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 Computational Biology, Emory University

School of Public Health
UNIVERSITY OF CALIFORNIA, BERKELEY

Overview & Motivations

1. We consider the problem of efficiently estimating the effect of a stochastic shift interventions for problem settings in which multi-stage sampling complicates the observed data structure.
2. We present a novel approach: an augmented targeted maximum likelihood estimator of a parameter defined as the outcome under a stochastic intervention with
   - consistency and efficiency guarantees even under multi-stage sampling, and
   - a form of multiple double robustness inherited from its constituent parts.
3. The proposed nonparametric estimation procedure provably attains fast convergence rates even when incorporating machine learning estimators.
4. A recent software implementation — the "txshift" R package — has been developed for applying this methodology very generally, including for causal inference and variable importance analyses.

Introduction & Data

- We illustrate the utility of our approach by applying the new method and software in an investigation of the effects of immune response biomarkers on HIV vaccine efficacy.
- *Question of interest:* **How does risk of HIV infection differ under posited shifts of the distribution of an immune response in the vaccine arm of an efficacy trial?**
- We simulate a data structure similar to that in the HVTN 505 HIV-1 efficacy trial:
  - About 2500 participants, with all observed cases matched to controls.
  - Background (W): sex, age, BMI, etc.
  - Intervention (A): immunomarkers (i.e., T-Cell profiles from ICS assay on HIV-1-stimulated PBMCs).
  - Outcome (Y): HIV-1 infection risk.
- *Takeaway:* **Variable importance measure for ranking multiple immune responses by their utility as immunogenicity study endpoints in future HIV vaccine trials.**

Methodology I

- Consider $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where no assumptions are placed on the statistical model containing $\mathcal{M}$ containing P_0.
- Rather than a deterministic intervention, consider a shift of the treatment (i.e., instead of $A = a$, consider $A = a + \delta$).
- To compare with the general linear model, the shift δ may be thought of as analogous to shifts in the slope of the regression line.
- To protect against positivity violations, make the shifting mechanism a function of the observed data: $d(a, w) = a + \delta$, if $a + \delta < u(w)$ and $d(a, w) = a$ otherwise.

Let's consider a simple statistical target parameter:

$$\Psi(P) = \mathbb{E}_P \bar{Q}(d(A, W), W), \tag{1}$$

for which the efficient influence function (EIF) is

$$D(P)(o) = H(a, w)y - \bar{Q}(a, w) + \bar{Q}(d(a, w), w) - \Psi(P) \tag{2}$$

Methodology II

- The second approach is non-parametric and uses Kaplan-Meier's estimator defined as

$$\widehat{S}(t) = \prod_{i:t(i)<t} \left(1 - \frac{d_i}{n_i}\right), \quad t \geq 0,$$

where d_i and n_i are the respective numbers of death and individual at risks at the ordered time $t^{(i)}$, $i = 1, \ldots, n$.

- Youlden et al. [?] only uses patients for whom no occurrence of a second melanoma is observed, in the estimation of S_1 and ignores the other patients, which causes a bias.
- Jewell corrects their estimator by including all the patients in the study.
- The ones that were excluded by Youlden et al. [?] still contain information about λ_1: those are censored observations at time U.

