Deep V: Moving Beyond Conventional IV in Causal Inference

Jamie Fogel, Andrew Kwong and Bernardo Modenesi
University of Michigan

Abstract

The central problem in empirical economics is to identify the causal effect of a "treatment" on outcomes of interest. This task is complicated, however, by potential correlations between latent variables and the outcome. Economists traditionally overcome this problem by use of instrumental variable analysis, which relies on strong assumptions of both linearity and homogeneity. We do away with these assumptions by using deep neural networks to estimate both treatments and outcomes, and validate our model by replicating the results of a seminal economics paper that studies the relationship between Chinese imports and American manufacturing employment.

1 Introduction

The central problem in empirical economics is to identify the causal effect of a "treatment" on outcomes of interest. This is equivalent to estimating the conditional expectation function for an outcome given variables that can, in principle, be manipulated by a policymaker or researcher. For example, one might be interested in the effect of changing U.S. trade policy (treatment) on wages or employment rates for American workers (outcome). Typical estimators of treatment effects in economics rely on strong linearity assumptions on the relationship of the variables and many also assume that the effect of treatment is homogenous across individuals. We depart from these strong assumptions in order to implement estimators of causal treatment effects that are robust to nonlinear conditional expectation functions and informative about richer heterogeneous treatment effects. In order to do so, we employ neural networks and Gaussian mixture models to perform treatment effect estimation in a nonlinear setting with heterogeneous treatment effects.

Most empirical research in economics is focused on out-of-sample prediction and as such requires assumptions about the unobservable data-generating process (DGP) in order to make progress. We are typically interested in the *causal* effect of some variable t on an outcome variable y. This means we want to know what would happen to y if we were to change t, *holding all other factors x that determine y fixed.* Mathematically, the parameter of interest is $\gamma = \mathbb{E}[y|t = T_1, x] - \mathbb{E}[y|t = T_0, x]$, for a desired treatment that would shift t from T_0 to T_1.

If we could observe the full set x of covariates that determine y, then this would be a simple task, however in social sciences we typically do not observe x, or at best we observe a subset of the vector x. For example, we might want to estimate the effect of education on earnings. In a researcher's ideal world, we would randomly assign different levels of education to individuals and then we could interpret differences in earnings outcomes across people with different levels of education as the causal effect of education on earnings, however this is rarely a feasible option. Alternatively we could take a sample of education and earnings data from the population and estimate the relationship between education and earnings. However, this is unlikely to represent a causal effect because people who have high levels of education may have had high earnings even with less education due to some other unobserved factors such as intelligence or family background.

In other words, people with low levels of education do not represent a good counterfactual for what the outcomes for people who received high education would have been had they instead received low education. Since we never observe the counterfactuals we are interested in[1], it is impossible to uncover the causal effect of interest without imposing more structure on the data.

One common method employed in economics to estimate causal effects in the presence of endogenous omitted variables is instrumental variables (IV), as discussed in the influential paper of Angrist (1990). IV estimation uses an "instrument," which affects the treatment variable, but only affects the outcome through its effect on the treatment, in order to uncover the causal effect. For example, suppose there was a lottery that randomly provided some students with college scholarships. The instrument would be an indicator for whether or not an individual won the lottery and received a scholarship. This instrument affects the level of education received and therefore affects earnings, if there is a non-zero effect of education on earnings. However, since the scholarship was allocated randomly, it should not affect earnings through any mechanism other than education. Therefore it would be a valid instrument. While IV estimation is commonly used in economics, it is typically implemented using the "two-stage least squares" (2SLS) procedure, which assumes a model linear in its parameters. We depart from the 2SLS framework by estimating treatment effects via IV using neural networks on both stages, that do not impose linearity in the estimation; which is new in the causal estimation literature.

