Instrument Validity Tests with Causal Trees:

With an Application to the Same-sex Instrument

Helmut Farbmacher Raphael Guber*

Munich Center for the Economics of Aging, Max Planck Society

September 29, 2019

Abstract

The use of instrumental variables (IVs) to identify causal effects is common practice in empirical economics, but it is fundamentally impossible to test their validity. However, assumptions sufficient for the identification of local average treatment effects (LATEs) jointly generate necessary conditions in the observed data that allow to refute an IV's validity. Suitable tests exist, but they may not be able to detect even severe violations of IV validity in practice. In this paper, we use the causal tree (CT) algorithm by Athey and Imbens (2016) to search the sample for violations of the LATE assumptions in a data-driven way. The new approach is applied to the sibling sex composition instrument in census data from China and the United States. We expect that, because of son preferences, the siblings sex instrument is invalid in the Chinese context. However, while existing IV validity tests are unable to detect violations, our CT based procedure does.

Keywords: instrument validity, machine learning, causal tree

JEL codes: C12, C18, C26

*guber@mea.mpisoc.mpg.de. We thank Martin Huber, Giovanni Mellace and participants at seminars in Konstanz, Odense, and the IAAE annual meeting 2018 for helpful comments. Guber acknowledges funding through the International Doctoral Program "Evidence-Based Economics" of the Elite network of Bavaria.

1 Introduction

Empirical research that tries to credibly estimate causal effects relies heavily on instrumental variables (IVs) (Angrist and Pischke, 2010). However, IVs have to fulfill strong assumptions in order to be valid. Given that it is fundamentally impossible to directly test the validity of IVs, discussions on threats to the key identifying assumptions and robustness checks on the latter constitute a crucial part of many empirical papers.

Recently, Kitagawa (2015), Huber and Mellace (2015) and Mourifié and Wan (2017) derived tests that allow to refute IV validity based on necessary conditions in the data that are generated by the joint assumptions sufficient for the identification of LATEs (Imbens and Angrist, 1994; Angrist et al., 1996).[1] The unifying idea across these three studies is that, given a treatment status, the estimated density of compliers, which is the unobserved subpopulation of individuals whose treatment status is causally affected by the instrument, must be non-negative at any point of the distribution of the outcome variable. These tests are inconsistent; that is, they cannot detect all violations, even in large samples. In practice, violations might only get detected when they are large enough.[2]

The underlying idea of this paper is that the presence or intensity of violations may vary across subpopulations defined by covariate values. For example, a direct effect of the IV on the outcome variable might be large enough in a subgroup such that existing tests can detect it, but in the full sample, this direct effect is watered down to the point where these tests cannot detect it anymore.

In this paper, we utilize the causal tree (CT) algorithm, developed by Athey and Imbens (2016), AI hereafter, to split the sample along covariate values with the aim to detect subpopulations where the LATE assumptions are violated. We assume that the outcome variable has finite discrete support. Thus, the testable

[1] Related studies include Arai et al. (2018), Kédagni and Mourifie (2015, 2016), Slichter (2015), Eckhoff Andresen et al. (2018) and Machado et al. (2019).

[2] To give an example, consider a binary outcome Y, treatment D and instrument Z. If all LATE assumptions hold, then 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, then it has to be larger (in absolute terms) then the share of compliers, which are shifted to the point $Y = 1$ by Z to be detected, which is not guaranteed.

implications reduce to a finite number of moment inequalities that consist of differences in joint probabilities. The original aim of the CT algorithm is to split the sample with respect to heterogeneity in conditional treatment effects, but it can easily be modified to search for heterogeneity in differences between conditional joint probabilities. This heterogeneity is conceptually restricted to non-negative values when the LATE assumptions hold, but may take on negative values if not. In the latter case, CT can find these violations when they occur in the full sample or within covariate cells. The proposed procedure can be easily implemented with existing software packages.[3] The CT algorithm puts no restrictions on the number and support of the covariates.

When violations are uniform across the sample and their severity does not vary with covariates, our test has less power than existing tests in finite samples, due to honest splitting. In this case, the CT algorithm will not split the sample or splits will occur due to noise only. In many applications, however, violations of IV validity are likely to vary within a sample. We presume that finding violations of the LATE assumptions in at least one subpopulation casts doubt on the IV in the whole sample, because researchers cannot rule out further undetected violations.[4] The null hypothesis of our test is formulated accordingly.

We also discuss the common case when conditioning on some or all covariates is necessary to guarantee unconfoundedness of the instrument and provide corresponding splitting criteria and test statistics. Again, these are easily implementable, requiring only an estimate of the instrument propensity score. In the present setting, the number of leaves of the trees (subpopulations), and thus the number of testable inequalities, is allowed to grow with the sample size.[5] Furthermore, every observation may contribute to multiple conditional moment inequalities. Therefore, a testing procedure that allows for any number of inequalities, even larger than the sample size, and correlation between them is necessary. To test the null hypothesis of at least one violation, we rely on Chernozhukov, Chetverikov, and Kato

[3]For implementation, we use the R package `causalTree`, which is available from Susan Athey's Github page `https://github.com/susanathey/causalTree` (accessed on 11/23/2017.). R code for the implementation of the test can be found on the author's website.

