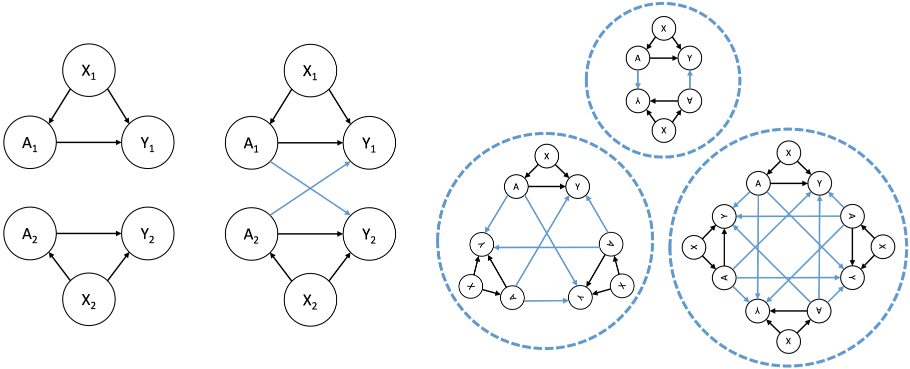
Efficient Nonparametric Estimation of Incremental Propensity Score Effects with Clustered Interference

Chanhwa Lee
Department of Biostatistics, UNC
Mar 22, 2023

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Interference



No interference **Interference** **Clustered Interference**

- COVID-19 vaccine protects family members from SARS-CoV-2 infections (Salo et al. 2022)

Causal questions

- Prevalence of COVID-19 in a city when 50% of citizens are vaccinated compared to when 30% of citizens are vaccinated
- Risk of COVID-19 when an individual is vaccinated versus not vaccinated when 50% of other individuals in the same city are vaccinated
- How to define intervention effects under clustered interference?
- How to estimate effects (what assumptions, what methods)?

Notation

- N_i: number of units in cluster $i \in \{1, \ldots, m\}$
- Unit $j \in \{1, \ldots, N_i\}$ in cluster i,
  $Y_{ij} \in \mathbb{R}$: outcome, $A_{ij} \in \{0, 1\}$: treatment, $\mathbf{X}_{ij} \in \mathbb{R}^p$: covariates
- $\mathbf{O}_i = (\mathbf{Y}_i, \mathbf{A}_i, \mathbf{X}_i, N_i)$: observed data for cluster i
  $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})$, $\mathbf{A}_i = (A_{i1}, \ldots, A_{iN_i})$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iN_i})$
- $\mathcal{A}(N_i)$: set of all length N_i binary vectors

Potential outcome under clustered interference

- $Y_{ij}(\mathbf{a}_i)$: potential outcome for unit j in cluster i when individuals in the cluster receives treatment assignment according to $\mathbf{a}_i \in \mathcal{A}(N_i)$
- $Y_{ij}(\mathbf{a}_i) = Y_{ij}(a_{ij}, \mathbf{a}_{i(-j)})$, $\mathbf{a}_{i(-j)} = (a_{i1}, \ldots, a_{i(j-1)}, a_{i(j+1)}, \ldots, a_{iN_i})$

No interference

- $Y_{ij}(a_{ij}, \mathbf{a}_{i(-j)}) = Y_{ij}(a_{ij}, \mathbf{a}'_{i(-j)})$

Treatment Allocation Policy

- ATE: Potential outcomes when treating everyone vs. no one (?)

$$\mathbb{E}\left\{N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}(1,\ldots,1)\right\} - \mathbb{E}\left\{N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}(0,\ldots,0)\right\}$$

- Stochastic treatment allocation policy (Muñoz and Van Der Laan 2012)
  - Counterfactual scenario: a cluster of size N_i and covariate $\mathbf{X}_i$ receives treatment $\mathbf{a}_i$ with probability $Q(\mathbf{a}_i|\mathbf{X}_i, N_i)$

Example

- Type B policy (Tchetgen Tchetgen and VanderWeele 2012)

$$Q(\mathbf{a}_i|\mathbf{X}_i, N_i) = \prod_{j=1}^{N_i} \alpha^{a_{ij}}(1-\alpha)^{1-a_{ij}}$$

