Instrument Validity Tests with Causal Trees:

With an Application to the Same-sex Instrument

Preliminary Version

Raphael Guber*

University of Munich & Munich Center for the Economics of Aging

January 2, 2018

Abstract

Tests to refute assumptions necessary for the identification of local average treatment effects (LATE) exist but may have low power in practice. In this paper, we employ recently developed machine learning tools as data-driven improvements for these tests. Specifically, we use the causal tree (CT) algorithm of Athey and Imbens (2016) to directly search the covariate space for violations of the LATE assumptions. The new approach is applied to the sibling sex composition instrument in China and the US. We find that the instrument violates at least one of the LATE assumption in China but not the US.

Keywords: Instrument validity, recursive partitioning, machine learning

*guber@mea.mpisoc.mpg.de.

1 Introduction

Empirical research that tries to credibly estimate causal effects of treatments on outcomes relies heavily on instrumental variables (IVs) (Angrist and Pischke, 2010). Since it is fundamentally impossible to directly test for the validity of IVs, discussions on threats to the key identifying assumptions and robustness checks on the latter constitute a crucial part of any empirical paper. Quantitative evidence in the form of statistical tests would be helpful in such debates. Recently, Kitagawa (2015), Huber and Mellace (2015b) and Mourifié and Wan (2017) derived such formal tests for empirically refutable implications that are generated by the joint assumptions necessary for the identification of local average treatment effects (LATE) by Imbens and Angrist (1994) and Angrist et al. (1996).[1] The unifying idea across these three studies is that the estimated share of compliers, the subpopulation of individuals whose treatment status is causally affected by the instrument, must be non-negative at any point of the distribution of the observed potential outcome variables (we discuss differences between these papers later). In practice, the tests can have low power even if one or more LATE assumptions are severely violated, because violations cancel out with non-violations.[2] The power

[1] Related studies include Kédagni and Mourifie (2015, 2016), Slichter (2015) and Machado et al. (2013).

[2] To give an example, consider a binary outcome Y, treatment D and instrument Z. If all LATE assumptions hold we must empirically observe that $P(Y=1, D=1|Z=1) - P(Y=1, D=1|Z=0) \geq 0$ when Z has a non-negative effect on treatment take-up. Say Z is also random but has a positive direct effect on the outcome. Then $P(Y=1, D=1|Z=1) \geq P(Y=1, D=1|Z=0)$ in any case and the violation cannot be detected. If the direct effect is negative,

of the tests decreases with the share of compliers, holding other features of the data generating process fixed. However, an instrument with a low compliance rate might produce an uninteresting LATE and estimation may suffer from weak IV bias. Violations may be more easily detectable within subpopulations defined by covariate values. For example, violations of the exclusion restriction might be stronger in certain populations, or the share of compliers in a population might be close to zero, such that violations are easier to detect.

In this paper, we propose to utilize covariates unaffected by the treatment and the instrument to substantially increase the power of these tests in practice. More specifically, we use the causal trees (CT) algorithm developed by Athey and Imbens (2016), AI hereafter, to recursively partition the sample along covariates with the aim to detect subpopulations in the sample where the LATE assumptions are violated.[3] When the outcome variable has finite discrete support, we formulate the target criterion such that the algorithm directly searches the covariate space for violations. In this case, a tree for every point in the observed support is grown. If the outcome is continuous, we propose two other approaches based on CTs that try to increase the power of existing tests. The first relies on searching the sample for subpopulation where the estimated share of compliers is low or even negative and then applying the procedure by Mourifié and Wan (2017). The

it has to be larger (in absolute terms) then the share of compliers which are shifted to the point $Y = 1$ by Z in order to get detected, which is not guaranteed.

[3] For implementation, we use the R package `causalTree`, available from Susan Athey's Github page `https://github.com/susanathey/causalTree`, accessed on 11/23/2017.

second uses no covariates, but using a preliminary estimate of the complier density, searches directly on the support of the outcome variable for LATE violations. We also discuss the case when conditioning on some or all covariates is necessary to guarantee unconfoundedness of the instrument and provide corresponding splitting criteria and test procedures.

A common problem in adaptive search methods is that extreme values of the splitting target tend to induce splits. In order to guarantee unbiased estimates of the quantities involved in the test statistic, we follow AI and conduct honest splitting.[4] For every tree, the sample is split into a training sample on which the tree is built and an estimation sample on which the actual tests are performed. To increase efficiency, the roles of the two samples are then swapped. To test the Null hypothesis we rely on Chernozhukov, Chetverikov, and Kato (2016)'s (henceforth CCT) many-moments method for unconditional moment inequalities. In the presented setting, the number of leaves, and thus the number of inequalities, is allowed to grow with the sample size. Therefore, a flexible method that allows for any number of and correlation between inequalities seems appropriate. More specifically, we use a hybrid version, which includes pre-selection of likely non-binding inequalities in a first step and bootstraping critical values in a second step. When the outcome is continuous, we follow Mourifié and Wan (2017) and use the intersection bounds method of Chernozhukov, Lee, and Rosen (2013) (henceforth

[4] To be precise, we use algorithm CT-H of AI.

CLR).

The proposed approach has two advantages over simply performing the LATE tests within arbitrarily defined covariate subgroups. First, it provides a data-driven way to detect violations of LATE assumptions. As will be presented in more detail later, the CT algorithm's original aim is to maximize heterogeneity in treatment effects, which corresponds to finding the largest possible violation to LATE assumptions in this setup (under consideration of expected estimation uncertainty). Thus, researchers can credible demonstrate that they searched for violations of key identifying assumptions and not only report the tests' results for subgroups where this was not the case. Second, honest splitting and the many-moments testing procedure safeguard the researcher from over-rejecting the Null hypothesis of no violation of LATE assumptions, even as the number of moment inequalities to test rises with the number of leaves the CT algorithm produces.

The new procedure is then applied to the sibling same-sex IV in US and Chinese census data. Angrist and Evans (1998) introduced the sex composition of the first two children as an instrumental variable for the number of children a women has, exploiting parental preferences for a mixed sex composition of their children. Rosenzweig and Wolpin (2000) doubt the exclusion restriction of the same-sex IV based on economies of scale: having two children of the same sex directly influences parental labor supply, as cloth- and room-sharing are more likely than for opposite sex children. Building on this critique, Bütikofer (2010) used household consumption data from rich and poor countries to show that, indeed, consumption equivalence scales differ between households with same sex and mixed sex children in some developing countries. Monotonicity may be violated if a subpopulation has preferences for a specific gender, such as son preferences prevalent in South Korea (Lee, 2008), India (Rosenzweig and Wolpin, 2000), or the US (Dahl and Moretti, 2008). As a further consequence of gender preferences, the first born child's sex may cause selection into a sample of mothers with at least two children. Dahl and Moretti (2008) find that mothers with first born daughters are less likely to marry and more likely to divorce, which in turn could decrease the probability of a second child for some mothers. On the other hand, families with first born daughters are larger on average. Hence, mothers with two daughters may systematically differ from mothers with two sons and mixed sex children with respect to unobserved variables related to their and the father's gender preferences. Thus, there exist reason to believe that the IV assumptions might be violated, in particular in developing countries, where economies of scale arguments and gender preferences are more prevalent. We therefore test the validity of the same-sex IV first in the 1980 US census sample of Angrist and Evans (1998) and then in a similarly drawn sample from the 2000 Chinese census. China is a particularly promising setting to detect violations, since its one-child policy and son-preferences create distorted incentives for the size and sex compostion of children. We find that in both samples, existing IV validity tests are not able to detect any violations. However, using our proposed CT-based procedure, we are able to reject at least on LATE assumption in the chinese data, but equal to Huber (2015), not in the US sample of Angrist and Evans (1998).