[4]In some cases, ex-post removal of problematic sub-groups might be justifiable but does not guarantee the validity of the IV in the remaining sample.

[5]This can be achieved by, for example, setting the minimum number of observations within a leaf to a constant value.

(2016)'s, CCK hereafter, many-moments method for unconditional moment inequalities. More specifically, we use a hybrid version, which includes pre-selection of likely non-binding inequalities in a first step and bootstrapping critical values in a second step.

The proposed approach has several advantages over simply performing the existing IV validity tests within arbitrarily defined covariate subgroups. First, it provides a data-driven way to detect violations of LATE assumptions. Researchers can credibly demonstrate that they have efficiently searched for violations of key identifying assumptions and they can show that the data does not reject their proposed IV. Second, the algorithm automatically chooses the covariates and covariate values to condition on. Third, honest splitting and the many-moments testing procedure safeguard the researcher from over-rejecting the null hypothesis of no violations as the number of moment inequalities to test rises with the number of subgroups. No further adjustment of p-values for multiple hypothesis testing are necessary. Fourth, it allows for a strong first stage effect, i.e. a large share of compliers in the full sample. Existing test procedures are ceteris paribus more powerful when the share of compliers is low. However, applied researchers want a strong IV in order 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 gains ceteris paribus more power when the share of compliers is low within a subpopulation.

The new procedure is then applied to the same-sex IV. Angrist and Evans (1998) introduced the sex composition of the first two children as an exogenous source of variation in the number of children that a women has, which is caused by parental preferences for a mixed sex composition of their children.

The validity of the same-sex IV can be questioned for several reasons. First, 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, because clothes and room-sharing are more likely between same than opposite sex children. Second, monotonicity may be violated by son preferences, which are prevalent in many countries (Lee, 2008; Rosenzweig and Wolpin, 2000; Dahl and Moretti, 2008). Parents who want to have two boys stop having further children when the first two are already boys, but they would have

had a third child if the first two were a girl and a boy. Third, son preferences may cause non-random sex compositions. For example, parents can "opt out" of a girl-girl constellation by aborting the fetus of the second child, until the mother is pregnant with a son.

Thus, there is reason to believe that the LATE assumptions might be violated, in particular in developing countries, where economies of scale arguments and gender preferences are more prevalent. However, Dahl and Moretti (2008) provide evidence that son preferences are also prevalent in the United States. Therefore, we test the validity of the same-sex IV in the 1980 US census sample of Angrist and Evans (1998), and then in a similarly drawn sample from the 2000 Chinese census. Because of its one-child policy and son-preferences, it is unlikely that researchers would have used the same sex IV in Chinese data. The question that we try to answer with this application is: given potentially strong violations of the LATE assumptions, are existing tests and/or our test procedure able to reject IV validity?

We find that in both samples, existing IV validity tests are unable to detect any violations. However, using our proposed CT-based procedure, we are able to reject at least one LATE assumption in the Chinese data. Similar to Huber (2015), we do not find significant violations in the US sample of Angrist and Evans (1998).

This paper also 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 and detect heterogeneity in the latter. The CT algorithm is extended to causal forests by Wager and Athey (2018), which was applied to two randomized control trials by Davis and Heller (2017). Athey et al. (2019) consider detecting heterogeneity of parameters in estimation equation models, while the approach by Asher et al. (2016) relies on 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 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. (2017) use Lasso-type estimators to detect treatment effect heterogeneity in job search programs. In the presence of IVs that violate the exclusion restriction, Kang et al. (2016) and Windmeijer et al. (2019) use variants

of Lasso to select valid IV's 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 review the existing LATE assumption tests. In Section 3, we briefly introduce the CT algorithm and highlight the key concepts. Section 4 contains the application of CT to the IV validity tests. We provide simulation results in Section 5. The empirical applications follow in Section 6. We conclude in Section 7.

2 Testing for violations of LATE assumptions

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 $d, z \in \{0,1\}$. Following Kitagawa (2015), three assumptions are sufficient for the identification of LATEs in this 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*

We will relax assumption 2 to hold conditional on covariates below. 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 $\mathcal{Y}$. Then, under assumptions 1 to 3, empirically it must hold that:

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

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

for every $B \in \mathcal{B}_{\mathcal{Y}}$. To see why, consider the four compliance types: compliers (C), always-takers (A), never-takers (N) and defiers (F) (Angrist et al., 1996), which are distributed in the data as shown in Table 1.

