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

Nima Hejazi

2018-10-21

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 (Rubin 1978) & (Rubin 1980).
- *Interference* is often defined through the loosening of this assumption (Hudgens and Halloran 2008).

Networks: Are you (still) on facebook too?

- In a population of causally connected units, several types of network structures may arise, each 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"* Basse and Airoldi (2018).

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 Basse and Airoldi (2018) focus on resolving issues in the second setting.

Network-correlated outcomes: We're not *that* different

- Most often studied in the context of observational studies.
- Basse and Airoldi (2018) focus on
- ...

G.W. Basse and E.M. Airoldi, 2018+, *Biometrika*

Goals 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.
- How can information about a network be used to inform randomization strategies for estimating causal effects?

Approach

- Use *model-assisted restricted randomization strategies*, leveraging a static network known pre-intervention.
- Restricted randomization has a long history in experimental design – Basse and Airoldi (2018) 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.
- In turn, focus on MSE suggests new notions of balance in network-based randomization (related to network degree statistics).

Findings

- Proposed approach maintains design unbiasedness of the difference-in-means estimator, even when the working model is misspecified (i.e., robustness).
- When the working model is correct, inference is improved through higher precision 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 $Y_i(Z_i)$:
  - $Y_i(1)$ for $Z_i = 1$ and
  - $Y_i(0)$ for $Z_i = 0$.

Assumptions

- *Stable Unit Treatment Value Assumption* (Rubin 1974) & (Rubin 1978).
    - i.e., $Y_i(Z) = Y_i(Z_i)$
    - explicitly disallows network interference
- Finite population setting: recall that potential outcomes $Y(Z)$ are unknown but constant quantities, given Z.
- *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, i.e., randomization distribution (Imbens and Rubin 2015).

Parameter of interest: ATE

- For illustration, focus on 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 proposed methodology requires that a network be known at the design stage (pre-specified).
- 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., $A_{ii} = 1$), and
  - the neighborhood of unit i be the index set $\mathcal{N}_i = \{j : A_{ij} = 1\}$.

A simplified model

- For illustrative purposes, assume 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.
- "The network induces correlation among the outcomes that are assigned to control because the mean of each $Y_i(0)$ is given by the sum of the covariate values X_j of units j in a neighborhood of i".

A simplified model

- Constant treatment effect model: τ is the difference between the potential outcomes $\{Y_i(0), Y_i(1)\}$.
- *Intuition*: in the absence of network connections and treatment $Z_i = 0$:
  - $Y_i(0)$ is a measure of an intrinsic property of the observational unit (e.g., time spent on social media), as determined by covariates X.
  - Network connections alter the natural value $Y_i(0)$ that would occur, through the induced network structure.
  - The intervention $\text{do}(Z_i = 1)$ induces a causal effect τ such that $Y_i(1) = Y_i(0) + \tau$.
- The *normal-sum* model is just a starting point...

Optimal treatment allocation

- To ascertain an optimal treatment allocation strategy, need a notion of error to define optimality.
- Basse and Airoldi (2018) propose the *conditional MSE*:
    1. fix a treatment allocation vector Z, then
    2. for the *normal-sum model*, $\mathrm{MSE}(\hat{\tau} \mid Z) \equiv \mathbb{E}\{(\hat{\tau} - \tau^*)^2 \mid Z\}$
- Now, an optimal treatment allocation $Z^* \in \mathcal{Z}$ is one that minimizes the conditional MSE.

Where are the networks?

- A decomposition of the conditional MSE is informative of network statistics:

$$\mathrm{MSE}(\hat{\tau} \mid Z) = \mu^2\{\delta_N(Z)\}^2 + \gamma^2\omega(Z)^T\omega(Z) + \sigma^2\omega(Z)^T A^T A\omega(Z)$$

- Each of the terms in the MSE decomposition is informative
  - Bias^2: $\mu^2\{\delta_N(Z)\}^2$
  - *Network-agnostic* variance component: $\gamma^2\omega(Z)^T\omega(Z)$
  - *Network-aware* variance component: $\sigma^2\omega(Z)^T A^T A\omega(Z)$
- Model-assisted restriction randomization strategies seek to minimize the conditional MSE, but tradeoffs occur in these components.

