MIT Libraries | DSpace@MIT

MIT Open Access Articles

Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Peysakhovich, Alexander and Dean Eckles. "Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables." Proceedings of the 2018 World Wide Web Conference on World Wide Web, April 2018, Lyon, France, ACM Press, 2018.

As Published: http://dx.doi.org/10.1145/3178876.3186151

Publisher: ACM Press

Persistent URL: https://hdl.handle.net/1721.1/129382

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution 4.0 International license

Massachusetts Institute of Technology

Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables

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

Dean Eckles
Massachusetts Institute of Technology
Cambridge, MA
deckles@mit.edu

ABSTRACT

Scientific and business practices are increasingly resulting in large collections of randomized experiments. Analyzed together multiple experiments can tell us things that individual experiments cannot. We study how to learn causal relationships between variables from the kinds of collections faced by modern data scientists: the number of experiments is large, many experiments have very small effects, and the analyst lacks metadata (e.g., descriptions of the interventions). We use experimental groups as instrumental variables (IV) and show that a standard method (two-stage least squares) is biased even when the number of experiments is infinite. We show how a sparsity-inducing l_0 regularization can (in a reversal of the standard bias–variance tradeoff) reduce bias (and thus error) of interventional predictions. We are interested in estimating causal effects, rather than just predicting outcomes, so we also propose a modified cross-validation procedure (IVCV) to feasibly select the regularization parameter. We show, using a trick from Monte Carlo sampling, that IVCV can be done using summary statistics instead of raw data. This makes our full procedure simple to use in many real-world applications.

CCS CONCEPTS

• **General and reference → Experimentation**; • **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Machine learning**;

KEYWORDS

causality, experimentation, instrumental variables, machine learning

ACM Reference Format:
Alexander Peysakhovich and Dean Eckles. 2018. Learning Causal Effects From Many Randomized Experiments Using Regularized Instrumental Variables. In *Proceedings of The 2018 Web Conference (WWW 2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3178876.3186151

1 INTRODUCTION

Randomized experiments (i.e. A/B tests, randomized controlled trials) are a popular practice in medicine, business, and public policy [6, 24]. When decision-makers employ experimentation they have a far greater chance of learning true causal relationships and making good decisions than via observation alone [20, 25, 28]. However, a single experiment is often insufficient to learn about the causal mechanisms linking multiple variables. Learning such multivariate causal structures is important for both theory building and making decisions [16, 21].

Consider the situation of a internet service for watching videos. The firm is interested in how watching different types of videos (e.g., funny vs. serious, short vs. long) affects user behaviors (e.g. by increasing time spent on the site, inducing subscriptions, etc.). Such knowledge will inform decisions about content recommendation or content acquisition. Even though the firm can measure all relevant variables, training a model on observational data will likely be misleading. Existing content recommendation systems and heterogeneous user dispositions will produce strong correlations between exposure and time spent or subscription, but the magnitude of this correlation will, in general, not match what would occur if the decision-maker *intervened* and changed the promotion or availability of various video types. Thus, we are interested not just in prediction but prediction under intervention [9, 10, 30].

The standard solution is to run a randomized experiment exposing some users to more of some type of video. However, a single test will likely change many things in the complex system. It is hard to change the number of views of funny videos without affecting the number of views of serious videos or short videos. This is sometimes called the problem of 'fat hand' because such interventions touch multiple causal variables at once and so the effect on a single variable is not identified. To solve this issue the company would need to experiment with several factors simultaneously, perhaps conducting new experiments specifically to measure effects via each mechanism [21].

However, because routine product experimentation is common in internet companies [5, 24, 38], this firm has likely already run many A/B tests, including on the video recommendation algorithm. The method proposed in this paper can either be applied to a new set of experiments run explicitly to learn a causal effect vector [as in, e.g., 13], or can be applied to repurpose already run tests by treating them as random perturbations injected into the system and using that randomness in a smart way.

Our contributions arise from adapting the econometric method of instrumental variables [IV; 1, 32, 40] to this setting. It is well known that a standard IV estimator — two-stage least squares (TSLS) — is biased in finite samples [3, 35]. For our case, it also has asymptotic bias. We show that this bias depends on the distribution of the treatment effects in the set of experiments under consideration.

Our main technical contribution is to introduce a multivariate l_0 regularization into the first stage of the TSLS procedure and show that it can reduce the bias of estimated causal effects. Because in

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW 2018, April 23–27, 2018, Lyons, France
© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.
ACM ISBN 978-1-4503-5639-8/18/04..
https://doi.org/10.1145/3178876.3186151

finite samples this regularization procedure reduces bias but adds variance, we introduce a method to trade these off and select a regularization parameter. We call this procedure *instrumental variables cross-validation* (IVCV). In an empirical evaluation that combines simulation and data from hundreds of real randomized experiments, we show that the l_0 regularization with IVCV outperforms TSLS and a Bayesian random effects model.

Finally, we show how to perform this estimation in a computationally and practically efficient way. Our regularization and cross-validation procedures only require summary statistics at the level of experimental groups. This is advantageous when using raw data is computationally or practically burdensome, e.g., in the case of internet companies. This means the computational and data storage complexities of the method are actually quite low. In addition, standard A/B testing platforms [5, 41] should already compute and store all the required statistics, so the method here can be thought of as an "upcycling" of existing statistics.

2 CONFOUNDING AND THE BASIC IV MODEL

Suppose we have some (potentially vector valued) random variable X and a scalar valued outcome variable Y. We want to ask: what happens to Y if we change some component of X by one unit, holding the rest constant? Formally, we study a linear structural (i.e. data generating) equation pair

$$X = U\psi + \epsilon_X$$

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

where U, ϵ_X, and ϵ_Y are independent random variables with mean 0, without loss of generality. Note that in A/B testing we are often interested in relatively small changes to the system, and thus we can just think about locally linear approximations to the true function. We can also consider basis expansions. We refer to X as the causal variables (in our motivating example this would be a vector of time spent on each video type), Y as the outcome variables (here overall user satisfaction), U as the unobserved confounders, ϵ as noise, and β as the causal effects.