- GLMM shift policy (Barkley et al. 2020, Papadogeorgou et al. 2019)

$$P(\mathbf{A}_i|\mathbf{X}_i, N_i) = \int \prod_{j=1}^{N_i}\{g(\mathbf{X}_{ij}^\top\beta+u)\}^{A_{ij}}\{1-g(\mathbf{X}_{ij}^\top\beta+u)\}^{1-A_{ij}}d\Phi(u)$$

$$Q(\mathbf{a}_i|\mathbf{X}_i, N_i) = \int \prod_{j=1}^{N_i}\{g(\mathbf{X}_{ij}^\top\beta'+u)\}^{a_{ij}}\{1-g(\mathbf{X}_{ij}^\top\beta'+u)\}^{1-a_{ij}}d\Phi(u)$$

Cluster Incremental Propensity Score (CIPS) Policy

- Incremental Propensity Score Interventions (Kennedy 2019)
- Shifting propensity score (PS) distribution such that counterfactual odds of treatment is $\delta \in (0, \infty)$ times observed odds of treatment
- CIPS(δ) policy

Advantages

1. Individual-level treatment assignment probability depends on individuals' covariates
2. PS model need not to be parametric
3. Only one parameter δ: extend to which the PS distribution was shifted such that receipt treatment was more (or less) likely

Cluster Incremental Propensity Score (CIPS) Policy (cont'd)

- $\pi_{ij} = \pi(j, \mathbf{X}_i, N_i) = \mathbb{P}(A_{ij} = 1 | \mathbf{X}_i, N_i)$: PS of unit j in cluster i
- $\pi_{ij,\delta} = \mathbb{P}_\delta(A_{ij} = 1 | \mathbf{X}_i, N_i) = \delta\pi_{ij}/(\delta\pi_{ij} + 1 - \pi_{ij})$: shifted PS
- Odds ratio of treatment in counterfactual versus factual

$$\frac{\pi_{ij,\delta}}{1 - \pi_{ij,\delta}} \Big/ \frac{\pi_{ij}}{1 - \pi_{ij}} = \delta$$

- CIPS(δ) policy distribution

$$\mathbb{P}_\delta(\mathbf{A}_i = \mathbf{a}_i | \mathbf{X}_i, N_i) = \prod_{j=1}^{N_i} (\pi_{ij,\delta})^{a_{ij}} (1 - \pi_{ij,\delta})^{1 - a_{ij}}$$

Causal Estimands

- EAPO when the odds were δ times the factual

$$\mu(\delta) = \mathbb{E}\left\{N_i^{-1}\sum_{j=1}^{N_i}\sum_{\mathbf{a}_i\in\mathcal{A}(N_i)} Y_{ij}(\mathbf{a}_i)P_\delta(\mathbf{A}_i = \mathbf{a}_i|\mathbf{X}_i, N_i)\right\}$$

- Overall effect

$$OE(\delta, \delta') = \mu(\delta) - \mu(\delta')$$

(e.g.) Difference in risks of COVID19 when the odds of vaccination were δ times the factual versus δ' times the factual

Causal Estimands

- EAPO when treated ($t = 1$) / untreated ($t = 0$) when the odds of others were δ times the factual

$$\mu_t(\delta) = \mathbb{E}\left\{ N_i^{-1} \sum_{j=1}^{N_i} \sum_{\mathbf{a}_{i(-j)} \in \mathcal{A}(N_i - 1)} Y_{ij}(t, \mathbf{a}_{i(-j)}) P_\delta(\mathbf{a}_{i(-j)} | \mathbf{X}_i, N_i) \right\}$$

- Direct effect

$$DE(\delta) = \mu_1(\delta) - \mu_0(\delta)$$

(e.g.) Risk of COVID19 when vaccinated versus not when odds of vaccination were δ times the factual

