Targeted Maximum Likelihood Estimation for Causal Inference with Network Data

Paul Zivich

Post-Doctoral Researcher
Department of Epidemiology
Causal Inference Research Laboratory
UNC Gillings School of Global Public Health

March 30, 2022

Acknowledgements

Supported by NIH T32-HD091058 and T32-AI007001.

Slides are based on my dissertation work. Special thanks to my dissertation committee: Allison Aiello (chair), M Alan Brookhart, Michael Hudgens, James Moody, David Weber.

Additional thanks to Betsy Ogburn, Stephen Cole, and Jessie Edwards for additional discussions.

Work in-progress, so any errors are mine.[1]

pzivich@unc.edu @PausalZ pzivich

[1] Footnotes are reserved asides for possible later discussion

Outline

Causal inference with potential outcomes

- Independent data
- Dependent data
- Parameter of interest

Assumptions

Network-TMLE

- Overview
- Detailed look at each step

Illustrative example

Notation

A: binary action of interest (e.g., treatment, exposure, etc.)
Y: outcome of interest (binary or continuous)
W: vector of baseline variables

$E[\dots]$: expected value function
$\Pr(\dots)$: probability function

Causal inference with potential outcomes

Causal inference

Primary concern will be estimation of causal effects

- What would have been the mean outcome if some units had taken an action
- Focus is on average of unit's outcomes, not the network
- Need to define what a 'causal effect' is

Potential outcomes

- Let $Y_i(a)$ be the potential outcome under action a
- The outcome i will have if received a

Causal effects

Population causal mean

$$E[Y(a=1)]$$

Policy (indicated by ω)

- Algorithm that assigns a for individuals
- Can think of function that assigns probabilities for a
- Here, the policy is: $\Pr^*(A_i = 1) = 1$

Stochastic causal effects

Previous policy was deterministic

- Assigned a fixed value of A to each unit

Generalization for stochastic policies

- Policy is: $0 \leq \Pr^*(A_i = 1|W_i) \leq 1$
- Example: $\Pr^*(A_i = 1) = 0.75$

Population causal mean

$$E\left[\sum_{a \in \mathcal{A}} Y(a) \overset{*}{\Pr}(A_i = a|W_i)\right]$$

Here, $\mathcal{A} = \{0, 1\}$.[2]

[2] To see why stochastic policies are a generalization, try plugging in the deterministic policy from the previous slide

Something is missing...

Previous causal means relied on an assumption

- Potential outcome only depended on a of i
- Formally, the assumption of no interference[3]

Questionable in a variety of contexts

- Examples: vaccination, behaviors
- Can lead bias when connections are ignored[4]

[3] The term 'interference' originates from Cox (1958). Unfortunately, the term implies that this is a nuisance and not of immediate interest.

[4] See Zivich et al. (2021) *AJE* for an example in observational data

Causal effects with interference

Let $Y_i(\mathbf{a}) = Y_i(a_i, a_{-i})$ be the potential outcome, where

$$\mathbf{a} = (a_1, a_2, ..., a_n)$$

$$a_{-i} = (a_1, a_2, ..., a_{i-1}, a_{i+1} ..., a_n)$$

Now potential outcome is uniquely defined by all $\mathbf{a}$

Parameters of possible interest

Unit-specific (direct) effect

$$E[Y(a_i = 1, a_{-i}) - Y(a_i = 0, a_{-i})]$$

Spillover (indirect) effect

$$E[Y(a_i = 0, a_{-i}) - Y(a_i = 0, a'_{-i})]$$

Total effect

$$E[Y(a_i = 1, a_{-i}) - Y(a_i = 0, a'_{-i})]$$

Overall effect

$$E[Y(\mathbf{a}) - Y(\mathbf{a}^*)]$$

Causal effects with interference

Problem: excessively large number of possible potential outcomes

- $n = 10$ means $2^{10} = 1024$ possible $Y_i(\mathbf{a})$
- $n = 20$ means $2^{20} = 1048576$ possible $Y_i(\mathbf{a})$

Will use some assumptions to restrict this

Types of interference

- Partial interference[5]
- General interference

[5] Not discussed further here. See Hudgens & Halloran (2008) or Halloran & Hudgens (2016) for details and approaches

General interference

In principle, allow any two units to 'interfere' with each other

- But only consider those connected in a network
- Adjacency matrix $\mathcal{G}$

Further assumptions to reduce the problem

- Only consider immediate contacts
  - j only matters for $Y_i(\mathbf{a})$ if edge between i and j
  - Refer to as weak dependence throughout

