A Causal Inference Algorithm for Heterogeneous Treatment Effects

Presentation to MIT Econometrics Lunch

Daniel Aronoff

MIT

October 14, 2021

Acknowledgement: Victor Orestes, MIT Economics and Violet Felt, MIT EECS contributed to this project

Roadmap

1. Motivation
2. Approach
3. Framework
4. The Learning Problem
5. Neyman Orthogonality
6. The Algorithm
7. Next Steps

Motivation

High - dimensional data enables more granular partitioning of a population

"Personalization" of treatment effects

Personalized Medicine Assign treatments for groups defined by e.g. genetic information, social interactions *in addition to* traditional markers

Personalized Pricing Offer 'deals' to consumers within groups defined by e.g. social media activity *in addition to* traditional markers

Personalized Attention Identify and exploit individualized social media habits associated with responsiveness to particular inducements (e.g. Instagram appealing to vulnerable girls - maybe not so good)

✓ How do we learn the partitions of the data that separate the population into homogeneous treatment response groups?

Motivation

The treatment effect parameter $\theta(X)$ will be highly expressive

That can lead to problems

- Over-fitting

- Local Imbalances

- Difficulty partitioning X into homogeneous treatment response groups

Motivation

It may be difficult to obtain a high quality estimator of the propensity function p_o

With a quasi - linear model

$$Y = \theta(X)T + f(X) + \epsilon_1$$

and a squared loss function

$$L_D(\theta, \{\mathbb{E}[Y|X], p_o\}) = \mathbb{E}[((Y - \mathbb{E}[Y|X]) - (\theta(X)(T - p(X))^2]$$

A low quality estimator of p_o implies that Neyman Orthogonality does not obtain

$$D_p D_\theta L_D(\theta, \{\mathbb{E}[Y|X], p_o\})|\theta' - \theta, p - p_o| \neq 0$$

Approach

We propose an algorithm that negotiates the tension in estimating p_o

On one hand, we want a $p(X)$ that predicts the treatment T
- This does not ensure good balance in the final outcome

On the other hand, we want a $p(X)$ that enforces Neyman Orthogonality
- This helps to enforce local balance, but does not ensure a good prediction of the treatment

Approach

We propose an algorithm that negotiates the tension in estimating p_o

On one hand, we want a $p(X)$ that predicts the treatment T
- This does not ensure good balance in the final outcome

On the other hand, we want a $p(X)$ that enforces Neyman Orthogonality
- This helps to enforce local balance, but does not ensure a good prediction of the treatment

$\hookrightarrow$ An adversarial algorithm

Player 1 Estimate $p(X)$ to predict T *subject to* a Neyman Orthogonality constraint

Player 2 For any choice of $p(X)$, choose the DGP that pushes the Neyman Orthogonality moment condition furthest from zero

Iterate over $\theta(X)$ until convergence

Approach

Intuition

I1 If our propensity model *is* correct, then the Neyman Orthogonality moment condition should be satisfied for any DGP in the class of admissible DGP's

I2 If our propensity model *is not* correct (i.e. the function class from which we choose $p(X)$ does not contain the true function) then imposing Neyman Orthogonality might lead us to choose the most efficient $p(X)$ in the class

Approach

There are questions we have not (yet) addressed

▷ What restrictions on the problem must obtain for $\hat{\theta}(X)$ convergence to hold?

▷ When will convergence imply that $\hat{\theta}(X)$ or $L_D(\hat{\theta}(X), \mathbb{E}[Y|X], p(X))$ is of high quality in some meaningful sense?