This paper contributes to a recent and fast growing literature that accommodates machine learning tools to the needs of applied econometricans who wish to estimate causal effects. The CT alogirthm is extended to random forests in Wager and Athey (2017) and more general models in Athey et al. (2017), see also Asher et al. (2016) for moment-based trees. 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 exist many possible instrument or control variables relative to the number of observations, see also Belloni et al. (2014a) for an overview. Wager et al. (2016), Bloniarz et al. (2016) and Athey et al. (2016) improve the efficiency of average treatment effect estimates in randomized experiments with Lasso-based balancing. Knaus et al. (2017) use Lasso type estimators to detect treatment effect heterogeneity in job search programmes. In the presence of IVs that violate the exclusion restriction, Kang et al. (2016) and Windmeijer et al. (2016) use variants of Lasso to select these invalid IV's in linear models.

This paper proceeds as follows. In next section, we briefly introduce the CT algorithm and highlight key concepts. In Section 3, we review the existing LATE assumption tests. Section 4 contains our application of CT to the LATE tests. The empirical applications follows in Section 5. We conclude in Section 6.

2 Recursive partitioning for heterogenous treatment effects

This section describes the CT algorithm and its implementation. AI modify Breiman et al. (1984)'s classification and regression trees (CART) algorithm to search for heterogeneity in causal effects across covariates $X \in \mathcal{X}$. Let the observed outcome Y have support $\mathcal{Y}$, the treatment status indicator $D \in \{0, 1\}$, where $D = 1$ indicates treatment, and the binary instrument $Z \in \{0, 1\}$. The potential outcomes are denoted with Y^{dz} and D^z, where $d, z \in \{0, 1\}$.

For illustrative purposes, consider the causal effect of Z on D (also known as the first stage effect) at some point $X = x$: $\alpha(x) = E[D(1) - D(0)|X = x]$. It is important to note that the CT algorithm (as implemented in the R package `causalTree`) will estimate any effect with the difference in conditional means of some target variable between treated and control units, i.e. with the sample analogues of $E[D|Z = 1, X = x] - E[D|Z = 0, X = x]$ for $\alpha(x)$. The algorithm is essentially a data mining tool that at every step evaluates a splitting criterion (an expected mean squared error) at every possible split of the data, i.e. at any covariate value of any covariate. CT will split the sample at that covariate with that value which delivers the largest heterogeneity in the causal effect between the newly formed subgroups, subject to an estimate of the variance of the causal effect. The algorithm is greedy, in the sense that it only tries to improve the splitting criterion at the next split, without considering possible future splits. The splitting ends after some conditions are met, which are discussed below. The result of the algorithm is a so called tree, denoted with Π, which is a collection of L leaves (l): $\Pi = \{l_1, ..., l_L\}$. Let $l(x, \Pi)$ denote the leaf from tree Π which contains the point x.

Before presenting the splitting criterion, we need to highlight an important challenge, which is inference within the resulting leaves. If there exists a subsample, e.g. at some $X_1 \leq x_1$, where the splitting criterion is large by chance alone, then the algorithm is likely to place a split there. However, the observed sample at that point may not be a random sample of the underlying population. Thus, any estimates formed from the observations used for placing the splits are likely to be biased. To solve this problem, AI propose so called honest estimation, implemented by a double tree approach. One randomly chosen part of the sample, called the training sample of size N^{tr}, is used to built the tree, while the remaining sample, called the estimation sample of size N^{est}, is used to make inference. Since we observe X also in the estimation sample, we know in which leaf a single observation falls, even if it was not used to built the tree.

The empirical splitting criterion of AI is

$$\begin{aligned} -\widehat{EMSE}_{\alpha}(S^{tr}, S^{est}, \Pi) = & \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\alpha}^2(X_i; S^{tr}, \Pi) \\ & - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{l \in \Pi} \left(\frac{S^2_{tr,treat}(l)}{p} + \frac{S^2_{tr,cont}(l)}{1-p} \right) \end{aligned} \qquad (2.1)$$

The first term measures the average heterogeneity in $\alpha(x)$, estimated in the training sample S^{tr} with tree Π. The second term is a penalty for within-leaf variance (S^2) of the outcome variable in the treated and untreated populations, weighted by $p = P(Z = 1)$. Given the same potentially achieved treatment effect heterogeneity, the algorithm prefers splits where the estimates will have lower variance.

Two more issues regarding implementation deserve further discussion. The first concerns honest splitting. Splitting the data in half and using only one part for the implementation of the test procedures presented above seems wasteful. To alleviate this inefficiency somewhat, we follow the idea of Chernozhukov et al. (2017) and swap roles of the samples. Instead of training and estimation, call the two random halves of the full sample S^A and S^B. For every tree proposed in the subsequent sections, grow the trees Π^{S^A} and Π^{S^B}, respectively. For every observation in S^B

save the information in which leaves of Π^{S^A} it would fall and in which leaves of Π^{S^B} observations from S^A would fall. Let $\Pi = \Pi^{S^A} \cup \Pi^{S^B}$. The test statistics generated from Π^{S^A} are then estimated in S^B and vice versa. The two trees may form leaves on different elements of X or values of the same X and have different sizes. Using two parallel tree structures increases the number of inequalities to be tested and thus increases the chance to detect violations of the LATE assumptions. In the remainder of the paper, we keep this sample swapping procedure implicit and by estimation mean estimation in the part of the data where the tree is not built.

The second issue concerns pruning of the tree. Growing a tree deep uncovers more heterogeneity and thus makes it more likely to find violations of the LATE assumptions. On the other hand, a larger tree means smaller sample sizes within new leaves, leading to noisier estimates of the test statistics. A classic solution to solve this bias-variance trade-off is to penalize tree size proportional to a constant, whose optimal level is determined via K-fold cross-validation (often K=10). We follow this practice and prune all trees. See Athey and Imbens (2015a) or e.g. Hastie et al. (2009) for more details on cross-validation. Note that when using honest estimation, cross-validation is only performed within the training sample. When growing the tree initially, limiting parameters may be a minimum number of observations from the treated and control group in new leaves, see the R package `causalTree` for all options.

3 Testing LATE assumptions

This section reviews the tests for IV validity poposed by Kitagawa (2015), Huber and Mellace (2015b) and Mourifié and Wan (2017). Following Kitagawa (2015), three assumptions allow the identification of LATEs in the present setup.

***Assumption* 1.** *(Exclusion restriction):* $Y^{d1} = Y^{d0} = Y^d$ *for* $d \in \{0, 1\}$ *wp1.*

***Assumption* 2.** *(Random Assignment):* $Z \perp (Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0)$

***Assumption* 3.** *(Monotonicity):* $D^1 \geq D^0$ *or* $D^0 \geq D^1$ *wp1*

Without loss of generality, assume $D^1 \geq D^0$ and assume that this is a priori known to the researcher. Let $\mathcal{B}_{\mathcal{Y}}$ be a collection of Borel sets from $\mathcal{Y}$.