Our primary point of departure in the literature is "Deep IV: A Flexible Approach for Counterfactual Prediction" by Hartford et al. (2017). This paper develops a theoretical framework for estimating treatment effects using neural networks when causal identification is provided by instrumental variables (IV). We use Hartford et al. (2017)'s method in order to replicate and extend Autor et al. (2013)'s seminal paper in economics on the effect of increased trade with China on American workers. Specifically, we use Autor et al. (2013)'s data and take their exogeneity assumptions on their instrument as valid in order to assess the degree to which their conclusions would change if the assumptions of linear and homogeneous treatment effects are relaxed.

The literature on nonlinear IV begins with Newey and Powell (2003), who specify conditions for identification of the parameter of interest in a completely nonlinear setting. In order to prove the consistency of their nonlinear IV estimators, the authors use sieve estimation. Hartford et al. (2017) use the same identification strategy as Newey and Powell (2003), but estimate the parameter of interest using Deep Neural Networks (DNN). These authors claim that, among machine learning techniques, the Neural Networks provides best prediction power for the IV. Then we depart from the connection of these papers in Economics and Computer Science, respectively, to shed light on the potential gains of implementing DNN for causal inference in social sciences.

We proceed by first explaining the theoretical framework underlying our paper. Next we show results from a simulation where we vary functional forms and dataset size, while comparing DNN IV and traditional IV estimation. Then we present our results for IV using the data in Hartford et al. (2017) using the DNN IV. Finally, we discuss areas for further study and conclude.

Our primary contribution will be in identifying and implementing a useful application of neural networks outside of computer science, in a causal inference framework.

[1] This is known as the fundamental problem of causal inference, as posed by Holland (1986).

2 General IV Empirical Framework

We assume that our outcome variable of interest y is generated by a structural equation with additive errors:

$$y = g(t, x) + \varepsilon \tag{1}$$

where t is the policy or treatment of interest (e.g., trade policy), and x is a vector of observable characteristics. $g(\cdot)$ is an unknown and possibly non-linear continuous function of x and t. ε is an unobservable additively separable error term which may be correlated with the treatment t, hence the need for instrumental variables.

Following Newey and Powell (2003), we define the counterfactual prediction function

$$h(t, x) \equiv g(t, x) + \mathbb{E}[\varepsilon|x],$$

which is the conditional expectation of y given the treatment t and other observable confounding variables x, *holding the distribution of ε constant as t is changed*, which is crucial to the causal estimation. Estimating the causal effect of t on y is therefore tantamount to estimating $h(t, x)$; we can infer the effect of changing from policy t_0 to t_1, by looking at the difference in mean outcomes $h(t_1, x) - h(t_0, x) = g(t_1, x) - g(t_0, x)$. The reason we focus on $h(t, x)$ rather than $g(t, x)$ is that since ε may be correlated with t, $g(t, x)$ does not estimate the effect of t on y *holding the distribution of ε constant as t is changed,* while $h(t, x)$ does hold the distribution of ε constant as t is changed.

Standard machine learning approaches fit $\mathbb{E}[y|t, x] = g(t, x) + \mathbb{E}[\varepsilon|t, x]$. In general, this estimand will be a biased estimate of the structural objective $h(t, x)$ because $\mathbb{E}[\varepsilon|t, x] \neq \mathbb{E}[\varepsilon|x]$ so $h(t, x) \equiv g(t, x) + \mathbb{E}[\varepsilon|x] \neq \mathbb{E}[y|t, x] = g(t, x) + \mathbb{E}[\varepsilon|t, x]$. Fortunately, this endogeneity problem can be solved using instrumental variables as long as a valid instrument can be found and observed.

The assumptions necessary for a valid instrument are the same in this paper as they are in standard linear 2SLS. The instruments z must satisfy[2]:

1. *Relevance:* $F(t|x, z)$, the distribution of t given x and z, must not be constant in z
2. *Exclusion:* z does not enter the structural equation $y = g(t, x) + \varepsilon$ — i.e., $z \perp\!\!\!\perp y|(x, t, \varepsilon)$
3. *Unconfounded Instrument:* z is conditionally independent of the error — i.e. $z \perp\!\!\!\perp \varepsilon|x$.

