Robust Nonparametric Inference for Stochastic Interventions Under Multi-Stage Sampling

Nima S. Hejazi, Mark J. van der Laan, and David C. Benkeser

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

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 augmented targeted maximum likelihood estimator of a parameter defined as the outcome under a stochastic 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.
- A new software contribution — the "*txshift*" R package [2] — implements these estimators and leverages state-of-the-art machine learning algorithms in the procedure.

Data: HIV Vaccine Trials

- We illustrate our approach by applying the method in an investigation 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]:
  - About 2500 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 future HIV-1 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 $\mathcal{M}$.
- Rather than a deterministic intervention, consider a shift of the treatment (i.e., consider a shift of the intervention so that $A = A + \delta$ for a user-specified δ).
- To protect against violations of the assumption of positivity, the shifting mechanism may be made a function of the observed data:

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

- We consider a simple causal target parameter, introduced in [4]:

$$\Psi(P)(X) = \mathbb{E}_P \overline{Q}(d(A, W), W), \tag{2}$$

  - for which Wald-style inference is attainable through the efficient influence function (EIF), given in [1]:

$$D(P)(o) = H(a, w)y - \overline{Q}(a, w) + \overline{Q}(d(a, w), w) - \Psi(P)(o), \tag{3}$$

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

$$H(a, w) = \mathbb{I}(a < u(w)) \frac{g_0(a - \delta \mid w)}{g_0(a \mid w)} + \mathbb{I}(a \geq u(w) - \delta). \tag{4}$$

Methodology II: Corrections for Two-Phase Sampling

- In the HVTN 505 HIV-1 trial, all infected individuals are matched to controls using a complex matching mechanism, which makes the observed data structure $O = (W, \Delta A, Y)$.
  - $\Delta = f(V) \in \{0, 1\}$ 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] provides an avenue to estimate the target parameter by inverse weighting.
- Improvements in the efficiency of the IPCW-TMLE may be attained through a more complex EIF:

$$0 = P_n \frac{\Delta}{\Pi_n^*(V)} D^F(P_{X,n}^*) - \left\{ \frac{\Delta}{\Pi_n^*(V)} - 1 \right\} \mathbb{E}_n(D^F(P_{X,n}^0) \mid \Delta = 1, V) \tag{1}$$

- This augmented estimator exhibits several desirable properties
  - *efficiency*, achieving the CR lower bound among all asymptotically linear estimators;
  - *multiple robustness*, consistency of the parameter estimate when any combination of (g, Q) and $(\Pi, \mathbb{E}_0(D^F(P^F) \mid V))$ is correctly estimated;
  - valid statistical inference even when Π_0 is estimated nonparametrically.

Results & Discussion

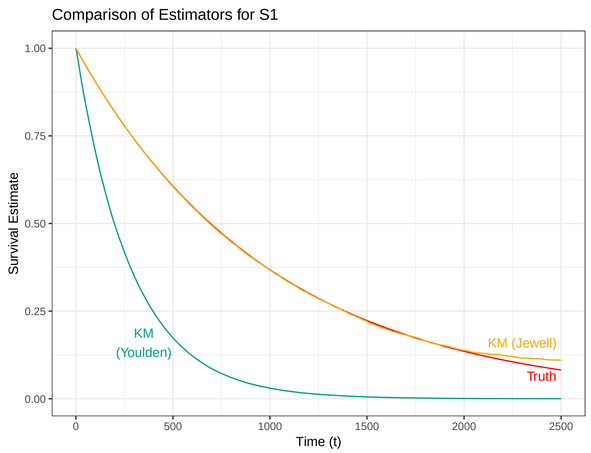

Figure 1: This figure demonstrates...

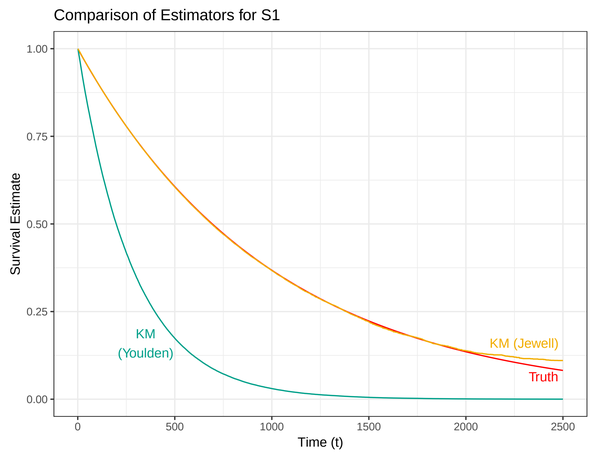

Figure 2: This figure demonstrates...

References

[1] Iván Díaz and Mark J van der Laan. Stochastic treatment regimes. In *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 167–180. Springer Science & Business Media, 2018.

[2] Nima S Hejazi, Mark J van der Laan, and David C Benkeser. *txshift: Targeted Minimum Loss-Based Estimation of the Causal Effect of Stochastic Interventions and Variable Importance Analysis*. URL `https://github.com/nhejazi/txshift`. R package version 0.2.0.

[3] 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.

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

[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.S. Hejazi**, Ph.D. student, Group in Biostatistics, NHEJAZI@BERKELEY.EDU
- **M.J. van der Laan**, Professor of Biostatistics & Statistics, LAAN@BERKELEY.EDU
- **D.C. Benkeser**: Assistant Professor of Biostatistics, BENKESER@EMORY.EDU