The key insight for all tests is that if Z monotonically shifts individuals into treatment, is random, and has no direct effect on Y, then the joint probability of observing Y in any $B \in \mathcal{B}_{\mathcal{Y}}$ under treatment and observing a complier has to be weakly greater under $Z = 1$ than under $Z = 0$ for all B. Conversely, the joint probability of observing Y in B under non-treatment and a complier has to be weakly greater under $Z = 0$ than under $Z = 1$. Put differently, the estimated share of compliers has to be non-negative at every point of the distribution of Y

given a realized treatment status:

$$P(Y^1 \in B, C) \geq 0 \tag{3.1}$$

$$P(Y^0 \in B, C) \geq 0 \tag{3.2}$$

where C denotes compliers who are characterized by $D^1 - D^0 = 1$. Empirically, it must hold that :

$$P(Y \in B, D = 1|Z = 1) - P(Y \in B, D = 1|Z = 0) \geq 0 \tag{3.3}$$

$$P(Y \in B, D = 0|Z = 0) - P(Y \in B, D = 0|Z = 1) \geq 0. \tag{3.4}$$

Testable implications (3.3) and (3.4) have first been proposed by Balke and Pearl (1997) and Heckman and Vytlacil (2005), but without suggesting a specific testing procedure. Kitagawa (2015), proposition 1.1 and Mourifié and Wan (2017), theorem 1, establish that (3.3) and (3.4) are sharp, in the sense that they are the strongest testable implications of assumption 1 to 3 given the observed data. Therefore, this paper does not advance the fundamental logic of these tests.

Kitagawa (2015) proposes to test (3.3) and (3.4) using a variance-weighted Kolmogorov-Smirnov type statistic, whose unknown distribution is bootstrapped to derive critical values. One problem with this approach is that if Y is continuous, (3.3) and (3.4) generate a large amount of unconditional inequalities with potentially poor finite sample properties. Discretizing Y into a finite number of arbitrarily chosen sets could lead to some estimated negative densities not getting detected, as they might average out with nearby positive densities of compliers in the same set. Nevertheless, discretizing might be a practical solution in many applications.

Mourifié and Wan (2017) solve this issue elegantly by rewriting (3.3) and (3.4) as

$$P(D=1,Z=1|Y\in B)P(Z=0)-P(D=1,Z=0|Y\in B)P(Z=1)\geq 0 \quad (3.5)$$

$$P(D=0,Z=0|Y\in B)P(Z=1)-P(D=0,Z=1|Y\in B)P(Z=0)\geq 0 \quad (3.6)$$

from which they derive the conditional moment inequalities

$$\theta(y,1)=E[P(Z=1)D(1-Z)-P(Z=0)DZ|Y=y]\leq 0 \quad (3.7)$$

$$\theta(y,0)=E[P(Z=0)(1-D)Z-P(Z=1)(1-D)(1-Z)|Y=y]\leq 0 \quad (3.8)$$

for all $y\in\mathcal{Y}$. They then test

$$H_0:\theta_0\equiv\sup_{y\in Y,d\in\{0,1\}}\theta(y,d)\leq 0, H_1:\theta_0>0. \quad (3.9)$$

using the intersection bounds approach of Chernozhukov et al. (2013). The moments in (3.7) and (3.8) are smoothed via local linear or series estimation over y, thereby avoiding the need for contact sets as the test of Kitagawa (2015) does.

However, the smoothing requires the choice of a tuning parameter. The basic idea behind the CLR method is to adjust the estimates of $\theta(y, d)$ by a critical value times their point-wise standard error and then apply the sup function. See Mourifié and Wan (2017) for the specifics of their implementation. Furthermore, Mourife and Wan relax assumption 3 to De Chaisemartin (2017)'s condition that at every potential outcome value there exist weakly more compliers than defiers. They call this assumption conditionally more compliers (CMC). That is, defiers are allowed, but there must exist at least as many compliers with the same potential outcome value to cancel them out. The testable implications remain the same.

Huber and Mellace (2015b) relax assumptions 1 and 2 to hold only in expectation, as this is sufficient to identify average effects. They derive four moment inequalities that must hold in this situation:

$$E[Y|Z=1, D=1, Y \leq y_{1-p}^{11}] \leq E[Y|D=1, Z=0] \leq E[Y|Z=1, D=1, Y \geq \; y_{p}^{11}] \tag{3.10}$$

$$E[Y|Z=0, D=0, Y \leq y_{1-q}^{00}] \leq E[Y|D=0, Z=1] \leq E[Y|Z=0, D=0, Y \geq \; y_{q}^{00}] \tag{3.11}$$

where

$$p = \frac{P(D = 1|Z = 1) - P(D = 1|Z = 0)}{P(D = 1|Z = 1)}$$
$$q = \frac{P(D = 0|Z = 0) - P(D = 0|Z = 1)}{P(D = 0|Z = 0)}$$

and y_k^{zd} denotes the k-th quantile of the distribution of Y in the group with $Z = z$ and $D = d$. Laffers and Mellace (2017) prove that (3.10) and (3.11) are the strongest testable implications when assumptions 1 and 2 hold in expectation and assumption 3 holds as stated above.

The idea behind these inequalities is that, point identified mean potential outcomes of always-takers ($E[Y|D = 1, Z = 0]$) and never-takers ($E[Y|D = 0, Z = 1]$) have to lie between trimmed means from the mixed population of always-takers and compliers ($Z = 1, D = 1$) and never-takers and compliers ($Z = 0, D = 0$), respectively, where the share of compliers within the respective mixed group (p and q respectively) is trimmed from the top or bottom of the observed mixed distribution. Huber and Mellace (2015b) suggest to increase the power of their mean inequalities by imposing equality and inequality assumptions on mean potential outcomes across complier types, which are heavily used in the literature on partial identification of indirect and direct effects (Flores and Flores-Lagunes, 2013; Chen and Flores, 2015; Chen et al., 2017; Huber and Mellace, 2015a; Huber et al., 2017). Huber and Mellace (2015b) further suggest four inequalties that must hold if the

LATE assumptions hold as defined in assumptions 1 to 3, which allow for overlapping sets B. Overlapping sets may increase test power in finite samples, but this advantage diminishes as the sample size increases. If the sets are non-overlapping, Huber and Mellace (2015b)'s inequalities reduce to Kitagawa's. Since there exists no guidance on the construction of overlapping sets, they are not presented here. For testing, Huber and Mellace suggest Chen and Szroeter (2014)'s smoothed indicator approach or Bennett et al. (2009) bootstrap based recentering method.

4 Improving tests with causal trees

The existing test procedures presented in the previous section recognize that conditioning on a vector of covariates X can justify assumption 2 (independence) or increase the power of the tests. The latter usually proceeds by defining subpopulations in which the tests are then applied separately. However, a common problem is the potential large dimensionality of X, which makes implementation of the above tests for all x practically unfeasible. The alternative, arbitrarily defining subgroups along X, is inefficient and introduces the need to correct for multiple testing. In this section, we use CTs to reduce the dimensionality of X and directly search for violations of (3.3) or (3.4) in the covariate space. We assume that the covariates are measured pre-treatment and are fully independent of the instrument:

***Assumption* 4.** *(Covariates):* $X \perp Z$

An extension to the case were assumption 2 holds conditional on some subset of X is presented below. Under assumptions 1 to 4 Kitagawa's inequalities (3.3) and (3.4) must hold at any point x in the covariate space $X \in \mathcal{X}$:[5]

$$P(Y \in B, D = 1|Z = 1, X = x) - P(Y \in B, D = 1|Z = 0, X = x) \geq 0 \quad (4.1)$$

$$P(Y \in B, D = 0|Z = 0, X = x) - P(Y \in B, D = 0|Z = 1, X = x) \geq 0. \quad (4.2)$$

The specific application of CT depends on the nature of $\mathcal{Y}$. We start with the case where Y is discrete.