- Spillover effect when treated ($t = 1$) / untreated ($t = 0$)

$$SE_t(\delta, \delta') = \mu_t(\delta) - \mu_t(\delta')$$

(e.g.) Difference in risks of COVID19 when vaccinated ($t = 1$) when the odds of others getting vaccinated were δ times the factual versus δ' times the factual

Assumptions and Identifiability

Assumption

(A1) *Consistency*: $Y_{ij} = \sum_{\mathbf{a}_i \in \mathcal{A}(N_i)} Y_{ij}(\mathbf{a}_i) \mathbb{1}(\mathbf{A}_i = \mathbf{a}_i)$

(A2) *Conditional Exchangeability*: $Y_{ij}(\mathbf{a}_i) \perp\!\!\!\perp \mathbf{A}_i | \mathbf{X}_i, N_i$ for all $\mathbf{a}_i \in \mathcal{A}(N_i)$

(A3) *Positivity*: $\mathbb{P}(A_{ij} = 1 | \mathbf{X}_i, N_i) \in (c, 1-c)$ for some $c \in (0,1)$

(A4) *Finite moments*: $\left| \mathbb{E}(Y_{ij} | \mathbf{A}_i, \mathbf{X}_i, N_i) \right| \leq C$ and $\mathbb{E}(Y_{ij}^2 | \mathbf{A}_i, \mathbf{X}_i, N_i) \leq C$ for some $C < \infty$

(A5) *Finite cluster size*: $\mathbb{P}(N_i \leq n_{\max}) = 1$ for some $n_{\max} \in \mathbb{N}$

Lemma (Identifiability of Causal Estimands)

$$\mu(\delta) = \mathbb{E}\left\{ \sum_{\mathbf{a}_i \in \mathcal{A}(N_i)} \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}(Y_{ij} | \mathbf{A}_i = \mathbf{a}_i, \mathbf{X}_i, N_i) \mathbb{P}_\delta(\mathbf{A}_i = \mathbf{a}_i | \mathbf{X}_i, N_i) \right\}$$

Methods

Inference procedure:

1. Find efficient influence function (Semiparametric efficiency theory)
2. Decide nuisance functions estimator (Super leaner ensemble estimator)
3. Implement sample splitting estimator
4. Investigate asymptotic property of the estimator

Why?

- Asymptotically normal and semiparmetric efficient
- Not restricting complexity of nuisance function estimators
- Super learner ensemble estimator asymptotically attains best performance of algorithms included in library

Efficient Influence Function (EIF)

Theorem (EIF of $\mu(\delta)$)

$$\varphi^*_{\mu(\delta)}(\mathbf{O}_i) = \sum_{\mathbf{a}_i \in \mathcal{A}(N_i)} \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}\big(Y_{ij} | \mathbf{A}_i = \mathbf{a}_i, \mathbf{X}_i, N_i\big) \times \mathbb{P}_\delta(\mathbf{A}_i = \mathbf{a}_i | \mathbf{X}_i, N_i)$$

$$\times \left[1 + \sum_{l=1}^{N_i} \frac{1}{\mathbb{P}_\delta(A_{il} = a_{il} | \mathbf{X}_i, N_i)} \frac{(2a_{il} - 1)\delta \{A_{il} - \pi_{il}\}}{\{\delta \pi_{il} + 1 - \pi_{il}\}^2} \right]$$

$$+ \left[\frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ Y_{ij} - \mathbb{E}\big(Y_{ij} | \mathbf{A}_i, \mathbf{X}_i, N_i\big) \right\} \right] \frac{\mathbb{P}_\delta(\mathbf{A}_i | \mathbf{X}_i, N_i)}{\mathbb{P}(\mathbf{A}_i | \mathbf{X}_i, N_i)} - \mu(\delta)$$