– Table 1 about here –

Then, (2.1) can be re-written as

$$
\begin{aligned}
&P(Y \in B, D = 1|Z = 1) - P(Y \in B, D = 1|Z = 0) \\
&= P(Y^{11} \in B|C, Z = 1)P(C|Z = 1) + P(Y^{11} \in B|A, Z = 1)P(A|Z = 1) \qquad (2.3) \\
&- P(Y^{10} \in B|A, Z = 0)P(A|Z = 0) - P(Y^{10} \in B|F, Z = 0)P(F|Z = 0) \\
&\overset{A1}{=} P(Y^{1} \in B|C, Z = 1)P(C|Z = 1) + P(Y^{1} \in B|A, Z = 1)P(A|Z = 1) \\
&- P(Y^{1} \in B|A, Z = 0)P(A|Z = 0) - P(Y^{1} \in B|F, Z = 0)P(F|Z = 0) \\
&\overset{A2}{=} P(Y^{1} \in B|C)P(C) + P(Y^{1} \in B|A)P(A) - P(Y^{1} \in B|A)P(A) - P(Y^{1} \in B|F)P(F) \\
&= P(Y^{1} \in B|C)P(C) - P(Y^{1} \in B|F)P(F) \\
&\overset{A3}{=} P(Y^{1} \in B|C)P(C) = P(Y^{1} \in B, C) \geq 0
\end{aligned}
$$

A similar derivation can be done for (2.2), which leads to

$$P(Y \in B, D = 0|Z = 0) - P(Y \in B, D = 0|Z = 1) \overset{A1,A2,A3}{=} P(Y^{0} \in B, C) \geq 0.$$

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. Testable implications (2.1) and (2.2) have first been proposed by Balke and Pearl (1997) and Heckman and Vytlacil (2005), but without suggesting a specific testing procedure. Kitagawa (2015) proposes to test (2.1) and (2.2) 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, then (2.1) and (2.2) generate an infinite amount of unconditional inequalities. Discretizing Y into a finite number of arbitrarily chosen sets 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. Nevertheless, discretizing might be a practical solution in many applications. Mourifié and Wan (2017) solve this issue elegantly by rewriting (2.1) and (2.2) as conditional moment inequalities (conditional on $Y = y$) and then applying the intersection bounds approach of Chernozhukov et al. (2013). Huber and Mellace (2015) relax assumptions 1 and 2 to hold only in expectation as this is sufficient to identify average effects. They propose a test based on four moment inequalities, which can handle both discrete

and continuous outcomes.

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 1 to 3 given the observed data. Laffers and Mellace (2017) prove that the inequalities proposed by Huber and Mellace (2015) are the strongest testable implications when assumptions 1 and 2 hold in expectation and assumption 3 holds as stated above.

The decomposition in (2.3) demonstrates why tests based on (2.1) and (2.2) are inconsistent. Consider for example the case when A2 and A3 hold, but not A1. Then (2.1) equals

$$P(Y^{11} \in B|C)P(C) + P(Y^{11} \in B|A)P(A) - P(Y^{10} \in B|A)P(A). \tag{2.4}$$

Only when the direct effect of the instrument is large enough among always takers will $P(Y^{11} \in B|A)$ sufficiently differ from $P(Y^{10} \in B|A)$ for some B in order for (2.1) to become negative. When A2 does not hold, $P(A|Z = 1)$ might not equal $P(A|Z = 0)$ but this violation will not be detected in a sample of any size when it is rather small compared to the probability mass $P(Y^1 \in B, C)$. It is also easy to see that violations of different assumptions can cancel each other out.

Now consider the availability of covariates X, which are measured pre-treatment and that are fully independent of the instrument:

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

Under assumptions 1 to 4, Kitagawa's inequalities (2.1) and (2.2) must hold at any point x in the covariate space $\mathcal{X}$:[6]

$$P(Y \in B, D = 1|Z = 1, X = x) - P(Y \in B, D = 1|Z = 0, X = x) \geq 0 \tag{2.5}$$

$$P(Y \in B, D = 0|Z = 0, X = x) - P(Y \in B, D = 0|Z = 1, X = x) \geq 0. \tag{2.6}$$

Returning to the decomposition in (2.3) shows that conditioning on X can be helpful in four ways. First, the share of compliers within some covariate cell,

[6]While rejection of (2.5) or (2.6) actually jointly tests assumptions 1 to 4, assumption 4 can be tested separately using, for example, balancing tables. If no imbalances are found, violations of (2.5) or (2.6) are then more likely due to violations of assumptions 1 to 3 which is the main interest here.

$P(C|X = x)$, might be lower than in the full sample or even zero, such that violations in that cell can be more easily detected, ceteris paribus. Second, the share of defiers might locally be larger. Third, the direct effect of the instrument might be stronger for some subpopulation, which makes finding violations of A1 easier. Fourth, a non-random instrument leads to different shares of always-takers between $Z = 1$ and $Z = 0$ and this imbalance is more pronounced among some subpopulation.