We seek direction and/or participation in addressing those questions

Literature

Orthogonal ML Approach to Causal Inference

Chernozhukov et.al (2018) *Double Machine Learning for Treatment and Causal Parameters*

Foster and Syrgkanis (2020) *Orthogonal Statistical Learning*

Estimating High Dimensional Heterogeneity

Semanova et. al.(2021) *Estimation and Inference on Heterogeneous Treatment Effects in High - Dimensional Dynamic Panels*

Farrell et. al. (2021) *Deep Learning for Individual Heterogeneity*

GAN Estimation with Regularization

Liang (2021) *How Well Generative Adversarial Networks Learn Distributions*

Framework

Foster & Syrgankis (2020) *Orthogonal Statistical Learning*

Provides a meta-algorithm for estimating treatment effects in causal inference models with guarantees on excess risk and/or convergence of the treatment parameter

The meta-algorithm assumes the existence of a high- quality black-box estimator of nuisance parameters

We attempt to fill in the black-box with an estimation algorithm of the nuisance parameters and the target parameter in a setting where the treatment function is highly expressive and the partitions of the data into homogeneous response groups needs to be learned

In this setting, the estimate of the propensity function may be biased

Our goal is to describe an algorithm that nevertheless achieves good generalization performance

The Learning Problem

Heterogeneous Treatment Model

$$\underbrace{Y}_{\text{target variable}} = \underbrace{\theta_0(X)}_{\text{target parameter}} \cdot T + f_0(X) + \epsilon_1, \quad \mathbb{E}[\epsilon(T, X)] = 0, \quad T \in \{0, 1\}$$

$T = p_0(X) + \epsilon_2$

$\theta(X)$, p_0 and f_0 are unknown. $g_0 = \{\mathrm{E}[Y|X], p_0\}$ are nuisance parameters

The Goal(S) :

▷ Estimate $\hat{\theta}$ as the minimizer of an expected loss function

$\theta_0 = argmin_{\theta \in \Theta} L_D(\theta; g_0)$

Where we need to estimate the nuisance functions $g_0 \in \mathcal{G} = \{\mathcal{Y}, \mathcal{P}\}$

▷ **Oracle Excess Risk** Given n samples, find $\hat{\theta}$ to minimize

$L_D(\hat{\theta}, g_0) - L_D(\theta_0; g_0) \leq R_n$

The Learning Problem

A solution was provided by Foster & Syrgkanis

Meta-Algorithm

Give n samples S, split into S_1 and $S_2 = S/S_1$

Stage 1 Estimate $\hat{g}$ on S_1 with estimation error rate bound

$$||\hat{g} - g_0||_{\mathcal{G}} \leq \text{Rate}_D(\mathcal{G}, S_1, \delta) \quad \text{wp} \;\; 1 - \delta$$

Stage 2 Estimate $\hat{\theta}$ on S_2. Plug in $\hat{g}$; obtain excess risk bound

$$L_D(\hat{\theta}, \hat{g}) - L_D(\theta_0; g) \leq \text{Rate}_D(\Theta, S_2, \delta; g) \quad \text{wp } 1 - \delta$$

$$L_D(\hat{\theta}, \hat{g}) = \mathbb{E}[(\tilde{Y} - \hat{\theta}(X)\,\tilde{T})^2] \quad \text{where}$$

$$\tilde{Y} = Y - \mathbb{E}[Y|X] = \epsilon_1, \quad \tilde{T} = T - p(X) = \epsilon_2$$

The Learning Problem

Theorem (F & S Thm 1 - Convex Loss Function)

Suppose there is some $\theta^* \in argmin_{\theta \in \Theta} L_D(\theta, g_0)$ *and regularity conditions are met. Then the Meta-Algorithm produces a* $\hat{\theta}$ *such that wp at least* $1 - \delta$

.

$$||\hat{\theta} - \theta^*||_\Theta^2 \leq \frac{4}{\lambda} Rate_D(\Theta, S_2, \delta/2; \hat{g}) + \frac{\beta_1}{2\lambda}(\frac{\beta_2^2}{\lambda} + 2\kappa) \cdot (Rate_D(\mathcal{G}, S_1, \delta/2))^4$$

.

