Valid Causal Inference with (Some) Invalid Instruments

Jason Hartford [1] **Victor Veitch** [2] **Dhanya Sridhar** [3] **Kevin Leyton-Brown** [1]

Abstract

Instrumental variable methods provide a powerful approach to estimating causal effects in the presence of unobserved confounding. But a key challenge when applying them is the reliance on untestable "exclusion" assumptions that rule out any relationship between the instrument variable and the response that is not mediated by the treatment. In this paper, we show how to perform consistent instrumental variable estimation despite violations of the exclusion assumption. In particular, we show that when one has multiple candidate instruments, only a majority of these candidates—or, more generally, the modal candidate–response relationship—needs to be valid to estimate the causal effect. Our approach uses an estimate of the modal prediction from an ensemble of instrumental variable estimators. The technique is simple to apply and is "black-box" in the sense that it may be used with any instrumental variable estimator as long as the treatment effect is identified for each valid instrument independently. As such, it is compatible with recent machine-learning based estimators that allow for the estimation of conditional average treatment effects (CATE) on complex, high dimensional data. Experimentally, we achieve accurate estimates of conditional average treatment effects using an ensemble of deep network-based estimators, including on a challenging simulated Mendelian randomization problem.

1. Introduction

Instrumental variable (IV) methods are a powerful approach for estimating treatment effects: they are robust to unobserved confounders and they are compatible with a variety of flexible nonlinear function approximators (see e.g. Newey & Powell, 2003; Darolles et al., 2011; Hartford et al., 2017; Singh et al., 2019; Bennett et al., 2019; Muandet et al., 2020; Dikkala et al., 2020), thereby allowing nonlinear estimation of heterogeneous treatment effects.

In order to use an IV approach, one must make three assumptions. The first, *relevance*, asserts that the treatment is not independent of the instrument. This assumption is relatively unproblematic, because it can be verified with data. The second assumption, *unconfounded instrument*, asserts that the instrument and outcome do not share any common causes. This assumption cannot be verified directly, but in some cases it can be justified via knowledge of the system; e.g. the instrument may be explicitly randomized or may be the result of some well understood random process. The final assumption, *exclusion*, asserts that the instrument's effect on the outcome is entirely mediated through the treatment. This assumption is even more problematic; not only can it not be verified directly, but it can be very difficult to rule out the possibility of direct effects between the instrument and the outcome variable. Indeed, there are prominent cases where purported instruments have been called into question for this reason. For example, in economics, the widely used "judge fixed effects" research design (Kling, 2006) uses random assignment of trial judges as instruments and leverages differences between different judges' propensities to incarcerate to infer the effect of incarceration on some economic outcome of interest (see Frandsen et al., 2019, for many recent examples). Mueller-Smith (2015) points out that exclusion is violated if judges also hand out other forms of punishment (e.g. fines, a stern verbal warning etc.) that are not observed. Similarly, in genetic epidemiology, "Mendelian randomization" (Davey Smith & Ebrahim, 2003) uses genetic variation to study the effects of some exposure on an outcome of interest. For example, given genetic markers that are known to be associated with a higher body mass index (BMI), we can estimate the effect of BMI on cardiovascular disease. However, this only holds if we are confident that the same genetic markers do not influence the risk of cardiovascular disease in any other ways. The possibility of such "direct effects"—referred to as "horizontal pleiotropy" in the genetic epidemiology literature—is regarded as a key challenge for Mendelian randomization (Hemani et al., 2018).

It is sometimes possible to identify *many* candidate instru-

[*]Equal contribution [1]University of British Columbia, Vancouver, Canada [2]University of Chicago, Illinois, USA [3]Columbia University, New York, USA. Correspondence to: Jason Hartford <jasonhar@cs.ubc.ca>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

ments, each of which satisfies the relevance assumption; in such settings, demonstrating exclusion is usually the key challenge, though in principle unconfounded instrument could also be a challenge. For example, many such candidate instruments can be obtained in both the judge fixed effects and Mendelian randomization settings, where individual judges and genetic markers, respectively, are treated as different instruments. Rather than asking the modeler to gamble by choosing a single candidate about which to assert these untestable assumptions, this paper advocates making a weaker assumption about the whole set of candidates. Most intuitively, we can assume *majority validity*: that at least a majority of the candidate instruments satisfy all three assumptions, even if we do not know which candidates are valid and which are invalid. Or we can go further and make the still weaker assumption of *modal validity*: that the modal relationship between instruments and response is valid. Observe that modal validity is a weaker condition because if a majority of candidate instruments are valid, the modal candidate–response relationship must be characterized by these valid instruments. Modal validity is satisfied if, as Tolstoy might have said, "All happy instruments are alike; each unhappy instrument is unhappy in its own way."

This paper introduces ModeIV, a robust instrumental variable technique. ModeIV allows the estimation of nonlinear causal effects and lets us estimate conditional average treatment effects that vary with observed covariates. It is simple to implement—it involves fitting an ensemble with a modal aggregation function—and is black-box in the sense that it is compatible with any valid IV estimator, which allows it to leverage any of the recent machine learning-based IV estimators. Despite its simplicity, ModeIV has strong asymptotic guarantees: we show consistency and that even on a worst-case distribution, it converges point-wise to an oracle solution at the same rate as the underlying estimators. We experimentally validated ModeIV using both a modified version of the demand simulation from Hartford et al. (2017) and a more realistic Mendelian randomization example modified from Hartwig et al. (2017). In both settings—even with data with a very low signal-to-noise ratio—we observed ModeIV to be robust to exclusion-restriction bias and accurately recovered conditional average treatment effects.

2. Related Work

Background on Instrumental Variables We are interested in estimating the causal effect of some treatment variable, t, on some outcome of interest, y. The treatment effect is confounded by a set of observed covariates, x, and unobserved confounding factors, ϵ, which affect both y and t. With unobserved confounding, we cannot rely on conditioning to remove the effect of confounders; instead we use an instrumental variable, z, to identify the causal effect.

