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 consider the problem of efficiently estimating the effect of a stochastic shift interventions in studies with two-phase sampling of the treatment.
- We present an estimator of the average counterfactual outcome under a stochastic shift intervention with
  - consistency and efficiency guarantees,
  - a multiple double robustness property.
- The proposed estimator is asymptotically normal with estimable variance, thereby allowing for the construction of confidence intervals and hypothesis tests.
- The *txshift* R package [2] implements these estimators and leverages state-of-the-art machine learning in the procedure.

Data: HIV Vaccine Trials

- Our approach is motivated by application to investigations of the effects of immune responses on HIV vaccine efficacy.
- *Question:* **How does risk of HIV infection differ under shifts of an immune response in the vaccine arm of an efficacy trial?**
- We simulate a data structure based on the HVTN 505 HIV-1 efficacy trial, as in [3]:
  - 2504 participants, with all observed cases matched to controls.
  - Background (W): sex, age, BMI, etc.
  - Intervention (A): immunobiomarkers (i.e., T-Cell profiles from ICS assays on preserved HIV-1-stimulated PBMCs).
  - Outcome (Y): HIV-1 infection status.
- *Takeaway:* **Variable importance measure for ranking immune responses by utility as immunogenicity study endpoints in all future HIV-1 vaccine efficacy trials.**

Methodology I: The Effect of a Stochastic Intervention

- Consider $X = (W, A, Y) \sim P_0^X \in \mathcal{M}$, where W is a set of baseline covariates, A a treatment, and Y an outcome of interest, with no assumptions placed on the statistical model $\mathcal{M}$.
- Consider a shift of the treatment (i.e., $d(A, W)$) so that $A = A + \delta$ for a user-specified shift δ.
- To protect against violations of the assumption of positivity, the shifting mechanism may be made a function of the observed data, where $u(w)$ is the maximum shift with support in the data:

$$d(a, w) = \begin{cases} a + \delta, & a + \delta < u(w) \\ a, & \text{otherwise} \end{cases}$$

- The causal parameter is identified by the observed data parameter [4]:

$$\Psi(P_0^X) = \mathbb{E}_{P_0} \overline{Q}(d(A, W), W), \tag{1}$$

  where $\overline{Q}(d(A, W), W)$ is the conditional mean of the outcome given $A = d(A, W)$ and W.
- The efficient influence function (EIF) of Ψ relative to a nonparametric model is

$$D^F(P_0^X)(x) = H(a, w)(y - \overline{Q}(a, w)) + \overline{Q}(d(a, w), w) - \Psi(P_0^X)(x), \tag{2}$$

  where the auxiliary term, $H(a, w)$, may be expressed as

$$H(a, w) = \mathbb{I}(a < u(w)) \frac{g(a - \delta \mid w)}{g(a \mid w)} + \mathbb{I}(a \geq u(w) - \delta). \tag{3}$$

Methodology II: Corrections for Two-Phase Sampling

- In the HVTN 505 HIV-1 trial, all infected participants and the matched subset of controls have A (immunobiomarkers) measured, thus making the observed data structure $O = (W, \Delta A, Y) \sim P_0$.
  - $\Delta \in \{0, 1\}$ is the missingness mechanism introduced by sampling, under which the observed immune response (ΔA) is arbitrarily set to 0 when unobserved.
  - We assume that, given $V := (W, Y)$, Δ is Bernoulli distributed with probability $\Pi_0(V)$.
- The IPCW-TMLE [5] estimates the target parameter by adding *inverse weights* to the loss function.
- Improvements in the efficiency are attainable using a TMLE based on the EIF:

$$D(P_0)(o) = \frac{\Delta}{\Pi(v)} D^F(P_0^X)(x) - \left(1 - \frac{\Delta}{\Pi(v)}\right) \mathbb{E}(D^F(P_0^X)(x) \mid \Delta = 1, V = v)$$

- This augmented estimator exhibits several desirable properties
  - *efficiency*, achieving the CR bound for the class of regular asymptotically linear estimators;
  - *multiple robustness*, consistency of the parameter estimate when one of (g, Q) and one of $(\Pi, \mathbb{E}_0(D^F(P_0^X)(X) \mid \Delta = 1, V))$ is consistently estimated;
  - valid statistical inference even when Π is estimated nonparametrically.

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] I. Díaz and M. J. van der Laan, "Stochastic treatment regimes," in *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media, 2018, pp. 167–180.

[2] N. S. Hejazi, M. J. van der Laan, and D. C. Benkeser, *txshift: Targeted Learning of Causal Effects under Stochastic Treatment Regimes in R*, 2018, r package version 0.2.0. [Online]. Available: https://github.com/nhejazi/txshift

[3] H. E. Janes, K. W. Cohen, N. Frahm, S. C. De Rosa, B. Sanchez, J. Hural, C. A. Magaret, S. Karuna, C. Bentley, R. Gottardo *et al.*, "Higher t-cell responses induced by dna/rad5 hiv-1 preventive vaccine are associated with lower hiv-1 infection risk in an efficacy trial," *The Journal of infectious diseases*, vol. 215, no. 9, pp. 1376–1385, 2017.

[4] I. D. Muñoz and M. J. van der Laan, "Population intervention causal effects based on stochastic interventions," *Biometrics*, vol. 68, no. 2, pp. 541–549, 2012.

[5] S. Rose and M. J. van der Laan, "A targeted maximum likelihood estimator for two-stage designs," *The International Journal of Biostatistics*, vol. 7, no. 1, pp. 1–21, 2011.

Contact Information

- **N. Hejazi**, Ph.D. candidate, NHEJAZI@BERKELEY.EDU
- **M. van der Laan**, Prof. of Biostatistics & Statistics, LAAN@BERKELEY.EDU
- **I. Diaz**: Asst. Prof. of Biostatistics, ILD2005@MED.CORNELL.EDU