- Assume impact of immediate contacts can be expressed via a summary measure
  - Denoted by A_i^s generally
  - Example: $A_i^s = \sum_{j=1}^{n} \mathcal{G}_{ij} A_j$

Parameter of interest

Hereafter, interested in following parameter

$$\psi = E\left[n^{-1}\sum_{i=1}^{n}\sum_{a\in\mathcal{A},a^s\in\mathcal{A}^s} Y_i(a,a^s)\,\overset{*}{\Pr}(A_i = a, A_i^s = a^s | W_i, W_i^s) \mid \mathbf{W}\right]$$

where $\mathbf{W} = W_1, W_2, ..., W_i, ..., W_n$

Intuition for Parameters

Imagine the following

- There are a large number of replications of the network $\mathcal{G}$.
- $\mathbf{W}$ is held fixed across replications

ψ is the expected mean of Y for the n units under the policy ω

- The tricky part is we only get to see a single $\mathcal{G}$
- Akin to having a single observation

Assumptions

Identification

Potential outcomes, $Y_i(a, a^s)$ are not observed

- Identify quantity given observable data?
- Can be given by design (randomization)
  - No progress since only observe *one* network
  - Still need something extra
- Make untestable assumptions

Identification Assumptions

Causal Consistency

$$Y_i = Y_i(a, a^s) \text{ if } a = A_i, a^s A_i^s$$

Exchangeability (no unobserved confounding)

$$E[Y(a, a^s)|W, W^s] = E[Y(a, a^s)|A, A^s, W, W^s]$$

Positivity[6]

$$\Pr(A, A^s|W, W^s) > 0 \text{ for all } \overset{*}{\Pr}(A, A^s|W, W^s) > 0$$

[6] There are two variations on the positivity assumption. Deterministic positivity is needed for identification (see Westreich & Cole (2010) for details)

Targeted maximum likelihood estimation (TMLE)

TMLE in general

Take two estimators

- g-formula: models Y as function of A and W[7]
- IPW: models A as a function of W

Combines them together in a smart way

- Combines them in a targeting model[8]
- Essentially, take predicted values of Y from the g-formula and shift them by η
  - Where η is estimated via the targeting model and IPW
- Has a number of advantages over the constituent methods
  - Double-robustness, semiparametric efficiency, variance estimator, machine learning[9]

[7] For an introduction to g-computation, see Snowden et al. (2011)
[8] For an intro to TMLE with IID data, see Schuler & Rose (2017)
[9] For advantages on the machine learning side, see Zivich & Breskin (2021)

TMLE in networks

Overview

- Estimate $E[Y|A, A^s, W, W^s]$
- Estimate the inverse probability weight
- Targeting step
- Point estimation of parameter of interest
- Variance estimation

Preliminary Data Prep

Determine summary measures for A^s, W^s

- Not trivial, need background information
- If incorrect, potential for bias

Calculate summary measures and setup data

- Bound Y to be $(0, 1)$

Y_i	A_i	A_i^s	W_i	W_i^s
0.99	0	3	1	2
0.50	1	2	1	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0.01	1	0	0	2

Estimate $E[Y|A, A^s, W, W^s]$

Model the outcome

$$E[Y_i|A_i, A_i^s, W_i, W_i^s; \beta] = \beta_0 + \beta_1 A_i + \beta_2 A_i^s + \beta_3 W_i + \beta_4 W_i^s$$

Predicted value of Y_i under A_i, A_i^s: $\hat{Y}_i$

Estimate the Weights

Need to construct the following inverse probability weights

$$\frac{\Pr^*(A_i, A_i^s | W_i, W_i^s)}{\Pr(A_i, A_i^s | W_i, W_i^s)}$$

First, the denominator

- Factor into $\Pr(A_i | W_i, W_i^s) \Pr(A_i^s | A_i, W_i, W_i^s)$
- Estimate $\Pr(A_i | W_i, W_i^s; \alpha)$ using a logit model
- Estimate $\Pr(A_i^s | A_i, W_i, W_i^s)$ using an appropriate model
- Multiply predicted probabilities from models

Estimate the Weights

Now for the numerator

- Problem: $\Pr^*(A_i, A_i^s | W_i, W_i^s)$ is hard to specify
  - Can easily make 'impossible' policies by accident
- Instead will specify policy as $\Pr^*(A_i | W_i, W_i^s)$

Monte Carlo Procedure

- Create k copies of the data
- Assign A_{ik}^* using $\Pr^*(A_i | W_i, W_i^s)$ in each copy
- Calculate A_{ik}^{s*} using $\mathcal{G}$
- Estimate models for factored probabilities as before, but *using all* k copies
- Predict probabilities using models and A_i, A_i^s