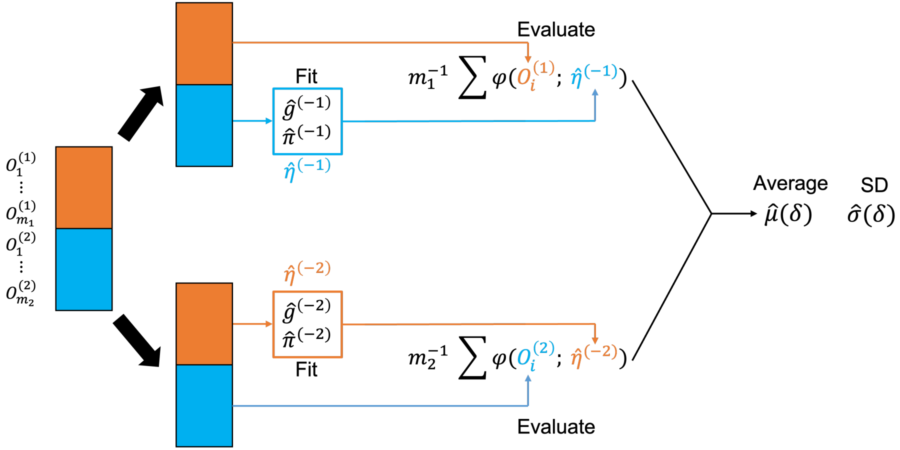
- $\eta = (g, \pi)$: nuisance functions

  (i) $g(j, \mathbf{a}_i, \mathbf{x}_i, n_i) = \mathbb{E}\big(Y_{ij} | \mathbf{A}_i = \mathbf{a}_i, \mathbf{X}_i = \mathbf{x}_i, N_i = n_i\big)$
  $= \mathbb{E}\big(Y_{ij} | A_{ij} = a_{ij}, \overline{\mathbf{A}}_{i(-j)} = \overline{\mathbf{a}}_{i(-j)}, \mathbf{X}_{ij} = \mathbf{x}_{ij}\big)$

  (ii) $\pi(j, \mathbf{x}_i, n_i) = \mathbb{P}(A_{ij} = 1 | \mathbf{X}_i = \mathbf{x}_i, N_i = n_i) = \mathbb{P}(A_{ij} = 1 | \mathbf{X}_{ij} = \mathbf{x}_{ij})$

- Plug-in estimator: $\widehat{\mu}(\delta) = m^{-1} \sum_i \varphi(O_i; \widehat{\eta})$

Sample Splitting Estimator (Chernozhukov et al. 2018)



Theoretical Results

Theorem

Let $\|f\| = \{\int f(\mathbf{o})^2 d\mathbb{P}(\mathbf{o})\}^{1/2}$ *denote the squared* $L_2(\mathbb{P})$ *norm. Also, let* $\delta \in \mathcal{D} = [\delta_l, \delta_u]$*, where* $0 < \delta_l < \delta_u < \infty$*. Assume the following:*

(B1) $\widehat{\pi}^{(-k)}(j, \mathbf{X}_i, N_i) \in (c, 1-c)$, $|\widehat{g}^{(-k)}(j, \mathbf{A}_i, \mathbf{X}_i, N_i)| \leq C < \infty$

(B2) $\left\| \sum_{j=1}^{N} |(\widehat{\pi}^{(-k)} - \pi)(j, \mathbf{X}, N)| \right\| = o_{\mathbb{P}}(m^{-p})$, $p \geq 0$.

(B3) $\left\| \sum_{\mathbf{a} \in \mathcal{A}(N)} \sum_{j=1}^{N} |(\widehat{g}^{(-k)} - g)(j, \mathbf{a}, \mathbf{X}, N)| \right\| = O_{\mathbb{P}}(m^{-q})$, $q \geq 0$

