Combining Rollout Designs and Clustering for Causal Inference under Low-order Interference

Changhao Shi

Renmin University of China, Statistics

May 15, 2024

- Cortez-Rodriguez, M., Eichhorn, M. and Yu, C.L., 2024. Combining Rollout Designs and Clustering for Causal Inference under Low-order Interference. *arXiv preprint arXiv:2405.05119*.

Review: classic causal inference

- Binary treatment $Z_i \in \{0, 1\}$
- Potential outcomes $Y_i(1)$ and $Y_i(0)$
- Causal effects: comparisons of potential outcomes
- Common choice: average causal effect (ACE)

$$
\begin{aligned}
\mathrm{ACE} &\overset{\text{def}}{=} E\{Y(1) - Y(0)\} \\
&= E\{Y(1)\} - E\{Y(0)\} \\
&\overset{(1)}{=} E\{Y(1) \mid Z = 1\} - E\{Y(0) \mid Z = 0\} \\
&\overset{(2)}{=} E\{Y^{\text{obs}} \mid Z = 1\} - E\{Y^{\text{obs}} \mid Z = 0\}
\end{aligned}
$$

  - (1) holds when $\{Y(1), Y(0)\} \perp\!\!\!\perp Z$
  - (2) holds when $Y^{\text{obs}} = Y(1)Z + Y(0)(1 - Z)$
- For many policy makers, ACE is the quantity of interest

Review: causal inference under interference

- Violation of SUTVA
- Common in advertising, epidemiology and educational studies
- Potential outcomes $Y_i(z)$, where $z \in \{0, 1\}^n$
- Causal effects of interest
  - total treatment effect (TTE)
    $$\mathrm{TTE} = \frac{1}{n} \sum_{i=1}^{n} \{Y(\mathbf{1}) - Y(\mathbf{0})\}$$
  - average direct effect (ADE)
    $$\mathrm{ADE} = \frac{1}{n} \sum_{i=1}^{n} E\big\{ Y_i(z_i = 1, Z_{-i}) - Y_i(z_i = 0, Z_{-i})\big\}$$
  - average indirect effect (AIE)
    $$\mathrm{AIE} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} E\{ Y_j(z_i = 1, Z_{-i}) - Y_j(z_i = 0, Z_{-i})\}$$
- Standard methods for ACE cannot be applied naively

Review: general framework for interference

- A social network
  - through which individuals interfere each other
  - observable and correctly measured
- An exposure mapping
  - determines the extent and intensity of the interference
  - technically reduces the number of potential outcomes
  - canonical examples (minor notation abuse)
    - (no interference) $Y_i(z) = Y_i(z_i)$
    - (neighborhood interference) $Y_i(z) = Y_i(z_{\mathcal{N}_i})$
    - (arbitrary interference) $Y_i(z) = Y_i(z)$
    - ("individualized" interference) $Y_i(z) = Y_i(?)$
- Estimators: ht, hajek, difference-in-means, etc
- Experimental designs $Z \sim P(z)$: complete randomization, Bernoulli randomization, cluster randomization, etc

Notation and framework for unobservable networks

- Structure of social network may be unavailable or costly to collect
- An unknown directed graph with edge set $E \subset [n] \times [n]$
- An edge $(j, i) \in E$ means i is affected by j's treatment
- In-neighborhood of i: $\mathcal{N}_i = \{j \in [n] : (j, i) \in E\}$
- Potential outcomes function: $Y_i : \{0, 1\}^n \to \mathbb{R}$
- Under assumption of consistency, one may see

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S} \subseteq [n]} a_{i,\mathcal{S}} \prod_{j \in \mathcal{S}} z_j \prod_{j' \in [n] \setminus \mathcal{S}} (1 - z_{j'}) = \sum_{\mathcal{S} \subseteq [n]} c_{i,\mathcal{S}} \prod_{j \in \mathcal{S}} z_j \qquad (1)$$

- Equation (1) means $Y_i(z)$ is a polynomial in z of degree at most n
- Estimand of interest: $\mathrm{TTE} := \frac{1}{n} \sum_{i=1}^{n} (Y_i(\mathbf{1}) - Y_i(\mathbf{0}))$

Standard assumptions

- (Neighborhood Interference) If $\mathbf{z}, \mathbf{z}'$ have $z_j = z_j' \; \forall \; j \in \mathcal{N}_i$, then $Y_i(\mathbf{z}) = Y_i(\mathbf{z}') \; \forall i \in [n]$
- (Bounded Potential Outcomes)

$$Y_{\max} := \max_{i\in[n]} \sum_{\mathcal{S}\subseteq\mathcal{N}_i, |\mathcal{S}|\le\beta} |c_{i,\mathcal{S}}|$$