Kitagawa (2015), Huber and Mellace (2015) and Mourifié and Wan (2017) also apply their tests within covariate cells. However, how to form these covariate cells such that they make finding violations of IV validity more likely is an open question, in particular when covariates are continuous. Arbitrarily defining subgroups is inefficient and introduces the need to correct for multiple testing. A further problem is the potential large dimensionality of X, which makes implementation of the above tests for all x practically unfeasible. Therefore, we now introduce the Casual Tree (CT) algorithm.

3 Recursive partitioning for heterogenous treatment effects

This section briefly describes the CT algorithm and its implementation. The algorithm is based Breiman et al. (1984)'s classification and regression trees (CART) algorithm and its aim is to search for heterogeneity in conditional average treatment effects. For illustrative purposes, consider the average 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 this effect with the difference in conditional means of some target variable between treated and control units; that is, 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 at every possible split of the data; that is, at any covariate value of any covariate. CT will split the sample at that covariate value which delivers the largest heterogeneity in the causal effect between the newly formed subgroups. 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, for example, 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 build the tree, while the remaining sample, called the estimation sample of size N^{est}, is used to estimate treatment effects which are ultimately reported. Since we also observe X 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 the honest causal tree algorithm (CT-H) is

$$\begin{aligned} -\widehat{EMSE}_{\alpha}(S^{tr}, S^{est}, \Pi) = & \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\alpha}^2(X_i; S^{tr}, \Pi) \\ & - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}} \right) \sum_{l \in \Pi} \left(\frac{S^2_{tr,treat}(l)}{pr} + \frac{S^2_{tr,cont}(l)}{1 - pr} \right) \end{aligned} \tag{3.1}$$

The first term on the right represents the average heterogeneity in $\alpha(x)$, as 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 $pr = P(Z = 1)$. Given the same potentially achieved treatment effect heterogeneity, the CT-H algorithm prefers splits where the estimates will have lower variance. This second term anticipates honest estimation and is missing in the adaptive version of the algorithm (CT-A). In the application and simulations below, we use the CT-A algorithm and found no substantial

differences in results when using either CT-H or CT-A.

Two more issues regarding implementation deserve further discussion. The first concerns honest splitting. Splitting the data in half and using only one part for estimation seems wasteful. To alleviate this inefficiency somewhat, we follow the idea of Chernozhukov et al. (2018) and swap the roles of the samples. Instead of training and estimation, we call the two random halves of the full sample S^A and S^B. For every tree proposed in the subsequent sections, we grow the trees Π^{S^A} and Π^{S^B}, respectively. For every observation in S^B, we 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 this 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 deeply uncovers more heterogeneity and, thus, makes it more likely to find violations of the LATE assumptions. A larger tree also 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, which is determined via K-fold cross-validation (often K=10). We follow this practice and prune all trees. See Athey and Imbens (2015a) or, for example, 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, the minimum number of observations from the treated and control group in new leaves is an important limiting parameter. In principle, one can set this parameter to one because the final tree will be pruned. In practice, computational speed will be increased by larger values.

4 Improving IV validity tests with causal trees

In this section, we use CTs to search for violations of (2.1) or (2.2) in the covariate space. To do this, define for any $B \in \mathcal{B}_\mathcal{Y}$ the pseudo outcome variables $R^z = \mathbb{1}\{Y \in B\}D^z$ and $Q^z = -\mathbb{1}\{Y \in B\}(1 - D^z)$. Furthermore, let

$$\begin{aligned} \sigma(1, x) &= E[R|Z = 1, X = x] - E[R|Z = 0, X = x] && (4.1) \\ &= E[R^1 - R^0|X = x] \\ \sigma(0, x) &= E[Q|Z = 1, X = x] - E[Q|Z = 0, X = x] && (4.2) \\ &= E[Q^1 - Q^0|X = x] \end{aligned}$$

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

$$\sigma(1, x) \geq 0 \tag{4.3}$$

$$\sigma(0, x) \geq 0 \tag{4.4}$$

for all B and x, which directly corresponds to inequalities (2.5) and (2.6). The CT algorithm is now used to split the sample at covariate values to find heterogeneity in $\sigma(d, x)$ for $d \in \{0, 1\}$. Under assumption 1 to 3, $\sigma(d, x)$ can only vary between 0 and 1 for any x. However, if the IV is invalid, $\sigma(d, x)$ can vary between -1 and 1. In this case, the CT algorithm is able to split the sample such that for some $X = x^*$ we may find that indeed $\sigma(1, x^*) < 0$.

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. Instead of considering $\sigma(1, x)$ and $\sigma(0, x)$ for a potentially infinite number of points in $\mathcal{X}$, we consider

$$\sigma(1, l) = E[R^1 - R^0|X \in l(x\,;\,\Pi_{B,1})] \tag{4.5}$$

$$\sigma(0, l) = E[Q^1 - Q^0|X \in l(x\,;\,\Pi_{B,0})] \tag{4.6}$$

for all $l(x\,;\,\Pi_{B,d})$ and B, which are a potentially large but finite number of inequal-

ities. Note that our procedure also tests the unconditional inequalities (2.1) and (2.2), in the root node of every tree.

