Learning causal models from existing randomized experiments: Meta-analysis using regularized instrumental variables

Alexander Peysakhovich
Facebook Artificial Intelligence Research
770 Broadway
New York, NY 10003
alexpeys@fb.com

Dean Eckles
MIT Sloan School of Management
Cambridge, MA 02139
dean@deaneckles.com

1. INTRODUCTION

Most of machine learning is the study of *prediction* [9]. On the other hand, the study of causality is the study of *prediction under intervention* [6, 5, 13, 11]. This often, but not always, requires different strategies depending on where the intervention occurs. For example, we may know that the number of ice cream cones sold in a given day predicts the number of drownings in swimming pools on that day. However, even knowing this correlation, if we want to make policies which reduce drownings, it is likely that banning ice cream will not achieve our goals. Thus, in this case, we want to learn which variables *cause* drownings, not merely learn drownings' conditional probability density.

In this extended abstract, we discuss how meta-analysis of many existing randomized experiments (which themselves were, perhaps, done for completely other purposes) can allow us to learn complex causal graphs. We then connect this idea to work on reinforcement learning and discuss how it can be a complement to those approaches.

A large corpus of statistical methods for learning to predict under interventions have grown up in the social and biomedical sciences over the last few decades ([11, 2]). These methods work with the following model of the world. There is a variable X which is the variable we are interested in later intervening on, often called the causal variable (this may be vector valued). This variable is described by the following structural equation:

$$X = \epsilon + U\psi.$$

There is a variable y which is the outcome we are trying to achieve which is generated by the linear structural equation

$$Y = X\beta + U\gamma + \eta.$$

The focus on the linear model is justified either because it is the simplest possible model to solve or because we are interested in relatively small interventions.[1]

[1]For this abstract assume all variables are centered and have well behaved distributions.

NIPS

DOI: NA

The focus of causal inference is to learn the parameter vector β.

Figure 1 shows the directed acyclic graph (DAG, [12]) which describes this system. If all variables in our system were observed, causal inference would be easy (we could simply run a linear regression). However, what makes causal inference hard in these situations is that the variable U is unobserved to the learning agent. This *omitted variable bias* means that simply applying a standard technique optimized for prediction can lead one astray (as in our ice cream and drowning example above). There are many methods developed to try to deal with this problem ([11, 2]), however, they are highly optimized for much smaller data sets and much smaller numbers of parameters than we are used to in modern machine learning.

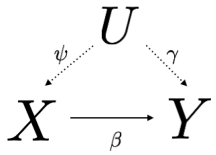

Figure 1: DAG representing our baseline structural equation models.

We now summarize our full paper on how to combine many randomized trials to learn causal models and especially those where X is high dimensional. Note that these trials can even be ones that were conducted with very different goals in mind, such as the A/B testing that internet companies routinely conduct to optimize various parts of their sites. We can think of this as 'upcycling' existing data.

2. LEARNING CAUSAL RELATIONSHIPS BY META ANALYSIS OF MANY RANDOMIZED EXPERIMENTS

We now introduce instrumental variables (henceforth IV, [1]) as a method for estimating β and then discuss the special properties of the estimator when we use it to meta-analyze existing randomized experiments. Suppose that we have some variable Z that has two properties.

First, Z is not caused by anything in the (X, U, Y) system.

Second, Z affects Y only via X. This latter assumption is known as the exclusion restriction.

Formally this modifies the structural equations to be

$$X = Z\mu + U\psi + V$$

and

$$Y = X\beta + U\gamma + \epsilon_X$$

with the DAG as below:

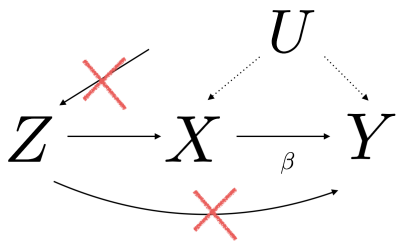

Figure 2: DAG representing our new structural equations with the IV added. Crosses represent causal relationships that are ruled out by the IV assumptions.