Instrumental variable estimation can be thought of as an inverse problem: we can directly identify the causal[1] effect of the instrument on both the treatment and the response before asking the inverse question, "what treatment–response mappings, $f : t \not\to y$, could explain the difference between these two effects?" The problem is identified if this question has a unique answer. If the true structural relationship is of the form, $y = f(t, x) + \epsilon$, one can show that, $E[y|x, z] = \int f(t, x)dF(t|x, z)$, where $E[y|x, z]$ gives the instrument–response relationship, $F(t|x, z)$ captures the instrument–treatment relationship, and the goal is to solve the inverse problem to find $f(\cdot)$. In the linear case, $f(t, x) = \beta t + \gamma x$, so the integral on the right hand side of reduces to $\beta E[t|x, z] + \gamma x$ and β can be estimated using linear regression of y on the predicted values of t given x and z from a first stage regression. This procedure is known as Two-Stage Least Squares (Angrist & Pischke, 2008). More generally, the causal effect is identified if the integral equation has a unique solution for f (Newey & Powell, 2003).

Nonlinear IV A number of recent approaches have leveraged this additive confounders assumption to extend IV analysis beyond the linear setting. Newey & Powell (2003) and Darolles et al. (2011) proposed the first nonparametric procedures for estimating these structural equations, based on polynomial basis expansions. These methods relax the linearity requirement, but scale poorly in both the number of data points and the dimensionality of the data. To overcome these limitations, recent approaches have adapted deep neural networks for nonlinear IV analyses. DeepIV (Hartford et al., 2017) fits a first-stage conditional density estimate of $\hat{F}(t|x, z)$ and uses it to solve the above integral equation. Both Bennett et al. (2019) and Dikkala et al. (2020) adapt generalized method of moments (Hansen, 1982) to the nonlinear setting by leveraging adversarial losses, while Singh et al. (2019) and Muandet et al. (2020) propose kernel-based procedures for estimation using two-stage and dual formulations of the problem, respectively. Puli & Ranganath (2020) showed conditions that allow IV inference with latent variable estimation techniques.

Inference with invalid instruments in linear settings Much of the work on valid inference with invalid instruments is in the Mendelian randomization literature, where violations of the exclusion restriction are common. For a recent survey, see Hemani et al. (2018). There are two broad approaches to valid inference in the presence of bias introduced by invalid instruments: averaging over the bias, or eliminating the bias with ideas from robust statistics. In the first setting, valid inference is possible under the assumption that each instrument introduces a random bias, but that the

[1] Strictly, non-causal instruments suffice but identification and interpretation of the estimates can be more subtle (see Swanson & Hernán, 2018).

mean of this process is zero (although this assumption can be relaxed (c.f. Bowden et al., 2015; Kolesár et al., 2015)). Then, the bias tends to zero as the number of instruments grow. Methods in this first broad class have the attractive property that they remain valid even if none of the instruments is valid, but they rely on strong assumptions that do not easily generalize to the nonlinear setting considered in this paper.

The second class of approaches to valid inference assumes that some fraction of the instruments are valid and then uses the fact that biased instruments are outliers whose effect can be removed by leveraging robust estimators. For example, by assuming *majority validity* and constant linear treatment effects[2], Kang et al. (2016) and Guo et al. (2018) show that it is possible to consistently estimate the treatment effect via a Lasso-style estimator that uses the sparsity of the $\ell\psi$norm to remove invalid instruments. Under the same linearity and constant effect assumptions, Hartwig et al. (2017) showed that one can estimate the treatment effect under *modal validity* by estimating the mode of a set of Wald estimators. In this paper, we use the same modal insight as Hartwig et al., but generalize the approach to a nonlinear setting, thereby removing the strong assumption of constant treatment effects. Finally, Kuang et al. (2020) recently showed that, under majority validity, it is possible to leverage structure learning techniques produce a "summary (valid) IV" that can be plugged into downstream estimators. They focus on a setting with binary instruments, responses and confounders whereas we aim for a generic method that places no constraints on the data generating process beyond those necessary for identification.

Ensemble models Ensembles are widely used in machine learning as a technique for improving prediction performance by reducing variance (Breiman, 1996) and combining the predictions of weak learners trained on non-uniformly sampled data (Freund & Schapire, 1995). These ensemble methods frequently use modal predictions via majority voting among classifiers, but they are designed to reduce variance. Both the median and mode of an ensemble of models have been explored as a way of improve robustness to outliers in the forecasting literature (Stock & Watson, 2004; Kourentzes et al., 2014), but we are not aware of any prior work that explicitly uses these aggregation techniques to eliminate bias from an ensemble.

Mode estimation If a distribution admits a density, the mode is defined as the global maximum of the density function. More generally, the mode can be defined as the limit of a sequence of modal intervals—intervals of width h that contains the largest proportion of probability mass—such that $x_{\text{mode}} = \lim_{h\to 0} \arg\max_x F([x - h/2, x + h/2])$. These two definitions suggest two estimation methods for estimating the mode from samples: either one may try to estimate the density function and the maximize the estimated function (Parzen, 1962), or one might search for midpoints of modal intervals from the empirical distribution functions. To find modal intervals, one can either fix an interval width, h, and choose x to maximize the number of samples within the modal interval (Chernoff, 1964), or one can solve the dual problem by fixing the target number of samples to fall into the modal interval and minimizing h (Dalenius, 1965; Venter, 1967). We use this latter Dalenius–Venter approach as the target number of samples can be parameterized by the number of valid instruments, thereby avoiding the need to select a kernel bandwidth h.

[2]That is, assuming that the true structural equation is some linear function of the treatment and invalid instruments, and that all units share the same treatment effect parameter, β.