Targeting Step

Estimate the following weighted, intercept-only logit model

$$\text{logit}(Y) = \eta + \text{logit}(\hat{Y})$$

where the weights are

$$\frac{\pi_i^*}{\pi_i} = \frac{\Pr^*(A_i, A_i^s | W_i, W_i^s)}{\Pr(A_i, A_i^s | W_i, W_i^s)}$$

Broadly, can think about η as a correction factor

- 'Corrects' the outcome model predictions for the Y via IPW
- Apply this correction in the estimation step

Point estimation

Process

- Predict outcomes using g-computation under the policy: $\hat{Y}_i^*$
- Update the predictions: $\tilde{Y}_i^* = \text{expit}(\text{logit}(\hat{Y}_i^*) + \hat{\eta})$
- Mean: $\hat{\psi} = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i^*$

Problem

- Stochastic policy has multiple values for A_i^*, A_i^{s*}
- Use Monte Carlo integration
  - Take the previous k copies
  - Predict outcomes under A_{ik}^*, A_{ik}^{s*} for each copy
  - Calculate $\hat{\psi}_k$
  - Mean: $\hat{\psi} = k^{-1} \sum_k \hat{\psi}_k$

Variance estimation

Influence-curve-based variance estimator

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \left(\frac{\pi_i^*}{\pi_i} (Y_i - \hat{Y}_i) \right)^2$$

Restrictive assumption

- Dependence between observations is solely due to direct-transmission
- Unlikely to be the case, since likely latent (unobserved) variables related to the outcome

Variance estimation

Alternative influence-curve-based variance estimator

$$\hat{\sigma}^2 = n^{-1} \sum_{i,j} \mathbb{G}_{ij} \left(\frac{\pi_i^*}{\pi_i} (Y_i - \hat{Y}_i) \times \frac{\pi_j^*}{\pi_j} (Y_j - \hat{Y}_j) \right)$$

- where $\mathbb{G}$ is $\mathcal{G}$ with the leading diagonal set to 1

Less restrictive assumption

- Valid for direct transmission
- Also allows for latent transmission up to 2 edges away

Illustrative example

Motivating Problem

What would have been the expected (mean) outcome among n individuals under the stochastic policy ω?[10]

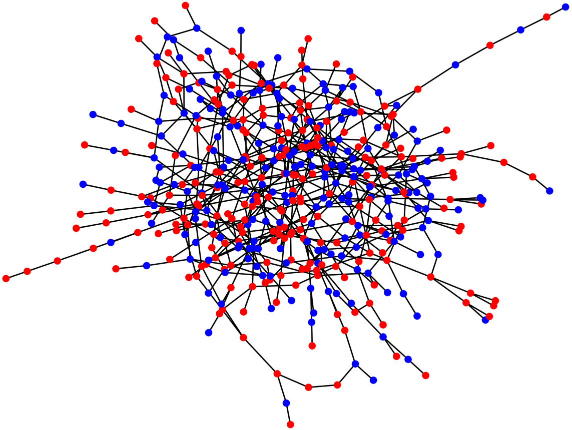
- Example
  - Incentive (action) and subsequent behavior adoption (outcome)
- A is binary action, Y is binary outcome
- Identification assumptions all assumed to be met
- $\Pr^*(A_i) = \omega$ where $\omega \in \{0.1, 0.2, 0.3, ..., 0.9\}$

Summary measures

$$A_i^s = \sum_{j=1}^{n} A_j \mathcal{G}_{ij} \quad W_i^s = \sum_{j=1}^{n} W_j \mathcal{G}_{ij}$$

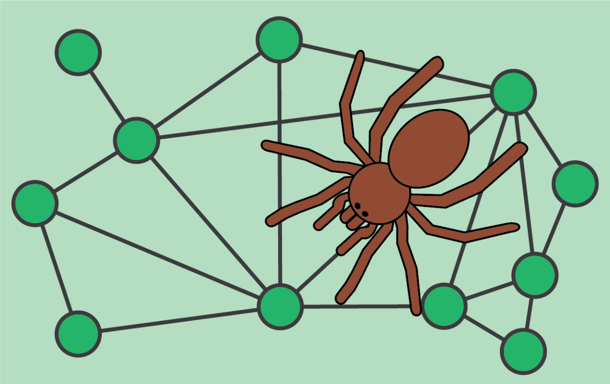
[10] Note: all data here is being simulated! This example is only meant as an illustration