4.1 Discrete outcome variable

Assume for now that Y is a discrete random variable. Define for all B the pseudo outcome variables $P^z = \mathbb{1}\{Y \in B\}D^z$ and $Q^z = -\mathbb{1}\{Y \in B\}(1 - D^z)$. Further-

[5] While rejection of (4.1) or (4.2) actually jointly test assumptions 1 to 4, assumption 4 can be tested separately using e.g. balancing tables. If no imbalances are found, violations of (4.1) or (4.2) are then more likely due to violations of assumptions 1 to 3 which is the main interest here.

more, let

$$\sigma(1,x) = E[P|Z=1,X=x] - E[P|Z=0,X=x] \tag{4.3}$$
$$= E[P^1 - P^0|X=x]$$
$$\sigma(0,x) = E[Q|Z=1,X=x] - E[Q|Z=0,X=x] \tag{4.4}$$
$$= E[Q^1 - Q^0|X=x]$$

be the conditional average causal effect of Z on P and Q respectively at $X = x$. Under assumptions 1 to 4:

$$\sigma(1,x) \geq 0 \tag{4.5}$$
$$\sigma(0,x) \geq 0 \tag{4.6}$$

for all B and x, which directly corresponds to inequalities (4.1) and (4.2). The CT algorithm is now used to recursively split the sample at covariate values in order to find heterogeneity in $\sigma(d,x)$ for $d \in \{0,1\}$. It thus directly searches for violations of (4.1) or (4.2) in the covariate space.

Let $\Pi_{B,d}$ denote the resulting tree for B and $l(x\,;\Pi_{B,d})$ the leaf from the tree $\Pi_{B,d}$ which contains the point x. The tree is a finite set of L such leaves: $\Pi_{B,d} = \{l_{1,d,B}, l_{2,d,B}, ..., l_{L,d,B}\}$, where L may vary with the chosen combination of B and d. We can think of CT creating $L-1$ dummy variables indicating membership of observation $i = 1, ..., N$ in leaf $l_1, ..., l_L$ for a chosen B and d. Instead of considering $\sigma(1, x)$ and $\sigma(0, x)$ for potentially infinite points x we consider

$$\sigma(1, l) = E[P^1 - P^0 | X \in l(x\,;\,\Pi_{B,1})] \tag{4.7}$$

$$\sigma(0, l) = E[Q^1 - Q^0 | X \in l(x\,;\,\Pi_{B,0})] \tag{4.8}$$

for all $l(x\,;\,\Pi_{B,d})$ and B, which are a potentially large but finite number of inequalities. Note that our procedure also tests the unconditional inequalities (3.3) and (3.4) as the root node (or stump) of every tree is included in the collection of leaves (albeit tested on a smaller sample). Furthermore, when $\Pi_{B,d}$ consists only of the root node after pruning, then our approach reduces to that of Kitagawa (2015).

Now in order to implement the CCT test procedure, we transform (4.7) and (4.8) into unconditional moment inequalities. Define

$$\zeta(1, l) = -\sigma(1, l)p_{L1} = -E[P\,Z/p_1 - P(1 - Z)/(1 - p_1), X \in l(x\,;\,\Pi_{B,1})] \tag{4.9}$$

$$\zeta(0, l) = -\sigma(0, l)p_{L0} = -E[QZ/p_0 - Q(1 - Z)/(1 - p_0), X \in l(x\,;\,\Pi_{B,0})] \tag{4.10}$$

where $p_d = P(Z = 1|X_i \in l(x\,;\,\Pi_{B,d}))$ and $p_{Ld} = P(X_i \in l(x\,;\,\Pi_{B,d}))$, which is simply the fraction of observations in leaf $l(x\,;\,\Pi_{B,d})$.[6] $\zeta(1, l)$ and $\zeta(0, l)$ can be

[6] We multiply the original σ functions with minus one to ease the implementation of the CCT method.

consistently estimated by

$$\hat{\zeta}(1,l) = \frac{1}{N}\sum_i \mathbb{1}\{Y_i \in B\}\,\mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\}D_i\,[(1-Z_i)/(1-\hat{p}_1) - Z_i/\hat{p}_1]$$
$$\hat{\zeta}(0,l) = \frac{1}{N}\sum_i \mathbb{1}\{Y_i \in B\}\,\mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\}(1-D_i)\,[Z_i/\hat{p}_0 - (1-Z_i)/(1-\hat{p}_0)]$$

where

$$\hat{p}_1 = \frac{1}{|i:\{X_i \in l(x\,;\,\Pi_{B,1}\}|}\sum_{i:\{X_i \in l(x\,;\,\Pi_{B,1}\}} Z_i$$
$$\hat{p}_0 = \frac{1}{|i:\{X_i \in l(x\,;\,\Pi_{B,0}\}|}\sum_{i:\{X_i \in l(x\,;\,\Pi_{B,0}\}} Z_i.$$

We then test the following null hypothesis:

$$H_0 : \zeta_0 \equiv \sup_{(B,d,l)\in\mathcal{B}_{\mathcal{Y}}\times\{0,1\}\times\Pi_{B,d}} \zeta(d,l) \leq 0, \text{ versus } H_1 : \zeta_0 > 0 \tag{4.11}$$

To be specific, for every B and $d = 0, 1$ combination a single tree is grown using the CT algorithm, where $\sigma(d, x)$ is target causal effect. For testing, (4.9) and (4.10) are estimated. For another set, say $G \in \mathcal{B}_{\mathcal{Y}} \neq B$ we do not test $\zeta(d, l) \leq 0$ using information on leaves of the tree $\Pi_{B,d}$, but only of the tree $\Pi_{G,d}$. One might worry that the above approach adds many inequalities which are non-binding. However, the CCT testing procedure presented below is relatively insensitive to the number of inequalities considered. Furthermore, it includes a pre-selection step, which screens out uninformative inequalities. We know describe the procedure.

4.2 Test procedure

Let $\hat{\zeta}_j$ denote the j-th of p inequalities we'd like to test and let $\hat{\sigma}_j^2$ be its estimated variance. CCT consider the test statistic

$$T = \max_{1 \leq j \leq p} \frac{\sqrt{N}\hat{\zeta}_j}{\hat{\sigma}_j}. \tag{4.12}$$

The goal is to find a critical value c which is used to reject H_0 if T is larger than c with test size α. Under the H_0 it must hold that

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

hence finding an upper bound for the $(1-\alpha)$ quantile of $\frac{\sqrt{N}(\hat{\zeta}_j - \zeta_j)}{\hat{\sigma}_j}$ is sufficient to keep the test's size at or below α. CCT show that under rather weak requirements, a critical value for T is

$$c(\alpha) = \frac{\Phi^{-1}(1-\alpha/p)}{\sqrt{1-\Phi^{-1}(1-\alpha/p)/N}}. \tag{4.13}$$

A great advantage of (4.13) in the presence of large p relative to N is that it grows only very slowly in p. However, $c(\alpha)$ is very conservative: As $N \to \infty$ and p fixed,

$c(\alpha)$ becomes $\Phi^{-1}(1-\alpha/p)$ which is identical to a Bonferroni adjusted critical value for testing p hypothesis with size α.

Therefore, in a first step, we reduce the number of inequalities to be tested by screening out those unlikely to be binding. We then use a bootstrap to construct a critical value. The bootstrap accounts for correlation among the $\hat{\zeta}_j$, which is likely given that single leaves are split from other leaves.