3. ModeIV

In this paper, we assume we have access to a set of k candidate variables, $\mathcal{Z} = \{z_1, \ldots, z_k\}$, which are 'valid' instrumental variables if they satisfy relevance, exclusion and unconfounded instrument, and are 'invalid' otherwise. Denote the set of valid instruments, $\mathcal{V} := \{z_i : z_i \not\perp\!\!\!\leftarrow t,\ z_i \perp\!\!\!\leftarrow \epsilon,\ z_i \perp\!\!\!\leftarrow y | x, t, \epsilon\}$, and the set of invalid instruments, $\mathcal{I} = \mathcal{Z} \setminus \mathcal{V}$. We further assume that each valid instrument identifies the causal effect. In the additive confounder setting, this amounts to assuming that the unobserved confounder's effect on y is additive, such that $y = f(t, x, z_{i:i\in\mathcal{I}}) + \epsilon\psi$for some function f and $E[y|x, z_{i:i\neq j}, z_j] = \int f(t, x, z_{i:i\neq j}) dF(t|x, z_{i:i\neq j}, z_j)$ has the same unique solution for all j in $\mathcal{Z}$.

The ModeIV procedure requires the analyst to specify a lower bound $V \geq 2$ on the number of valid instruments and then proceeds in three steps.

1. Fit an ensemble of k estimates of the conditional outcome $\{\hat{f}_1, \ldots, \hat{f}_k\}$ using a non-linear IV procedure applied to each of the k instruments. Each $\hat{f}$ is a function mapping treatment t and covariates x to an estimate of the effect of the treatment conditional on x.
2. For a given test point (t, x), select $[\hat{l}, \hat{u}]$ as the smallest interval containing V of the estimates $\{\hat{f}_1(t, x), \ldots, \hat{f}_k(t, x)\}$. Define $\hat{\mathcal{I}}_{\text{mode}} = \{i : \hat{l} \leq\!\!\!\leftarrow \hat{f}_i(t, x) \leq \hat{u}\}$ to be the indices of the instruments corresponding to estimates falling in the interval.
3. Return $\hat{f}_{\text{mode}}(t, x) = \frac{1}{|\hat{\mathcal{I}}_{\text{mode}}|} \sum_{i\in\hat{\mathcal{I}}_{\text{mode}}} \hat{f}_i(t, x)$

Figure 1 shows this procedure graphically. The idea is that the estimates from the valid instruments will tend to cluster around the true value of the effect, $E[y|\text{do}(t), x]$. We assume that the most common effect is a valid one; i.e., that

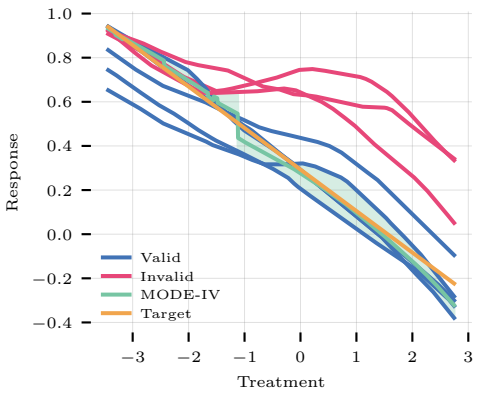

Figure 1. Example of the ModeIV algorithm with 7 candidates (4 valid and 3 invalid) from the biased demand simulation (see Section 4). The 7 estimators shown in the plot are each trained with a different candidate, and at every test point t, the mode of the 7 predictions is computed point-wise. The region highlighted in green contains the 3 predictions that formed part of the modal interval for each given input. The ModeIV prediction—the mean of the 3 closest prediction—is shown in solid green.

the modal effect is valid. To estimate the mode, we look for the tightest cluster of points which, by definition, are the points contained in $\hat{\mathcal{I}}_{\text{mode}}$. Intuitively, each estimate in this interval should be approximately valid and hence approximates the modal effect. Finally, we average these estimates to reduce variance.

The next theorem formalizes this intuition by showing that ModeIV asymptotically identifies and consistently estimates the causal effect.

Theorem 1. *Fix a test point (t, x) and let $\hat{\beta}_1, \ldots, \hat{\beta}_k$ be estimators of the causal effect of t at x corresponding to k (possibly invalid) instruments. E.g., $\hat{\beta}_j = \hat{f}_j(t, x)$. Denote the true effect as $\beta\psi = E[y|do(t), x]$. Suppose that*

1. *(consistent estimators) $\hat{\beta}_j \to \beta_j$ almost surely for each instrument. In particular, $\beta_j = \beta\psi$ whenever the jth instrument is valid.*

2. *(modal validity) At least v of the instruments are valid, and no more than $v - 1$ of the invalid instruments agree on an effect. That is, v of the instruments yield the same estimand if and only if all of those instruments are valid.*

Let $[\hat{l}, \hat{u}]$ be the smallest interval containing v of the instruments and let $\hat{\mathcal{I}}_{\text{mode}} = \{i : \hat{l} \le \hat{\beta}_i \le \hat{u}\}$. Then,

$$\sum_{i \in \hat{\mathcal{I}}_{\text{mode}}} \hat{w}_i \hat{\beta}_i \to \beta\psi$$

almost surely, where $\hat{w}_i, w_i$ are any non-negative set of weights such that each $\hat{w}_i \to w_i$ a.s. and $\sum_{i \in \hat{\mathcal{I}}_{\text{mode}}} w_i = 1$.

We defer all proofs to the supplementary material.

Of course, the ModeIV procedure can be generalized to allow estimators of the mode that are different from the one used in Steps 2 and 3. The key advantage of the Dalenius–Venter modal estimator is that the optimal choice for its only hyper-parameter, V, does not depend on the distribution of the estimators at a given test point. By contrast, kernel density-based modal estimators require tuning a length-scale parameter, where the optimal choice may vary as a function of the test point, (t, x). It is also straightforward to implement[3], and relatively insensitive to the choice of V. The procedure as a whole is, however, k times more computationally expensive than running single estimation procedure at both training and test time.