In general, we are interested in estimating the causal effect β because we are interested in intervention, e.g., one which will change our data-generating model to

$$X = U\psi + \epsilon_X + a.$$

In the presence of unobserved confounders trying to learn causal relationships using predictive models naively can lead us astray [9, 10, 30, 33]. Suppose that we have observational data of the form (X, Y) with U completely unobserved. If we use this data to estimate the causal effect β we can, due to the influence of the unobserved confounder, get an estimate that is (even in infinite samples) larger, smaller or even the opposite sign of the true causal effect β.

To see this consider the linear structure equation model above and suppose that we only observe (X, Y) where both are scalar. Since the underlying model is linear, we can try to estimate it using a linear regression. However, not including the confounder U in the regression yields the estimator:

$$\hat{\beta}_{\text{obs}} = (X'X)^{-1}(X'Y)$$

When all variables are scalar algebra yields

$$\mathbb{E}[\hat{\beta}_{\text{obs}}] = \beta + \gamma \frac{\text{Cov}(X, U)}{\text{Var}(X)}.$$

Thus, the best linear predictor of Y given X ($\hat{\beta}_{\text{obs}}$) may not be lead to a good estimate of what would happen to Y if we *intervened* (β).

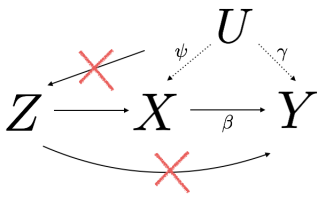

Figure 1: DAG representing our structural equations, in which the relationship between X and Y is confounded by U, and including the instrumental variable Z. Crosses represent causal relationships that are ruled out by the IV assumptions.

We now discuss instrumental variable (IV) estimator as a method for learning the causal effects. Suppose that we have some variable Z that has two properties (see Figure 1 for the directed acyclic graph which represents these assumptions):

(1) Z is not caused by anything in the (X, U, Y) system; that is, Z is as good as randomly assigned.
(2) Z affects Y only via X. This is often called the exclusion restriction or complete mediation assumption [3]

In terms of the structural equations, this modifies the equation for X to be

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

with the appropriate independence assumptions.

The standard IV estimator for β is two-stage least squares (TSLS) and works off the principle that the variance in X can be broken down into two 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).

TSLS can be thought of as follows: 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. It is straightforward to show that as n approaches infinity this estimator converges to the true causal effect β [39, Theorem 5.1].

All IV methods make a full rank assumption. In order to estimate the effect of each variable X_j on Y with the other X's held constant it must be the case that Z is such that it causes independent variation in all dimensions of X. This implies that we must, at least, have as many instruments as the dimension of β for TSLS to work. An interesting and fruitful direction for future work is what to do when some subspace of X is well spanned by our instruments but some some subspace is not.

3 IV WITH TEST GROUPS WITHOUT METADATA

In our setting of interest, randomly assigned groups from a large collection of experiments are the instruments.

Formally, here the IV is a categorical variable indicating which of K test groups a unit (e.g., user) was assigned to in one of many experiments. For simplicity of notation, we assume that each treatment group $g \in \{1, ..., K\}$ has exactly $n_g = n_{\text{per}}$ units assigned to it at random.

3.1 Computational Properties

The way to represent the first stage regression of the TSLS is to use the *one-hot representation* (or dummy-variable encoding) of the group which each unit is assigned to, such that Z_i is a K-dimensional vector of 0s and a single 1 indicating the randomly assigned group.

In this setup the TSLS estimator has a very convenient form. The first stage regression of X on Z simply yields estimates that are group level means of X in each group. This means that if each group has the same number of units (e.g., users) and the same error variance, the second stage has a convenient form as well: we can recover β by simply regressing group level averages of X on Y [3, section 4.1.3].

Thus, to estimate causal effects from large meta-analyses practitioners do not need to retain or compute with the raw data (which can span millions or billions of rows in the context of A/B testing at a medium or large internet company), but rather can retain and compute with sample means of X and Y in each A/B test group (this is now just thousands of rows of data). These are quantities that are recorded already in the most automated A/B testing systems [5, 41]. Working with summary statistics simplifies computation enormously and allows us to reuse existing data.

3.2 Asymptotic Bias in the Grouped IV Estimator

There are now multiple ways to think about the asymptotic properties of this "groups as IVs" estimator. Either we increase the size of each experiment ($n_{\text{per}} \to \infty$) or we get more experiments ($K \to \infty$). The former is the standard asymptotic sequence, but for meta-analysis of a growing collection of experiments, the latter is the more natural asymptotic series, so we fix n_{per} but we raise K.

We fix ideas with the case where X, Y, Z, U are scalar. We denote the group level means of our variables with bars (e.g., $\bar{X}$ to be the random variable that is the group-level means of X). Recall that our TSLS is, in the group case, a regression of $\bar{Y}$ on $\bar{X}$.

Decompose the causal variable group level average into

$$\bar{X} = \bar{Z} + \bar{U}\psi + \epsilon_{\bar{X}},$$

where

$$\bar{Z} \equiv Z\mu = \mathbb{E}[X|Z]$$

is the true first stage of the IV model (i.e. what we are trying to learn in the first stage of the TSLS). In the case of experiments as instruments this term has a nice interpretation: it is the true average value of the causal variables when assigned to that experimental group. If we assume (without loss of generality) that the mean of X is 0 then this first-stage can also be interpreted as the true treatment effect of the experiment.

While we are not considering asymptotic series where n_{per} goes to infinity, n_{per} will generally also be large enough that so that we can use the normality of sample means guaranteed by the central limit theorem. Thus, $\bar{U}$ and $\bar{\epsilon}_X$ are normal with mean 0 and variance proportional to ${n_{per}}^{-1}$.