- (β-Order Interactions) $c_{i,\mathcal{S}} = 0$ for all $|\mathcal{S}| > \beta$
  - using the notation $\mathcal{S}_i^\beta := \{\mathcal{S} \subseteq \mathcal{N}_i \; : \; |\mathcal{S}| \le \beta\}$, the TTE is

$$\mathrm{TTE} = \frac{1}{n}\sum_{i=1}^n \big(Y_i(\mathbf{1}) - Y_i(\mathbf{0})\big) = \frac{1}{n}\sum_{i=1}^n \sum_{\mathcal{S}\in\mathcal{S}_i^\beta\setminus\varnothing} c_{i,\mathcal{S}}$$

- ("Time-Invariant" Potential Outcomes)

$$Y_{i,t}^{\mathbf{obs}} = Y_i(\mathbf{z}^t) + \varepsilon_{i,t}, \quad \varepsilon_{i,t} \overset{iid}{\sim} N(0, \sigma^2)$$

Literature review

- Yu et al. (2022) first proposed a class of linear interpolation estimators (with prior baseline information) for Heterogeneous Additive Network Effects Models

$$Y_i(\mathbf{z}) = \alpha_i + \beta_i z_i + \sum_{k \in [n]} \gamma_{ki} z_k$$

- Cortez et al. (2022) extended this approach to a class of polynomial interpolation estimators (with staggered rollout designs) for Low-order Interference Models

$$Y_i(\mathbf{z}) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i, |\mathcal{S}| \leq \beta} c_{i,\mathcal{S}} \cdot \mathrm{I}\big(\mathcal{S} \text{ treated}\big) = \sum_{\mathcal{S} \subseteq \mathcal{N}_i, |\mathcal{S}| \leq \beta} c_{i,\mathcal{S}} \prod_{j \in \mathcal{S}} z_j$$

- Cortez et al. (2024) combined their method with network clustering techniques to further reduce variance, albeit at the expense of slightly increased bias (bias-variance trade-off)

Staggered Rollout Design (Cortez et al., 2022)

- Treatment is incrementally given to random subsets of individuals
  - treatment is assigned to individuals in T stages
  - individuals' outcomes are measured $T+1$ times
    - a baseline measurement before treatment
    - a measurement after each treatment round
- Treatment assignment in round t: $\mathbf{z}^t$
  - each entry z_i^t is monotone increasing with t
- Bern(β, p) Rollout Design:
  - model degree β
  - treatment budget p
  - $\mathbf{z}^t$ for $t \in \{0, \ldots, \beta\}$
  - $u_i \overset{iid}{\sim} U(0,1)$, for each $i \in [n]$
  - $z_i^t = 1(u_i \leq \frac{t}{\beta} p)$, for each $t \in \{0, \ldots, \beta\}$

Graph Agnostic Estimator (Cortez et al., 2022)

- By the β-order interactions assumption, one may see

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z}^t)\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}}\cdot\mathbb{E}\left[\prod_{j\in\mathcal{S}} z_j^t\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^\beta} c_{i,\mathcal{S}}\cdot\left(\frac{tp}{\beta}\right)^{|\mathcal{S}|} =: F\left(\frac{tp}{\beta}\right)$$

- In each round t, a noisy measurement $\widehat{F}(\frac{tp}{\beta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{z}^t)$ of $F(\frac{tp}{\beta})$ is observed
- Target is $\mathrm{TTE} = F(1) - F(0)$
- Viewing the estimation problem as a polynomial interpolation problem gives rise to the following unbiased estimator for the TTE

$$\widehat{\mathrm{TTE}}_{\mathrm{PI}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{\beta}\Big(\ell_{t,\mathbf{p}}(1) - \ell_{t,\mathbf{p}}(0)\Big) Y_i(\mathbf{z}^t), \qquad \ell_{t,\mathbf{p}}(x) = \prod_{\substack{s=0\\ s\neq t}}^{\beta} \frac{\beta x - ps}{pt - ps}.$$

- $\mathrm{Var}(\widehat{\mathrm{TTE}}) = O\big(\frac{d^2\beta^{2\beta}}{np^{2\beta}}\big)$, the multiplier $(\beta/p)^{2\beta}$ is not satisfied

Preview: PL v.s. 2-Stage

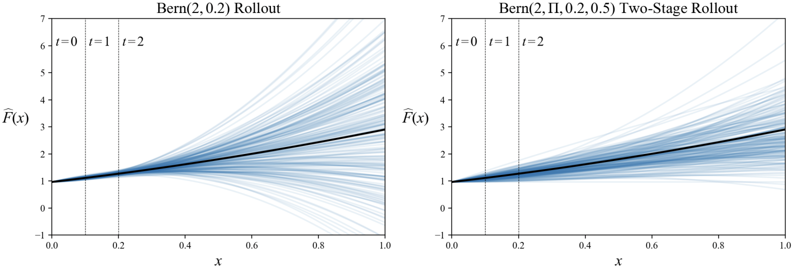

Figure: Visualization of extrapolated polynomials used to estimate TTE across 200 runs of a rollout experiment on an $\mathrm{SBM}(200, 10, 0.25, 0.05)$ instance with $\beta = 2$. The left plot uses a Bernoulli rollout, as in (Cortez et al., 2022), while the right plot uses a clustered 2-stage rollout. While the sampling error at the observation points $x = 0.1, 0.2$ is less in the Bernoulli experiment, the extrapolation error leads to a higher overall variance.