Despite its simplicity, ModeIV has strong point-wise worst-case guarantees. Theorem 2 shows that if each estimate is bounded,[4] then even in the worst case where $v - 1$ invalid candidates all agree on an effect, ModeIV converges at the same rate as the underlying estimators to the solution of an oracle that uniformly averages the valid instruments. In particular, if the estimators achieve the parametric rate, $1/\sqrt{n}$, in the number of instances n, then ModeIV also converges at $1/\sqrt{n}$.

Theorem 2. *For some test point (t, x), let $\mathcal{Z} = \{\hat{\beta}_1, \ldots, \hat{\beta}_k\}$ be k estimates of the causal effect of t at x. Assume,*

***[Bounded estimates]** Each estimate is bounded by some constants, $[a_i, b_i]$*

***[Convergent estimators]** Each estimator converges in mean squared error at a rate n^{-r} (where $r = \frac{1}{2}$ if the estimator achieves the parametric rate), and hence each estimator has finite variance, $Var(\hat{\beta}_i) = \frac{\sigma_i}{n^{-2r}}$ for some σ_i.*

Then, if $\sigma\psi = \max_{i \in \mathcal{V}} \sigma_i$ there exists a, C, such that $E[(\text{ModeIV}(\mathcal{Z}) - \beta)^2 - (\frac{1}{v} \sum_{i \in \mathcal{V}} \hat{\beta}_i - \beta)^2] \le 9kC\sigma n^{-r}$.

4. Experiments

We studied ModeIV empirically in two simulation settings. First, we investigated the performance of ModeIV for non-linear effect estimation as the proportion of invalid instruments increased for various amounts of direct effect bias. Second, we applied ModeIV to a realistic Mendelian randomization (MR) simulation to estimate heterogeneous treat-

[3] See the appendix for an efficient Pytorch (Paszke et al., 2019) implementation

[4] Boundedness is benign as long as we are not extrapolating too far outside of the range of data we observe: standard estimators do not typically make predictions for $E[y|\text{do}(t), x]$ outside of the range $[\min_i y_i, \max_i y_i]$ of observed y_i's.

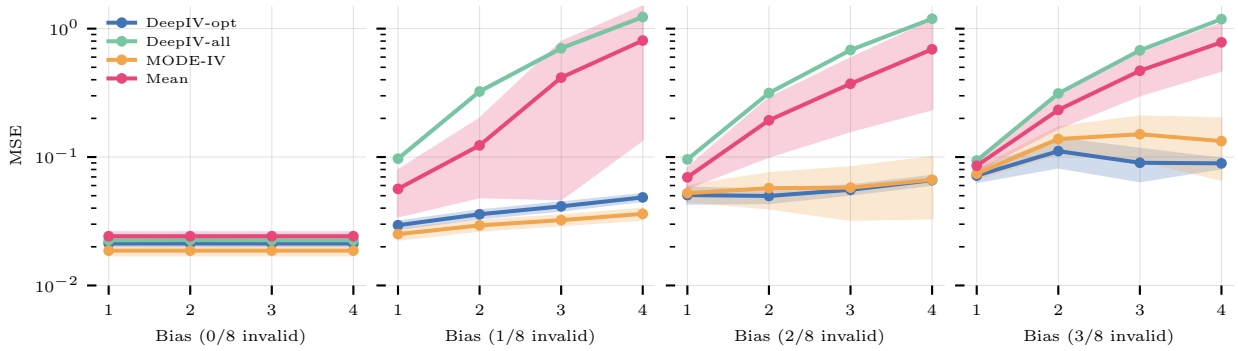

Figure 2. ModeIV is insensitive to the amount of exclusion violation bias. This figure shows performance on the biased demand simulation for various numbers of invalid instruments. The x-axis shows the amount of exclusion violation bias (introduced by scaling the $\gamma\psi$ parameter in the response equation below).

ment effects. For all experiments, we use DeepIV (Hartford et al., 2017) as the nonlinear estimator. The existing methods for addressing bias from invalid instruments are designed for the linear setting, so as baselines we compare to DeepIV with oracle access to the set of valid instruments (DeepIV-opt); the ensemble mean (Mean) which tests whether any performance improvements that we observe are driven by variance reduction from ensembling; and a naive approach that fits a single instance of DeepIV treating all instruments as valid (DeepIV-all). For the MR experiments we also compare to Guo et al. (2018). The heterogeneous effects in our MR simulation violates Guo et al.'s linearity assumption, but their method is designed with MR in mind so the comparison illustrates the effect of incorrectly assuming linearity in this setting. In the appendix, we evaluated ModeIV on Guo et al.'s linear MR data generating process; we found that as long as a large enough sample is used, it accurately recovered the true effect.

Biased demand simulation We evaluated the effect of invalid instruments on estimation by modifying the low dimensional demand simulation from Hartford et al. (2017) to include multiple candidate instruments. The demand simulation models a scenario where the treatment effect varies as a function[5], , of time, x_0, and other observed covariates x.

$$z_{1:k}, \nu\psi \sim \mathcal{N}(0,1) \quad x_0 \sim \text{unif}(0,10) \quad e \sim \mathcal{N}(\rho\nu, 1-\rho^2),$$
$$t = 25 + (z^T\beta^{(zt)} + 3) \quad (x_0) + \nu\psi$$
$$y = 100 + 10x_{1:d}^T\beta^{(x)} \quad (x_0) + \underbrace{(x_{1:d}^T\beta^{(x)} \quad (x_0) - 2)t}_{\text{Treatment effect}} + \underbrace{\gamma 60 \sin(z^T\beta^{(zy)})}_{\text{Exclusion violation}} + e$$

[5] $(x_0) = 2\left((x_0-5)^4/600 + e^{-4(x_0-5)^2} + x_0/10 - 2\right)$. See the appendix for a plot of the function and full details of the simulation.