Taking the expectation of both sides of the structural equation $y = g(p, x) + \varepsilon$ conditional on $[x, z]$ and imposing the IV assumptions the identification strategy will use the following:

$$\begin{aligned}\mathbb{E}[y|x, z] &= \mathbb{E}[g(t, x)|x, z] + \mathbb{E}[\varepsilon|x] \\ &= \int h(t, x) dF(t|x, z) \end{aligned} \tag{2}$$

where the above uses assumption 3): $\mathbb{E}[\varepsilon|x] = \mathbb{E}[\varepsilon|x, z]$ and $dF(t|x, z)$ is the conditional distribution of the treatment given covariates x and the instrument z. Since both $\mathbb{E}[y|x, z]$ and $dF(t|x, z)$ are observable and therefore can be estimated from the data, $h(\cdot)$ is identified by equation (2). Specifically, $h(\cdot)$ is estimated by solving

$$\min_{\hat{h} \in H} \sum_{t=1}^{T} \left(y_t - \int \hat{h}(t, x) dF(t|x_t, z_t) \right)^2 \tag{3}$$

[2] As similar to Angrist (1990).

where H is a arbitrarily flexible function space. In 2SLS H is the class of linear functions of x and z. We will use neural networks instead of linear regression in order to consider a much larger function space.

In practice, the treatment distribution $F(t|x_t, z_t)$ is unknown so we estimate it with $\hat{F}(t|x_t, z_t) \approx F(t|x_t, z_t)$, often called the *nuisance function.* Equation (3) is the second stage of the IV estimation while the estimation of $F(t|x_t, z_t)$ is the first stage.

Standard approaches to IV rely on a linearization of both $\hat{h}$ and $\hat{F}$. This requires the strong assumptions of linearity (both the first and second stage linear regressions are correctly specified) and homogeneity (the policy affects all individuals in the same way).[3] Newey and Powell (2003) develop nonparametric extensions of 2SLS, showing that the linear regressions can be replaced with a linear projection onto a series of known basis functions, a class of *sieve estimators.* While this is an effective strategy for introducing flexibility and heterogeneity with low-dimensional inputs, such estimators become computationally infeasible for anything beyond trivially low-dimensional $[x, z]$. Instead, we follow Hartford et al. (2017) by learning F and h using a deep neural network.

To summarize, the goal is to estimate a counterfactual prediction function $h(t, x) \equiv g(t, x) + \mathbb{E}[\varepsilon|x]$ which traces out the effect of a change in the treatment t on the outcome of interest y, holding the distribution of the error term ε unchanged. This function can be found as the solution to (3), which is analogous to the second stage in 2SLS estimation. However, $F(t|x_t, z_t)$ is not directly observable so it must be approximated with $\hat{F}(t|x_t, z_t)$, which is analogous to the first stage of 2SLS because it predicts the treatment t as a function of covariates x and the instruments z. In the next section we discuss the details of our neural network implementation.

3 Estimation and Neural Network Implementation

3.1 First Stage Estimation: $F(t|x, z)$

The first stage models $F(t|x, z)$ as a mixture of Gaussian distributions

$$F_\phi(t|x, z) = \sum_k \pi_k(x, z; \phi)\mathcal{N}\left(\mu_k(x, z; \phi), \sigma_k^2(x, z; \phi)\right),$$

where component weights $\pi_k(x, z; \phi)$ and parameters $[\mu_k(x, z; \phi), \sigma_k^2(x, z; \phi)]$ form the final layer of a neural network parametrized by ϕ. We model F as a mixture of Gaussian distributions because the treatment is continuous and therefore its conditional distribution must be continuous as well. Hartford et al. (2017) note that in the case of a discrete treatment, one could instead use a multinomial mixture network.