(B4) $\mathbb{E}(Y_{ij}^4 | \mathbf{A}_i, \mathbf{X}_i, N_i) \leq C^* < \infty$

Then,

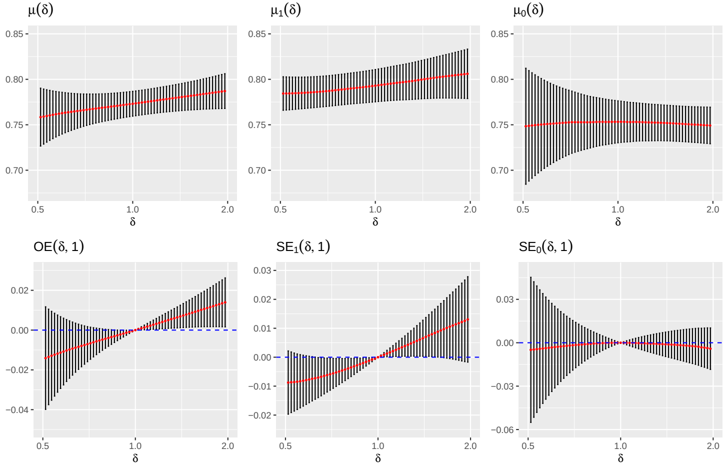
1. *Under (B1) and (B2),* $\widehat{\mu}(\delta) = \mu(\delta) + o_{\mathbb{P}}(m^{-p})$.
2. *Under (B1)* $\sim$ *(B3) with* $p, q \geq 1/4$, $\sqrt{m}\{\widehat{\mu}(\delta) - \mu(\delta)\} \rightsquigarrow \mathbb{G}(\delta)$ in $\ell^{\infty}(\mathcal{D})$.
   Here, $\mathbb{G}(\cdot)$*: mean 0 GP with covariance* $\mathbb{E}\{\mathbb{G}(\delta_1)\mathbb{G}(\delta_2)\} = \mathbb{E}\{\varphi^*_{\mu(\delta_1)}\varphi^*_{\mu(\delta_2)}\}$
3. *Under (B1)* $\sim$ *(B4) with* $p, q \geq 1/4$, $\widehat{\sigma}^2(\delta) \xrightarrow{p} \sigma^2(\delta)$.
   Therefore, $\sqrt{m}\{\widehat{\mu}(\delta) - \mu(\delta)\}/\widehat{\sigma}(\delta) \xrightarrow{d} N(0, 1)$.

Application to Senegal DHS

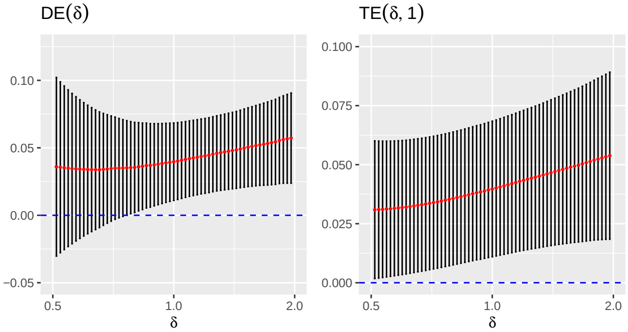
Senegal Demographic and Health Survey (DHS) 2015-2019

- Cluster: Census block ($i = 1, \ldots, 1074$)
- Unit: Household ($j = 1, \ldots, N_i$), max(N_i) = 12, mean(N_i) = 4.25
- $Y_{ij} = \mathbb{1}$(All children in household were diarrhea-free)
- $A_{ij} = \mathbb{1}$(Household had private water source or flushable toilet)
- $X_{ij} =$ {household size, children age, parents employment, block size, block location, etc}
- Clean water facility in a household may decrease risk of diarrhea in other household (Park et al. 2021)
- Proposed method with $\delta \in [0.5, 2]$. $\delta = 1$ (factual) served as comparison for $SE_1(\delta, 1)$, $SE_0(\delta, 1)$, $OE(\delta, 1)$
- Nuisance functions estimator: Ensemble of penalized logistic regression, spline regression, GAM, GBM, random forest, and ANN using super learner

Application to Senegal DHS (cont'd)



Application to Senegal DHS (cont'd)



1. Having private water source or flushable toilet decreases the risk of diarrhea among children ($DE(\delta) > 0$ and increasing)
2. Children from clean water households may receive an additional protective spillover effect from neighboring clean water households ($SE_1(\delta) > 0$ and increasing)

Acknowledgement

Funding:

- NIH grant R01 AI085073

Advisors:

- Dr. Michael G. Hudgens, Department of Biostatistics, University of North Carolina at Chapel Hill
- Dr. Donglin Zeng, Department of Biostatistics, University of North Carolina at Chapel Hill

Reference I

Barkley, Brian G et al. (2020). “Causal inference from observational studies with clustered interference, with application to a cholera vaccine study”. In: *The Annals of Applied Statistics* 14.3, pp. 1432–1448.