We highlight the differences between this data generating process and the original in red: here we have k instruments whose effect on the treatment is parameterized by $\beta^{(zt)}$, instead of a single instrument in the original; we include an exclusion violation term which introduces bias into standard IV approaches whenever $\gamma\psi$ is non-zero. The vector $\beta^{(zy)}$ controls the direct effect of each instrument: invalid instruments have nonzero $\beta_i^{(zy)}$ coefficients, while valid instrument coefficients are zero.

We fitted an ensemble of k different DeepIV models that were each trained with a different instrument z_i. In Figure 2, we compare the performance of ModeIV with three baselines: DeepIV with oracle access to the set of valid instruments (DeepIV-opt); the ensemble mean (Mean); and a naive approach that fit a single instance of DeepIV treating all instruments as valid (DeepIV-all). The x-axis of the plots indicates the scaling factor γ, which scales the amount of bias introduced via violations of the exclusion restriction. All methods performed well when all the instruments were valid. Once the methods had to contend with invalid instruments, Mean and DeepIV-all performed worse than ModeIV because of both methods' sensitivity to the biased instruments. ModeIV's mean squared error closely tracked that of the oracle method as the number of biased instruments increased, and the raw mean squared errors of both methods also increased as the number of valid instruments in the respective ensembles correspondingly fell.

Sensitivity When using ModeIV, one key practical question that an analyst faces is choosing V, the lower bound on the number of valid instruments. We evaluated the importance of this choice in Figure 3 by testing the performance of ModeIV across the full range of choices for V with different numbers of biased instruments. We found that, as expected, the best performance was achieved when V was the true number of valid instruments, but also that similar levels

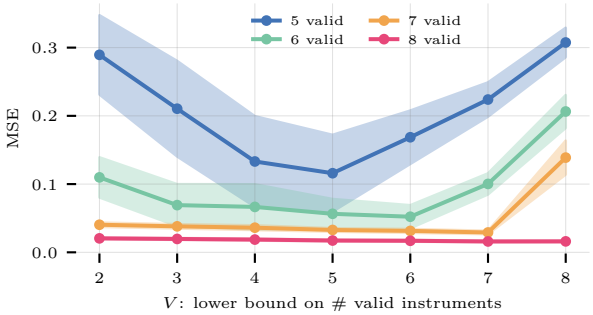

Figure 3. ModeIV's sensitivity to the choice of number of valid instruments parameter V. Best performance is achieved when V is equal to the true number of valid instruments, but the method is relatively insensitive to more conservative choices of V.

of performance could be achieved with more conservative choices of V. That said, with only 5 valid instruments, ModeIV tended to perform worse when V was set too small. For an illustration of why this occurs, consider Figure 1 which visualizes the ModeIV procedure. In the figure, there are a number of regions of the input space where the invalid instruments agreed by chance (e.g. $t \in [-1, 0.]$), so these regions bias ModeIV for small mode set sizes. Overall, we observed that setting $V = \lfloor k/2 \rfloor$ (where k is the number of instruments) tended to work well in practice.

Asymptotically, ModeIV remains consistent when fewer than half of the instruments are valid, but when this is the case there are far more ways that Assumption 2 of Theorem 1 can be violated. This is illustrated in Figure 1 which shows that there are a number of regions where the bias instruments agree by chance. Because of this, we recommend only using ModeIV when one can assume that the majority of instruments are valid, unless one has prior knowledge to justify *modal validity* without assuming the majority of instruments are valid.[6]

Bootstrap inference When using deep learning-based estimators, one typically does not have closed form expressions for confidence intervals or knowledge of the joint distribution over estimators, so we evaluated the performance of ModeIV with bootstrap confidence intervals. Table 1 summarizes the results. On this simulation we found that bootstrap confidence intervals with ModeIV were reasonably accurate as long as V was set low enough: with $V = 2$ or 3, coverage was above 90% for 95% confidence intervals. With larger settings of V, we found worse performance as the narrower intervals did not account for occasional selection of biased instruments. Figure 4 shows a plot of the bootstrap confidence intervals for both the average dose-response curve and various conditional averages. The intervals are narrow enough that they show the true dose-response curve, while still providing reasonable coverage.

[6]For example, if direct effects are strictly monotone and disagree, chance agreements among invalid instruments can only occur in a finite number of locations.

	4/7 valid	5/7 valid	6/7 valid
ModeIV-2	90.26%	94.79%	94.27%
ModeIV-3	87.82%	92.13%	92.79%
ModeIV-4	85.30%	90.10%	91.16%
ModeIV-5	72.80%	85.88%	88.06%
ModeIV-6	51.87%	70.63%	83.80%
ModeIV-7	34.12%	45.05%	66.44%
DeepIV-All	16.85%	18.48%	19.89%
DeepIV-Opt	95.07%	95.85%	95.79%
Ens-Mean	34.37%	45.29%	66.65%

Table 1. ModeIV attains reasonable point-wise coverage for bootstrap 95% confidence intervals on the biased demand simulation.

Mendelian randomization simulation For the second experiment, we evaluated our approach on simulated data adapted from Hartwig et al. (2017), which is designed to reflect violations of the exclusion restriction in Mendelian randomization studies.