A standard way to use the existence of an IV to recover causal effects is to use least squares (TSLS). In the first stage we regress X on Z. We then replace X by the predicted values from the regression. In the second stage, we regress Y on these fitted values. This can be performed in one shot by taking the projection matrix

$$P = Z(Z'Z)^{-1}Z'$$

and computing

$$\hat{\beta}_{TSLS} = (X'PX)^{-1}(X'PY).$$

It is straightforward to show that as n approaches infinity this estimator converges to the true causal effect β ([1], [2]).

The mathematical intuition behind the IV estimator is as follows: the variance in X can be broken down into 2 components. The first component is confounded with the true causal effect (i.e. comes from U). The second component, on the other hand, is independent of U. Thus, if we could regress Y only on the random component, we could recover the causal effect β. Knowing Z allows us to do exactly this (i.e. by using only the variation in X caused by Z not U).

Traditional IV analyses in the social sciences use randomly assigned shocks (eg. which month someone was born in combined with the lottery for the draft) to study causal relationships of interest (eg. the effect of education on wages [2]). However, these kinds of analyses focus on small numbers of causal variables and small numbers of instruments.

A great potential instrument is actual randomized experimentation. These are generally randomly assigned (satisfying IV assumption 1) and with a little bit of thinking, we can write down the exclusion restrictions (for example, experiments on a spam prevention algorithm can only affect user sentiment via the channel of which emails the user sees). Indeed some modern research uses randomized experiments as IVs (eg. [8] study peer effects by looking at the assignment of peers into experimental conditions as an instrument).

However, it is often hard to change just one X at a time and in addition the dimensionality of X is large then a learning agent which designs specialized explorations for each parameter (as a reinforcement learning agent would) would require a massive amount of time and data to learn what they need to.

For this reason, we advocate digging into the pile of *already completed* experiments which we can pool together and meta-analyze. For example, a company interested in learning the effect of different kinds of spam on user sentiment (and using that to prioritize different kinds of spam fighting initiatives) can meta-analyze a large collection of prior A/B tests of their spam algorithms. Often there may be hundreds, or thousands of tests available for such a meta-analysis.

To implement this analysis, for the variable Z we can use the treatment assignment of each row in the data set. Note that this one-hot encoding of treatments means that we don't need any meta-data about what each trial actually did, we only need to know who was assigned where. One nice thing about such an approach is that the projection stage of the TSLS procedure described above becomes just a regression of X on these one-hot encodings which amounts to simply taking within-group means of each causal variable. The second stage then is just a regression of group level means of X and group level means of Y with precision weighting. This means that learning agents don't need to store huge amounts of data, rather summary statistics are sufficient for all estimation. Standard A/B testing platforms [3], [19] should already compute and store all the required statistics, so the method here can be thought of as an "upcycling" of existing data.

Note, however, that while this is a consistent estimator (in the sense of converging to β citeangrist1996identification) as n goes to infinity, there are actually 2 dimensions to consider here. We can either consider the asymptotic sequence where the number of units per treatment arm goes to infinity (n_{per}) or the case where the number of treatments in the analysis goes to infinity (we call this K). The natural situation is to raise K, this corresponds to the data generating process where we get more potential trials to analyze but each trial continues to be of a fixed finite size.[2]

In this case, however, the TSLS estimator is a biased estimate of the true causal effect. Indeed, it can be shown that it biased towards the observational (confounded) estimate. To see this, note that we can write structural equations for our TSLS procedure as

$$\bar{X} = \bar{\epsilon} + \bar{U}\psi + Z$$

and

$$\bar{Y} = \bar{X}\beta + \bar{U}\gamma + \bar{\eta}$$

where bars indicate group level means of X and Y variables. Note that as n_{per} goes to infinity, $\bar{U}, \bar{\epsilon}, \bar{\eta}$ converge to 0 (because they are themselves means of iid random variables with mean 0 we can invoke the central limit theorem) but for any finite n_{per} they are non-degenerate distributions. We can use some algebra to see that the regression of $\bar{Y}$ on $\bar{X}$

