Identifying Valid Instruments via Effect Agreement

Jason Hartford
Department of Computer Science
University of British Columbia
Vancouver, Canada
jasonhar@cs.ubc.ca

Kevin Leyton-Brown
Department of Computer Science
University of British Columbia
Vancouver, Canada
kevinlb@cs.ubc.ca

Abstract

Instrumental variable methods are powerful, but rely on strong and untestable assumptions. In particular, it can be difficult to defend the exclusion restriction, which requires that there exist no unblocked direct paths from instrument to outcome, because there may be many plausible but unlikely ways for the instrument to affect the outcome directly. We consider settings in which there exist multiple candidate instruments but only a subset are valid. Our main result shows how to identify the set of valid instruments under an "effect agreement" assumption that requires that direct instrument effects do not offset the instrument's indirect effects that are mediated by the treatment. Leveraging this result, we give a practical backward-selection algorithm for estimating a set of valid instruments and show empirically that (1) we can identify sets of valid instruments with the number of false positives decreasing with data set size and (2) that the resulting parameter estimates compare favourably with those of an oracle that knows in advance which of the instruments are valid.

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 and Powell, 2003; Darolles *et al.*, 2011; Hartford *et al.*, 2017; Lewis and Syrgkanis, 2018; Singh *et al.*, 2019; Bennett *et al.*, 2019], thereby allowing nonlinear estimation of heterogeneous treatment effects.

However, some of the key assumptions that support IV approaches are not testable. IV methods all assume:

1. *Relevance:* the treatment is not independent of the instrument.
2. *Exclusion:* the instrument's effect on the outcome is entirely mediated through the treatment (i.e., there are no unblocked direct paths from instrument to outcome).
3. *Unconfounded instrument:* the instrument and outcome do not share any common causes (i.e., there are no unblocked back-door paths between the instrument and outcome).

Relevance is the only one of these assumptions that is testable. The unconfounded instrument assumption is often justified by appealing to knowledge of the system (e.g. the instrument may be explicitly randomized or may be the result of some well understood random process), but the exclusion restriction assumption is often difficult to defend.[1]

[1] Discrete treatment variables also require a monotonicity assumption to estimate 'local average treatment effects' (see [Angrist and Pischke, 2008, Section 4.4.1] for details). The methods presented here have the same requirements and causal interpretation.

"Do the right thing": machine learning and causal inference for improved decision making workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Weaker assumptions are required in the presence of multiple potential instruments, some of which are valid and some of which fail to satisfy the exclusion restriction. In this case, exclusion becomes testable [Frandsen *et al.*, 2019], and there exist methods that give consistent estimation of average treatment effects under weaker assumptions than exclusion. For example, the mode of treatment effect estimates will typically be unbiased because the subset of valid instruments will all estimate the same treatment effect [Hartwig *et al.*, 2017]; similarly, assuming independence of direct and indirect instrument effects enables consistent treatment effect estimates [Kolesár *et al.*, 2015].

These approaches are appealing because they offer the promise of automated instrument variable methods: one can search for relevant instruments among a set of variables that we can assume to meet the unconfounded instrument assumption and then simply assume some "reasonable" fraction of them meet the exclusion restriction in order to estimate casual effects. Indeed, this approach is already growing in popularity in the epidemiology literature: Mendelian randomization studies use different genetic variations as instruments and apply this automated approach [Hemani *et al.*, 2017]. However, these existing approaches treat each instrument estimate independently, and test implications and estimate parameters by leveraging the fact that only the valid instruments will agree on the true relationship. In this work we ask a stronger question: given multiple potentially valid instruments, can we identify which are valid?

In this work, we show that the subset of valid instruments is asymptotically identified for linear models[2] if we make an "effect agreement" assumption which says that direct instrument effects share the same sign as indirect treatment effects. To see the implications of this assumption, consider the following examples from economics and epidemiology where multiple potentially valid instruments are available,

- **Judge fixed effects** research designs use random assignment of trial judges as an instrument and leverage differences between different judges's propensity to incarcerate to infer the effect of incarceration on some outcome of interest. 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. The "effect agreement" assumption in this setting amounts to assuming that all of these latent acts by judges amount to different forms of punishment having the same effect direction on the outcome as the treatment.

- **Mendelian randomization** studies use genetic variation to study the effects of some exposure on an outcome. For example, some genes are associated with heavier smoking. If exclusion holds, we could use expression of these genes as an instrument to test the effect of smoking on lifespan. However, if some of these genes also affect lifespan directly (for example, if they are also associated with other forms of substance addiction that are not observed), the effect agreement assumption requires that their direct effect also leads to shorter lifespans. Conversely, if the same genes are also associated with an unobserved addiction to exercise which offsets the negative effect of smoking, then effect agreement is not satisfied.

