Hidden Confounder with Noisy Ordinal Measurements

Chase Mathis

June 2024

Overview

I have been thinking of the following two ideas, closely related to Unobserved Confounding:

- Bootstrapping two proxy confounders from only one
- Under linear constraints + finite samples, Tchetgen Tchetgen's Least Squares estimator admit large values quite often. (A good overview of this is in this paper on page 13).
- How Cross-Fitting is important for these estimators in finite samples.

Bayesian Boostrapping: Basic Idea

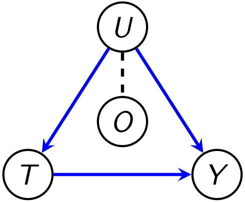


Figure: Observed Ordinal Data

- What is the *ATE* of T on Y.
- We *only* have ordinal data. [1]

[1] I will focus on orindal data, but I perhaps this method can extend to other noisy measurements.

Tchetgen and Proximal Causal Inference

Tchetgen and Miao has made some theoretical work with Proxy Latent Variables.

The basic (and fascinating) idea of theirs is that one can “do away” with the unobserved confounder, in the prescence of two other *independent* proxy variables [4].

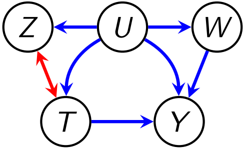


Figure: General Framework: (1) Edges between Z and T can be bi-directed. (2) Edges between (Z, T) and (W, Y) can be omitted.

Proximal Causal Inference with Do Calculus

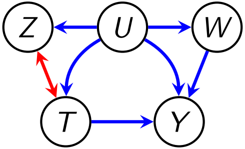


Figure: General Framework

- Tchetgen shows non-parametrically, that you can recover $p(y \mid \text{do}(T))$ [4].

Quick Proof of Linear Case

Taken from: [3] Assumptions:

$$
\begin{aligned}
&E[Y \mid A, Z, U] = \beta_0 + \beta_a A + \beta_u U \\
&E[W \mid A, Z, U] = \alpha_0 + \alpha_u U \\
&|E[U \mid A, Z]| < \infty
\end{aligned}
$$

$$
\begin{aligned}
E[W \mid A, Z] &= \alpha_0 + \alpha_u E[U \mid A, Z] \\
E[Y \mid A, Z] &= \beta_0 + \beta_a A + \beta_u E[U \mid A, Z]
\end{aligned}
$$

Equations (27) and (28) imply Result 1 in the main text as follows:

$$
E[Y \mid A, Z] = \beta_0^* + \beta_a^* A + \beta_u^* E[W \mid A, Z]
$$

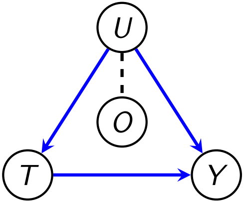
$$
\text{where } \beta_0^* = \beta_0 - \beta_u \frac{\alpha_0}{\alpha_u}, \beta_a^* = \beta_a, \beta_u^* = \frac{\beta_u}{\alpha_u}
$$

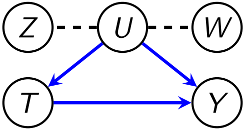
Proximal control variable S is a linear transformation of $E[U \mid A, Z]$, i.e.

$$
S = E[W \mid A, Z] = \alpha_0 + \alpha_u E[U \mid A, Z] \propto_L E[U \mid A, Z]
$$

“Bootstrapping another Proxy”

- One needs two proxy variables to identify $p\,(y \mid do(t))$.
- Can we “add” another independent Proxy?
- Can we go from the left to the right?