Instruments, z_i, represent SNPs—locations in the genetic sequence where there is frequent variation among people—modeled as random variables drawn from a Binomial$(2, p_i)$ distribution corresponding to the frequency with which an individual gets one or both rare genetic variants. The treatment and response are both continuous functions of the instruments with Gaussian error terms. The strength of the instrument's effect on the treatment, α_i, and direct effect on the response, δ_i, are both drawn from Uniform$(0.01, 0.2)$ distributions for all i. For all experiments we used 100 candidate instruments and varied the number of valid instruments from 50 to 100 in increments of 10; we set δ_i to 0 for all valid instruments. More formally,

$$z_i \sim \text{Binomial}(2, p_i) \quad \beta(x) := \text{round}(x^T \gamma^{(xt)}, 0.1).$$

$$t := \sum_{j=1}^{K} \alpha_j z_j + \rho u + \epsilon_x, \;\; y := \beta(x)t + \sum_{j=1}^{K} \delta_j z_j + u + \epsilon_y$$

In the original Hartwig et al. simulation, the treatment effect β was fixed for all individuals. Here, we make the treatment effect vary as a function of observable characteristics to model a scenario where treatments may affect different subpopulations differently. We simulate this by making the treatment effect, $\beta(x)$, a sparse linear function of observable characteristics, $x \in \mathbb{R}^{10}$, where 3 of the 10 coefficients, $\gamma_i^{(xt)}$ were sampled from $U(0.2, 0.5)$ and the remaining $\gamma_i^{(xt)}$ were set to 0. We introduce non-linearity by rounding to the nearest 0.1, which makes the learning problem harder,

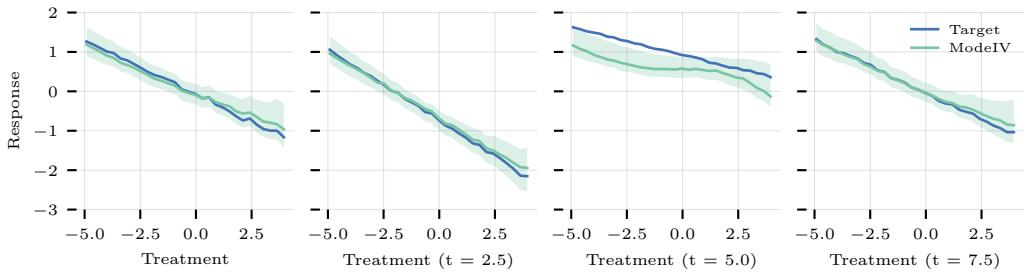

Figure 4. Bootstrap 95% confidence interval for the (conditional) dose-response curve of ModeIV with $V = 3$ for the biased demand simulation. The left plot shows the unconditional curve, and the remaining three curves are conditioned on the time variable, t.

while making it easier to visually show differences between the fitted functions and their targets.

Mendelian randomization problems tend to have low signal-to-noise ratios because typical response variables tend to be influenced by a large number of unobserved factors; in this simulation, the treatment explains only 1-3% of the response variance. This makes the setting challenging for neural networks, which tend to perform best on low-noise regimes. To address this, we leveraged the inductive bias that the data is conditionally linear in the treatment effect, by using a neural network to parameterize the slope of the treatment variable rather than outputting the response directly. So, for these problems, we defined $\hat{f}(t, x) = g(\phi(x))t + h(\phi(x))$, where $g(\cdot)$ and $h(\cdot)$ are linear layers that act on a shared representation $\phi(x)$.

Among the DeepIV-based benchmarks, the general trends that we observed on the Mendelian randomization simulation—summarized in Table 2—were similar to those we observed in the biased demand simulation: DeepIV-all performed poorly and ModeIV closely tracked the performance of our oracle, DeepIV-opt. On this simulation the mean ensemble (Mean) achieved stronger performance, but still did not match ModeIV.

Aside from the heterogeneity induced by $\beta(x)$, this data generating process is linear so we can use it to evaluate the effect of heterogeneity on methods that assume a constant linear treatment effect. Guo et al.'s Two-Stage Hard Thresholding (TSHT) accurately recovered the average treatment effect (ATE) on this problem (see Table 5 in the appendix), but as Table 2 shows, it was not able to match the performance of ModeIV in predicting $E[y|\text{do}(t), x]$. Note that this is a challenging baseline: with the sample size used in this simulation (400 000), Guo et al.'s method has very few false positives in identifying the valid instruments (see Table 5 in the appendix), so it is essentially running two-stage least squares on a linear model with a 'random' (from the perspective of the model) coefficient $\beta(x)$. Linear models are optimal in this setting, so ModeIV can only outperform TSHT by leveraging the interaction between x and β. That said, there is a trade-off: TSHT was unbiased in predicting the ATE, but both DeepIV-Opt and ModeIV picked up some bias: DeepIV-Opt over-estimated the conditional average treatment effect by 0.035 and ModeIV by 0.045 for true effect sizes that range between -0.3 and 0.3 (see Table 3 in the appendix). This bias is small but significant, and is also reflected in lower coverage from bootstrap confidence intervals. We found that when targeting a 90% confidence interval, DeepIV-Opt achieved only 80% coverage and ModeIV only managed 60% (Table 4 in the appendix).

Conditional average treatment effects and bootstrap inference Figure 5 shows the predicted dose–response curves for a variety of different levels of the true treatment effect. The six plots correspond to six different subspaces of x that all have the same true conditional treatment effect. Each of the light blue lines shows ModeIV's prediction for a different value of x. The model is not told that the true $\beta\psi$ is constant for each of these sub-regions, but instead has to learn that from data so there is some variation in the slope of each prediction. Despite this, the majority of predicted curves match the sign of the treatment effect for each subgroup of x and accurately predicted the relative differences between the subgroups.

