0.980	0.923	0.980
DGP5 (global direct effects, varying signs)	0.593	0.006	0.003	0.000
Process 2				
DGP0 (exog but uninformative)	0.036			
DGP1 (exog and relevant)	0.002			
DGP2 (local defiers)	0.658			
DGP3 (local direct effects)	0.767			
DGP4 (global direct effects)	0.972			
DGP5 (global direct effects, varying signs)	0.716			

Table displays rejection frequencies (based on 1000 replications) for significance level of 5% and $n = 3000$. FGK refers to our test, K to Kitagawa (2015)'s test, MW to Mourifié and Wan (2017), and HM to Huber and Mellace (2015)'s test based on full independence using Chen and Szroeter (2014)'s smooth indicator method with tuning parameters as defined in section 3 of Huber and Mellace (2015). Huber and Mellace (2015)'s test based on mean independence performs worse in our setting with respect to size and power and is, therefore, not reported. For the FGK test we regress out the marginal effects that X has on $Q_{J,d}$ and Z using random forests.

Table 2: Test results of the validity tests for the Vietnam draft IV

	p	T	$c(\alpha)$	p-value
FGK	6	1.49	2.39	0.412
Kitagawa (2015) test				1.000
H&M (2015) full independence				1.000
H&M (2015) mean independence				1.000

Data extract from 1979 and 1981-85 March CPS. Sample size is 20,606.

Table 3: Test results of the validity tests for the sibling sex IV

	p	T	$c(\alpha)$	p-value
Original Outcome				
FGK	20	2.28	2.81	0.228
Kitagawa (2015) test				1.000
H&M (2015) full independence				0.999
H&M (2015) mean independence				1.000
Synthetic Outcome				
FGK	24	9.08	2.87	0.000
Kitagawa (2015) test				1.000
H&M (2015) full independence				0.999
H&M (2015) mean independence				1.000

Data from 1980 US census. Sample size is 220,502. Mourifié and Wan (2017) test is based on a random subsample of 26,342 observations due to computational constraints. It does not reject at the 10% level for both outcomes. We discretize Y into four intervals using an equidistant grid from the minimum to the maximum value of Y. To allow for a faster computation of our test, we use a subsample fraction equal to 0.05 when we grow the random forests (for the orthogonalization) and the causal forests. Moreover, we set the minbucket parameter of `rpart` to 1600 when we build the pruned trees.

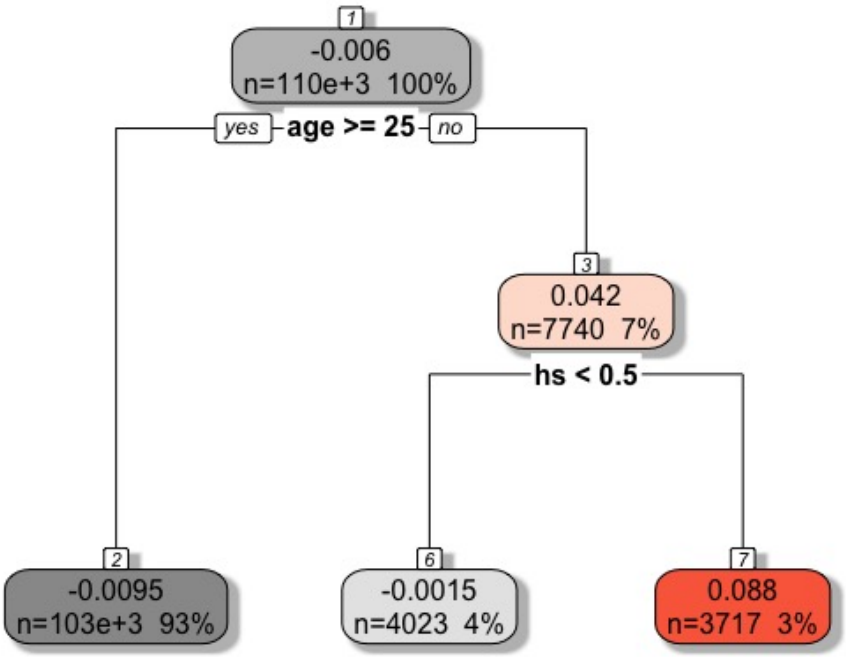

Figure 1: Data from 1980 US census (synthetic outcome). Pruned regression tree leading to the maximum t-statistic. The first value in every leaf indicates the effect heterogeneity in the training sample. Positive values indicate violations of the null hypothesis, which still need to be confirmed in the estimation sample. The second line shows the absolute and relative size of each leaf. The text beneath the leaf shows the variable and value on which the leaf was split next.

A Details about the R package `LATEtest`

In this appendix, we provide further information about the implentation of our testing procedure and provide pseudo code of the R package `LATEtest` (available at `https://github.com/farbmacher/LATEtest`). We employ `regression_forest` from the `grf` package to estimate $e(x)$ and $\mu_{J,d}(x)$. To account for potential confounding, we construct residuals using the out-of-bag predictions, which are computationally more convenient to obtain than the K-fold predicted values. We employ `grf` (1.1.0) (available at `https://github.com/grf-labs/grf`) using these residuals to obtain the CF estimates $\widehat{\tau}_{J,d}(x)$. To partition the data, we grow a single tree on each score $\widehat{\Gamma}_{J,d}$ using the CART algorithm. For this task, we use `rpart` (4.1-15) (available at `https://github.com/bethatkinson/rpart`). In our package we set the minbucket parameter of `rpart`, which defines the minimum number of observations ultimately required to be in each leaf, to 200 by default. Note that different procedures can be used to partition the data. One could, alternatively, analyze the trees of the causal forest to find the "best" tree (for an implementation, see `https://gist.github.com/msegar/c15af2bfc067e5319eecfd67d785fde5`). Moreover, following Chernozhukov et al. (2018) we could group based upon the predicted CATEs of the CF.