With finite n_{per} we can show that, even as $K \to \infty$, TSLS will be biased [cf. 2, 7]. Suppose for intuition that $\bar{Z}$ has mean 0 and finite variance $\sigma^2_{\bar{Z}}$ this bias has the closed form which can be derived as follows. First, denote $\bar{A}$ as the group level mean of variable A. From the structural equations we know that:

$$\bar{X} = \bar{Z} + \bar{U}\psi + \epsilon_{\bar{X}}$$
$$\bar{Y} = \bar{X}\beta + \bar{U}\gamma + \epsilon_{\bar{Y}}$$

Since the TSLS estimator in this case is a regression of $\bar{X}$ on $\bar{Y}$ we can use the equation derived above for the scalar case to rewrite

$$\mathbb{E}[\beta_{TSLS}] = \beta + \gamma \frac{\text{Cov}(\bar{X}, \bar{U})}{\text{Var}(\bar{X})}.$$

$$\operatorname*{plim}_{K\to\infty} \hat{\beta}_{\text{TSLS}} = \beta + \frac{\gamma\psi \frac{\sigma^2_U}{n_{\text{per}}}}{\psi^2 \frac{\sigma^2_U}{n_{\text{per}}} + \frac{\sigma^2_{\epsilon_X}}{n_{\text{per}}} + \sigma^2_{\bar{Z}}}.$$

To understand where this bias comes from, think about the case where $\bar{Z}$ is always 0. The instrument does nothing, however the group-level averages still include group-level confounding noise; that is, for finite n_{per}, $\bar{U}$ has positive variance.

Thus, we simply recover the original observational estimate that we have already discussed as including omitted variable bias. When Z is not degenerate, $\bar{X}$ and $\bar{Y}$ include variation from both $\bar{U}$ and $\bar{Z}$. As n_{per} increases the influence of $\bar{U}$ decreases and so $\hat{\beta}_{\text{TSLS}}$ is consistent for β.

While in many cases, where variation induced by instrumental variables is large, this bias can be safely ignored, in the case of online A/B testing this is likely not the case. Since much of online experimentation involves hill climbing and small improvements (on the order of a few percent or less) that add up, the TSLS estimator can be quite biased in practice (more on this below).

4 BIAS-REDUCING REGULARIZATION

We now introduce a regularization procedure that can decrease bias in the TSLS estimator. We show that, in this setting a l_0-regularized first stage is computationally feasible and can help reduce this bias under some conditions on the distribution of the latent treatment effects.

4.1 Intuition via a Mixture Model

There are many types of A/B tests conducted — some are micro-optimizations at the margin and some are larger explorations of the action space. Consider the stylized case with two types of tests calling the smaller variance type 'weak' tests while the larger variance ones are 'strong.' Here we can model the first stage $\bar{Z}$ as being drawn from a two-component mixture distribution:

$$\bar{Z} = \mathbb{E}[X|Z] \sim \begin{cases} \mathcal{N}(0, \sigma^2_{weak}) & \text{with probability } p \\ \mathcal{N}(0, \sigma^2_{strong}) & \text{with probability } (1-p) \end{cases}$$

If we knew which group was drawn from which component and ran two separate TSLS procedures using only groups whose $\bar{Z}$ is drawn from the same component, we would asymptotically get two estimators:

$$\operatorname*{plim}_{K\to\infty} \hat{\beta}_{\text{TSLS},j} = \beta + \gamma \frac{\psi \frac{\sigma_U^2}{n_{\text{per}}}}{\psi^2 \frac{\sigma_U^2}{n_{\text{per}}} + \frac{\sigma_{\epsilon X}^2}{n_{\text{per}}} + \sigma_j^2}$$

Here $j \in \{\text{weak}, \text{strong}\}$ representing the component from which a particular group's $\bar{Z}$ is drawn. Because $\sigma^2_{\text{strong}} > \sigma^2_{\text{weak}}$ we will have that $\hat{\beta}_{\text{TSLS, strong}}$ is a less asymptotically biased (and thus asymptotically better) estimator than $\hat{\beta}_{\text{TSLS, weak}}$. Thus, if we could choose, we would choose to only use strong tests for our estimation of the causal effect.

In reality, we would likely not know which component each group is drawn from and if simply ran a TSLS on the full data set, this estimator will be a weighted combination of the two estimators.

Within this discrete mixture model, we are limited to how much we can reduce bias (since $\text{plim}_{K\to\infty} \hat{\beta}_{\text{TSLS, strong}} \neq \beta$). However if the treatment effects are drawn from a distribution which is an infinite mixture of normals that has full support on normals of all variances (for example a t distribution) then we can asymptotically reduce the bias below any ϵ by using only observations which come from components with arbitrarily large variances.

We now introduce a regularization procedure that tries to perform this selection. Because using this regularization effectively decreases our dataset size, decreasing the bias increases the variance. Thus, afterwards we will turn to a procedure to set the regularization parameter to make good finite sample bias–variance tradeoffs.

4.2 Formalizing First Stage Regularization

Consider a data set $(\bar{X}_g, \bar{Y}_g)$ of vectors of group-level averages. Let

$$p(x) = \Pr(|\bar{U}_g + \bar{\epsilon}_{x,g}| > |x|)$$

be the p-value for a group-level observation x under a 'no intervention' null with $Z = 0$. Given that under no-intervention $\bar{X}$ is distributed normally computing p is straightforward and requires simply the observational (within-condition) covariance matrix of X.