Thus effect agreement amounts to assuming that direct effects do not offset the indirect effect of the instrument via the treatment. We note that this assumption can be viewed as as stronger version of the faithfulness assumption [Spirtes *et al.*, 2000] which is used in the causal discovery literature. In this context, faithfulness amounts to assuming that the direct and indirect effects don't cancel each other out exactly (rather than just agreeing on signs); here we need the stronger effect agreement assumption because we allow for unobserved confounding.

Given our identification result, we present a practical, greedy backward-selection algorithm for estimating a set of valid instruments. We show experimentally (1) that our algorithm identifies sets of valid instruments with the number of false positives decreasing with data set size and (2) that the parameter estimates that result from running two-stage least squares using our estimated instruments compare favourably with those of an oracle that knows in advance which of the instruments are valid.

[2]Linearity is convenient but we believe that similar approaches will be possible under weaker assumptions.

2 Methodology

We assume that our data are drawn from the following data generating process,

$$
\begin{aligned}
&z_i \sim \mathcal{D}_i \quad \text{for } i \text{ in } [1 \dots K], \quad u, \epsilon_x, \epsilon_y \sim \mathcal{D}_i \quad \text{for } i \text{ in } [u, x, y] \\
&x \leftarrow \sum_{j=1}^{K} \alpha_j z_j + \rho u + \epsilon_x \\
&y \leftarrow \beta x + \sum_{j=1}^{K} \delta_j z_j + u + \epsilon_y = \sum_{j=1}^{K} (\beta\alpha_j + \delta_j) z_j + (1+\rho) u + \epsilon_y \\
&\delta_i = 0 \text{ for all } i \in \mathcal{V},\ \delta_i \neq 0 \text{ for all } i \in \mathcal{I},\ \mathcal{V} \cup \mathcal{I} = \{1, \dots, K\}, |\mathcal{V}| \geq 1 \\
&\text{Effect agreement: } (\beta\alpha_j)(\delta_j) \geq 0 \text{ for all } j,
\end{aligned} \tag{1}
$$

where the instruments z_i and error terms ϵ_x, ϵ_y are each generated independently according to some distribution $\mathcal{D}_i$. The unobserved confounder u induces a correlation between x and y parameterized by ρ. $\mathcal{V}$ denotes the index set of valid instruments; $\mathcal{I}$ denotes the invalid instruments; and the true causal effect to be estimated is β. Finally, Equation 1 formalizes the effect agreement assumption that direct effects do not offset indirect effects.

Now consider the regression sum of squares statistic that we evaluate after regressing each of x and y on a candidate set of instruments, $\mathcal{C}$,

$$
s(y, Z; \mathcal{C}) = \frac{1}{n} \sum_{i=1}^{n} \left[(\bar{y} - y_i)^2 - \left(\sum_{j \in \mathcal{C}} \hat{\alpha}_j^y z_{i,j} - y_i \right)^2 \right],
$$

where $s(x, Z; \mathcal{C})$ is defined analogously for the treatment x. Call $\phi(\mathcal{C}) = \frac{s(y,Z;\mathcal{C})}{s(x,Z;\mathcal{C})}$ the *score* of a candidate set of instruments.

Proposition 1. *Under the assumptions given above and in the infinite data limit, if $\hat{\alpha}_j^x$ and $\hat{\alpha}_j^y$ are fit using ordinary least squares, $\phi(\mathcal{C})$ is minimized when $\mathcal{C}$ is any non-empty subset of the valid instruments, $\mathcal{V}$.*

Proof. In the infinite data limit, $\hat{\alpha}_j^x \xrightarrow{p} \alpha_j$ and $\hat{\alpha}_j^y \xrightarrow{p} (\beta\alpha_j + \delta_j)$ converge to their expected values, and the expected value of each of $s(x, Z; \mathcal{C})$ and $s(y, Z; \mathcal{C})$ scores are given by

$$
E[s(x, Z; \mathcal{C})] = \sum_{i \in \mathcal{C}} (\alpha_i)^2 \sigma_{z_i}^2 \qquad E[s(y, Z; \mathcal{C})] = \sum_{i \in \mathcal{C}} (\beta\alpha_i + \delta_i)^2 \sigma_{z_i}^2.
$$