5. Limitations

Local average treatment effects. The key assumption that ModeIV relies on is that each valid instrument consistently estimates the same function, $f(t, x)$. In settings with discrete treatments, one typically only identifies a "(conditional) local average treatment effect" (CLATE / LATE respectively) for each instrument. The LATE for instrument i can be thought of as the average treatment effect for the sub-population that changes its behavior in response to a change in the value of instrument i; if the LATEs differ across instruments, this implies that each instrument will result in a different estimate of $E[\hat{f}_i(t, x)]$ regardless of

Model	50% valid	60% valid	70% valid	80% valid	90% valid	100% valid
DeepIV (valid)	0.035 ± (0.001)	0.035 ± (0.001)	0.034 ± (0.001)	0.034 ± (0.001)	0.032 ± (0.0)	0.024 ± (0.001)
MODE-IV 30%	0.037 ± (0.001)	0.037 ± (0.001)	0.038 ± (0.001)	0.039 ± (0.001)	0.041 ± (0.001)	0.032 ± (0.001)
MODE-IV 50%	0.037 ± (0.001)	0.037 ± (0.001)	0.038 ± (0.001)	0.039 ± (0.001)	0.04 ± (0.001)	0.032 ± (0.001)
Mean	0.041 ± (0.001)	0.041 ± (0.001)	0.043 ± (0.001)	0.043 ± (0.001)	0.045 ± (0.001)	0.036 ± (0.001)
DeepIV (all)	0.099 ± (0.004)	0.116 ± (0.003)	0.149 ± (0.005)	0.149 ± (0.005)	0.142 ± (0.003)	0.025 ± (0.0)
TSHT	0.089 ± (0.005)	0.075 ± (0.003)	0.073 ± (0.003)	0.073 ± (0.003)	0.072 ± (0.003)	0.072 ± (0.003)

Table 2. Performance on the Mendelian randomization simulation for various proportions of valid instruments. The ensemble methods performed far better than the DeepIV model, which treated all instruments as valid, and ModeIV, which gave significantly better performance than the mean ensemble, was close to the performance of DeepIV on the valid instruments.

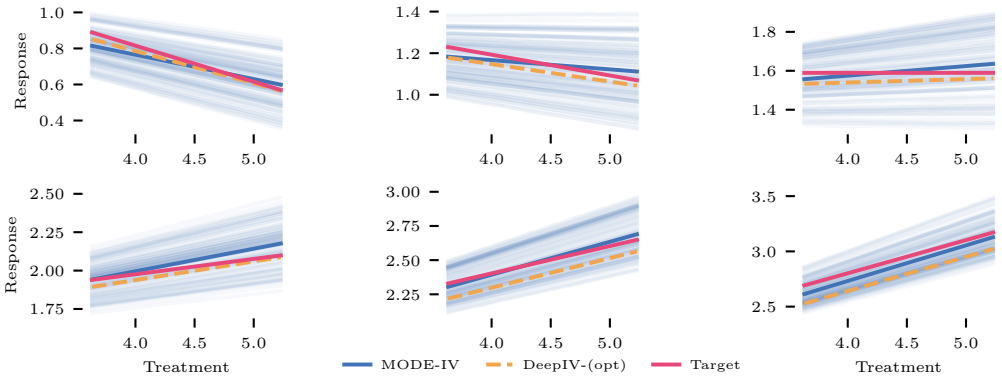

Figure 5. Estimated conditional dose–response curves for the Mendelian randomization simulation. Each light blue curve shows ModeIV's estimate $f(t, x)$ for some IID sample of x; each figure's dark curve represents the average over all samples of x. The six plots show the different subsets of the range, x, where true slope $\beta(x)$ is *(left to right)* $-0.2, -0.1, 0., 0.1, 0.2$ and 0.3 respectively.

whether any of the instruments are invalid. In such settings, ModeIV will return the average of the V closest $\hat{f}_i(t, x)$'s, but one would need additional assumptions on how these estimates cluster relative to biased estimates to apply any causal interpretation to this quantity. The alternative is the approach that we take here: assume that a common function $f(t, x)$ is shared across all units and allow for heterogeneous treatment effects by allowing the treatment effect to vary as a function of observed covariates x. This shared heterogeneous effect assumption is weaker than prior work on robust IV, which requires a "constant effect" effect assumption that every individual responds in exactly the same way to the treatment via a parameter, β.

Selecting instruments? ModeIV constitutes a consistent method for making unbiased predictions but, somewhat counter-intuitively, it does not directly offer a way of inferring the set of valid instruments. For example, one might imagine identifying the set of candidates that most often form part of the modal interval $\hat{\mathcal{I}}_{\text{mode}}$. The problem is that while candidates that fall within the modal interval $\hat{\mathcal{I}}_{\text{mode}}$ tend to be close to the mode, the interval can include invalid instruments that yielded an effect close to the mode by chance. Since these invalid estimates are close to the truth, they do not hurt the estimate. We can see this in Figure 1 where invalid instruments form part of the modal interval in the region $t \in [-3.5, -2]$, without introducing bias.

Relaxing independence of instrumental variables. We assume that each of the valid instruments is unconfounded. This is easiest to achieve in settings where each instrument is independent. This setting is shown in Figure 6 (*left*) where we have k candidates, $\{z_i : i \in 1, \ldots, k\}$, some of which are valid, and some of which are invalid (e.g. z_k shown in pink has a direct effect on the response). This independent candidates setting is most common where the instruments are explicitly randomized: e.g. in judge fixed effects where the selection of judges is random.