For a given threshold $q \in (0, 1]$, let

$$\bar{X}_g^q \equiv \begin{cases} \bar{X}_g & \text{if } p(\bar{X}_g) < q \\ 0 & \text{otherwise.} \end{cases}$$

We then define the regularized IV estimator as

$$\hat{\beta}_q = (\bar{X}^{q\prime} \bar{X}^q)^{-1} (\bar{X}^{q\prime} \bar{Y}).$$

Thus, this procedure is equivalent to an l_0 regularization in the first stage of the TSLS regression. In particular, when $\bar{U}_g + \bar{\epsilon}_{x,g}$ has a normal distribution, as in the present case, then this is equivalent to l_0-regularized least squares.

Recall that in the binary mixture example above, this regularization would preferentially retain groups that come from the higher variance (strong) component. This extends to infinite mixtures, such as the t, where this procedure will preferentially set $\bar{X}_g$ to zero for groups where $\bar{Z}_g$ is drawn from a lower variance component.

So far we have focused on scalar X. This procedure naturally extends to multidimensional settings. Just as with the single dimension we compute the 'null' distribution from no-intervention conditions. We then compute $p(\bar{X}_g)$ for each group and threshold all dimensions of the experimental group g; that is, if this probability is above a threshold q we set the whole vector $\bar{X}_g$ to 0. This gives us the multi-dimensional, group-based l_0 regularizer which we will apply in our experiments.

This group-l_0 regularization can be inefficient in certain regiments of treatment effects — for example, in a regime where each A/B test explicitly only moves a single dimension of X (i.e. 'skinny hand' interventions). We show how this can go wrong in our synthetic data experiment but we also see that real A/B tests appear not to fall into this regime.

5 CAUSAL CROSS-VALIDATION

We now turn to an important practical question: because there is a bias–variance tradeoff how should one set the regularization parameter when K is finite to optimize for prediction under intervention?

First, let us suppose that we have access to the raw data where a row is a (X_i, Z_i, Y_i) which is a unit i's, X, Y and treatment assignment Z. We propose a procedure to set our hyperparameter q. We describe 2-fold version as it conveys the full intuition, but extension to k-folds is straightforward.

Instrumental variables cross-validation algorithm (IVCV):

(1) Split *each treatment* in the data set into 2 folds, call these new data sets $\{(X_i^1, Y_i^1, Z_i^1)\}$ and $\{(X_i^2, Y_i^2, Z_i^2)\}$.
(2) Compute treatment level averages $\{(\bar{X}_g^1, \bar{Y}_g^1)\}$ and $\{(\bar{X}_g^2, \bar{Y}_g^2)\}$ as described above where j now indexes experimental groups.
(3) Compute $\hat{\beta}_q$ for a variety of thresholds q using $\{(\bar{X}_g^1, \bar{Y}_g^1)\}$.
(4) Compute treatment level predictions of Y using fold 1 for each level of q: $\hat{Y}_g^q = \bar{X}^1 \hat{\beta}_q$.
(5) Choose q which minimizes $\text{IVCV}(q) = \sum_j (\bar{Y}_g^2 - \hat{Y}_g^q)^2$.

The intuition behind IVCV is similar to the main idea behind IV in general. Recall that our objective is to use variation in X that is not caused by U. The IVCV algorithm uses the X value from fold 1 and compares the prediction to the Y value in fold 2 because fold 1 and fold 2 share a Z but differ in U (since U is independent across units but Z is the same within group). This intuition has been exploited in split-sample based estimators [2, 18, 22].

We can demonstrate the importance of using the full causal loss by comparing the IVCV procedure to other two candidates. The first is simply applying naive CV in the second stage (i.e., splitting each group into 2, training a model on fold 1 and computing the CV loss naively as $\|Y_2 - X_2 \hat{\beta}_q\|^2$). The second is stagewise, in which the regularization parameter is chosen to minimize MSE in the first stage, and then the second stage is fit conditional on the selected model [as in 8, 19]. We compare these approaches in a simple linear model with scalar X, such that

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

$$\bar{X} = \bar{Z} + \bar{U}.$$

We set the first stage to have fat tails

$$\bar{Z} = \mathbb{E}[X \mid Z] \sim t(df = 3, scale = .4).$$

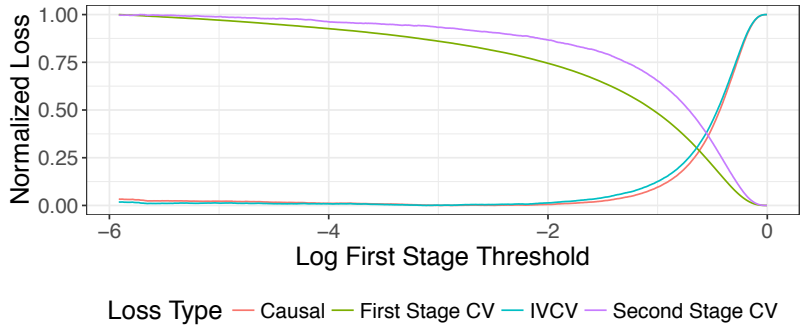

Figure 2: Comparison of stagewise vs. IVCV method. X-axis is the strength of regularization (lower p-value implies stronger regularization). Optimizing for stagewise loss would imply using almost no regularization whereas optimizing for IVCV loss implies strong regularization. Causal loss coincides much more with IVCV loss than stagewise loss.

So there is confounding in the observational case set $\gamma = 10$, $n_{\text{per}} = 100$ and $K = 2500$.

Recall that our goal is to find a hyperparameter (regularization threshold) which gives us the best prediction $\hat{\beta}_q$ of the causal parameter. Formally, we write this as

$$\text{CausalLoss}(q) = \sum_k (\hat{\beta}_{qk} - \beta_k)^2.$$

The causal loss is an unobserved quantity in the IV problem and thus we need to choose a surrogate loss to try to approximate it in our cross-validation procedure. A good choice of CV procedure is one whose loss function tracks closely with the true causal loss function.