Now, notice that for any set of valid instruments $\mathcal{V}' \subseteq \mathcal{V}$, $\phi(\mathcal{V}') = \frac{E[s(y,Z;\mathcal{V}')]}{E[s(x,Z;\mathcal{V}')]} = \beta^2$ since $\delta_i = 0$ for all i in $\mathcal{V}$. Furthermore, for any invalid instrument $i \in \mathcal{I}$, $\phi(\{i\}) > \beta^2$ since

$$
\phi(\{i\}) = \frac{(\beta\alpha_i + \delta_i)^2}{\alpha_i^2} > \frac{(\beta\alpha_i)^2}{\alpha_i^2} \text{ because } |\beta\alpha_i + \delta| > |\beta\alpha_i| \text{ by effect agreement.}
$$

Hence for any $\mathcal{C} \not\subseteq \mathcal{V}$, $\phi(\mathcal{C}) > \phi(\mathcal{V}')$ for all $\mathcal{V}' \subseteq \mathcal{V}$ as required. □

Corollary 1.1. *The set of valid instruments, $\mathcal{V}$, is uniquely identified as the largest subset of instruments that minimizes $\phi(\cdot)$. That is, $\mathcal{V} = \arg\max_{\mathcal{C} \text{ s.t. } \phi(\mathcal{C}) = \min_{\mathcal{C}'} \phi(\mathcal{C}')} |\mathcal{C}|$.*

Corollary 1.1 establishes that the valid set of instruments can be identified, but only in the asymptotic limit of the $\phi(\mathcal{C})$ for all $2^K - 1$ non-empty subsets $\{1, \dots, K\}$.

Greedy backward IV selection. Algorithm 1 gives a practical alternative to the asymptotic idea presented in Corollary 1.1. We replace $\phi(\cdot)$ with its empirical analogue, $\hat{\phi}(\cdot)$, and we perform the combinatorial optimization procedure implied by the $\arg\min$ by running greedy backward selection on $\hat{\phi}(\cdot)$. At each step, the algorithm greedily removes the candidate instrument that leads to the largest reduction in the empirical scoring function. It terminates either when no further progress can be made or when the algorithm is left with a prespecified minimum number of instruments.

Algorithm 1: Greedy backward IV selection

Input : (Z, x, y) a set of candidate instruments, treatment and response, γ min valid instruments
Output : A set of instruments estimated to be valid

$s^* \leftarrow \hat{\phi}(x, y, Z)$
while $|Z| > \lceil |Z| * \gamma \rceil$ **do**
  forall $i \in Z$ **do**
    $s_i \leftarrow \hat{\phi}(x, y, Z \setminus i)$
  end
  if $\min_i s_i > s^*$ **then**
    return Z
  else
    $s^* = \min_i s_i$
    $Z \leftarrow Z \setminus j$ where $j = \arg\min_i s_i$
  end
end
return Z

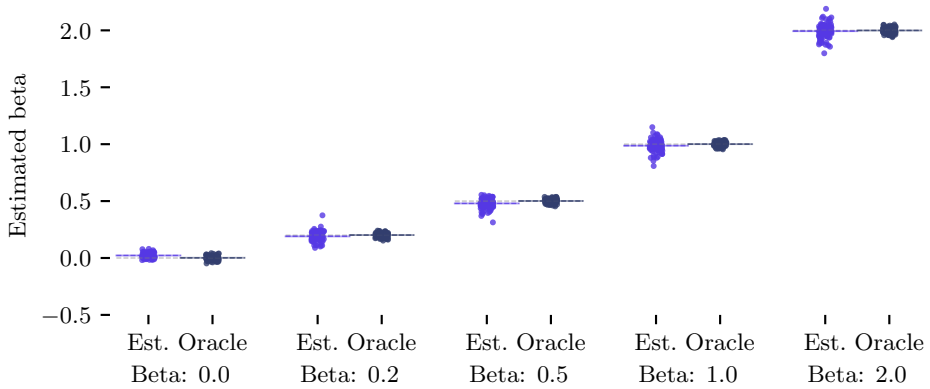

Figure 1: Estimated β from two stage least squares using the instruments returned by Algorithm 1 (labeled 'Est') and the true valid instruments (labeled 'Oracle') on 100 000 data points. For this example only 10 out of the 30 instruments were valid and the γ hyper-parameter in Algorithm 1 was set such that the minimum number of valid instruments was 5.

3 Simulated experiments

We evaluate our approach on the simulated data from Hartwig *et al.* [2017], which is designed to reflect violations of the exclusion restriction in Mendelian randomization studies. We refer the reader to Hartwig *et al.* for full details of the simulation, only noting important details here. The data generating process matches the structural equations given in Section 2. Instruments, z_i, are discrete random variables drawn from a Binomial$(2, p)$ distribution with $p \sim$ Uniform$(0.1, 0.9)$. The treatment and response are both continuous functions of the instruments with Gaussian error terms. Both α_i and δ_i are drawn from Uniform$(0.1, 0.9)$ distributions for all i. For all experiments we use 30 candidate instruments and vary the number of valid instruments from 5 to 25 in increments of 5; we set δ_i to 0 for all valid instruments.

