Nonparametric-efficient Causal Mediation Analysis for Stochastic Interventions

Nima Hejazi, Mark van der Laan, and Iván Díaz

Graduate Group in Biostatistics & Dept. of Statistics, UC Berkeley

Division of Biostatistics, Dept. of Healthcare Policy & Research, Weill Cornell Medicine

Overview & Motivations

- Using stochastic interventions, we present a decomposition of the *population intervention effect* into direct and indirect effects.
  - Define causal contrasts of effects of continuous and categorical exposures
  - ...
- We propose estimators of these direct and indirect effects:
  - *Classical parametric*: substitution and re-weighted (IPW) estimators
  - *Nonparamtric-efficient*: one-step and TMLE using data adaptive regression
- Our efficient estimators are asymptotically linear under a condition requiring $n^{1/4}$-consistency of certain regression functions.

Software Implementation

- The `medshift` R package [3] implements these estimators and leverages state-of-the-art machine learning in the procedure.
  - Construction of all estimators via the eponymous `medshift()` function.
  - Uses the `sl3` R package to provide machine learning facilities.
- Construction of TML estimators using tools from the `tlverse` software ecosystem.
- ...

Construction of Nonparametric-Efficient Estimators

- To avoid entropy conditions on initial estimators, we rely on cross-fitting [6, 1]. Let $\hat{\eta}_j$ be the estimator of $\eta = (g, m, e, \phi)$ and $j(i)$ the index of the validation set containing observation i.
- A one-step estimator [4] may be constructed by augmenting the substitution estimator with the efficient influence function:

$$\hat{\theta}_{\text{OS}}(\delta) = \frac{1}{n}\sum_{i=1}^{n} D_{\hat{\eta}_{j(i)},\delta}(O_i) = \frac{1}{n}\sum_{i=1}^{n}\left\{D^{Y}_{\hat{\eta}_{j(i)},\delta}(O_i) + D^{A}_{\hat{\eta}_{j(i)},\delta}(O_i) + D^{Z,W}_{\hat{\eta}_{j(i)},\delta}(O_i)\right\}.$$

  - ...
  - ...
- A targeted minimum loss-based estimator may be constructed by using the efficient influence function as an estimating equation, updating estimates of nuisance components:

$$\hat{\theta}_{\text{TMLE}}(\delta) = \ldots,$$

  where
  - Unlike the one-step estimator, the TMLE is a substitution estimator.
  - Use unversal least favorable submodels for one-step estimation [5].

Stochastic Population Intervention (In)Direct Effects

- Consider $O = (W, A, Z, Y) \sim P_0 \in \mathcal{M}$, where W is a set of baseline covariates, A an intervention, Z a mediator between A and outcome, and Y the outcome, with no assumptions on model $\mathcal{M}$.
- We may decompose the PIE in terms of a *population intervention direct effect (PIDE)* and a *population intervention indirect effect (PIIE)*:

$$\psi(\delta) = \overbrace{\mathbb{E}\{Y(g,q) - Y(g_\delta,q)\}}^{\text{PIDE}} + \overbrace{\mathbb{E}\{Y(g_\delta,q) - Y(g_\delta,q_\delta)\}}^{\text{PIIE}}.$$

- We show the causal parameter $\mathbb{E}\{Y(g_\delta, q)\}$ is identified by the observed data parameter [2]:

$$\theta(\delta) = \int m(a,z,w) g_\delta(a \mid w) p(z,w) d\nu(a,z,w).$$

- Letting $\eta = (g, m, e, \phi)$, the efficient influence function for $\theta(\delta)$ in the nonparametric model M is $D^{Y}_{\eta,\delta}(o) + D^{A}_{\eta,\delta}(o) + D^{Z,W}_{\eta,\delta}(o) - \theta(\delta)$, where

$$D^{Z,W}_{\eta,\delta}(o) = \int m(z,a,w) g_\delta(a \mid w) d\kappa(a),$$

$$D^{A}_{\eta,\delta}(o) = \frac{g_\delta(a \mid w)}{g(a \mid w)}\left\{\phi(a,w) - \int \phi(a,w) g_\delta(a \mid w) d\kappa(a)\right\}$$

$$D^{Y}_{\eta,\delta}(o) = \frac{g_\delta(a \mid w)}{e(a \mid z,w)}\{y - m(z,a,w)\},$$

where $\phi_0(a,w) = \mathbb{E}\left\{\frac{g(A|W)}{e(A|Z,W)} m(Z,A,W) \mid A = a, W = w\right\}$.

Results & Discussion

- All estimators approx. unbiased in large samples; however, inefficient TMLE with HAL has bias not converging at $n^{-\frac{1}{2}}$.
- Fitting Π with HAL or GLM, efficient TMLE has lower variance than the inefficient.

References

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen *et al.*, "Double machine learning for treatment and causal parameters," *arXiv preprint arXiv:1608.00060*, 2016.

[2] I. Díaz and N. S. Hejazi, "Causal mediation analysis for stochastic interventions," *in revision*, 2019. [Online]. Available: https://arxiv.org/abs/1901.02776

[3] N. S. Hejazi and I. Díaz, *medshift: Causal mediation analysis for stochastic interventions in R*, 2019, R package version 0.0.8. [Online]. Available: https://github.com/nhejazi/medshift

[4] J. Pfanzagl and W. Wefelmeyer, "Contributions to a general asymptotic statistical theory," *Statistics & Risk Modeling*, vol. 3, no. 3-4, pp. 379–388, 1985.

[5] M. van der Laan and S. Gruber, "One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels," *The international journal of biostatistics*, vol. 12, no. 1, pp. 351–378, 2016.

[6] W. Zheng and M. J. van der Laan, "Cross-validated targeted minimum-loss-based estimation," in *Targeted Learning*. Springer, 2011, pp. 459–474.

But wait, there's more!

- **N. Hejazi**: `nhejazi@berkeley.edu`; **M. van der Laan**: `laan@berkeley.edu`; **I. Díaz**: `ild2005@med.cornell.edu`
- `https://arxiv.org/abs/1901.02776`
- Check out Iván's talk tomorrow morning!