This gives us 3 candidate cross-validation loss functions to compare to the true casual loss in our simulations

(1) The second stage CV loss

$$\sum_g (\bar{Y}_g^2 - \bar{X}_g^2 \hat{\beta}_q)^2$$

(2) The first stage CV loss

$$\sum_g (\bar{X}_g^2 - \hat{\bar{X}}_g^{2q})^2$$

(3) The IVCV loss

$$\sum_g (\bar{Y}_g^2 - \bar{X}_g^1 \hat{\beta}_q)^2$$

Figure 2 shows these losses as a function of the parameter q averaged over 500 simulations of the model above. We see that both the first stage loss curve and the second stage loss curve look very different from the causal loss curve. However, the IVCV loss curve matches almost exactly. Thus, either stage error naively yields a very different objective function from minimizing the causal error. In particular, we see that making the bias–variance tradeoffs for the first stage need not coincide with a desirable bias-variance tradeoff for inferring β.

The l_0-regularized IV estimator only requires the kinds of summary statistics per experimental group that are already recorded in the course of running A/B tests, which has practical and computational utility. However, the cross-validation procedure above requires the use of raw data. We now turn to the following question: if the raw data is unavailable, but summary statistics are, can we use these summary statistics to choose a threshold q?

Suppose that we have access to summary means $\{(\bar{X}_g, \bar{Y}_g)\}$ for each treatment j and the covariance matrix of $(\bar{X}, \bar{Y})$ conditional on $Z = 0$ which we denote by τ. We note that τ can be estimated very precisely from observational data or, in the case of the experimental meta-analysis just looking at covariances among known control groups. We assume that n_{per} is large enough such that the distributions of U and ϵ in groups of size $\frac{n_{\text{per}}}{2}$ are well approximated by the Gaussian $\mathcal{N}(0, \frac{\sigma_i^2}{\frac{n_{\text{per}}}{2}})$.

To perform IVCV under these assumptions, we use a result from the literature on Monte Carlo [29, ch. 8]. If some vector X is distributed

$$X \sim \mathcal{N}(\mu, \Sigma)$$

then any linear combination $T = \theta X$ has a normal distribution. Moreover, conditional on $T = t$ the distribution of X is also normal and can be written explicitly as

$$X|T = t \sim \mathcal{N}(\mu + \Sigma\Theta'(t - \theta\mu)), \Sigma - \Sigma\Theta'(\Theta\Sigma\Theta')^{-1}\Theta\Sigma).$$

This means if we know the observational covariance matrix τ then for every group g we can take the group level averages $(\bar{X}_g, \bar{Y}_g)$ and sample using the equation above to get $\bar{X}_g^1$ and $\bar{X}_g^2$ such that $\bar{X}_g^1 + \bar{X}_g^2 = 2\bar{X}_g$. Since by the central limit theorem the generating Gaussian model is approximately correct, this procedure simulates the split required by IVCV without having access to the raw data. The algorithm is as follows:

Summary statistics instrumental variables cross-validation algorithm (sIVCV):

(1) Start with data comprising of treatment group means $\{(\bar{X}_g, \bar{Y}_g)\}$.
(2) Use the covariance matrix to perform Monte Carlo sampling to simulate groups

$$\{(X_i^1, Y_i^1, Z_i^1)\}$$

$$\{(X_i^2, Y_i^2, Z_i^2)\}$$

(3) Use the IVCV algorithm to set the hyperparameter using the simulated splits.

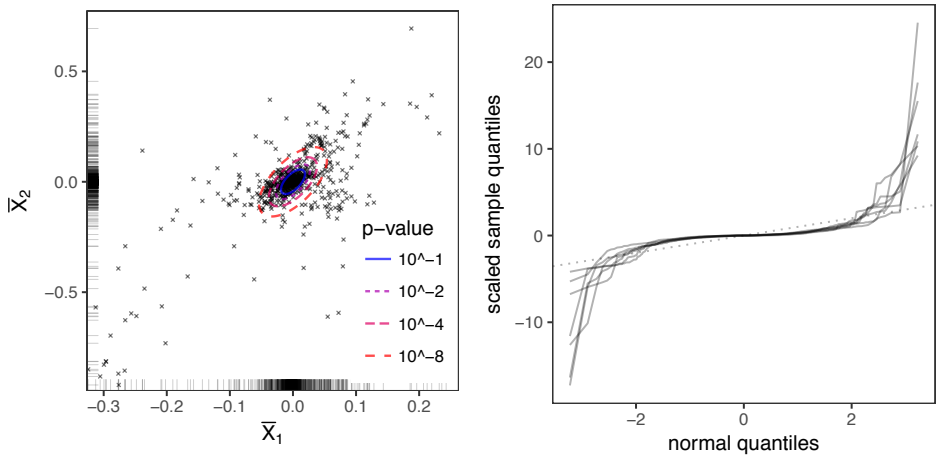

Figure 3: Left: Two dimensions of the multivariate means for sampled test groups ($\bar{X}_g$). The joint distribution exhibits many more outliers than a multivariate normal would suggest. Right: QQ-plots for the dimensions of the sampled test groups ($\bar{X}_g$). For each variable the marginal distributions are notably non-normal.

(4) Estimate β using the selected hyperparameters on the full data set.

6 EVALUATION

We now evaluate the IVCV procedure empirically. True causal effects in real IV data are generally unobservable, so comparisons of methods usually lack a single number to validate against. Examples of the kinds of evaluations usually done in work on IV include comparing different procedures and showing that one yields estimates which are more 'reasonable.'

Simulations allow us to know the true causal effects, but can lack realism. We strike a middle ground by using simulations where we set the causal effects ourselves but use real data to generate distributions for other variables. In our simulations we use a model given by

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

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

Thus, in this case all the variance in X that is not driven by our instruments is confounding variance.