Network



MossSpider

Available for Python 3.6+[11]



```
python -m pip install mossspider
```

[11] Maintained at https://github.com/pzivich/MossSpider

Analysis with MossSpider

```python
from mossspider import NetworkTMLE
```

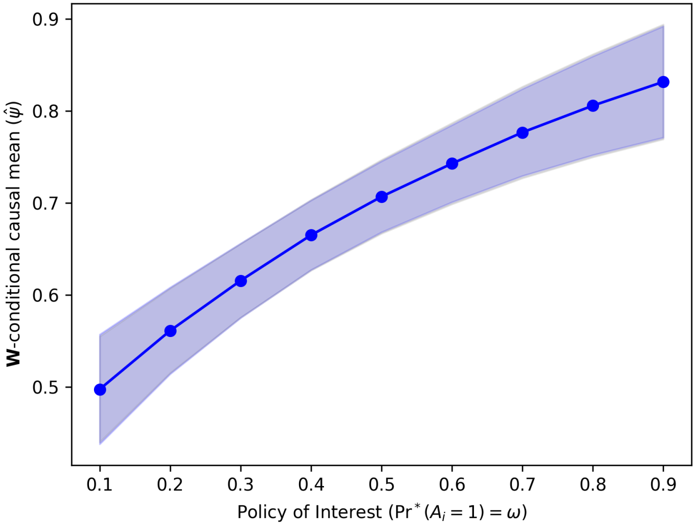
```python
# Initialize NetworkTMLE
ntmle = NetworkTMLE(network=H,
                    exposure="A",
                    outcome="Y")
# Model for Pr(A | W, W^s; \delta)
ntmle.exposure_model(model="W + W_sum")
# Model for Pr(A^s | A, W, W^s; \gamma)
ntmle.exposure_map_model(model="A + W + W_sum",
                         measure="sum",
                         distribution="poisson")
# Model for E[Y | A, A^s, W, W^s; \alpha]
ntmle.outcome_model(model="A + A_sum + W + W_sum")
```

Analysis with MossSpider

```python
# Policies to evaluate
policy = [0.1, 0.2, 0.3, 0.4, 0.5,
          0.6, 0.7, 0.8, 0.9]

# Evaluating each policy
for p in policy:
    ntmle.fit(p=p,                  # Policy
              samples=200,          # replicates
              seed=20220316)        # random seed
```

Results



0.9
0.8
0.7
0.6
0.5
W-conditional causal mean ($\hat{\psi}$)
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Policy of Interest ($\Pr^*(A_i = 1) = \omega$)

Summary

Summary

Causal inference with network data is difficult

- Unverifiable assumptions
- Simplifications of interference processes

Network-TMLE

- One modern approach
- Overview of how it works
- Implementation available in `mossspider`

References

Network-TMLE readings

- Ogburn EL, Sofrygin O, Diaz I, & Van Der Laan MJ. (2017). "Causal inference for social network data". *arXiv preprint* arXiv:1705.08527.
- Sofrygin O, & van der Laan MJ. (2017). "Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population". *Journal of Causal Inference*, 5(1).
- van der Laan MJ. (2014). "Causal inference for a population of causally connected units". *Journal of Causal Inference*, 2(1), 13-74.

Other resources

- Halloran ME, & Hudgens MG. (2016). "Dependent happenings: a recent methodological review". *Current Epidemiology Reports*, 3(4), 297-305.
- Hudgens MG, & Halloran ME. (2008). "Toward causal inference with interference". *Journal of the American Statistical Association*, 103(482), 832-842.
- Schuler MS, & Rose S. (2017). "Targeted maximum likelihood estimation for causal inference in observational studies". *American Journal of Epidemiology*, 185(1), 65-73.
- Snowden JM, Rose S, & Mortimer KM. (2011). "Implementation of G-computation on a simulated data set: demonstration of a causal inference technique". *American Journal of Epidemiology*, 173(7), 731-738.
- Westreich D, & Cole SR. (2010). "Invited commentary: positivity in practice". *American Journal of Epidemiology*, 171(6), 674-677.
- Zivich PN, & Breskin A. (2021). "Machine learning for causal inference: on the use of cross-fit estimators". *Epidemiology*, 32(3), 393-401.
- Zivich PN, Volfovsky A, Moody J, & Aiello AE. (2021). "Assortativity and Bias in Epidemiologic Studies of Contagious Outcomes: A Simulated Example in the Context of Vaccination". *American Journal of Epidemiology*, 190(11), 2442-2452.