4.1 Test procedure

The CCK test procedure is developed for unconditional moment inequalities, where a moment is described by the mean of a random variable. To implement their approach in our setting, we define the random variables

$$t_i(1,l) = \mathbb{1}\{Y_i \in B\}\, \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\} D_i\, [(1-Z_i)/\hat{p}_{01} - Z_i/\hat{p}_{11}]$$
$$t_i(0,l) = \mathbb{1}\{Y_i \in B\}\, \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\} (1-D_i)\, [Z_i/\hat{p}_{10} - (1-Z_i)/\hat{p}_{00}]$$

for all $l(x\,;\,\Pi_{B,d})$ and B where

$$\hat{p}_{11} = \frac{1}{N}\sum_i Z_i\, \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\} \tag{4.7}$$

$$\hat{p}_{01} = \frac{1}{N}\sum_i (1-Z_i)\, \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\} \tag{4.8}$$

$$\hat{p}_{10} = \frac{1}{N}\sum_i Z_i\, \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\} \tag{4.9}$$

$$\hat{p}_{00} = \frac{1}{N}\sum_i (1-Z_i)\, \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\}. \tag{4.10}$$

Let $\zeta(d,l) = E[t_i(d,l)]$ denote the expected values of the $t_i(d,l)$ variables. Let $p = |\zeta(d,l)|$ denote the total number of inequalities that we would like to test. We test the following null hypothesis:

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

That is, we test whether at least one moment inequality is violated. 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 the target causal effect. For testing, we create L variables $t_i(d,l)$, one for every leaf of the tree $\Pi_{B,d}$. The collection of such created variables from all trees are then used in the test procedure below.

For ease of notation, let t_{ij} denote the j-th from a total of p variables that we constructed. Furthermore, let $\hat{\zeta}_j = \frac{1}{N}\sum_i t_{ij}$ denote the sample mean of the j-th variable and $\hat{\sigma}_j^2 = \frac{1}{N}\sum_i (t_{ij} - \hat{\zeta}_j)^2$ the sample variance. CCK 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. CCK 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 while 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 the first step, we reduce the number of inequalities to be tested by screening out those unlikely to be binding,, i.e. where (2.5) and (2.6) are clearly positive. We then use an empirical bootstrap instead of (4.13) to construct a critical value. The bootstrap accounts for correlation among the $\hat{\zeta}_j$, which is helpful given that single leaves are split from other leaves.

The moment selection step follows the first part of CCK'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 t^*_{ij} for all $j \in \widehat{J}$ and calculate

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

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 $\sum_{\hat{J}} \mathbb{1}\{W_{\hat{J}} \geq T\}/\text{B}$. Note that the bootstrap does not involve the tree building procedure in every replicate. Instead, the t_i variables are constructed using information on leaf membership in the trees from the original sample.

4.2 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.** *(Unconfoundedness):*
$Z \perp (Y^{11}, Y^{10}, Y^{01}, Y^{00}, D^1, D^0, X)|W$ *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. When constructing the variables $t_i(d, l)$ we divide by the estimated joint probabilities $P(Z = 1, X_i \in l(x\,;\,\Pi_{B,d})) = P(Z = 1|X_i \in l(x\,;\,\Pi_{B,d}))P(X_i \in l(x\,;\,\Pi_{B,d}))$ or $P(Z = 0, X_i \in l(x\,;\,\Pi_{B,d}))$. Thus, we already condition estimation of $P(Z = 1)$ (or $P(Z = 0)$) to be within leaves, which are a function of the covariates. However, although the trees motivated above may use some elements of W for splitting, it is not guaranteed that all will be used because the tree is built to detect heterogeneity in $\sigma(d, x)$ and not to enable assumption 5. AI suggest inverse probability weighting with the estimated (instrument) propensity score to remove bias for estimation within leaves. The propensity score estimates are re-normalized within leaves and $Z = 1$ and $Z = 0$ groups to increase efficiency (Hirano et al., 2003). Denote the

propensity score of the instrument with $e(w) = P(Z = 1|W = w)$ and an estimate of it as $\hat{e}(w)$. Then, we modify the estimated joint probabilities in (4.7) to (4.10):

$$\hat{p}_{11} = \frac{\hat{e}(w)}{\sum_{i:X_i \in l(x\,;\,\Pi_{B,1}),Z=1} \hat{e}(w)} \frac{1}{N} \sum_i \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\} \tag{4.14}$$

$$\hat{p}_{01} = \frac{(1-\hat{e}(w))}{\sum_{i:X_i \in l(x\,;\,\Pi_{B,1}),Z=0} (1-\hat{e}(w))} \frac{1}{N} \sum_i \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,1})\} \tag{4.15}$$

$$\hat{p}_{10} = \frac{\hat{e}(w)}{\sum_{i:X_i \in l(x\,;\,\Pi_{B,0}),Z=1} \hat{e}(w)} \frac{1}{N} \sum_i \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\} \tag{4.16}$$

$$\hat{p}_{00} = \frac{(1-\hat{e}(w))}{\sum_{i:X_i \in l(x\,;\,\Pi_{B,0}),Z=0} (1-\hat{e}(w))} \frac{1}{N} \sum_i \mathbb{1}\{X_i \in l(x\,;\,\Pi_{B,0})\}. \tag{4.17}$$