3.2 Second Stage Estimation: $h(t, x)$

In the second stage, we use a deep neural network with a single real-valued output to approximate the counterfactual prediction function h. Following (3), network parameters θ are optimized to minimize the mean-squared error (MSE) loss function, on our training data,

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_t \left(y_t - \int h_\theta(p, x_t)d\hat{F}_\phi(p|x_t, z_t)\right)^2. \tag{4}$$

[3] This can be relaxed somewhat in order to uncover another parameter of interest in the literature, namely LATE.

Note that $\hat{F}_\phi(p|x_t, z_t)$ is the estimated treatment distribution from the first stage. Since we do not have a closed-form solution for the integral

$$\int h_\theta(p, x_t) d\hat{F}_\phi(p|x_t, z_t)$$

we follow Hartford et al. (2017) in performing Monte Carlo integration. Specifically, we replace the integral with an average over values of h evaluated at B different draws of the treatment t_b from the estimated probability distribution $d\hat{F}_\phi(p|x, z)$: $\int h(t)dF(t|x, z) \approx \frac{1}{B}\sum_{b=1}^{B} h(t_b)$. Following Hartford et al. (2017)'s recommendation, we use two draws ($B = 2$). The neural network is then trained using stochastic gradient descent on both stages.

3.3 Neural Network Architecture Selection

We select our choice of hyper-parameters for the artificial neural network by first starting from Hartford et al. (2017)'s choice of hyper-parameters, and then build upon them following heuristic rules.

Hartford's architecture makes use of three hidden layers, containing 128, 64, and 32 neurons respectively. We follow the rule of thumb (Demuth et al. (2014)):

$$N_h = \frac{N_s}{(\alpha * (N_i + N_o))} \tag{5}$$

where N_h is the upper bound for the number of hidden neurons before over fitting occurs, $N_s = 1444$ is the number of training samples, N_i and N_o are the number of input and output layers respectively, and α is a scaling factor between 2 and 10. Thus, when $\alpha = 2$,

$$N_h = \frac{1444}{2 * (2 + 1))} = 244.66$$

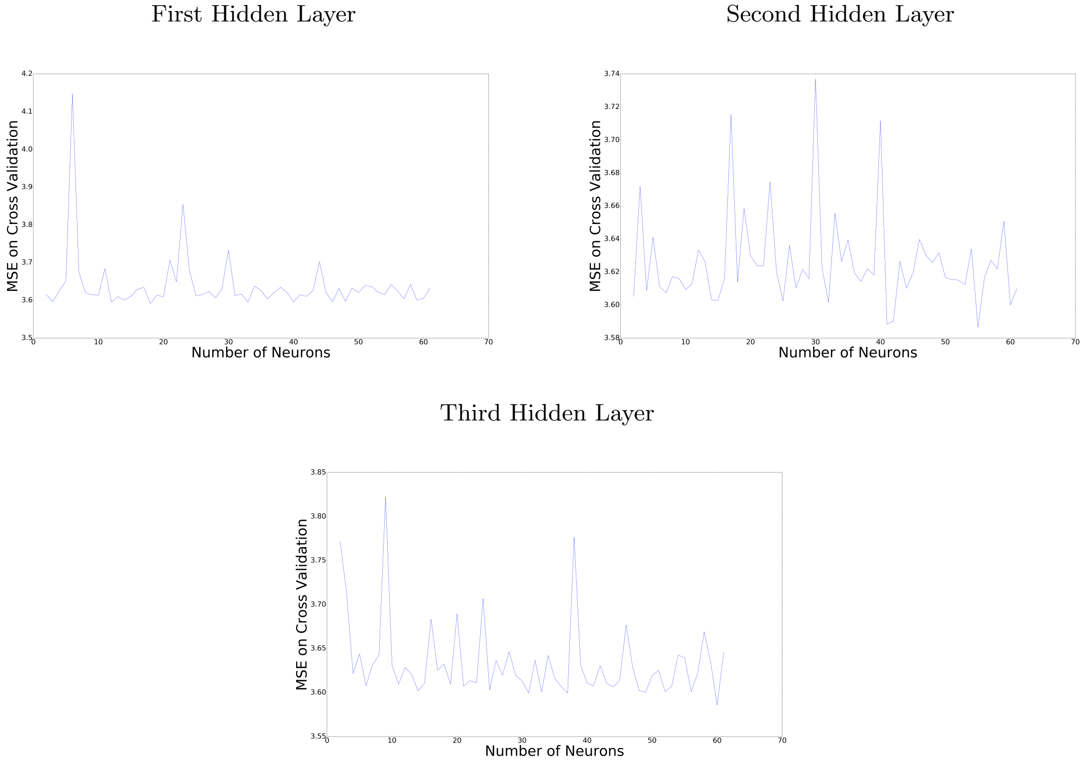
and so we are under the upper bound for small values of α .