6.1 Real A/B Tests

The multivariate case is made difficult and interesting when U has a non-diagonal covariance matrix and $\bar{Z}$ has some unknown underlying distribution, so we generate these distributions from real data derived from 798 randomly assigned test groups from a sample of A/B tests run on a recommendation algorithm at Facebook. We define our endogenous, causal Xs as 7 key performance indicators (i.e. intermediate outcomes examined by decision-makers and analysts); we standardize these to have mean 0 and variance 1. As the distribution of U we use the estimated covariance matrix among these outcomes in observational data. Third, we take the experiment-level empirical means of the Xs as the true $\bar{Z}$, to which we add the confounding noise according to the distribution of U.

We show a projection of these $\bar{Z}$ onto 2 of the X dimensions in Figure 3(A). We see that the A/B tests appear to have correlated effects but do span both dimensions independently, many groups are retained even with strong first stage regularization, and the distribution has much more pronounced extremes than would be expected under a Gaussian model. Figure 3(B) compares the observed and Gaussian quantiles, illustrating that all dimensions are notably non-normal (Shapiro–Wilk tests of normality reject normality for all the variables at $ps < 10^{-39}$).

We set β as the vector of ones and γ as a diagonal matrix with alternating elements 1 and -1, so that there is both positive and negative confounding. For each simulated data set, we compute the *causal MSE loss* for β; that is, the expected risk from intervening on one of the causal variables at random. If $\hat{\beta}$ is our estimated β vector then recall that this is given by

$$\text{CausalLoss}(\hat{\beta}) = \sum_k (\hat{\beta}_k - \beta_k)^2.$$

6.2 Results

In addition to the l_0-regularized IV method and TSLS, we examine a Bayesian random effects model, as in Chamberlain and Imbens [12] but with a t, rather than Gaussian, distribution for the instruments. Formally we let

$$\bar{Z} \sim t(d)$$

with a standard choice of prior

$$d \sim \text{Gamma}(2, .2).$$

To tilt the evaluation against our procedure we also give the Bayesian model the true covariance matrix for $\bar{U}$. To fit the Bayesian model we use Stan [11]. We compare the Bayesian random effects model and our regularized IV model to the infeasible Oracle estimator where the estimate of the first stage $\mathbb{E}[\bar{X} \mid \bar{Z}]$ is known with certainty.

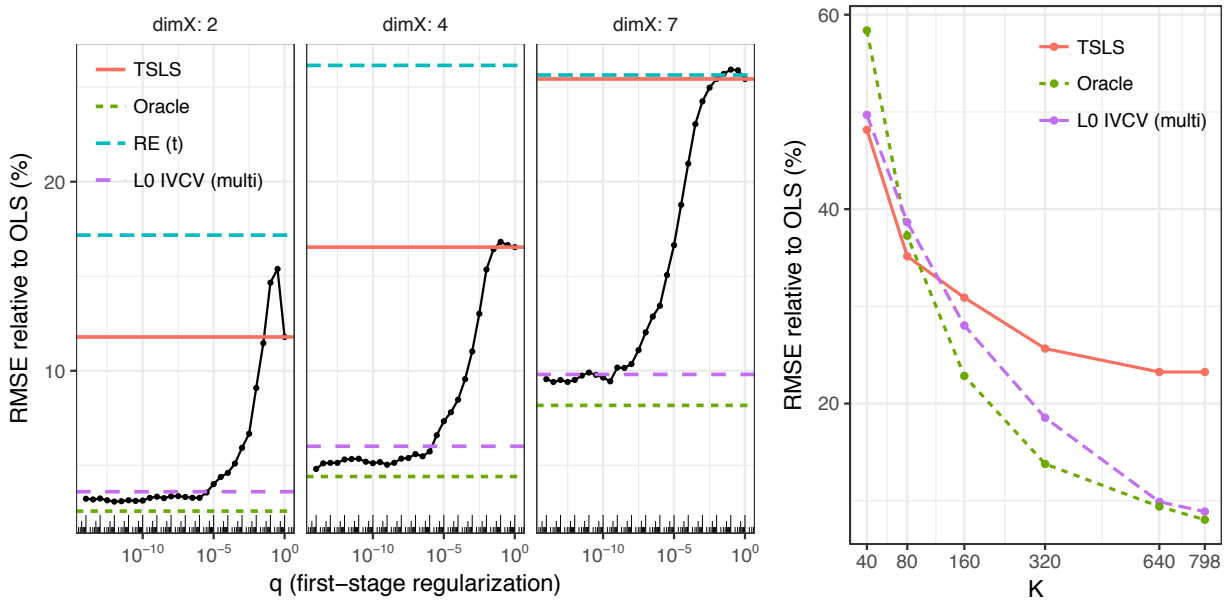

Figure 4: Left: Causal error (relative to a naive observational estimator) for the full l_0-regularization path (solid black), TSLS (solid red), IVCV selected parameters (dashed purple) and Bayesian random effects model (dashed teal). IVCV outperforms all other estimation techniques. Right: Error in estimating causal effects for varying numbers of test groups K. IVCV is useful even with a relatively small meta-analysis, while TSLS exhibits asymptotic bias. With a very small number of test groups, the Oracle can actually underperform TSLS because of near collinearity.

Figure 4(A) shows the results for various dimensions of X for 1,000 simulations. Because of the high level of confounding in the observational data, the observational (OLS) estimates of the causal effect are highly biased, such that even the standard TSLS decreases our causal MSE by over 70%.

We see that the l_0-regularization path (black line) reduces error compared with TSLS and, with high regularization, approaches the Oracle estimator. Furthermore, feasible selection of this hyperparameter using IVCV leads to near optimal performance (purple line). The Bayesian random effects model can reduce bias, but substantially increases variance and thus MSE.

We also look at how large the collection of experimental groups needs to be to see advantages of a regularized estimator relative to a TSLS procedure.