With respect to splitting, there are two options. The first is to ignore confounding of the instrument, which may result in a weakened ability of CT to find violations. The second is to modify the splitting target to build a transformed outcome tree (TOT). To implement the latter, we define the pseudo outcome variables

$$R^{z*} = P^z \frac{Z - \hat{e}(w)}{\hat{e}(w)(1-\hat{e}(w))}$$
$$Q^{z*} = Q^z \frac{Z - \hat{e}(w)}{\hat{e}(w)(1-\hat{e}(w))}.$$

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

$$E[R^{z*}|X = x] = E[R^1 - R^0|X = x] = \sigma(1, x)$$
$$E[Q^{z*}|X = x] = E[Q^1 - Q^0|X = x] = \sigma(0, x)$$

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 R^{z*} and Q^{z*} as outcome variables.
The disadvantage of the TOT approach is that it does not use information on Z directly, making it a less efficient estimator of $\sigma(d, x)$, see AI. Consequently, the tree is likely to be pruned stronger than necessary and leaves with potentially helpful

heterogeneity are cut off. Thus, there exists a trade-off for an unconfounded instrument: 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. Which splitting rule is preferable will depend on the degree of confoundedness of the instrument, but it is beyond the scope of this paper to derive the condition where switching to one approach is optimal.

5 Simulations

To test the finite sample performance of our new procedure, we set up the following data generating process, which is close to that of **?**:

$$
\begin{aligned}
Y^* &= 2D + \gamma Z + \sum_{k=1}^{3} \beta_k X_k + U \\
D &= \mathbb{1}\{\alpha Z + \epsilon\} \\
\text{where} X_k &\sim N(0,1),\ \beta_k \sim \text{Uniform}(0,1)\ \text{,for } k = 1,2,3, \\
(U,\epsilon) &\sim N(0,5), \text{cov}(U,\epsilon) = 0.5, Z \sim \text{Bernoulli}(0.5).
\end{aligned}
$$

The three covariates X_1, X_2 and X_3 predict the outcome Y^*. We set the coefficient γ to either 0 (DGP1), 1 (DGP2) or 5 if $X_1 < 0, X_2 > 0$, and $X_3 > 0$, and 0 otherwise (DGP3). In DGPs 1 to 3 the first stage coefficient α is either 0.2 or 0.6 leading to complier shares of 2% and 10%, respectively. For DGP4 we set γ to 0 but α is equal to -10 if $X_1 < 0$ and $X_2 > 0$, and 3 otherwise. DGP1 represents the valid IV case, DGP2 is a global violation of the exclusion restriction, while DGP3 shows a strong local violation of the exclusion restriction with $E[\gamma] = 0.625$. DGP4 corresponds to a local violation of monotonicity, with a globally estimated share of compliers of 24%. For implementation, Y^* is discretized into four equally large subsets. We compare our procedure to the Huber and Mellace (2015) probability test that uses Chen and Szroeter (2014)'s smooth indicator method[7]. The sample size is either 1,000 or 3,000 observations.

Table 2 shows rejection frequencies from 10,000 simulations. We find that for

[7]The bootstrap in Chen and Szroeter (2014)'s smooth indicator approache uses 999 replications.

both sample and complier population sizes, our procedure does not over-reject the Null of no violation of LATE assumptions at the 5% level under DGP1 (no violation case). When the violation of the exclusion restriction is global (DGP2), the probability based test of Huber and Mellace (2015) performs better than our procedure under all scenarios. This was to be expected, because splitting the data by covariates cannot improve the power of the test and honest splitting leads to a lower sample size when testing for global violations (in the root node of the tree). However, the difference in power between Huber and Mellace (2015) and our test is not very large for N=3,000. Under DGP3 the exclusion restriction is only locally violated. Consequently, splitting the sample by covariates can lead to an improvement in test power. In this case, our procedure has a larger power than the test of Huber and Mellace (2015) for both sample sizes and complier shares. In case of DGP4, a local violation of the monotonicity assumption, our procedure again outperforms Huber and Mellace (2015).