The *algorithm* on the next page shows the main steps of our proposed test procedure in pseudo code. Note that the *selection step* in line 14 can be extended. For instance, we additionally compare each leaf with its left or right pair and use only the leaves that have a larger t-statistic in the training data than its pair. To incorporate the variance of the estimates, we rely on t-statistics for the selection of the promising subgroups. We do not intend to perform any testing in this step but only use the selection as an explorative tool to find promising leaves.

Algorithm 1 LATEtest

Input: n training examples of the form (Y_i, D_i, Z_i, X_i), where Y_i is the response, D_i the treatment indicator, Z_i the instrument indicator and X_i are the covariates. A minimum leaf size k, a significance level α and the number of subsets to form J.

1: Split the sample into two subsamples of equal size A and B.
2: On both samples separately train a regression forest with Z_i as outcome and X_i as covariates to obtain the leave-one-out estimates $\widehat{e}^{(-i)}(X_i)$.
3: **for** each J **do**
4: **for** $d = 0, 1$ **do**
5: Construct the pseudo outcome variables $Q_{J,d,i}$.
6: **for** both samples separately **do**
7: Train a regression forest with $Q_{J,d,i}$ as outcome and X_i as covariates to obtain the leave-one-out estimates $\widehat{\mu}^{(-i)}_{J,d}(X_i)$.
8: Train a causal forest with $Q_{J,d,i} - \widehat{\mu}^{(-i)}_{J,d}(X_i)$ as outcome, $Z_i - \widehat{e}^{(-i)}(X_i)$ as treatment and X_i as covariates to obtain the leave-one-out estimates $\widehat{\tau}^{(-i)}_{J,d}(X_i)$.
9: Construct the estimates $\widehat{\Gamma}_{J,d,i}$ as defined in (3.2).
10: **end for**
11: Use sample A to fit a CART tree with $\widehat{\Gamma}_{J,d,i}$ as outcome, X_i as covariates, minimal leaf size k and apply cost complexity pruning.
12: **for** each leaf $l = 1, \ldots, l_{\max}$ **do**
13: Calculate the t-statistic $t^{(A)}_{J,d,l}$ over observations $\widehat{\Gamma}_{J,d,i}$ in sample A contained in leaf l.
14: **if** $t^{(A)}_{J,d,l} > -\Phi^{-1}(1 - 0.05/l_{\max})$ **then**
15: Calculate the t-statistic $t^{(B)}_{J,d,l}$ over observations $\widehat{\Gamma}_{J,d,i}$ in sample B contained in leaf l and store the values in a vector T_{vec}.
16: **end if**
17: **end for**
18: Repeat lines 11-17 with swapped roles for samples A and B.
19: **end for**
20: **end for**
21: **if** $\max(T_{vec}) > \Phi^{-1}(1 - \alpha/|T_{vec}|)$ **then**
22: Reject the null hypothesis.
23: **end if**

B Additional simulation results

Table B.1: Discretizing Y

Number of intervals	2	4	6
DGP0 (exog but uninformative):			
FGK	0.043	0.036	0.035
K	0.044	0.060	0.068
DGP2 (local defiers):			
FGK	0.896	0.749	0.503
K	0.000	0.005	0.009
HM	0.000	0.002	0.000
DGP3 (local direct effects):			
FGK	0.855	0.743	0.518
K	0.013	0.057	0.050
HM	0.096	0.002	0.000
DGP4 (global direct effects):			
FGK	0.978	0.907	0.821
K	0.940	0.980	0.987
HM	1.000	0.980	0.937
DGP5 (global direct effects, varying signs):			
FGK	0.543	0.593	0.451
K	0.000	0.006	0.002
HM	0.000	0.000	0.000

Table displays rejection frequencies (based on 1000 replications) for significance level of 5%, $n = 3000$ and process 1. FGK refers to our test, K to Kitagawa (2015)'s test, and HM to Huber and Mellace (2015)'s test based on full independence using Chen and Szroeter (2014)'s smooth indicator method. Huber and Mellace (2015)'s test based on mean independence performs worse in our setting with respect to size and power and is, therefore, not reported. We discretize Y using an equidistant grid of intervals from the minimum to the maximum value of Y (number of intervals vary by columns).

Table B.2: Effectiveness of the selection step

	before		after	
	mean #	RF	mean #	RF
DGP0 (exog but uninformative)	18.81	0.033	15.92	0.036
DGP1 (exog and relevant)	18.81	0.001	12.81	0.002
DGP2 (local defiers)	23.63	0.605	15.85	0.658
DGP3 (local direct effects)	24.20	0.712	13.73	0.767
DGP4 (global direct effects)	20.63	0.976	11.06	0.972
DGP5 (global direct effects, varying signs)	26.22	0.643	15.30	0.716

Table displays mean number of hypotheses obtained from pruned trees and rejection frequencies (based on 1000 replications) for significance level of 5%, $n = 3000$ and process 2. We discretize Y using an equidistant grid of four intervals from the minimum to the maximum value of Y, and regress out the marginal effects that X has on $Q_{J,d}$ and Z using random forests.