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

- We present a decomposition of the *population intervention effect*, defined through stochastic interventions.
- Population intervention effects provide a generalized framework in which a variety of interesting causal contrasts can be defined, including effects for continuous and categorical exposures.
- We propose estimators of these direct and indirect effects:
  - substitution (G-computation), re-weighted (IPW), and
  - efficient estimators based on flexible regression techniques (one-step, TMLE).

Software Implementation

- Our efficient estimators are asymptotically linear under a condition requiring $n^{1/4}$-consistency of certain regression functions.
- The efficient estimators are asymptotically normal with estimable variance, thereby allowing for the construction of confidence intervals and hypothesis tests.
- The `medshift` R package [2] implements these estimators and leverages state-of-the-art machine learning in the procedure.
- `tlverse` plug: `https://tlverse.org`
- ...

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)*:

$$\psi(\delta) = \overbrace{\mathbb{E}\{Y(g,q) - Y(g_\delta, q)\}}^{\text{PIDE}} + \overbrace{\mathbb{E}\{Y(g_\delta, q) - Y(g_\delta, q_\delta)\}}^{\text{PIIE}}. \tag{2}$$

- The causal parameter is identified by the observed data parameter [**?**]:

$$\theta(\delta) = \int m(a,z,w) g_\delta(a \mid w) p(z,w) d\nu(a,z,w). \tag{3}$$

- Let $\eta = (g, m, e, \phi)$. 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

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

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

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

Construction of Nonparametric-Efficient Estimators

- To avoid entropy conditions on the initial estimators, we rely on cross-fitting [3, 1] — denote by $\hat{\eta}_j$ the estimator of $\eta = (g, m, e, \phi)$, obtained by training the corresponding prediction algorithm using only data in the sample $\mathcal{T}_j$. Further, let $j(i)$ denote the index of the validation set which contains observation i. The estimator is thus defined as:

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

- TMLE!

Results & Discussion

- All estimators approx. unbiased in large samples; however, inefficient TMLE with HAL has bias not converging at $n^{-\frac{1}{2}}$.
- Fitting II 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] 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

[3] 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. Diaz**: `ild2005@med.cornell.edu`

- Pre-print of our original paper: `https://`