In order to tune the number of neurons within each layer with finer granularity, we iteratively search values within 30 of their initial values for one layer at a time; an exhaustive grid search of permutations is far too time consuming. In Figure 1, we see no substantial change in performance as we vary the number of neurons.

We also consider our network's number of hidden layers. Hartford's model has an input layer of size 8, and makes use of 3 hidden layers, while we only have an input layer of size 2. Since we are performing regression, our output layer is of course of size 1. If we follow the rule that the number of hidden layers should fall between the sizes of the input and output layers, then three hidden layers are unnecessary. We test this empirically by validating against a hold-out set 20% the size of the entire data set, and find no meaningful difference in performance when removing the 1st or 3rd hidden layers.

We attribute the model's lack of sensitivity to both the number of hidden layers and the number of neurons within each to the usage of L2 regularization, which pushes the weight value towards zero, and dropout regularization, wherein neurons are randomly left out and ignored both during the forward pass and back-propogation. Since equation 4 describes an upper bound, and our model only satisifes it for small values of α, we hypothesize that the model is actually overfitting; due to the L2 and dropout regularization, however, having an additional hidden layer and/or an excessive number of hidden neurons is within the margin of overfitting.

Figure 1: MSE as a function of neurons for each layer


First Hidden Layer

Second Hidden Layer

Third Hidden Layer

4 Simulations

In this section we run simulations in order to assess the power of the DNN IV method for estimating causal treatment effects in a nonlinear framework, as opposed to conventional IV estimation, i.e. 2SLS method. Since we know the true DGP in the simulations, we can compute the true prediction error and understand under what circumstances each of the estimators prevail over another. In this simulation particularly, we try to mimic the dataset which will be used below for the empirical estimations, in order to create data as realistic as possible. The mimicking happens by copying the dataset moments and using densities as close as possible to the ones observed in the data. It consists in the following:

$$y = g(x,t) + v \qquad \text{1st stage}$$

$$t = f(x,z) + u \qquad \text{2nd stage}$$

$$\begin{bmatrix} v \\ u \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix} \right), \qquad x \sim \text{Ber}(.5), \qquad (z + .73) \sim \text{Gamma}(\alpha = 2.18, \beta = .879)$$

$$f(x,z) = \begin{cases} 0.8 - 0.8x + 0.7z + 0.5zx + u & \text{(linear)} \\ 2 + 3x + 8\sin(0.3z) + [5\sin(0.5z) - 8\sin(0.3z)]\,x + u & \text{(non-linear)} \end{cases}$$

$$g(x,t) = \begin{cases} -0.7 + 0.7x - 0.3t - 0.6tx + v & \text{(linear)} \\ -1 - [0.5(1-x) + 1.1x]\log(\text{abs}(t-1) + 1)\text{sign}(t-1) + v & \text{(non-linear)} \end{cases}$$