The moment selection step follows the first part of CCT's two-step self-normalizing sums method. Let $0 < \beta_N < \alpha/2$ be a constant which is allowed to go to zero as $N \to \infty$. Calculate $c(\beta_N)$ using (4.13). Then the set of pre-selected moment inequalities $\widehat{J} \subset \{1, ..., p\}$ is given by

$$\widehat{J} = \{j \in \{1, ..., p\} : \sqrt{N}\hat{\zeta}_j/\hat{\sigma}_j > -2c(\beta_N)\}$$

For the second step, generate B i.i.d bootstrap samples ζ^*_{ij} for all $j \in \widehat{J}$ and calculate

$$W_{\hat{J}} = \max_{j \in \widehat{J}} \frac{\sqrt{N}E[\zeta^*_{ij} - \hat{\zeta}_j]}{\hat{\sigma}_j}$$

The critical value $c^*(\alpha)$ is then given by the $(1-\alpha+2\beta_N)$-quantile of $W_{\hat{J}}$. We can also calculate a pseudo p-value by $\sum_{\hat{J}}\{W_{\hat{J}} \geq T\}/\text{B}$.

4.3 Continuous outcome variable

When Y is a continuous variable, it is not possible to create trees for every point in the support of Y. In this case, we propose to built only one tree and then apply the testing procedure by Mourifié and Wan (2017) using information from the tree. More specifically, instead of splitting the sample with respect to heterogeneity in $\sigma(d, x)$ the new parameter of interest becomes the causal effect of Z on D, i.e. the first stage effect

$$\alpha(x) = E[D^1 - D^0 | X = x].$$

The motivation behind splitting the sample with respect to the share of complier is that in subsamples where this share is low, the likelihood that the inequalities for some y in the test by Mourifié and Wan (2017) become binding increases.[7] Denote the resulting tree with Π_D and its size with L_D. With covariates, equations

[7] Note that $E[D^1 - D^0] > 0$ but $E[D^1 - D^0 | X = x] < 0$ for some x does violate assumption 3 (if assumption 2 holds), but still allows identification of the LATE when there exist more complier than defiers for every potential outcome (De Chaisemartin, 2017). In other words, finding one subset of the sample where there exist significantly more defiers than compliers does not prohibit the identification of LATE a priori.

(3.5) and (3.6) become

$$\begin{aligned}
&P(D=1,Z=1|Y\in B,X=x)P(Z=0|X=x) && (4.14)\\
&-P(D=1,Z=0|Y\in B,X=x)P(Z=1|X=x)\geq 0 &&\\
&P(D=0,Z=0|Y\in B,X=x)P(Z=1|X=x) && (4.15)\\
&-P(D=0,Z=1|Y\in B,X=x)P(Z=0|X=x)\geq 0. &&
\end{aligned}$$

Instead of conditioning on all possible x, we condition on a finite set of $V = L_D - 1$ indicator functions, which indicate whether observation i belongs into leaf $l_v(x;\, \Pi_D)$ for $v = 1, ..., V$. Define these indicators as $J_v = \mathbb{1}\{X_i \in l_v(x;\, \Pi_D)\}$. For implementation, we pull the leaf indicators into the expectation of the inequalities (3.7) and (3.8) to create the following $2 \times V$ moment inequalities

$$\begin{aligned}
\sigma(y,1,v) &= E[J_v\, D\, [p_v(1-Z)-(1-p_v)\, Z]|Y=y]\leq 0\\
\sigma(y,0,v) &= E[J_v\, (1-D)[(1-p_v)\, Z\; - p_v\, (1-Z)]|Y=y]\leq 0
\end{aligned}$$

for all $y \in \mathcal{Y}$ and $v = 1, ..., V$, where $p_v = P(Z = 1|J_v = 1)$. The null hypothesis then becomes

$$H_0 : \sigma_0^c \equiv \sup_{y\in Y, d\in\{0,1\}, v\in(1,...,V)} \sigma(y,d,v) \leq 0, H_1 : \sigma_0^c > 0. \qquad (4.16)$$

The implementation follows as in Mourifié and Wan (2017) with $2 \times V$ conditional moment inequalities in the intersection bounds approach of CLR.

5 Unconfoundedness of the instrument

In many applications, assumption 2 will only hold when conditioning on a set of covariates. Denote this set as W, which may include some or all elements of X:

***Assumption* 5.** *(Conditional Random Assignment):*
$Z \perp (Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0, X) | W$ *where* $W \subseteq X$

This section modifies the splitting criterion and estimation of the test statistics under assumptions 1,3 and 5. Consider estimation first. In (4.9) and (4.10), we already condition estimation of $P(Z = 1)$ on functions of the covariates, i.e. within the leafs. However, although the trees motivated above may use some elements of W for splitting, it is not guaranteed that all will be used, since the tree is built to detect heterogeneity in $\sigma(d, x)$ or $\alpha(x)$ and not to enable assumption 5. Athey and Imbens (2016) suggest inverse probability weighting with the estimated propensity score to remove bias for estimation within leaves. Denote the propensity score of the instrument with $e(w) = P(Z = 1|W = w)$. Then, $\zeta(1, x)$ and $\zeta(0, x)$ can be consistently estimated by inverse probability weighting with the estimated propensity score, which is re-normalized within leaves:

$$
\begin{aligned}
\hat{\zeta}(1,l) &= \frac{1}{N}\sum_i \mathbb{1}\{Y_i \in B\}\,\mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\}D_i \\
&\quad \cdot \left[\frac{(1-Z_i)/(1-\hat{e}(w))}{\sum_{\{i:X_i\in l(x\,;\,\Pi_{B,1})\}}(1-Z_i)/(1-\hat{e}(w))} - \frac{Z_i/\hat{e}(w)}{\sum_{\{i:X_i\in l(x\,;\,\Pi_{B,1})\}} Z_i/\hat{e}(w)}\right] \\
\hat{\zeta}(0,l) &= \frac{1}{N}\sum_i \mathbb{1}\{Y_i \in B\}\,\mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\}(1-D_i) \\
&\quad \cdot \left[\frac{Z_i/\hat{e}(w)}{\sum_{\{i:X_i\in l(x\,;\,\Pi_{B,0})\}} Z_i/\hat{e}(w)} - \frac{(1-Z_i)/(1-\hat{e}(w))}{\sum_{\{i:X_i\in l(x\,;\,\Pi_{B,0})\}}(1-Z_i)/(1-\hat{e}(w))}\right]
\end{aligned}
$$

When Y is continuous, the moment inequalities for testing change to

$$
\begin{aligned}
\sigma(y,1,v) &= E[e(w)D(1-Z)J_v - (1-e(w))DZJ_v|Y=y] \leq 0 \\
\sigma(y,0,v) &= E[(1-e(w))(1-D)ZJ_v - e(w)(1-D)(1-Z)J_v|Y=y] \leq 0.
\end{aligned}
$$

for all $y \in \mathcal{Y}$ and $v = 1,...,V$.