$$L_D(\hat{\theta}, g_0) - L_D(\theta_0; g_0) \leq \frac{2\beta_1}{\lambda} Rate_D(\Theta, S_2, \delta/2; \hat{g}) + \frac{1}{\lambda}(\frac{\beta_2^2}{\lambda} + 2\kappa) \cdot (Rate_D(\mathcal{G}, S_1, \delta/2))^4$$

Theorem 1 says (roughly) that, if the loss function is smooth and we have high quality estimators of the nuisance parameters and Neyman Orthogonality holds, we can achieve a fast rate of convergence to the 'best' target parameter estimate and the minimum loss

The Learning Problem

Theorem (F & S Thm 1 - Convex Loss Function)

Suppose there is some $\theta^* \in argmin_{\theta \in \Theta} L_D(\theta, g_0)$ *and regularity conditions are met. Then the Meta-Algorithm produces a* $\hat{\theta}$ *such that wp at least* $1 - \delta$

.

$$||\hat{\theta} - \theta^*||_\Theta^2 \leq \frac{4}{\lambda} Rate_D(\Theta, S_2, \delta/2; \hat{g}) + \frac{\beta_1}{2\lambda}\left(\frac{\beta_2^2}{\lambda} + 2\kappa\right) \cdot (Rate_D(\mathcal{G}, S_1, \delta/2))^4$$

.

$$L_D(\hat{\theta}, g_0) - L_D(\theta_0; g_0) \leq \frac{2\beta_1}{\lambda} Rate_D(\Theta, S_2, \delta/2; \hat{g}) + \frac{1}{\lambda}\left(\frac{\beta_2^2}{\lambda} + 2\kappa\right) \cdot (Rate_D(\mathcal{G}, S_1, \delta/2))^4$$

Theorem 1 says (roughly) that, if the loss function is smooth and we have high quality estimators of the nuisance parameters and Neyman Orthogonality holds, we can achieve a fast rate of convergence to the 'best' target parameter estimate and the minimum loss

Key Point The Meta-Algorithm assumes (i) the existence of *black - box* algorithms to approximate the nuisance parameters and (ii) Neyman Orthogonality

✓ What if these conditions are not met?

Neyman Orthogonality

For $L_D(\theta, \{\mathbb{E}[Y|X], p_0\}) = \mathbb{E}[((Y - \mathbb{E}[Y|X]) - (\theta(X)(T - p(X))^2]$ we obtain for the nuisance parameter $p(X)$

$D_p D_\theta L_D(\theta, \{\mathbb{E}[Y|X], p_0\})|\theta' - \theta, p - p_o|$

$= 2 \cdot \mathbb{E}[((Y - \mathbb{E}[Y|X]) - \theta(X)(T - p_0(X)))(\theta'(X) - \theta(X)))(p(X) - p_0(X))]$

$-2 \cdot \mathbb{E}[(\theta(X)(T - p_0(X))(\theta'(X) - \theta(X))(p(X) - p_0(X)))]$

The expression equals zero when, for any x,

$\mathbb{E}[\theta(X)(T - p_0(X))|X = x] = 0$

Neyman Orthogonality

For $L_D(\theta, \{\mathbb{E}[Y|X], p_0\}) = \mathbb{E}[((Y - \mathbb{E}[Y|X]) - (\theta(X)(T - p(X))^2]$ we obtain for the nuisance parameter $p(X)$

$$D_p D_\theta L_D(\theta, \{\mathbb{E}[Y|X], p_0\})|\theta' - \theta, p - p_o|$$

$$= 2 \cdot \mathbb{E}[((Y - \mathbb{E}[Y|X]) - \theta(X)(T - p_0(X)))(\theta'(X) - \theta(X)))(p(X) - p_0(X))]$$

$$-2 \cdot \mathbb{E}[(\theta(X)(T - p_0(X))(\theta'(X) - \theta(X))(p(X) - p_0(X)))]$$

The expression equals zero when, for any x,

$$\mathbb{E}[\theta(X)(T - p_0(X))|X = x] = 0$$

$\hookrightarrow \hat{p} \sim p_0$ is a sufficient condition for Neyman Orthogonality