We repeat the TSLS, Oracle, and l_0-regularization with IVCV analyses in 100 simulations with smaller K (Figure 4(B)) for the case of the 7 dimensional X. Intuitively, what is important is the relative size of the tails of the distribution of the latent treatment effects $\bar{Z}$. As the tails get fatter, fewer experiments are required to get draws from the more extreme components of the mixture. We see that in this realistic case where $\bar{Z}$ is determined using a sampled set of Facebook A/B tests, feasible selection of the l_0-regularization hyperparameter using IVCV outperforms TSLS substantially for many values of K. Thus, meta-analyses of even relatively small collections of experiments can be improved by the first-stage l_0 regularization.

7 FULLY SIMULATED DATA

In addition to the evaluation using a real data set. We also consider the IVCV procedure in several completely synthetic data sets. The completely synthetic experiments allows us to elucidate the important assumptions for our procedure to work while the real data-based experiment shows that these assumptions are indeed satisfied in real world conditions.

We consider the same exact model as in the previous section except that we generate the first stage effects $\bar{Z}$ from a known parametric distribution and let U be normal. First, we consider

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

and we vary the distribution that $\bar{Z}$ is drawn from. We consider

(1) $\bar{Z}$ drawn from independent t with 3 degrees of freedom
(2) $\bar{Z}$ drawn from t with 3 degrees of freedom and covariance matrix drawn from inverse Wishart (a standard prior for covariance matrices) with $10 \times dim(X)$ degrees of freedo
(3) $\bar{Z}$ generated by first drawing σ^2 from an inverse Gamma distribution and then $\bar{Z}$ drawn from multivariate normal with mean 0 and covariance matrix $\sigma^2 I$

Note that in case 1 effects are axis aligned while in the next two larger values of one dimension can predict more extreme values of Z (and X) on another dimension. The final setup corresponds to the case where components are mean-uncorrelated but have correlated variance. This is the multivariate analog of our motivating example where some A/B tests are strong explorations of the parameter spaces and others are micro-optimizations at the margin. Note that

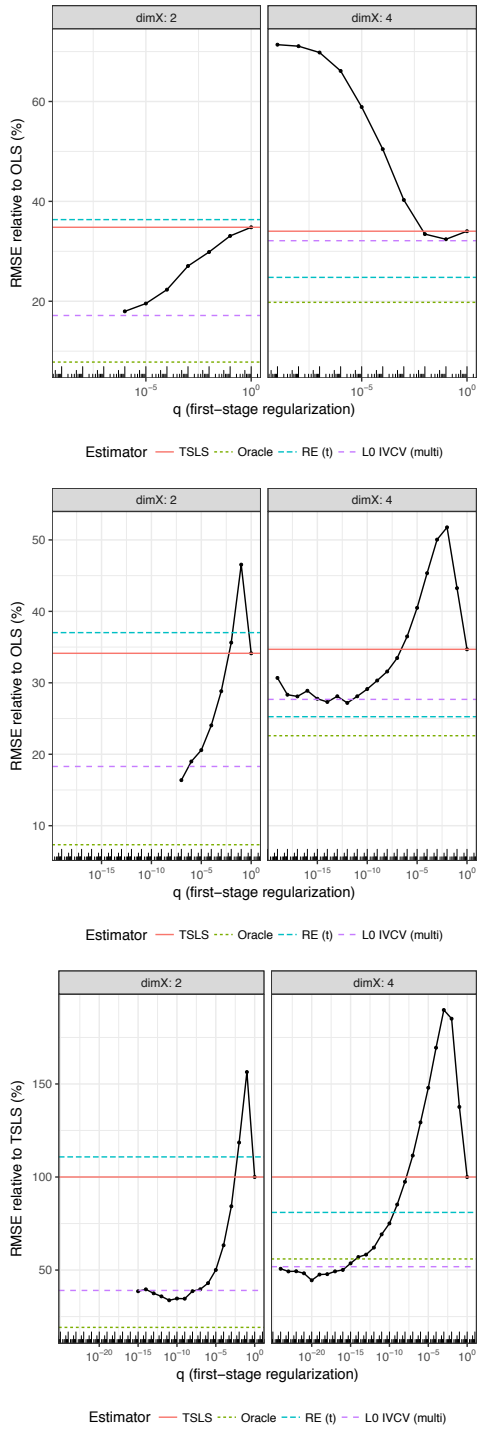

Figure 5: Performance of various IV estimation techniques under various first stage data generating assumptions (top = independent t, middle = Wishart t, bottom = correlated variances). We see that when the Z induced components of X are independent even for moderate dimensionality that the l_0 regularization performs less well. However, as soon as there is any correlation the IVCV procedure performs much better than TSLS and can both under or over-perform the Bayesian random effects model.

the marginal distribution for each dimension is, in all cases, a t distribution with 3 degrees of freedom (since the t can be written as a mixture of normals drawn from the inverse gamma).

Figure 5 shows the results of applying IVCV to the data generating processes above (top = independent t, middle = Wishart t, bottom = correlated variances). We restrict to $\dim(X) \in \{2, 4\}$ because it is sufficient to illustrate our main points. We see that in the independent t case the IVCV procedure (and indeed our multivariate l_0 regularization) can underperform the Bayesian random effects model fail to substantially improve on TSLS. This happens because in the independent t case there is a high probability that a single dimension is extreme enough to pass the regularization threshold and thus even strong regularization does not necessarily remove bias. On the other hand, when outcomes are correlated (or their variances are) we see that multivariate IVCV performs well because being extreme in one X component predicts having extreme outcomes in other components. Note that the fact that IVCV performs well in the distribution generated by real A/B tests suggests that real world A/B test effect variance is correlated within A/B test - ie. if a test moves one metric by a large amount, it likely moves others by a large amount.

8 CONCLUSION

Most analysis of randomized experiments, whether in academia, business, or public policy tends to look at each test in isolation. When meta-analyses of experiments are conducted, the goal is usually either to pool data about multiple instances of the same intervention or to find heterogeneity in the effects of interventions across settings. We instead propose that combining many experiments can help us learn richer causal structures. IV models give a way of doing this pooling. We have shown that in such situations using easily-implemented l_0 regularization in the first stage can lead to much better estimates of the causal effects, and thus better predictions about interventions, than using standard TSLS methods.