With respect to splitting, there exist two options. The first is to ignore confounding of the instrument and use the CT algorithm to split with respect to heterogeneity in differences of conditional means. The second is to modify the splitting target to build a transformed outcome tree (TOT). In the case where Y is discrete, define the pseudo outcome variables

$$
\begin{aligned}
P^{z*} &= P^{z}\frac{Z - e(w)}{e(w)(1 - e(w))} \\
Q^{z*} &= Q^{z}\frac{Z - e(w)}{e(w)(1 - e(w))}.
\end{aligned}
$$

The advantage of considering P^{z*} and Q^{z*} as outcome variables is that they are unbiased estimators of $\sigma(1, x)$ and $\sigma(0, x)$:

$$
\begin{aligned}
E[P^{z*}|X = x] &= E[P^1 - P^0|X = x] = \sigma(1, x) \\
E[Q^{z*}|X = x] &= E[Q^1 - Q^0|X = x] = \sigma(0, x)
\end{aligned}
$$

See Athey and Imbens (2015b), proposition 1, for a proof of the first equalities. Thus, instead of using the CT algorithm, the classic CART algorithm for regression trees can be used off the shelf with P^{z*} and Q^{z*} as outcome variables.

When Y is continuous, the transformed outcome becomes

$$
D^* = D\frac{Z - e(w)}{e(w)(1 - e(w))}
$$

where $E[D^*|X = x] = E[D^1 - D^0|X = x] = \alpha(x)$.

The disadvantage of the TOT approach is that it does not use information on Z directly, making it a less efficient estimator of $\sigma(d, x)$ and $\alpha(x)$. As a consequence, the tree is likely to be pruned stronger than necessary and leaves with potentially helpful heterogeneity are chipped away. Thus, there exists a trade-off: Using the CT algorithm for splitting may blunt the search for violations of the LATE assumptions. Using the TOT approach accounts for confounding, but distorts the pruning process due to increased variance in the pseudo outcomes. Which splitting rule is preferable will depend on the degree of confoundedness of the instrument. In one extreme case, the instrument is unconditionally random, then causal trees are definitely superior to TOTs for splitting. In the other extreme case when instrument assignment is strongly confounded by covariates, causal trees may set very distorted split points and the potential gain in power of testing (3.3) and (3.4) with covariates decreases.

6 Discretizing the outcome without covariates

In this section we use CT to directly search for IV validity violations on the support of the continuous outcome variable. Rewrite (3.3) and (3.4) as

$$P(D=1|Z=1,Y=y)P(Y=y|Z=1)-P(D=1|Z=0,Y=y)P(Y=y|Z=0)\geq 0$$
$$\Leftrightarrow E[D\, f_Y(y|Z=1)|Z=1,Y=y]-E[D\, f_Y(y|Z=0)|Z=0,Y=y]\geq 0 \quad (6.1)$$

$$P(D=0|Z=0,Y=y)P(Y=y|Z=0)-P(D=0|Z=1,Y=y)P(Y=y|Z=1)\geq 0$$
$$\Leftrightarrow E[(1-D)\, f_Y(y|Z=0)|Z=0,Y=y]-E[(1-D)\, f_Y(y|Z=1)|Z=1,Y=y]\geq 0, \quad (6.2)$$

where $f_Y(y|Z=z)$ is the conditional density of Y at $z\in\{0,1\}$. Denote its kernel density estimate with $\hat{f}_Y(y|Z=z)$. To facilitate implementation, define the two pseudo outcome variables $\hat{f}^1 = D\{\hat{f}_Y(y|Z=1)Z+\hat{f}_Y(y|Z=0)(1-Z)\}$ and $\hat{f}^0=-(1-D)\{\hat{f}_Y(y|Z=0)(1-Z)+\hat{f}_Y(y|Z=1)Z\}$. We set up the following splitting targets

$$\phi^1(y)=E[\hat{f}^1|Z=1,Y=y]-E[\hat{f}^1|Z=0,Y=y]\geq 0 \quad (6.3)$$
$$\phi^0(y)=E[\hat{f}^0|Z=1,Y=y]-E[\hat{f}^0|Z=0,Y=y]\geq 0,. \quad (6.4)$$

With the above formulation, CT directly searches for subsets of Y where (6.1) and

(6.2) may be violated. Obviously this is an imprecise approach since the estimated densities are treated as known. The CT splitting criterion (2.1) presented above will only consider sampling variation of $\hat{f}^d$ when placing splits, not underlying estimation uncertainty. It might therefore be preferable to studentize the point estimates before splitting and split on

$$\phi^{1*}(y) = E\left[\frac{\hat{f}^1}{\hat{\sigma}^1}\middle| Z=1, Y=y\right] - E\left[\frac{\hat{f}^1}{\hat{\sigma}^1}\middle| Z=0, Y=y\right] \geq 0 \tag{6.5}$$

$$\phi^{0*}(y) = E\left[\frac{\hat{f}^0}{\hat{\sigma}^0}\middle| Z=0, Y=y\right] - E\left[\frac{\hat{f}^0}{\hat{\sigma}^0}\middle| Z=1, Y=y\right] \geq 0, \tag{6.6}$$

where $\hat{\sigma}^1 = \sqrt{\hat{\text{Var}}(\hat{f}_Y(y|Z=1))}Z + \sqrt{\hat{\text{Var}}(\hat{f}_Y(y|Z=0))}(1-Z)$ and $\hat{\sigma}^0 = \sqrt{\hat{\text{Var}}(\hat{f}_Y(y|Z=1))}Z + \sqrt{\hat{\text{Var}}(\hat{f}_Y(y|Z=0))}(1-Z)$ use the estimated point-wise standard errors from the kernel density estimate. Denote the resulting trees from either version of $\phi^1(y)$ and $\phi^0(y)$ with Π_{Y^1} and Π_{Y^0} respectively.

Since the CT algorithm will search only over one variable, the terminal nodes of Π_{Y^1} and Π_{Y^0} describe compact intervals of the outcome variable under treatment and non-treatment respectively. Now instead of testing (6.1) and (6.2) directly, we can use e.g. Kitagawa's variance weighted KS statistic or Chen and Szroeter

(2014)'s smoothed indicator approach, to test

$$P(Y \in l^*(y; \Pi_{Y^1}), D = 1 | Z = 1) - P(Y \in l^*(y; \Pi_{Y^1}), D = 1 | Z = 0) \geq 0 \quad (6.7)$$

$$P(Y \in l^*(y; \Pi_{Y^0}), D = 0 | Z = 0) - P(Y \in l^*(y; \Pi_{Y^0}), D = 0 | Z = 1) \geq 0 \quad (6.8)$$

for all $l^*(y; \Pi_{Y^1}) \in \Pi_{Y^1}$ and $l^*(y; \Pi_{Y^0}) \in \Pi_{Y^0}$ where the l^* denote the end nodes of the corresponding tree. When assumption 2 is replaced by assumption 5, conditional independence of Z given W, estimation of $f_Y(y|Z = z, W = w)$ may become infeasible when the dimensionality of W is too high. A combination of the approaches presented in this paper are possible. For example, for continuous Y, first apply the above discretizing procedure, then use covariates X to perform the tests proposed in section 4.1 on the new outcome variable.

7 Application

In this section we apply our methodology for a discrete outcome variable to a commonly used instrumental variable for fertility decisions and sibship size, the same-sex instrument.[8] Building on the observation that parents prefer a mixed sex composition of their children, Angrist and Evans (1998) (henceforth AE) propose to use the occurence of same-sex siblings as an IV for the number of children. As

[8] Recent applications of the same-sex IV in various contexts include Dehejia et al. (2015), Brinch et al. (2017), Cools and Hart (2017), Fitzsimons and Malde (2014) and Aaronson et al. (2017).

already laid out in the introduction, there exist several reasons why this IV may be invalid, in particular in developing countries. Huber (2015) was the first to test this IV's validity using the approach of Huber and Mellace (2015b) in the 1980 US Census sample employed by AE. He finds no violation in the full sample and very few violations across 22 subgroups and concludes that the IV's validity in these data cannot be refuted. We first replicate Huber (2015)'s results in the AE data before turning to a sample drawn of the 2000 Chinese census.