Staggered rollout designs with clustering

Definition ($\mathrm{Bern}(\beta, \Pi, p, q)$ Two-Stage Rollout Design)

Given a model degree β, clustering Π, treatment budget p, and effective treatment threshold q, the treatment assignments $\mathbf{z}^t$ for $t \in \{0, \ldots, \beta\}$ in a $\mathrm{Bern}(\beta, \Pi, p, q)$ two-stage rollout design are computed as follows:

1. Determine the set of experimental units $\mathcal{U} = \{i \in [n] \colon W_{\pi(i)} = 1\}$, where $W_\pi \overset{\text{iid}}{\sim} Bern(\frac{p}{q})$.
2. Use a $\mathrm{Bern}(\beta, q)$ rollout to assign treatment to individuals in $\mathcal{U}$. For all $i \notin \mathcal{U}$, set $z_i^t = 0$.

- Two-Stage Estimator

$$\widehat{\mathrm{TTE}}_{\text{2-Stage}} := \frac{q}{np} \sum_{i=1}^{n} \sum_{t=0}^{\beta} \Big(\ell_{t,q}(1) - \ell_{t,q}(0)\Big) \cdot Y_i(\mathbf{z}^t), \quad \ell_{t,q}(x) = \prod_{\substack{s=0 \\ s \neq t}}^{\beta} \frac{\beta x - qs}{qt - qs}$$

Staggered rollout designs with clustering

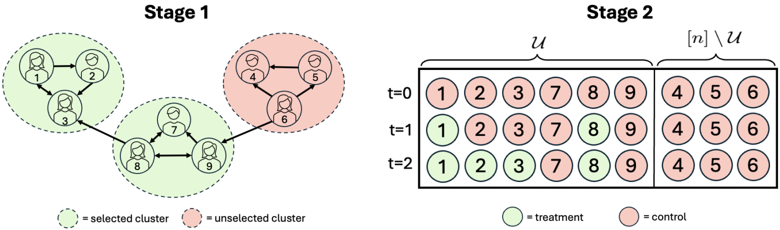

Figure: An illustration of a two-stage rollout design. In the first stage, we select a subset of the clusters as the experimental group $\mathcal{U}$. In the second stage, we carry out a rollout design on the units in $\mathcal{U}$ and deterministically assign the remaining units to control.

Intuition of estimators

"Under this two-stage design, any unit not in $\mathcal{U}$ should have relatively few treated neighbors and thus, their treatment effect estimate will be close to 0. Since $\mathbb{E}\left[|\mathcal{U}|\right] = \frac{np}{q}$, we scale the sum of the treatment effect estimates by $\frac{q}{np}$ in our final estimate of the TTE."

- $\mathcal{U} = \{i \in [n] \colon W_{\pi(i)} = 1\}$

$$\widehat{\mathrm{TTE}}_{\text{2-Stage}} := \frac{q}{np} \sum_{i=1}^{n} \sum_{t=0}^{\beta} \Big(\ell_{t,q}(1) - \ell_{t,q}(0)\Big) \cdot Y_i(\mathbf{z}^t), \quad \ell_{t,q}(x) = \prod_{\substack{s=0 \\ s \neq t}}^{\beta} \frac{\beta x - qs}{qt - qs}$$

Bias-Variance trade-off

$$\mathrm{MSE}(\widehat{\mathrm{TTE}}) = \underbrace{\{\mathop{\mathbb{E}}_{\mathcal{U},\mathbf{z}}[\widehat{\mathrm{TTE}} - \mathrm{TTE}]\}^2}_{} + \underbrace{\mathop{\mathrm{Var}}_{\mathcal{U}}\left(\mathop{\mathbb{E}}_{\mathbf{z}}[\widehat{\mathrm{TTE}} \mid \mathcal{U}]\right)}_{\text{sampling variance: orange}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{U}}\left[\mathop{\mathrm{Var}}_{\mathbf{z}}(\widehat{\mathrm{TTE}} \mid \mathcal{U})\right]}_{\text{extrapolation variance: green}}$$