Bias

- The bias term admits the decomposition

$$\mu \cdot \delta_{\mathcal{N}} = \mu \cdot \left(\frac{1}{N_1} \sum_{(i:Z_i=1)} |\mathcal{N}_i| - \frac{1}{N_0} \sum_{(i:Z_i=0)} |\mathcal{N}_i| \right)$$

- The bias is proportional to the *average degree* of each of the experimental arms (treatment and control groups).
- This is the difference in the average neighborhood sizes of the treated and untreated units – i.e., balance!
- Desirable treatment allocation vectors $\mathcal{Z}^b$ will minimize this difference in neighborhood sizes.

Network-agnostic variance term

- The first part of the variance term may be decomposed

$$\gamma^2 \omega^T \omega = \gamma^2 \left(\frac{1}{N_1} + \frac{1}{N_0} \right)$$

- Similar to the previous term, this term is minimized when $N_1 = N_0$.
- Thus, this term penalizes a difference in the size of treatment and control units, and is satisfied through *balance*.
- This is similar to prior work in balanced randomizations outside of the context of network-correlated outcomes.

Network-aware variance term

- The second part of the variance term may be written

$$
\begin{aligned}
\sigma^2 \cdot \omega^T A^T A \omega = & \frac{\sigma^2}{N_1^2} \cdot \sum_{i,j: Z_i = Z_j = 1} |\mathcal{N}_i \cap \mathcal{N}_j| \\
& + \frac{\sigma^2}{N_0^2} \cdot \sum_{i,j: Z_i = Z_j = 0} |\mathcal{N}_i \cap \mathcal{N}_j| \\
& - \frac{2\sigma^2}{N_1 \cdot N_0} \cdot \sum_{i,j: Z_i = 1 \text{and} Z_j = 0} |\mathcal{N}_i \cap \mathcal{N}_j|
\end{aligned}
$$

- Minimize contribution of this term to the MSE by
    1. assigning units with shared neighbors to different groups, and
    2. avoiding assigning treatment or control to clusters of densely connected units.

Classical randomization

- ...
- ...

Restricted randomization

- ...
- ...

Model-assisted restricted randomization

- ...
- ...

Model-assisted restricted randomization

- ...
- ...

Model-based optimal treatment allocation

- ...
- ...

Restricted randomization and rerandomization

- ...
- ...

Fisher intervals and inference

- ...
- ...

Key properties of the approach

- ...
- ...

Towards generalized network models

- The *normal-sum model* we discussed is just a simple case of a much broader family of models

$$Y_i(0) \mid X \sim^{ind} N(g[\{X_j\}_{j \in \mathcal{N}_i}], \gamma^2)$$

- Need regularity conditions on g to ensure that $\mathbb{E}(g[\{X_j\}_{j \in \mathcal{N}_i}] \mid \{X_j\}_{j \in \mathcal{S}})$ is well-behaved for any subset of nodes $\mathcal{S} \subset \mathcal{N}_i$.

Lessons for good designs

- Decrease the number of neighbors shared within treatment groups.
- Increase the number of units shared between treatment groups.
- Balance the size of the groups and the distribution of neighborhood sizes.

I’ve talked enough

Discussion

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

References

Basse, Guillaume W, and Edoardo M Airoldi. 2018. "Model-Assisted Design of Experiments in the Presence of Network Correlated Outcomes." *arXiv Preprint arXiv:1507.00803*.

Hudgens, Michael G, and M Elizabeth Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103 (482). Taylor & Francis: 832–42.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5). American Psychological Association: 688.

———. 1978. "Bayesian Inference for Causal Effects: The Role of