In both cases, Z indicates a same-sex composition of the first two children and D is an indicator equal to one if mothers have three or more children at the time of the census. The selection criteria for both samples are as follows. Mothers have to be of age 21 to 35, had their first child after the age of 14, their first child is younger than 18, and have at least two children.[9] We choose $\alpha = 5\%$, 5-fold cross validation for pruning, and a minimum number of observations with $Z = 1$ and $Z = 0$ in every leaf of 2,000.

7.1 US Census 1980

We refer to AE and Huber (2015) for a description of the sample. To allow for a faster computation, we take a random half of the original 394,840 observations, such that the working sample size is 197,420. The outcome variable is the number of weeks mothers worked last year, which takes on 53 unique values. As covariates,

[9] In the Chinese data mothers have to additionally reside in one-family households.

we consider mother's current age, age at first birth, sex of the first and second child, educational attainment (four levels), and dummies for hispanic ethnicity and race (black, other race). The effect of the same-sex instrument on the probability to have more than two children is about 6%, which compares to a baseline probability of 37%.

The results are shown in the first row Table 1. Both, the Huber and Mellace (2015b) and Kitagawa (2015) tests clearly do not reject the Null of no LATE assumption violations. Our CT based test procedure produces 326 moment inequalities, of which 324 are chosen after the pre-selection step. The test statistic of 2.41 is smaller than the critical value of 3.37 and the bootstrap p-value is 0.666. As a result, we do not reject the validity of the same-sex IV in the original AE sample and thereby replicate the finding of Huber (2015).

– Table 1 about here –

7.2 Chinese Census 2000

This sample is built on the 1 percent Chinese census of 2000, provided by IPUMS-I (Minnesota Population Center, 2017). After applying the selection criteria the sample size is 357,995 mothers.

The outcome variable is the number of days worked last week and as covariates, we consider mother's current age, age at first birth, sex of the first and second child, educational attainment, literacy status, and ethnic group.[10]

The sample is described in Table 2. We find that 57% of first born children are female. This was to be expected, as families with son-preferences are larger on average when the first child is female (and parents don't split up). Before restricting the sample to mothers with at least 2 children, only about 40% of mothers have 2 or more children and less than 8% have three or more. These figures compare to 66% and 27% in the US Census data, respectively. The effect of the same-sex instrument on the probability to have more than two children is 0.14 (t-statistic 102) unconditionally and 0.13 (t-statistic 97) after controlling for the mentioned covariates, which is a relative increase of roughly 100%. Thus, the effect size of this instrument is much larger in absolute and relative terms compared to the AE application. Again, we take a random half of the sample, such that the working sample size is 178,997. We chose $\beta_N = 0.0001$ for the moment pre-selection part of the test procedure.

– Table 2 about here –

The result of our IV validity test for the Chinese sample are presented in the second row of Table 1. As was the case for the US census sample, the tests of Huber and Mellace (2015b) and Kitagawa (2015) are not able to detect any violations. However, our CT-based approach clearly rejects the Null of no LATE

[10] Ethnic groups which make up less than 0.5% of the sample are assigned to the other or unknown group.

violations. The test statistic of 3.76 is greater than the critical value of 2.99 and the bootstrapped p-value is below 1%. We also find that the pre-selection step considerably decreases the number of tested inequalities: the tree building phase generates 150 moment inequalities, which are reduced to 88 after pre-selection.

To illustrate our CT-based procedure, Figure 1 shows the tree for $B = 7$ and $d = 0$. It consists of 17 leaves. The first value in every leaf is the estimate of $\sigma(0, x)$ in the training sample. The second value shows the fraction of all observations belonging to that leaf. The text beneath the leaf shows the variable and value on which the leaf was split next. We find that, as expected, the first child's gender is the most important predictor of heterogeneity in the difference $P(Y = 7, D = 0|Z = 0) - P(Y = 7, D = 0|Z = 1)$. When the first child's sex is male, the estimated probability of not having a further child and to work a full week is larger for mothers with same-sex children (i.e. two boys) than with mixed sex children. One explanation for this finding is that if parents have preferences for boys and in addition are restricted in their family size, already having two boys makes them keep their family size. If they have one boy and one girl, some want another boy. Figure 1 shows that, conditional on the first child's gender, there exist further heterogeneity in $\sigma(0, x)$ across mothers age at birth, educational attainment and ethnicity.

– Figure 1 about here –

8 Conclusion

The identification of local average treatment effects using instrumental variables is common in empirical research. The validity of instruments is a focal point of debate in such studies. Fortunately, the LATE framework generates empirically testable implications on the IV's validity. In this paper we proposed a machine learning based approach to improve the power of existing IV validity tests in a data driven way. The procedure uses the Causal Tree algorithm by Athey and Imbens (2016) to split the sample along covariate values in order to form subsamples of the data were violations are most likely. For testing, we use Chernozhukov et al. (2016)'s many moments test procedure. We consider the case when the outcome is discrete or continuous and modify the splitting criteria for unconfoundedness of the instrument after conditiong on a set of covariates. The approach can be implemented using existing software packages.

We find that our new procedure is able to detect violations of the LATE framework in a setting where we strongly expect them, but where existing tests fail to reject the Null of no violations. Specifically, we tested the validity of Angrist and Evans (1998)'s same-sex instrument in US and Chinese census data. We cannot reject validity in the original data of Angrist and Evans (1998) with either the existing or our new test procedure. However, the latter is able to reject IV validity in the 2000 Chinese census, which the former is not.

Our procedure also has its limitations. First, it requires rather large datasets, because in order to avoid bias from adaptive searching for violations we need to conduct honest splitting: one half of the sample is used to build trees while the other half is used to estimate the test statistic. Although we switch the roles of the samples to generate more moment inequalities, honest splitting implies that one can never use all observations to estimate the same moment inequality. Second, the new procedure requires the presence of covariates which are unaffected by the treatment. This paper thus also calls for the collection of rich pre-treatment variables: combined with machine learning methods they enable to detect treatment effect heterogeneity (Athey and Imbens, 2016), improve the precision of treatment effect estimates (Athey et al., 2016), and can help to refute instrumental variable assumptions.

References

Aaronson, D., R. H. Dehejia, A. Jordan, C. Pop-Eleches, C. Samii, and K. Schulze (2017). The effect of fertility on mothers' labor supply over the last two centuries.

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 J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Asher, S., D. Nekipelov, P. Novosad, and S. P. Ryan (2016). Classification trees for heterogeneous moment-based models. Technical report, National Bureau of Economic Research.

Athey, S. and G. Imbens (2015a). Recursive partitioning for heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*.

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. and G. W. Imbens (2015b). Machine learning methods for estimating heterogeneous causal effects. *Unpublished manuscript*.

Athey, S., G. W. Imbens, and S. Wager (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

Athey, S., J. Tibshirani, and S. Wager (2017). Generalized random forests. *arXiv preprint arXiv:1610.01271*.

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, 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. *The Review of Economic Studies 81*(2), 608–650.

Bennett, C. J. et al. (2009). Consistent and asymptotically unbiased minp tests of multiple inequality moment restrictions. *Vanderbilt University Department of Economics Working Papers 908*.

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.