The simulations seek to compare the estimation methods varying two dimensions: sample size and the linearity of the 1st and 2nd stages. The sample size varies roughly to the power of 10,

Table 1: Simulation Results: MSE of traditional IV (2SLS) and DeepIV, varying sample size and linearity of the 1st and 2nd stages

	n=1,444		n=10,000		n=100,000	
	2SLS	DeepIV	2SLS	DeepIV	2SLS	DeepIV
Lin-Lin	0.000	0.135	0.000	0.167	0.000	0.167
NL-Lin	0.001	0.639	0.001	0.166	0.000	0.137
Lin-NL	0.072	0.148	0.065	0.087	0.062	0.097
NL-NL	0.024	0.065	0.023	0.024	0.033	0.026

$n = 1,144$ (as in our dataset), $n = 10,000$ and $n = 100,000$. As listed above, $f(\cdot)$ and $g(\cdot)$ have both linear and non-linear functional forms, that will be adopted in order to generate data, each at a time. As a result, we get Table 1.

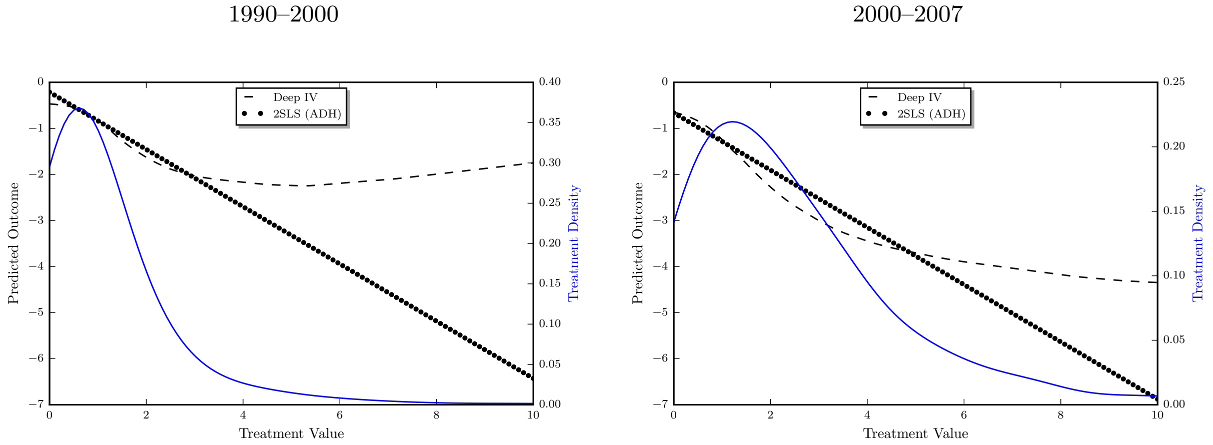
As shown in table (1), when the outcome is a linear function of the treatment, 2SLS does much better than DeepIV (top two rows). However, when the outcome is a nonlinear function of the treatment (bottom two rows) the two methods are much closer. 2SLS is still the better performer for smaller sample sizes, while Deep IV has the lower MSE for 100,000 observations. Therefore, we conclude that Deep IV provides little value when sample sizes are small and the DGP is approximately linear, but may be preferable to 2SLS as sample sizes grow and the DGP becomes more complex and nonlinear.

5 Empirical Application

Our empirical application is an extension of Autor et al. (2013). Autor et al. (2013) study the effect of increased competition from Chinese imports on American manufacturing employment. Essentially, they are testing the hypothesis that American workers facing greater exposure to Chinese imports (meaning that they produced goods that were close substitutes for Chinese imports) are less likely to be employed and, if employed, likely to earn lower wages, as a result of the rise of imports rom China. They divide the U.S. into two time periods and 722 community zones (CZs), which can be thought of as local labor markets, and test the hypothesis that CZs with a greater density of industries subject to competition from China subsequently had lower employment in manufacturing. A naive approach would be to simply estimate a linear regression of changes in manufacturing employment on changes in Chinese import competition, possibly including some linear control variables. However, it is likely in this context that there are unobserved factors correlated with both manufacturing employment and import competition, which would cause the naive regression to yield biased estimates. Faced with this concern, Autor et al. (2013) implement an IV estimation strategy.