We expand on the literature which uses multi-condition experiments as instruments [13, 15]. Such analyses usually feature a smaller number of experimental groups and a single causal variable. Our work contributes to the growing literature merging experimental techniques with methods from machine learning to allow deeper analyses than has been traditionally possible [4, 9, 17, 31, 33**?**]. In addition, our procedure is intended to be simple to implement and not require strong assumptions about the data-generating process.

Our work is also related to research on IV estimation with weak instruments [34–36]. In addition, we also contribute to existing research on regularized IV estimation [8, 12, 18]. In the case of univariate X and disjoint groups as instruments, the post-lasso method in Belloni et al. [8] coincides with the proposed l_0 regularization, however in the case where X is a vector, it does not.

Recently, active learning in the form of bandit optimization [14, 27] and reinforcement learning [37] have become quite popular in the AI community. Such approaches can be used in many of the same contexts as the IV analysis we have discussed here, and so may appear to be substitutes. However, we note there are many important differences the largest of which is that RL and bandit approaches try to perform policy optimization rather than learning of a causal graph. For this reason, often *why* RL/bandit estimated

policies work can be hard to understand. This is in contrast to explicit causal models (e.g., the linear model described above) which can be stated explicitly and in terms that are more natural to human decision-makers [23]. On the other hand, bandit/RL approaches have advantages in that they are explicitly online whereas many causal inference procedures (including the one we have described here) are 'batch' procedures that assume that data collection is a passive enterprise, separate from analysis. There is growing interest in combining these approaches [26] and we think that future work would benefit greatly from this fusion of thought.

REFERENCES

[1] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91, 434 (1996), 444–455.

[2] Joshua D Angrist and Alan B Krueger. 1995. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics* 13, 2 (1995), 225–235.

[3] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton university press.

[4] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[5] E. Bakshy, D. Eckles, and M. S. Bernstein. 2014. Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd ACM conference on the World Wide Web*. ACM.

[6] Abhijit Banerjee and Esther Duflo. 2012. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* PublicAffairs.

[7] Paul A Bekker. 1994. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society* (1994), 657–681.

[8] Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80, 6 (2012), 2369–2429.

[9] Léon Bottou. 2014. From machine learning to machine reasoning. *Machine Learning* 94, 2 (2014), 133–149.

[10] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.

[11] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2016. Stan: A probabilistic programming language. *Journal of Statistical Software* (2016).

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

[13] Dean Eckles, René F Kizilcec, and Eytan Bakshy. 2016. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7316–7322.

[14] John C Gittins. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* (1979), 148–177.

[15] Mathew Goldman and Justin M Rao. 2014. Experiments as Instruments: Heterogeneous Position Effects in Sponsored Search Auctions. *Available at SSRN 2524688* (2014).

[16] Donald P Green, Shang E Ha, and John G Bullock. 2010. Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose. *The Annals of the American Academy of Political and Social Science* 628, 1 (2010), 200–208.

[17] Justin Grimmer, Solomon Messing, and Sean J Westwood. 2014. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Unpublished manuscript, Stanford University, Stanford, CA* (2014).

[18] Christian Hansen and Damian Kozbur. 2014. Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics* 182, 2 (2014), 290–308.

[19] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2016. Counterfactual Prediction with Deep Instrumental Variables Networks. *arXiv preprint arXiv:1612.09596* (2016).

[20] Lars G Hemkens, Despina G Contopoulos-Ioannidis, and John PA Ioannidis. 2016. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: Meta-epidemiological survey. *British Medical Journal* 352 (2016).

[21] Kosuke Imai, Dustin Tingley, and Teppei Yamamoto. 2013. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176, 1 (2013), 5–51.

[22] Guido Imbens, Joshua Angrist, and Alan Krueger. 1999. Jackknife Instrumental Variables Estimation. *Journal of Applied Econometrics* 14, 1 (1999).

[23] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690* (2017).

[24] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1168–1176.

[25] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* (1986), 604–620.

[26] Finnian Lattimore, Tor Lattimore, and Mark D Reid. 2016. Causal Bandits: Learning Good Interventions via Causal Inference. In *Advances in Neural Information Processing Systems*. 1181–1189.

[27] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. ACM, 661–670.

[28] Michelle N Meyer. 2015. Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *J. on Telecomm. & High Tech. L.* 13 (2015), 273.

[29] Art B. Owen. 2016. *Monte Carlo Theory, Methods and Examples.* http://statweb.stanford.edu/~owen/mc/

[30] Judea Pearl. 2009. *Causality.* Cambridge University Press.

[31] Alexander Peysakhovich and Jeffrey Naecker. forthcoming. Machine learning and behavioral economics: Evaluating models of choice under risk and ambiguity. *Journal of Economic Behavior and Organization* (forthcoming).

[32] Olav Reiersöl. 1945. *Confluence analysis by means of instrumental sets of variables.* Ph.D. Dissertation. Stockholm College.

[33] Uri Shalit, Fredrik Johansson, and David Sontag. 2016. Bounding and Minimizing Counterfactual Error. *arXiv preprint arXiv:1606.03976* (2016).

[34] Douglas Staiger and James H Stock. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica* (1997), 557–586.

[35] James H Stock, Jonathan H Wright, and Motohiro Yogo. 2012. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* (2012).

[36] James H Stock and Motohiro Yogo. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge University Press, 80–108.

[37] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction.* Vol. 1. MIT press Cambridge.

[38] Hal Varian. 2016. Intelligent Technology. *Finance and Development* 53, 3 (2016).

[39] Jeffrey M Wooldridge. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

[40] Philip Green Wright. 1928. *The Tariff on Animal and Vegetable Oils.* The Macmillan Co.

[41] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. 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*. ACM, 2227–2236.