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

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond late with a discrete instrument. *Journal of Political Economy 125*(4), 985–1039.

Bütikofer, A. (2010). Sibling sex composition and cost of children. Technical report, mimeo.

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

Chen, X. and C. A. Flores (2015). Bounds on treatment effects in the presence of sample selection and noncompliance: the wage effects of job corps. *Journal of Business & Economic Statistics 33*(4), 523–540.

Chen, X., C. A. Flores, and A. Flores-Lagunes (2017). Going beyond late: Bounding average treatment effects of job corps training. *Journal of Human Resources*, 1015–7483R1.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Chernozhukov, V., D. Chetverikov, and K. Kato (2016). Testing many moment inequalities. Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

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.

Cools, S. and R. K. Hart (2017). The effect of childhood family size on fertility in adulthood: New evidence from iv estimation. *Demography 54*(1), 23–44.

Dahl, G. B. and E. Moretti (2008). The demand for sons. *The Review of Economic Studies 75*(4), 1085–1120.

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

Dehejia, R., C. Pop-Eleches, and C. Samii (2015). From local to global: External validity in a fertility natural experiment.

Fitzsimons, E. and B. Malde (2014). Empirically probing the quantity–quality model. *Journal of Population Economics 27*(1), 33–68.

Flores, C. A. and A. Flores-Lagunes (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business & Economic Statistics 31*(4), 534–545.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.

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., L. Laffers, and G. Mellace (2017). Sharp iv bounds on average treatment effects on the treated and other populations under endogeneity and noncompliance. *Journal of Applied Econometrics 32*, 56–79.

Huber, M. and G. Mellace (2015a). Sharp bounds on causal effects under sample selection. *Oxford Bulletin of Economics and Statistics 77*(1), 129–151.

Huber, M. and G. Mellace (2015b). 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.

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.

Kédagni, D. and I. Mourifie (2015). Generalized instrumental inequalities: Testing iv independence assumption.

Kédagni, D. and I. Mourifie (2016). Empirical content of the iv zero-covariance assumption: Testability, partial identification.

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

Knaus, M., M. Lechner, and A. Strittmatter (2017). Heterogeneous employment effects of job search programmes: A machine learning approach. *arXiv preprint arXiv:1709.10279*.

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

Lee, J. (2008). Sibling size and investment in children's education: An asian instrument. *Journal of Population Economics 21*(4), 855–875.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2013). Instrumental variables and the sign of the average treatment effect.

Minnesota Population Center (2017). Integrated public use microdata series, international: Version 6.5 [dataset].

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

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

Slichter, D. (2015). Testing instrument validity and identification with invalid instruments.

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

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* (just-accepted).

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.

Windmeijer, F., H. Farbmacher, N. Davies, and G. Davey Smith (2016). On the use of the lasso for instrumental variables estimation with some invalid instruments. Technical report, Department of Economics, University of Bristol, UK.

Tables and Figures

Table 1: Results of the IV validity tests

			Causal trees IV validity test				
Data	H&M(2015)	K(2015)	p	$\lvert\widehat{J}\rvert$	T	$c^*(\alpha)$	Pseudo p-value
US 1980 Census	0.539	0.985	326	324	2.41	3.37	0.666
China 2000 Census	1.000	0.986	150	88	3.76	2.99	0.002

H&M(2015) is the p-value of the Huber and Mellace (2015b) IV validity test and K(2015) of Kitagawa (2015)'s test. p is the number of inequalities to be tested before moment selection. $|\widehat{J}|$ is the number of actually tested inequalities after the pre-selection step with $\beta_N = 0.0001$. T is the test-statistic and based on Chernozhukov, Chetverikov, and Kato (2016)'s empirical bootstrap method. The critical value is denoted with $c^*(\alpha)$. All bootstraps involved use 2000 replications.

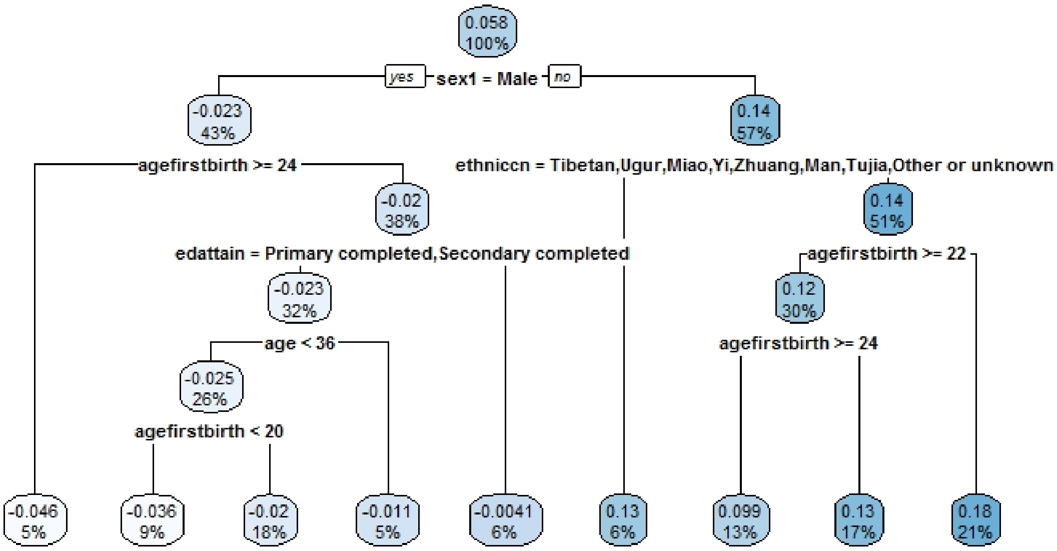

Figure 1: Tree $\Pi_{7,0}$ from the Chinese sample. The first value in every leaf is the estimate of $\sigma(0, x)$ in the training sample. The second value shows the fraction of all observations belonging to that leaf. The text beneath the leaf shows the variable and value on which the leaf was split next.

Table 2: Chinese data description

	Mean	SD	Min	Max
Outcome:				
Days worked last week	5.027	2.550	0	7
Treatment:				
More than two children	0.196	0.397	0	1
Instrument:				
Same-sex	0.422	0.494	0	1
Covariates:				
Age	31.677	2.802	21	35
Age at first birth	21.933	2.309	15	35
1st child is a girl	0.571	0.495	0	1
2nd child is a girl	0.422	0.494	0	1
Ethnicity: Han	0.873	0.333	0	1
Ethnicity: Mongol	0.005	0.071	0	1
Ethnicity: Hui	0.011	0.106	0	1
Ethnicity: Tibetan	0.006	0.075	0	1
Ethnicity: Ugur	0.013	0.113	0	1
Ethnicity: Miao	0.014	0.117	0	1
Ethnicity: Yi	0.014	0.119	0	1
Ethnicity: Zhuang	0.018	0.131	0	1
Ethnicity: Man	0.006	0.075	0	1
Ethnicity: Tujia	0.009	0.094	0	1
Ethnicity: Other/unknown	0.032	0.176	0	1
Literate	0.933	0.249	0	1
Education: Less than primary	0.149	0.356	0	1
Education: Primary	0.820	0.385	0	1
Education: Secondary or university	0.032	0.175	0	1

Sample drawn from the 1 percent 2000 Chinese census made available by Minnesota Population Center (2017) as IPUMS file. Sample is restricted to mother that reside in one-family households, are of age 21 to 35, had their first child after the age of 14,their first child is younger than 18, and have at least two children. N=178,997.