Results & Discussion

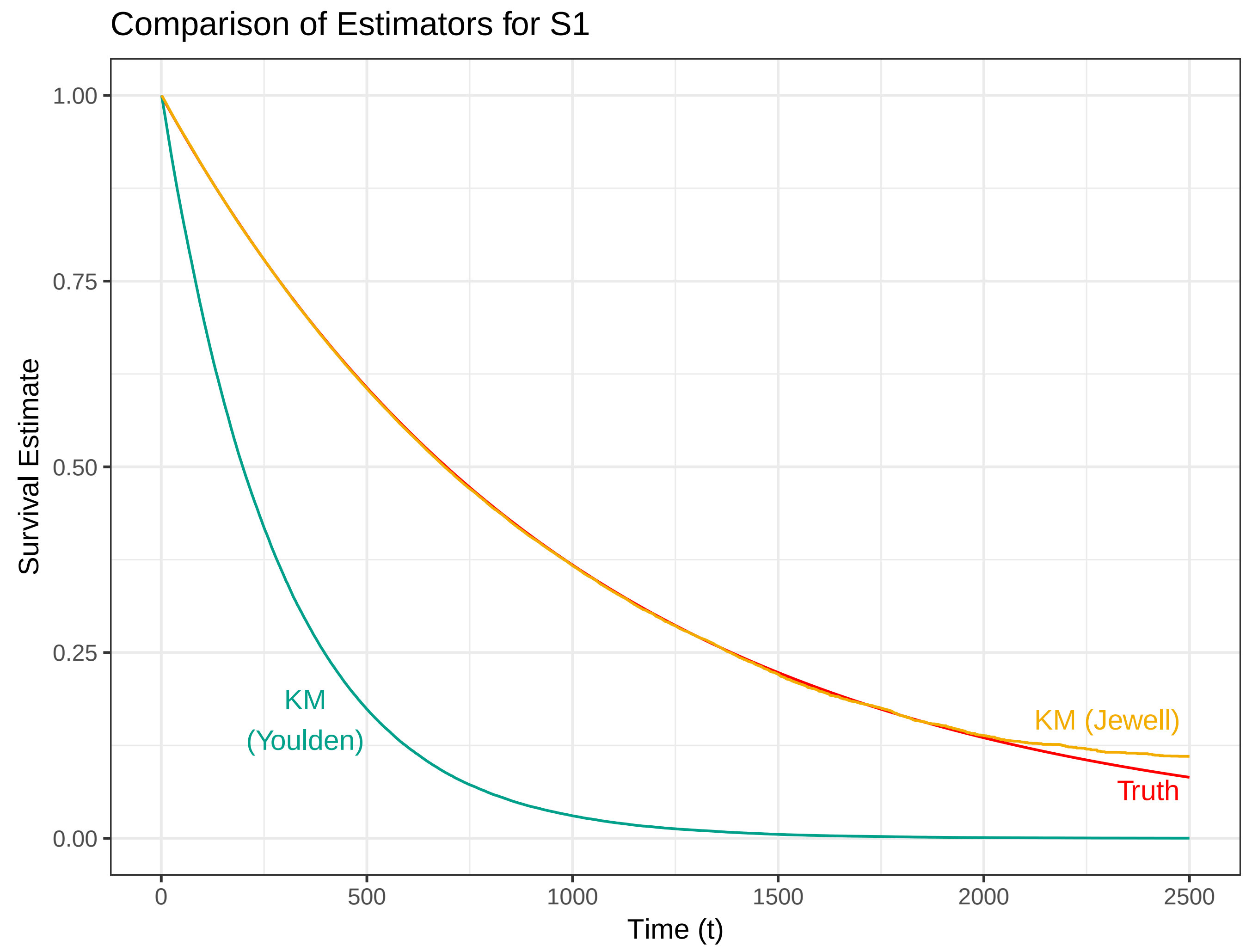

Figure 1: Average performance of estimators for S_1 for a sample of size $n = 1000$, over about 300 simulations.

- The Kaplan-Meier estimator proposed by Youlden displays obvious bias.
- The estimates of the survival curve produced by Cox regression and the Kaplan-Meier estimator with the Jewell correction show no such bias.
- Under the Cox model, Cox regression will outperform other estimators — it draws upon information across both subject groups over all time points.
- The Kaplan-Meier estimator exhibits a slight divergence from the truth in the right tail due to a well-studied finite-sample bias caused by censored observations.
- We display results for $n = 1000$ since this sample size is closest to that from the observational medical study we analyze.

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

[3] 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.

[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.

Contact Information

N.S. Hejazi: NHEJAZI@BERKELEY.EDU
M.J. van der Laan: LAAN@BERKELEY.EDU
D. C. Benkeser: BENKESER@EMORY.EDU