– Table 2 about here –

6 Application

In this section, we apply our procedure to the same-sex instrument, a widely used instrumental variable for endogenuous fertility decisions and sibship size.[8] Building on the observation that some parents prefer a mixed sex composition of their children, Angrist and Evans (1998) (henceforth, AE) propose to use the occurrence of same-sex siblings as an IV for the number of children. As already laid out in the introduction, there are 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 (2015) in the 1980 US Census sample that was 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

[8]Recent applications of the same-sex IV in various contexts include: Dehejia et al. (2019), Brinch et al. (2017), Cools and Hart (2017), Fitzsimons and Malde (2014) and Aaronson et al. (2017).

before turning to a sample drawn of the 2000 Chinese census. As discussed in the introduction, we expect the IV to be invalid in the Chinese setting.

In both data sets, 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 aged 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 $\alpha = 0.05$, 5-fold cross validation for pruning, $\beta_N = 0.0001$ for the moment pre-selection part, and a minimum number of observations with $Z = 1$ and $Z = 0$ in every leaf of 2,000.

6.1 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 final sample size is 197,420. The outcome variable is the number of weeks that the mothers worked last year, which takes on 53 unique values. As covariates, we consider the 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 (white, black, other). The effect of the same-sex instrument on the probability to have more than two children is about 6 percentage points, which compares to a baseline probability of 37%.

The results are shown in the first row of Table 3. Both the Huber and Mellace (2015) and Kitagawa (2015) tests clearly do not reject the null of no LATE assumption violations. Our CT based test procedure generates 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. Consequently, 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 3 about here –

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

6.2 Chinese Census 2000

This sample is built on the 1% Chinese census of 2000, provided by IPUMS-I (Minnesota Population Center, 2017). After applying the selection criteria, the sample size is 357,995 mothers. Again, we take a random half of the sample, such that the final sample size is 178,997.

The outcome variable is the number of days worked last week, which ranges from zero to seven. 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 4. We find that 57% of first born children are female. This was to be expected because families with son-preferences are larger on average when the first child is female (Dahl and Moretti, 2008). Before restricting the sample to mothers with at least two children, only about 40% of mothers have two or more children and less than 8% have three or more. This pattern is likely to be linked to China's (now past) one-child policy. 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 14 percentage points (t-statistic 102) unconditionally and 13 percentage points (t-statistic 97) after controlling for the mentioned covariates, which is an increase of roughly 100%. Thus, the effect size of this instrument is much larger in absolute and relative terms compared to the AE application.

– Table 4 about here –

The result of our IV validity test for the Chinese sample are presented in the second row of Table 3. As was the case for the US census sample, the tests of Huber and Mellace (2015) and Kitagawa (2015) are unable to detect any violations. However, our CT-based approach clearly rejects the null of no LATE violations. The test statistic of 5.26 is greater than the critical value of 3.08 and the bootstrapped p-value is 0.1%. We also find that the pre-selection step considerably decreases the number of tested inequalities: the tree building phase generates 162 moment inequalities, which are reduced to 102 after pre-selection.

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

To illustrate our CT-based procedure, Figure 1 shows the unpruned tree for $B = 7$ and $d = 0$, which consists of 17 leaves. The first value displayed 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, then 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. This observation is strongly in line with the existence of son preferences that lead to a violation of the monotonicity assumption. Figure 1 also shows that, conditional on the first child's gender, there is further heterogeneity in $\sigma(0, x)$ across mothers age at birth, educational attainment and ethnicity.

– Figure 1 about here –

7 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 of the IV's validity. In this paper, we proposed a machine learning based approach to improve existing IV validity tests in a data driven way. The procedure uses the CT algorithm by Athey and Imbens (2016) to split the sample along covariate values to form subsamples of the data were violations are most likely. For testing, we use Chernozhukov et al. (2016)'s many moments test procedure. Our 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 the validity of this instrument in the data of Angrist and Evans (1998) with either the existing or with our new test procedure. However, the latter is able to reject

IV validity in the 2000 Chinese census, while the former is not.

Our procedure also has some limitations. First, the new procedure requires the presence of covariates that are unaffected by the treatment. Second, it has less power in small samples than existing tests when violations do not vary with covariates, because we need to conduct honest splitting to avoid bias from adaptive searching for violations: 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. Third, our study is not directly applicable to continuous outcome variables. One possibility to use CT in this case may be to split the sample with respect to the share of compliers (the effect of Z on D) and then apply the tests of Huber and Mellace (2015) and Mourifié and Wan (2017) within the resulting subsamples and correct for multiple testing. The reasoning is that, as mentioned above, if the share of compliers is low, IV violation are easier to detect.[11] A second solution would be to repeatedly and randomly divide a continuos Y into a random number of subsets, use our splitting procedure for every permutation and then test the Null of any IV violations from any such permutations under the Chernozhukov et al. (2016) framework, which allows for a large number of moment inequalities.

[11] Splitting the data with respect to the first stage effect can also be used to test for assumptions 2 and 3 separately. This may also be useful to show whether monotonicity holds in the sample.