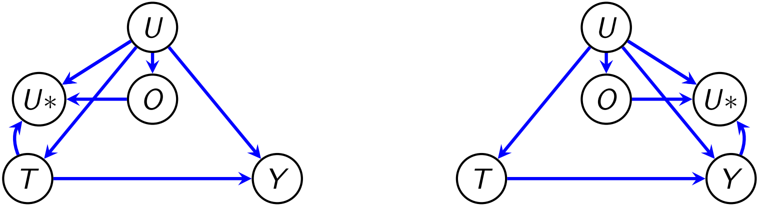




Multivariate Normal with Ordinal O

Let $(T, Y, U) \sim N(\theta, \Sigma)$. [2] And let O be ordinal data. Using Bayesian Probit Regression, we can simulate values of U, call them $U*$ given (T, Y, O) [2].

If we use only T or only Y, we can recover the following graph:



[2]While proxy causal learning works non-parametrically, I focus on multivariate normal. When linear regression is uses, Tchetgen dubs this method: Proximal Two Stage Least Square (P2SLS)

Issues with Conditional Independence

In either case, because we are using one proxy to generate the new proxy $U*$, we have that

$$U* \not\perp\!\!\!\perp O \mid U.$$

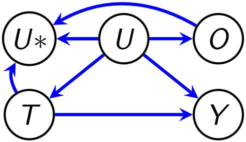
The remarkable thing is that using Bayesian sampling, empirically

$$U* \not\perp\!\!\!\perp U \mid O.$$

We actually *do* get new information not in U'.

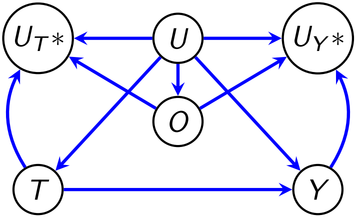
How to fix Issues of Dependence

If we could force or transform $U* \perp\!\!\!\perp O \mid U$, then we can use Tchetgen's Proximal 2 stage least squares.

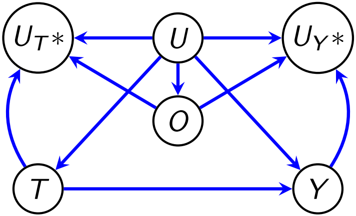


One idea: Using Both $U*$ simuations

Instead of using either dependence, structure, we can simply use both:



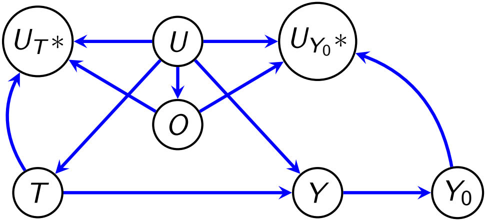
Issues with this Causal Diagram



While it is ok for $T \in \text{pa}_{U_{T*}}$, if $Y \in \text{pa}_{U_{Y*}}$, Tchetgen's estimator is biased.

One Solution: Add another noisy proxy

Instead of sampling directly using Y, we can add a noisy proxy Y_0



Then we can estimate β_{Y_0} using the backdoor criterion: Regress

$$U_{Y_0}* \sim Y_0 + Y$$

And residualize away the direct effect by setting:

$$U_{Y_0}* = U_{Y_0} * - \hat{\beta_{Y_0}} \times Y_0$$

An issue with this is that we're only removing the linear component. We hope and expect that Y_0 has a non-linear relationship with $U_{Y_0}*$

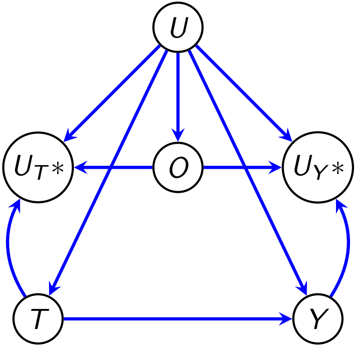
Simulation Method

Let $(U, T, Y) \sim N_3(\mu, \Sigma)$.