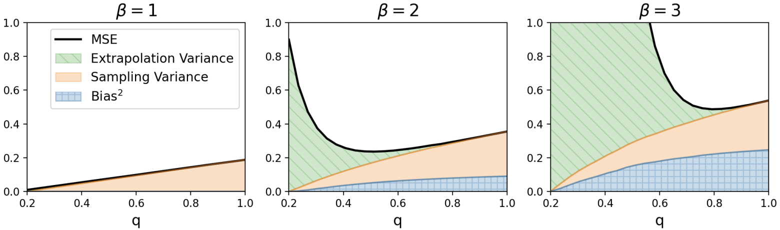

Figure: The MSE (black) of the two-stage estimator for different values of q on SBM instances under β-order potential outcomes models with $\beta \in \{1, 2, 3\}$. The shading indicates three components: squared bias (blue), variance from sampling (orange), and variance from extrapolation (green).

Theoretical results

- Bias

$$\mathbb{E}\left[\widehat{TTE}_{PI}\right] - TTE = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathcal{S}\in\mathcal{S}_i^{\beta}\setminus\varnothing} c_{i,\mathcal{S}}\left[\left(\frac{p}{q}\right)^{|\Pi(\mathcal{S})|-1} - 1\right]$$

- Extrapolation Variance

$$\mathop{\mathbb{E}}_{\mathcal{U}}\left[\mathop{Var}_{\mathbf{z}}\left(\widehat{TTE}\right)\right] \leq \frac{1}{q^{2(\beta-1)}} \cdot \frac{Y_{\max}^2 d^2 \beta^{2\beta}(\beta+1)^2}{np^2}$$

- Sampling Variance

$$\mathop{Var}_{\mathcal{U}}\left(\mathop{\mathbb{E}}_{\mathbf{z}}\left[\widehat{TTE}\right]\right) \leq q \cdot \max_{\pi\in\Pi}|\pi| \cdot \frac{Y_{\max}^2 d^2}{np}$$

- (Cortes et al., 2022) $\mathrm{Var}(\widehat{TTE}) = O\left(\frac{d^2\beta^{2\beta+2}}{np^{2\beta}} Y_{\max}^2 + \frac{\sigma^2\beta^{2\beta+1}}{np^{2\beta}}\right)$.

How to perform clustering?

- Clustering with full graph knowledge
  - using the METIS clustering library
- Clustering with covariate knowledge (features)
  - when each vertex is assigned to one feature, these assignments are used as the clustering
  - When vertices may have multiple features, an undirected weighted feature graph is formed, where the weight of edge (i, j) is the number of feature labels shared by i and j. Then cluster this weighted graph using METIS clustering library
- Clustering without network knowledge
  - randomly partition the vertices into the designated number of evenly-sized clusters

Simulation results

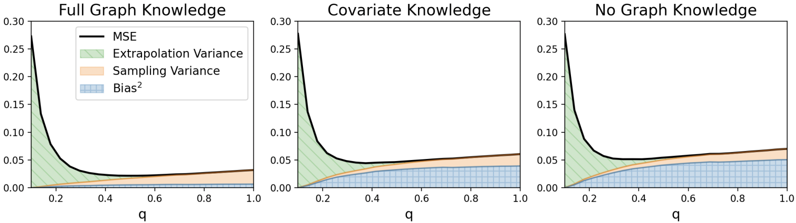

Figure: Mean Squared Error of the Two-Stage TTE estimator for three clustering methods of the AMAZON network, for a β-degree potential outcomes model with $\beta = 2$.

Simulation results

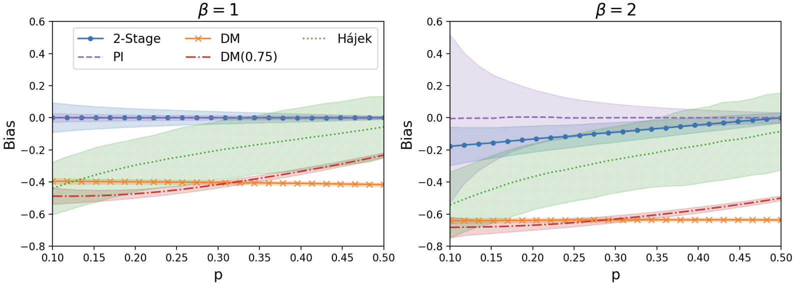

Figure: Performance of different estimators on the AMAZON network for various values of p. The bold line indicates the mean over 10,000 replications. The shading indicates the experimental standard deviation, calculated by taking the square root of the experimental variance over all replications.

Discussion

- Contribution
  - a finite sample bias-variance trade-off
- Future work
  - misspecification of potential outcomes models (hopeless)
  - time-varying network and potential outcomes (hopeless)
  - model selection of β (cool)
  - if K experiments are permitted and the network structure is known, we ask: To what extent can existing methods be improved?

Thank you!