Robust inference on the causal effects of stochastic interventions in two-phased vaccine efficacy trials

Nima Hejazi, David Benkeser, and Mark van der Laan

Graduate Group in Biostatistics & Department of Statistics, University of California, Berkeley
Department of Biostatistics and Bioinformatics, Emory University

Overview & Motivations

- We consider estimating the effect of a stochastic intervention in vaccine trials with two-phase sampling of the treatment.
- We present efficient estimators of the mean counterfactual outcome under a stochastic intervention with
  - consistency and efficiency guarantees,
  - a *multiple* double robustness property.
- The proposed estimators are asymptotically normal with a variance estimator based on a well-studied efficient influence function.
- Our `txshift R` package [3] implements these estimators and leverages state-of-the-art machine learning in the procedure.
- Our `haldensify R` package [2] estimates the propensity score, a conditional density.

HIV Vaccine Efficacy Trials

- Question: How does HIV infection risk vary with the shifting of an immunogenic response profile in a trial's vaccine arm.
- Analysis of data from the HIV Vaccine Trials Network 505 efficacy trial, as in [4]:
  - RCT with $n = 2504$; only $n = 189$ from vaccine arm in second-stage sample.
  - Background covariates (W): sex, age, BMI, behavioral HIV risk score.
  - Treatment (A): immunogenic response profiles, post-vaccination.
  - Outcome (Y): HIV-1 infection status at week 28 of trial.
  - All individuals with HIV-1 infections by week 28 in second-stage sample.
- Goal: Develop a ranking of immunogenic responses as study endpoints for future HIV-1 vaccine efficacy trials.

Efficient Corrections for Two-Phase Sampling

- In the HVTN 505 HIV-1 trial, immunogenic responses A are sequenced for infected individuals and a matched subset of controls, making $O = (W, \Delta, \Delta A, Y)$ the observed data structure.
  - $\Delta \in \{0, 1\}$ is the sampling mechanism introduced by two-phase sampling, under which the observed immunogenic response (ΔA) is arbitrarily set to 0 when unobserved.
  - Given $V := (W, Y)$, we assume that $\Delta \sim \text{Bern}(\pi_0(V))$.
- The IPCW-TMLE [5] estimates the target parameter by adding *inverse weights* to the loss function.
- Efficiency improvements are attainable by constructing estimators based on the augmented 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 augmentation allows the construction of estimators that
  - achieve the efficiency bound for the class of regular asymptotically linear estimators;
  - are *multiply robust*, with consistency of parameter estimates when one of $\{g, Q\}$ and one of $\{\pi_0(V), \mathbb{E}_0(D^F(P_0^X)(X) \mid \Delta = 1, V)\}$ are consistently estimated;
  - allow valid statistical inference to be derived even when π is estimated nonparametrically.

The Effect of a Stochastic Treatment Regime

- 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 $d(A, W) = A + \delta$ for a given shift δ. To protect against violations of the positivity assumption, make the shifting mechanism a function of the observed data, where $u(w)$ is the maximum shift supported in the observed data:

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

- The causal parameter has been shown to be identified by a functional of the observed data [1]:

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

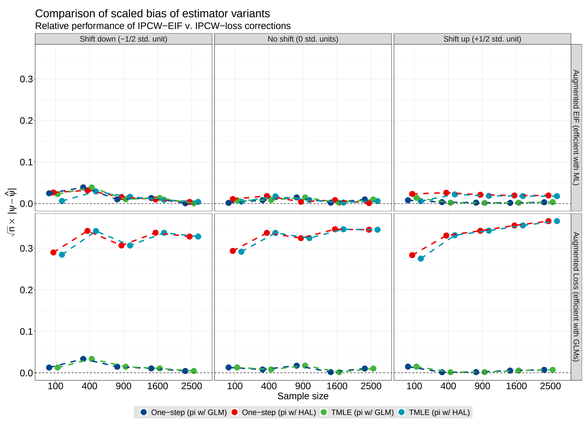
  where $\overline{Q}(d(A, W), W)$ is the conditional mean of the outcome given $A = d(A, W)$ and W.
- The efficient influence function (EIF) of Ψ relative to the nonparametric model $\mathcal{M}$ is

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

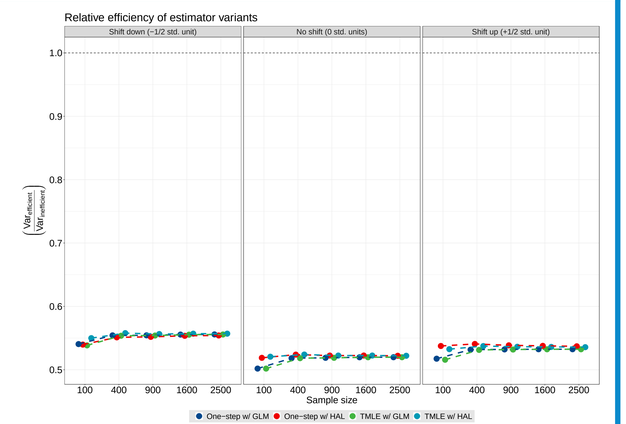
  where the auxiliary covariate term $H(a, w)$ may be expressed as

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

Numerical Study & Results



(a) Most estimator variants unbiased in large samples; inefficient variants with HAL fail to converge.



(b) Fitting π with HAL or GLM, efficient estimators show relative efficiency similar to simpler variants.

References

[1] Iván Díaz and Mark J van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

[2] Nima S Hejazi and David C Benkeser. `haldensify`: *Nonparametric conditional density estimation with the highly adaptive lasso in R*, 2019. URL `https://github.com/nhejazi/haldensify`. R package version 0.0.3.

[3] Nima S Hejazi and David C Benkeser. `txshift`: *Targeted Learning of the Causal Effects of Stochastic Interventions in R*, 2019. URL `https://github.com/nhejazi/txshift`. R package version 0.2.7.

[4] Holly E Janes, Kristen W Cohen, Nicole Frahm, Stephen C De Rosa, Brittany Sanchez, John Hural, Craig A Magaret, Shelly Karuna, Carter Bentley, Raphael 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*, 215(9):1376–1385, 2017.

[5] Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*, 7(1):1–21, 2011.

Contact Information

- **N. Hejazi**, Graduate Group in Biostatistics, UC Berkeley, `nhejazi@berkeley.edu`
- **D. Benkeser**: Asst. Prof. of Biostatistics, Emory University, `benkeser@emory.edu`
- **M. van der Laan**, Prof. of Biostatistics, UC Berkeley, `laan@berkeley.edu`