But the necessary condition is less strict. In the sample it is for $\mathbb{E}_n[\hat{\theta}(X)(T - \hat{p}(X))]$ to be zero

The Algorithm

We propose an algorithm to estimate the nuisance parameters and the target parameter for the heterogeneous partially linear treatment model with an MSE loss function

We assume it is possible to obtain a high quality estimate of $\mathbb{E}[Y|X]$ directly from the data

We use iterations to enforce the moment condition $\mathbb{E}_n[\hat{\theta}(X)(T - \hat{p}(X))] = 0$

This requires an initialization step because we must initiate an estimate of p_o before we can estimate $\theta(X)$

Conjecture If our initial estimate of p_o is not 'high quality', possibly because the function class $\mathcal{P}$ does not contain p_0, we might still satisfy Theorem 1 (F & S) by enforcing Neyman Orthogonality in the sample

That is, when we choose the best predictor of T subject to an orthogonality constraint, can we recover a "high quality" estimate of $\theta(X)$?

The Algorithm

Step 1[Initialization]

Estimate the nuisance parameters on data S_1

Estimate $\mathbb{E}[Y|X]$ using an ML method

$\tilde{Y} = Y - \mathbb{E}_n[Y|X]$

The Algorithm

Step 1 - continued

Estimate $\mathbb{E}[T|X]$ using an ML method with $[\epsilon, 1-\epsilon]$ loss function. One possible function is a (bounded) kernel Logistic function

$$p(X=x) = \sigma(\omega_1\phi(x)+\omega_0) = \frac{\epsilon^{\omega_1\phi(x)+\omega_0}}{1+\epsilon^{\omega_1\phi(x)+\omega_0}}$$

Estimate ω by solving the negative Log likelihood moment with an L_2 regularizer

$$\hat{\omega}_1, \hat{\omega}_0 = \text{argmin} L(\omega) =$$
$$\mathbb{E}_n[\underbrace{-\log\{\sigma(\omega_1\phi(x)+\omega_0)\cdot 1\{T=1\} + \log(1-\sigma(\omega_1\phi(x)+\omega_0)\cdot 1\{T=0\}\}}_{\text{Negative Log likelihood}} +$$
$$\underbrace{\lambda||\omega||^2]}_{L_2 \text{ regularizer}}$$

The Algorithm

Step 2

Estimate the treatment function on data S_2

Estimate $\hat{\theta}(X)$ by minimizing the MSE loss function using the estimates of $\tilde{Y}$ and $\tilde{T}$ obtained in Step 1, and cross validating

$$\hat{\theta}(X) = \text{argmin}_{\theta(X)} \mathbb{E}_n[(\tilde{Y} - \underbrace{\underbrace{\theta(X)}_{\text{target parameter}} \quad (T - \hat{p}(X))^2}_{\text{estimator}}] \tag{1}$$

The Algorithm

Step 2

Estimate the treatment function on data S_2

Estimate $\hat{\theta}(X)$ by minimizing the MSE loss function using the estimates of $\tilde{Y}$ and $\tilde{T}$ obtained in Step 1, and cross validating

$$\hat{\theta}(X) = \text{argmin}_{\theta(X)} \mathbb{E}_n[(\tilde{Y} - \underbrace{\underbrace{\theta(X)}_{\text{target parameter}} \quad (T - \hat{p}(X))^2}_{\text{estimator}}] \tag{1}$$

If $0 < p(x) < 1 \; \forall x \in X$ and the estimator is orthogonal to X, then $\hat{\theta}(X)$ is a consistent estimate of $\theta_0(X)$

$\hookrightarrow$ If $p(X)$ is either "low" quality or biased, $\hat{\theta}(X)$ might be low quality or biased...