Every experiment was run with 100 random seeds. We tested data set sizes of 1000, 10 000 and 100 000 and let β values vary between 0 and 2.

Average treatment effect estimation In order to evaluate the effect of the approach on the estimate of β, we compared the performance of two stage least squares using the instruments returned by Algorithm 1 to the true valid instruments. Figure 1 gives typical performance for β values from 0 to 2. We observed that while the estimated instruments exhibited larger variance, they were unbiased with respect to the true β. For smaller values of n we observed a small negative bias. Figure 2 *(left)* summarizes the performance we observed vs number of observations for different numbers of valid

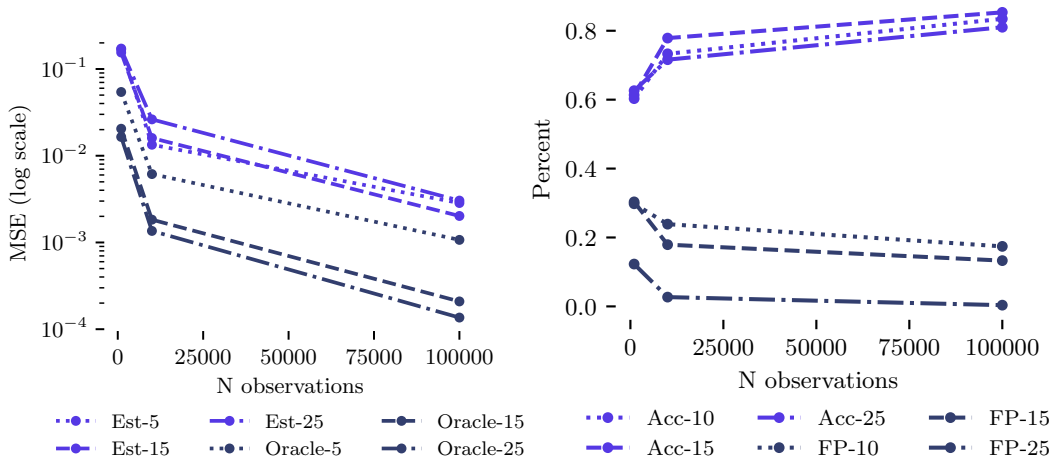

Figure 2: *(left)* Mean squared error between estimated $\hat{\beta}$ and true β as a function of data set size. The numbers in the legend indicate the number of valid instruments (out of 30). *(right)* Accuracy and false positive rates for the identification of instruments. Notice that while false positive rates were not driven to zero at the sample sizes we considered, invalid instruments introduced relatively small bias.

instruments. While our method gave rise to clearly worse MSE scores than the oracle, its performance was relatively invariant to the number of biased instruments (e.g., it achieved similar performance with 5 valid instruments and with 25 valid instruments).

Instrument identification Figure 2 *(right)* shows the performance of Algorithm 1 for identifying instruments. As the sample size grew from 1000 to 100 000 observations, accuracy increased from about 60% to about 80%. While we observed false positive rates above zero for problems with a large number of instruments, we found that these biased instruments did not bias the estimated treatment effect much. Intuitively, this occurred because these instruments were selected by the algorithm precisely because they had small values of δ; invalid instruments with small direct effects have similar scores as valid instruments with no direct effects, but the former also introduce correspondingly small amounts of bias to the parameter estimate.

4 Summary and conclusions

In this extended abstract we presented an approach for estimating causal effects given a candidate set of instruments, only some of which are valid. We leveraged an effect agreement assumption and showed that under this assumption, the set of valid instruments is identified. Clearly the next step for this work is extending these results beyond the linear case to see how they may be combined with more flexible nonlinear IV methods.

References

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

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *arXiv preprint arXiv:1905.12495*, 2019.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Brigham R Frandsen, Lars J Lefgren, and Emily C Leslie. Judging judge fixed effects. Technical report, National Bureau of Economic Research, 2019.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1414–1423. JMLR. org, 2017.

Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, 46(6):1985–1998, 2017.

Gibran Hemani, Jack Bowden, Philip C Haycock, Jie Zheng, Oliver Davis, Peter Flach, Tom R Gaunt, and George Davey Smith. Automating mendelian randomization through machine learning to construct a putative causal map of the human phenome. *BioRxiv*, page 173682, 2017.

Michal Kolesár, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

Michael Mueller-Smith. The criminal and labor market impacts of incarceration. *Unpublished Working Paper*, 18, 2015.

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

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *arXiv preprint arXiv:1906.00232*, 2019.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.