Chernozhukov, Victor et al. (Jan. 2018). “Double/debiased machine learning for treatment and structural parameters”. In: *The Econometrics Journal* 21.1, pp. C1–C68. ISSN: 1368-4221.

Kennedy, Edward H. (2019). “Nonparametric Causal Effects Based on Incremental Propensity Score Interventions”. In: *Journal of the American Statistical Association* 114.526, pp. 645–656.

Muñoz, Iván Dıaz and Mark Van Der Laan (2012). “Population intervention causal effects based on stochastic interventions”. In: *Biometrics* 68.2, pp. 541–549.

Reference II

- Park, Chan et al. (2021). “Optimal Allocation of Water and Sanitation Facilities To Prevent Communicable Diarrheal Diseases in Senegal Under Partial Interference”. In: *arXiv preprint arXiv:2004.08950*.
- Salo, Jussipekka et al. (2022). “The indirect effect of mRNA-based COVID-19 vaccination on healthcare workers’ unvaccinated household members”. In: *Nature Communications* 13.1, pp. 1–7.
- Tchetgen Tchetgen, Eric J and Tyler J VanderWeele (2012). “On causal inference in the presence of interference”. In: *Statistical Methods in Medical Research* 21.1, pp. 55–75.

Nuisance function estimation

- $\eta = (g, \pi)$: nuisance functions

$$
\begin{aligned}
\text{(i) } g(j, \mathbf{a}_i, \mathbf{x}_i, n_i) &= \mathbb{E}\big(Y_{ij} | \mathbf{A}_i = \mathbf{a}_i, \mathbf{X}_i = \mathbf{x}_i, N_i = n_i\big) \\
&= \mathbb{E}\big(Y_{ij} | A_{ij} = a_{ij}, \overline{\mathbf{A}}_{i(-j)} = \overline{\mathbf{a}}_{i(-j)}, \mathbf{X}_{ij} = \mathbf{x}_{ij}\big) \\
&= g^*(\mathbf{a}_{ij}, \overline{\mathbf{a}}_{i(-j)}, \mathbf{x}_{ij}), \\
\text{(ii) } \pi(j, \mathbf{x}_i, n_i) &= \mathbb{P}(A_{ij} = 1 | \mathbf{X}_i = \mathbf{x}_i, N_i = n_i) \\
&= \mathbb{P}(A_{ij} = 1 | \mathbf{X}_{ij} = \mathbf{x}_{ij}) \\
&= \pi^*(\mathbf{x}_{ij}),
\end{aligned}
$$

where $\overline{\mathbf{A}}_{i(-j)} = (\sum_{k \neq j} A_{ik})/(N_i - 1)$ and $\overline{\mathbf{a}}_{i(-j)} = (\sum_{k \neq j} a_{ik})/(n_i - 1)$

- Super learner ensemble of nonparametric and ML methods to estimate $\eta = (g, \pi)$
- $\varphi_{\mu(\delta)}(\mathbf{O}_i; \eta) = \varphi^*_{\mu(\delta)}(\mathbf{O}_i; \eta) + \mu(\delta)$: uncentered EIF
(i.e.) $\mathbb{E}\{\varphi_{\mu(\delta)}(\mathbf{O}_i; \eta)\} = \mu(\delta)$

Sample Splitting Estimator (Chernozhukov et al. 2018)