References

Aaronson, D., R. 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. National Bureau of Economic Research, w23717.

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.

Arai, Y., Y.-C. Hsu, T. Kitagawa, I. Mourifie, and Y. Wan (2018). Testing identifying assumptions in fuzzy regression discontinuity design. cemmap Working Paper, CWP50/18.

Asher, S., D. Nekipelov, P. Novosad, and S. P. Ryan (2016). Classification trees for heterogeneous moment-based models. National Bureau of Economic Research, w22976.

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 (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.

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

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., 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.

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 (2016). Testing many moment inequalities. cemmap Working Paper, CWP42/16.

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.

Davis, J. and S. B. Heller (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review Papers & Proceedings 107*(5), 546–550.

Dehejia, R., C. Pop-Eleches, and C. Samii (2019). From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics, forthcoming*.

Eckhoff Andresen, M., M. Huber, et al. (2018). Instrument-based estimation with binarized treatments: Issues and tests for the exclusion restriction. FSES Working Papers 492.

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

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.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

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.

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. SSRN.

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 (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics 212*(2), 522–555.

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. *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) 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.

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: Compliance types

	Z=1	Z=0
D=1	C,A	F,A
D=0	F,N	C,N

Table 2: Simulation results (10,000 simulations)

N=1,000	$\alpha = 0.2$		$\alpha = 0.6$	
	CT	HM	CT	HM
DGP 1	0.028	0.034	0.006	0.000
DGP 2	0.548	0.895	0.313	0.548
DGP 3	0.445	0.092	0.238	0.002
DGP 4	0.481	0.000		
N=3,000				
DGP 1	0.014	0.006	0.001	0.000
DGP 2	0.983	1.000	0.798	0.972
DGP 3	0.988	0.151	0.937	0.247
DGP 4	0.999	0.000		

Table displays rejection frequencies for significance level of 5%. CT refers to our test and HM refers to test by Huber and Mellace (2015). The minimum number of observations with either $Z = 1$ or $Z = 0$ within every leaf is set to 50. For DGP4 $\alpha = -10$ if $X_1 < 0$, $X_2 > 0$ and 3 otherwise.

Table 3: 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	1.000	0.936	326	324	2.41	3.37	0.666
China 2000 Census	1.000	1.000	162	102	5.26	3.08	0.001

H&M(2015) is the p-value of the Huber and Mellace (2015) IV validity test using Chen and Szroeter (2014)'s smooth indicator approach 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.

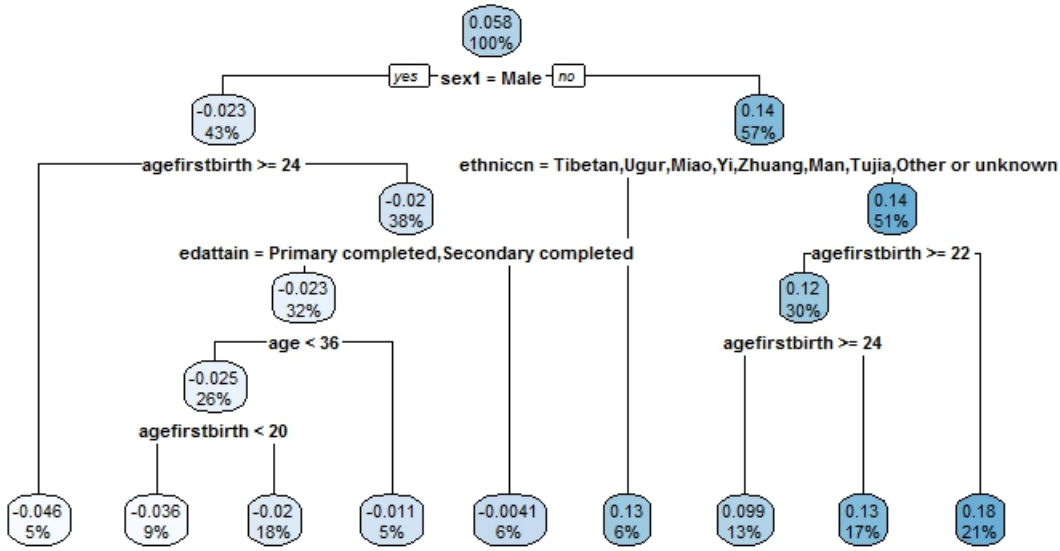

Figure 1: Tree $\Pi_{7,0}$ from the Chinese sample. The first value in every leaf is the estimate of $\sigma(0, x)$ in the training sample. The second value shows the fraction of all observations belonging to that leaf. The text beneath the leaf shows the variable and value on which the leaf was split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

Table 4: 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% 2000 Chinese census made available by Minnesota Population Center (2017) as IPUMS file. The sample is restricted to mothers that reside in one-family households, are aged 21 to 35, had their first child after the age of 14, their first child is younger than 18, and who have at least two children. N=178,997.