- Generate U, T, Y by sampling edge coefficients uniformly at random in $[-5, 5] \subset \mathbb{R}$
- I sample dispersion parameters uniformly at random in $[1, 10] \subset \mathbb{Z}$.
- Generate U' by (with probability 0.5)
  - binning it into groups of 5[3] or
  - $U' \sim \text{Binomial}(5, \text{expit}(U))$
- Calculate estimates with 100,000 simulations; 50,000 using cross-fitting; 50,000 without cross-fitting

[3] Likert scale influence

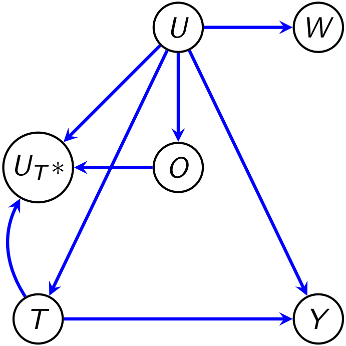
Simulation (Complete Bootsrap)



Graph 1: one ordinal proxy

- Naive: Regress $Y \sim \beta_O O + \beta_T * T$
- Boot: $\hat{W} \equiv \mathbb{E}[U_{Y*} \mid T, U_{T*}, O]$, then regress

Simulation (Partial Bootstrap with One Normal Proxy)

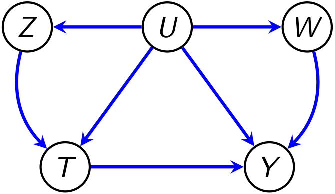


Graph 2: one ordinal proxy, one normal proxy

$\hat{W} \equiv$

- Naive: $\mathbb{E}[W \mid T, O]$
- Boot: $\mathbb{E}[W \mid T, U_{T*}, O]$

Simulation (baseline)



Simulation Results

	Slide 1		Slide 2		Slide 3	Oracle
crossfit	ATE Boot	ATE Naïve	ATE 1 Boot	ATE 1 Naïve	P2SLS	Oracle
No	0.64	0.94	0.06	6.83	7.59	0.03
Yes	0.67	0.94	0.72	1.2	1.95	0.03

Discussion

There are two things I notice:

- Generally, the Bayesian Bootstrap is a better method when there are no proxies.
- Tchetgen's P2SLS is oftentimes better (low median MSE; not shown), but sometimes is quite bad high Mean MSE.

Why is P2SLS sometimes Unstable?

The issue is that in the first stage of regression, the control proxy variable has a beta estimate near zero.
In P2SLS,

$$\hat{w} = a * T + b * U2$$

and sometimes $\mid b \mid \approx 0$.
In the second stage, we have

$$Y = \beta * T + \alpha * \hat{w} + \varepsilon$$

If $\hat{\beta} = \frac{Cov(T,Y|\hat{w})}{\mathrm{var}(T|\hat{w})}$, then $Var(T \mid \hat{w}) \approx 0$, which leads to unstable results.
If we can describe the probability that $b = 0$, then we can find the probability ATE estimate is off by a lot. [4]

---

[4] I am thinking about this. If it was simply $\frac{Cov(Z,Y)}{Cov(Z,T)} = \hat{\beta}_{IV}$, then you $Cov(Z, T) \sim$ variance gamma This is more difficult, though.

Cross-Fitting to help Unstable P2SLS

Cross-fitting, used in Double Machine learning [1] helps to produce better estimates for large machine learning models. I found that applying cross-fitting to P2SLS has the same result.

Questions

- Could this method work for more than just ordinal data?
- How does this method scale? I am unsure if a Bayesian estimator is consistent.
- How can we residualize away more than just the linear part?

References

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins.
Double/Debiased Machine Learning for Treatment and Causal Parameters, December 2017.
arXiv:1608.00060 [econ, stat].

Peter D. Hoff.
Latent variable methods for ordinal data.
In Peter D. Hoff, editor, *A First Course in Bayesian Statistical Methods*, pages 209–223. Springer, New York, NY, 2009.

Jiewen Liu, Chan Park, Kendrick Li, and Eric J. Tchetgen Tchetgen.
Regression-Based Proximal Causal Inference, June 2024.
arXiv:2402.00335 [stat].

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao.
An Introduction to Proximal Causal Learning, September 2020.