Model-assisted design of experiments in the presence of network correlated outcomes (G.W. Basse & E.M. Airoldi, 2018+)

Nima Hejazi

2018-10-16

Introduction

Interference: When people have friends

- Observational units are connected – so far, we've been dealing with causal analyses *in a vacuum*.
- Sometimes, it's reasonable to assume that units do not affect one another; often, it's not.
- A central assumption in causal models, necessary for identification results, is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1978) & (Rubin 1980).

Networks: Are you on instafacetweet too?

- In a population of causally connected units, several types of network structures may arise, each bringing its posing unique challenges for statistics.
- Broadly, the central statistical challenge is *"how to account for the presence of connections, or network data, observed pre-intervention, possibly with uncertainty, and often missing."*

Networks: Two perspectives

- Two main problem settings have been discussed in the causal inference literature
  1. *Network interference*: When the potential outcomes of a given unit are a function of its assigned treatment and that of others.
  2. *Network-correlated outcomes*: When the potential outcomes of units in a network are related through their baseline covariates.
- The first problem has been the subject of much attention in the literature, so (**???**) focus on the second setting.

Network interference: ...

- Mostly studied in the setting of randomized experiments
- something
- cite Eckles
- cite Mark
- cite Ogburn and Vanderweele

Network-correlated outcomes: ...

- Mostly studied in the setting of observational studies
- ...
- ...

Basse and Airoldi, 2018+, *Biometrika*

Goal and Motivation

- *The problem*: "how to assign treatment in a randomized experiment, when the correlation among the outcomes is informed by a network available at the design stage."
- Identify and estimate the causal effect of interference in the presence of confounding induced by correlated outcomes

Approach

- Use *model-assisted restricted randomization strategies*, leveraging a static network known pre-intervention.
- Restricted randomization has a long history in experimental design – (**???**) build off of this, using strategies that balance covariates properly.

Approach

- Posit a working model for the potential outcomes, conditional on the network known pre-intervention.
- Restrict the set of allowed randomization strategies such that the estimator of interest achieves low MSE.

Findings

- In turn, the focus on the MSE suggests new notions of balance in network-based randomization (related to network degree statistics).
- Proposed approach maintains design unbiasedness of the difference-in-means estimator, even when the working model is misspecified (cf, double robust?)
- When the working model is correct, inference is improved through higher precision (lower variance) of the estimator of interest.

Notation

- N observational units, indexed $i = 1, \ldots, n$.
- Binary treatment Z, where $Z_i = 1$ denotes assignment to treatment arm
- Real-valued outcome Y_i, with potential outcomes $Y_i(1)$ for $Z_i = 1$ and $Y_i(0)$ for $Z_i = 0$.

Assumptions

- Assume *SUTVA* (Rubin 1978): $Y_i(Z) = Y_i(Z_i)$, explicitly disallowing network interference.

ADD CITATION FOR RUBIN 1974

- Finite population setting: recall that potential outcomes $Y(Z)$ are unknown but constant quantities, given Z (not RVs).
- Randomized experiment: only source of variation is the allocation of treatment to units (controlled by experimenter).
- Treatment allocated based on distribution on the space of all binary vectors of length N, the randomization distribution. CITE IMBENS+RUBIN

Parameter of interest: ATE

- For illustration, authors focus on the ATE as the inferential target
- With the notation previously given, the ATE is defined as

$$\tau^* = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}$$

- Focus also on the difference-in-means estimator for the ATE:

$$\hat{\tau}(Y \mid Z) = \frac{\sum_{i=1}^{N} Z_i Y_i}{\sum_{i=1}^{N} Z_i} - \frac{\sum_{i=1}^{N}(1 - Z_i) Y_i}{\sum_{i=1}^{N}(1 - Z_i)}$$

An undirected network

- The approach requires that a network be known at the design stage.
- Let the network be an undirected graph $\mathcal{G}$ over N units, where
- $\mathcal{G}$ is simply an $N \times N$ binary adjacency matrix A, where all diagonal entries are unary (i.e., $A_{ii} = 1$).
- Let neighborhood of unit i be the index set $\mathcal{N}_i = \{jstA_{ij} = 1\}$

The Normal Sum Model

$$X_j \sim_{iid} N(\mu, \sigma^2)$$
$$Y_i(0) \mid X \sim_{ind} N(\sum_{j \in \mathcal{N}_i} X_j, \gamma^2)$$
$$Y_i(1) = Y_i(0) + \tau$$

- Observations in the same group are taken to have originated from a Normal distribution with the same mean. (Don't group with same mean then look the same? Group overlap?)
- Constant treatment effect model: τ is the difference between the potential outcomes $\{Y_i(0), Y_i(1)\}$.

Methodology

- ...
- ...
- ...

Findings

- ...
- ...
- ...

I've talked enough

Discussion

- ...
- ...
- ...

References

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics*. JSTOR, 34–58.

———. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75 (371). JSTOR: 591–93.