1. Randomly partition $(\mathbf{O}_1, \ldots, \mathbf{O}_m)$ into K disjoint groups
   - $S_i \in \{1, \ldots, K\}$: group membership for cluster i
   - $m_k = \sum_{i=1}^{m} \mathbb{1}(S_i = k)$: size of group $k \in \{1, \ldots, K\}$
2. For group $k \in \{1, \ldots, K\}$
   2.1 Fit $\widehat{\eta}^{(-k)} = (\widehat{g}^{(-k)}, \widehat{\pi}^{(-k)})$ trained on $\{\mathbf{O}_i : S_i \neq k\}$
   2.2 Obtain $\mathbb{P}_m^k\{\varphi_{\mu(\delta)}(\mathbf{O}; \widehat{\eta}^{(-k)})\} = m_k^{-1} \sum_{i:S_i=k} \varphi_{\mu(\delta)}(\mathbf{O}_i; \widehat{\eta}^{(-k)})$
3. Proposed estimator is average of estimated group-averaged uncentered EIFs:

$$\widehat{\mu}(\delta) = K^{-1} \sum_{k=1}^{K} \mathbb{P}_m^k\{\varphi_{\mu(\delta)}(\mathbf{O}; \widehat{\eta}^{(-k)})\}$$

4. Variance of $\widehat{\mu}(\delta)$ can be estimated by

$$\widehat{\sigma}(\delta)^2 = K^{-1} \sum_{k=1}^{K} \mathbb{P}_m^k\left[\left\{\varphi_{\mu(\delta)}(\mathbf{O}; \widehat{\eta}^{(-k)}) - \widehat{\mu}(\delta)\right\}^2\right]$$

Simulations

- $D = 1000$ simulations, each consisted of $m = 500$ clusters
- $N_i \overset{iid}{\sim} Unif\{5, 6, \ldots, 20\}, i = 1, \ldots, m$
- $C_i \sim \mathrm{N}(0, 1)$: one cluster-level covariate
- $X_{ij1} \sim \mathrm{N}(0, 1)$, $X_{ij2} \sim$ Bernoulli(0.5): individual-level covariates
- $A_{ij} \sim$ Bernoulli(p_{ij}^A): treatment status, $Y_{ij} \sim$ Bernoulli(p_{ij}^Y): outcome
  $p_{ij}^A = \text{expit}(0.1 + 0.2|X_{ij1}| + 0.2|X_{ij1}|X_{ij2} + 0.1\mathbb{1}(C_i > 0))$
  $p_{ij}^Y = \text{expit}(3 - 2A_{ij} - \overline{\mathbf{A}}_{i(-j)} - 1.5|X_{ij1}| + 2X_{ij2} - 3|X_{ij1}|X_{ij2} - 2\mathbb{1}(C_i > 0))$

		Nonparametric					Parametric				
Estimand	Truth	Bias	RMSE	ASE	ESE	Cov %	Bias	RMSE	ASE	ESE	Cov %
$\mu(2)$	0.300	-0.003	0.013	0.013	0.013	94.9%	-0.010	0.017	0.013	0.014	87.1%
$\mu_1(2)$	0.224	-0.004	0.014	0.014	0.014	93.7%	-0.017	0.022	0.014	0.015	75.9%
$\mu_0(2)$	0.507	0.003	0.018	0.017	0.017	94.5%	0.025	0.031	0.019	0.019	75.6%
$DE(2)$	-0.283	-0.007	0.019	0.019	0.018	94.2%	-0.042	0.046	0.021	0.020	45.6%
$SE_1(2,1)$	-0.018	-0.002	0.010	0.010	0.010	94.3%	-0.004	0.012	0.011	0.012	93.4%
$SE_0(2,1)$	-0.022	-0.002	0.012	0.012	0.012	94.4%	-0.006	0.015	0.014	0.014	92.5%
$OE(2,1)$	-0.063	-0.003	0.009	0.009	0.009	93.9%	-0.010	0.015	0.010	0.010	77.7%
$TE(2,1)$	-0.306	-0.009	0.017	0.015	0.015	92.1%	-0.047	0.050	0.017	0.017	22.7%

Discussion

- New set of causal estimands under clustered interference based on policy which shifts PS distribution, allowing for units' probability of receiving treatment to vary by their covariates and are not based on parametric model
- Efficient nonparametric sample splitting estimator which uses a variety of data-adaptive methods, reducing the risk of model mis-specification compared to parametric estimators
- Application to the Senegal DHS data:
  1. Having private water source or flushable toilet decreases the risk of diarrhea among children
  2. Children from clean water households may receive an additional protective spillover effect from neighboring clean water households