...which motivates the next step

The Algorithm

Step 3

Estimate the propensity function on data S_1

Next, we re-estimate the propensity function $p(X)$, regularizing it with an orthogonality constraint computed with $\hat{\theta}(X)$ obtained in Step 2 (the functional form of p(X) is defined as in Step 1)

To obtain consistency in our estimate of $\theta(X)$ we require that the estimator $\hat{\theta}(X)\tilde{T}$ be (as close as possible to being) orthogonal to X

Since we observe only a sample, we enforce orthogonality with an adversarial selection of a distribution $q(X)$, which ideally is highly expressive

The Algorithm

Step 3 - continued

The estimation problem can be expressed as a minimization of logistic loss constrained to distributions of $p(X)$ that induce an estimator that is orthogonal to X...under the most adverse distribution of X

We split S_1 into 2 sub-samples, S_{1p} and S_{1q}. On S_{1p} we compute

$$\text{argmin}_{p(X)\in\mathcal{P}} - \mathbb{E}_q[\underbrace{\log(p(X)) \cdot 1\{T=1\} + \log(1-p(X)) \cdot 1\{T=0\}}_{\text{Logistic Log likelihood}}] \quad (2)$$

Subject to (on S_{1q})

$$\text{argmax}_{q(X)\in\mathcal{Q}} |\mathbb{E}_q[\underbrace{\hat{\theta}(X)(T-p(X))}_{\text{estimator}}]| \leq \xi \quad (3)$$

Where $p(X) \in \mathcal{P}$ is a function class and $q(X) \in \mathcal{Q}$, an expressive class of distributions

The Algorithm

Step 3 - continued

For computation, we solve the Lagrangian representation of the estimation problem

It is an adversarial min-max problem

$$\hat{p}(X) = \text{argmin}_{p(X)\in\mathcal{P}}\{\text{argmax}_{q(X)\in\mathcal{Q}}$$

$$\mathrm{E}_q\Big[-\Big(\underbrace{\log\{p(X))\cdot 1\{T=1\} + \log(1-p(X))\cdot 1\{T=0\}}_{\text{Logistic Log likelihood}}\Big) + \gamma \underbrace{\hat{\theta}(X)(T-p(X))}_{\text{orthogonal regularizer}}\Big] \quad (4)$$

Where the left object is computed on S_{1p} and the right object is computed on S_{1q}

The Algorithm

Step 4

Index each iteration of $\hat{\theta}(X)$ and $\hat{p}(X)$, $i = 1, 2...$

for $i = 1$ return to Step 2

For $i > 1$

Choose τ

Check if $||p_i(X) - p_{i+1}(X)||_2^2 = \delta_{p_{i+\mathbf{1}}}$, $||\theta_i(X) - \theta_{i+1}(X)||_2^2 = \delta_{\theta_{i+\mathbf{1}}}$. If $\delta_{p_{i+\mathbf{1}}} > \tau$ or $\delta_{\theta_{i+\mathbf{1}}} > \tau$ continue to next round

Otherwise stop

Open Issues/Next Steps

Algorithm Design

The algorithm performs well in simulations without high dimensional data (where $\mathcal{Q}$ is a multivariate normal distribution). This does not imply good generalization to the high dimensional case

A key challenge is how to choose an appropriate class $\mathcal{Q}$, so that we do not have "anything goes" (i.e, point mass distributions) and still maintain tractability

✓ We do not want to have complementarity between the adversaries

Can we apply other methods such as a Generative Adversarial Network in step 3?

We must integrate the iterative procedure into the theoretical analysis

Open Issues/Next Steps

Algorithm Guarantees

Does the algorithm converge to a fixed point of $\hat{\theta}(X)$?

Does the/a fixed point of the algorithm approximately coincide with the true $\theta(X)_0$ or the minimizer of the Oracle excess risk rate bound?

What is the effect on generalization bounds and convergence if the function class $\mathcal{P}$ is mis-specified but the estimate of $\theta(X)(T - p(X))$ meets Neyman Orthogonality?

Does each iteration of the algorithm tighten the convergence bounds of the estimators for $\theta(X)$ and $p(X)$?

What are the limitations on underlying class of distributions and functional forms under which the algorithm works?

Thank You!