A more complex setting is shown in Figure 6 (*right*). Here, the candidates share a common cause, u. In this scenario, if u is not observed, each of the previously valid instruments (e.g. z_1, z_2 and z_{k-1} in the figure) are no longer valid

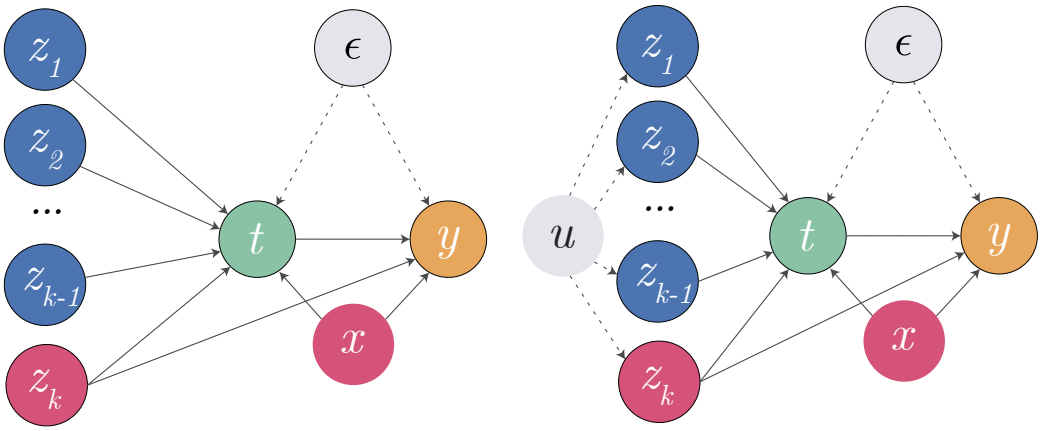

Figure 6. When each of the instruments is independent *(left)*, the unconfounded instrument assumption is easily satisfied for the valid instruments so ModeIV applies. When the instruments share a common cause *(right)*, care is needed to ensure that we block the path $z_{\text{valid}} \quad u \to z_{\text{invalid}} \to y$. See the limitations discussion above for details.

because they fail the unconfounded instrument assumption via the backdoor path $z_1 \quad u \to z_k \to y$. However, if we condition on all the candidates that have a direct effect on y and treat them as observed confounders, we block this path which allows for valid inference. Of course we do not know which of the candidates have a direct effect, so when building an ensemble, for each candidate z_i, we treat all $z_{j \neq i}$ as observed confounds to block these potential backdoor paths. This addresses the issue as long as there is not some z_{k+1} which is not part of our candidate set, but nevertheless opens up a backdoor path $z_1 \quad u \to \not{z}_{k+1} \to \not{y}$. If u is observed, we can simply control for it. This suggests a natural alternative approach would be to try to estimate u and control for its effect, using an approach analogous to Wang & Blei (2019).

6. Discussion

The conventional wisdom for IV analysis is: if you have many (strong) instruments and sufficient data, you should use all of them so that your estimator can maximize statistical efficiency by weighting the instruments appropriately. This remains true in our setting—indeed, DeepIV trained on the valid instruments typically outperformed any of the ensemble techniques—but of course requires a procedure for identifying the set of valid instruments. In the absence of such a procedure, falsely assuming that all candidate instruments are valid can lead to large biases, as illustrated by the poor performance of DeepIV-all. ModeIV gives up some efficiency by filtering instruments, but it gains robustness to invalid instruments with strong worst case asymptotic guarantees, and in practice we found that the loss of efficiency was negligible. Of course, that empirical finding will vary across settings. A useful future direction would find a procedure for recovering the set of valid instruments to further reduce the efficiency trade-offs.

Acknowledgements

This work was supported by Compute Canada, a GPU grant from NVIDIA, an NSERC Discovery Grant, a DND/NSERC Discovery Grant Supplement, a CIFAR Canada AI Research Chair at the Alberta Machine Intelligence Institute, and DARPA award FA8750-19-2-0222, CFDA# 12.910, sponsored by the Air Force Research Laboratory.

References

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, pp. 3564–3574, 2019.

Bowden, J., Davey Smith, G., and Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 06 2015.

Breiman, L. Bagging predictors. *Machine Learning*, 24(2): 123–140, 1996.

Chernoff, H. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41, December 1964.

Dalenius, T. The Mode–A Neglected Statistical Parameter. *Journal of the Royal Statistical Society. Series A (General)*, 128(1):110, 1965.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. Nonparametric instrumental regression. *Econometrica*, 79(5): 1541–1565, 2011.

Davey Smith, G. and Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, 2003.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, 2020.

Frandsen, B. R., Lefgren, L. J., and Leslie, E. C. Judging judge fixed effects. Technical report, National Bureau of Economic Research, 2019.

Freund, Y. and Schapire, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pp. 23–37. Springer, 1995.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054, 1982.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1414–1423. JMLR. org, 2017.

Hartwig, F. P., Davey Smith, G., and Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, 46(6):1985–1998, 2017.

Hemani, G., Bowden, J., and Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics*, 27(2):195–208, 2018.

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

Kling, J. R. Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876, 2006.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015.

Kourentzes, N., Barrow, D. K., and Crone, S. F. Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, 41(9):4235–4244, July 2014.

Kuang, Z., Sala, F., Sohoni, N., Wu, S., Córdova-Palomera, A., Dunnmon, J., Priest, J., and Re, C. Ivy: Instrumental variable synthesis for causal inference. volume 108 of *Proceedings of Machine Learning Research*, pp. 398–410, Online, 26–28 Aug 2020. PMLR.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression, 2020.

Mueller-Smith, M. The criminal and labor market impacts of incarceration. *Unpublished Working Paper*, 18, 2015. URL https://sites.lsa.umich.edu/mgms/wp-content/uploads/sites/283/2015/09/incar.pdf.

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5): 1565–1578, 2003.

Parzen, E. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3): 1065–1076, September 1962.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Puli, A. M. and Ranganath, R. Generalized control functions via variational decoupling. In *Advances in Neural Information Processing Systems*, 2020.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pp. 4593–4605, 2019.

Stock, J. H. and Watson, M. W. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, September 2004.

Swanson, S. A. and Hernán, M. A. The challenging interpretation of instrumental variable estimates under monotonicity. *International journal of epidemiology*, 47(4): 1289–1297, 2018.

Venter, J. H. On Estimation of the Mode. *The Annals of Mathematical Statistics*, 38(5):1446–1455, October 1967.

Wang, Y. and Blei, D. M. The blessings of multiple causes. *Journal of the American Statistical Association*, pp. 1–71, 2019.