[2]This is what makes much of this analysis different from other work on IV estimation with weak instruments [16], [15], [17] and existing research on regularized IV estimation ([4], [**?**], [7].

gives the coefficient

$$\beta + \frac{\gamma Cov(\bar{X}, \bar{U})}{\psi^2 Var(\bar{U}) + Var(\bar{Z}) + Var(\bar{\epsilon})}.$$

The second term is the committed variable bias and note that we can drive that to 0 in one of two ways. Either we can decrease the numerator (which happens as we raise n_{per}) or we can increase the denominator. Notice that the denominator's variance is proportional to $Var(Z)$ which is also the variance of treatment effects on X in our collection of experiments.

There is a nice takeaway here: suppose that the experiments we are meta-analyzing were drawn from a mixture distribution (eg. some are 'strong' explorations of the parameter space, some are 'weak' optimizations at the margin). Intuitively, if we knew which experiment was of which type and we ran TSLS restricted to those experiments, we would get differentially biased estimates (since weak optimizations have smaller variance of treatment effects and thus more bias). This means that when K is large, we can use simple l_0 regularization to isolate, just from observing X, experiments which come from the strong mixture component.

In the case of the finite mixture, we are limited to how much we can reduce the bias of our estimator but when Z is drawn from an infinite mixture which includes all variances (eg. a t distribution) we show in the full version of this paper that, with large K, we can completely remove finite sample IV bias by using first stage regularization. In the full version of the paper we also show why standard cross-validation cannot be used to set the hyperparameter (loosely speaking because cross-validation will try to optimize for the best predictive, not the beast causal model) but we also show how to select the regularization hyperparameter using a type of cross-validation.

3. REINFORCEMENT LEARNING

Before we close this abstract, we turn to another important recent development in the field of artificial decision-making: reinforcement learning (RL) [18]. We view reinforcement learning and causal inference as two complementary approaches to decision-making and we now discuss how we imagine they can be integrated.

RL is very powerful because it can learn very complex policies especially when combined with deep neural networks to represent complex state spaces [14]. In addition, because RL models have an explicit value model, they automatically are able to make decisions like explore vs. exploit. On the other hand, causal inference models are static and make no explicit suggestions about what to do about pathways that the model simply doesn't have the data to estimate.

However, RL has two key disadvantages. First, it can only learn *policies* not causal relationships. Thus, standard RL cannot make predictions about the effects of an action that it has not tried in the past (or maybe has not existed in the past). Second, standard RL does not have the ability to generalize from observation or from the past experiences of others (though we note that variants of these algorithms which can do that have been around in game theory for a long time [10]).

Each approach to making decisions appears to cover the weaknesses of the other, so we believe there is much to be gained from understanding and fusing RL and causal modeling.

4. REFERENCES

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

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

[3] E. Bakshy, D. Eckles, and M. S. Bernstein. Designing and deploying online field experiments. In *Proceedings of the 23rd ACM conference on the World Wide Web*. ACM, 2014.

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

[5] L. Bottou. From machine learning to machine reasoning. *Machine learning*, 94(2):133–149, 2014.

[6] L. Bottou, J. Peters, J. Q. Candela, D. X. Charles, M. Chickering, E. Portugaly, D. Ray, P. Y. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[7] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306, 2004.

[8] D. Eckles, R. F. Kizilcec, and E. Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322, 2016.

[9] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[10] D. Fudenberg and D. K. Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[11] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[12] J. Pearl. *Causality*. Cambridge university press, 2009.

[13] U. Shalit, F. Johansson, and D. Sontag. Bounding and minimizing counterfactual error. *arXiv preprint arXiv:1606.03976*, 2016.

[14] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[15] D. O. Staiger and J. H. Stock. Instrumental variables regression with weak instruments, 1994.

[16] J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 2012.

[17] J. H. Stock and M. Yogo. Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 2005.

[18] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[19] Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236. ACM, 2015.