Autor et al. (2013)'s main specifications use lagged Chinese import growth in other high-income countries as an instrument for Chinese import growth to the U.S. They perform extensive analysis to support the validity of their instrument so we assume it is valid and proceed by assessing the degree to which their results would change if they relaxed the assumption of a linear model underlying their two-stage least squares estimation strategy and instead used the flexible neural network-based estimation approach proposed by Hartford et al. (2017). Figure 2 compares ADH's estimated regression function (dotted line) with our Deep IV (dashed line) for their most basic specification with no covariates and a separate intercept for each time period. The horizontal axis represents the size of the import competition shock while the left vertical axis is the annual change in manufacturing employment as a share of the working age population, multiplied by ten. The blue line and the right vertical axis present a kernel density plot for the distribution of the treatment (import competition). Figure 2 shows that our Deep IV yields quite similar results to ADH's linear specification for import competition shocks less than 3 in 1990–2000 and less than 5 in 2000–2007. Deep IV suggests that the linear approximation for the treatment effect performs poorly for larger values of the treatment, however as shown by the blue lines, values of the treatment large enough for the linear regression to be a poor approximation occurred only rarely in the data. Therefore, at least in this context, we conclude that Deep IV provides qualitatively similar results to ADH's linear approach. Since we do not know the true underlying DGP it is impossible to say which method is closer to the truth.

Figure 2: Comparing Deep IV to ADH's 2SLS



6 Conclusion

This project investigates the problem of adapting machine learning techniques to causal inference tasks of interest to economists and social scientists, among others. We build on the Deep IV framework proposed by Hartford et al. (2017) and apply and extend their work in four key ways. First, we conduct rigorous analysis in order to choose a neural network architecture appropriate to our context. Second, we compare the performance of Deep IV and linear 2SLS on data sets that we simulated to have similar moments to the real data set we use for our empirical application. Third, we apply Deep IV to a seminal paper in economics to assess how the results would change if the assumption of a linear model is relaxed. Finally, in our empirical application we show how to summarize the results of Deep IV by plotting conditional expectation functions.

We find that Deep IV provides little improvement over 2SLS in our empirical application and actually performs worse than 2SLS on our simulated data, especially for small sample sizes. This is likely because we only consider an application with a single covariate and a relatively small sample size, making this an application in which Deep IV is likely to have less value-added than in other contexts. However, Deep IV's relative performance is likely to improve in analyses with more covariates and more observations, so moving beyond our simple application is an important and likely fruitful area for future exploration.

An important area of further research will be to better understand when 2SLS performs adequately, and when methods like Deep IV can provide further insight. For example, we envision a statistical test in which a neural network such as the Deep IV framework would reject the null hypothesis of linear and homogeneous treatment effects if the true treatment effects are sufficiently nonlinear and heterogeneous. Finally, we hope to develop methods for statistical inference for methods that combine machine learning and causal inference.

References

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, 1990, *80* (3), 313–336.

Autor, David H., David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, October 2013, *103* (6), 2121–68.

Demuth, Howard B, Mark H Beale, Orlando De Jess, and Martin T Hagan, *Neural network design*, Martin Hagan, 2014.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy, "Deep IV: A Flexible Approach for Counterfactual Prediction," in Doina Precup and Yee Whye Teh, eds., *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research* PMLR International Convention Centre, Sydney, Australia 06–11 Aug 2017, pp. 1414–1423.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, 1986, *81* (396), 945–960.

Newey, Whitney K. and James L. Powell, "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 2003, *71* (5), 1565–1578.