Robust Nonparametric Inference for Stochastic Interventions Under Multi-Stage Sampling

for the UC Berkeley Biostatistics Seminar Series,
given 02 April 2018

Nima Hejazi

Group in Biostatistics
University of California, Berkeley
stat.berkeley.edu/~nhejazi

nimahejazi.org
twitter/@nshejazi
github/nhejazi

slides: goo.gl/Vq6v5o

This slide deck is for a seminar-length talk (about 50 minutes) on a new approach to causal inference and nonparametric variable importance in the context of parameters defined as treatment shifts. Here, we introduce an additive treatment shift parameter, extensions for censored data (including a multiple double robustness property), new statistical software for applying our approach, and applications to a vaccine efficacy trial examining HIV. This talk was most recently given at a meeting of the Biostatistics Seminar Series at the University of California, Berkeley.

Source: `https://github.com/nhejazi/talk_txshift`
Slides: `https://goo.gl/LAoDUJ`
With notes: `https://goo.gl/Vq6v5o`

Preview: Summary

- The evaluation of vaccine efficacy is a high-impact scientific problem that leads to numerous statistical challenges.
- Stochastic interventions provide a flexible framework through which these statistical problems may be viewed from the perspective of causal inference.
- Standard targeted minimum loss-based estimation may be augmented to handle multi-stage sampling designs, like those common in efficacy trials.
- Statistical software is now readily available for deploying these types of techniques in a number of settings. We apply these methods in efficacy trials.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Motivation: Let's meet the data

- HIV Vaccine Trials Network (HVTN) 505 HIV-1 vaccine efficacy trial.
- 2504 participants, with all observed cases matched to controls after collection of endpoints of interest.
- Background quantities (W): sex, age, BMI, etc.
- Variables of interest (A): biomarkers of immune response (e.g., T-Cell response).
- Outcome of interest (Y): HIV-1 infection risk.
- **Question**: How would changes in the immune response profile impact risk of HIV-1 infection?

- A vaccine effective at preventing HIV-1 acquisition would be a cost-effective and durable approach to halting the worldwide epidemic.
- Identifying vaccine-induced immune-response biomarkers that predict a vaccine's ability to protect individuals from HIV-1 infection is a high priority.
- The study was halted on 22 April 2013 due to absence of vaccine efficcacy. There was no significant effect of the vaccine on the primary infection end- point of HIV-1 infection between week 28 and month 24.

Preventive Vaccines for HIV

- Substantial heterogeneity is present in the genetic characteristics of HIV.
- Preventive HIV vaccines constructed using only several antigens (out of a great many).
- **Success:** Protect well against infection caused by virus strains *similar* to the source strain.
- **Failure:** Don't protect against disease caused by strains antigenically *dissimilar* to source strain.

- HIV is a high-impact public health issue but numerous attempts to develop vaccines have met with only mild success.
- The complexity of the disease mechanism makes it quite challenging to study the numerous factors that contribute to a possible mitigation of infection risk.

Sieve Analysis: A Brief History

- The study of whether and how the efficacy of a vaccine varies with the virus' characteristics.
- Why "sieve"? Vaccine as a barrier against select strains, but dissimilar strains break through.
- Identification of sieve effects guides decisions for future development of multivalent vaccines.
- Sieve analysis is usually performed within a *competing risks* framework.

- The reliance on competing risks leads to the use of nonparametric estimators like Aalen-Johnson or semiparametric methods like the Cox model.
- Within this framework, could evaluate instantaneous risks of infection (i.e., hazard) or cumulative incidence. The latter could be more interesting from a public health perspective.

Immune Response and Vaccine Efficacy

- A 12-color intracellular cytokine staining (ICS) assay was performed.
- Cryopreserved peripheral blood mononuclear cells were stimulated with synthetic HIV-1 peptide pools.
- Immune responses of interest were
  1. Total magnitude of the $CD4^+$ T-cell response.
  2. COMPASS Env-specific $CD4^+$ T-cell polyfunctionality score.
  3. Total magnitude of the $CD8^+$ T-cell response.
  4. COMPASS Env-specific $CD8^+$ T-cell polyfunctionality score.
  5. $CD4^+$ and $CD8^+$ T-cell $\log_{10}$-transformed total magnitude variables.

- For a complete description of the immune responses of interests and how these were collected, consult the supplemental materials of HE Janes (2017).
- This class of data is difficult and expensive to collect, which begins to provide motivation for why it might be undesirable to restrict the types of analyses performed to classical semiparametrics.
- Such classical analyses severely restrict the scope of the scientific questions we're able to ask.

Immune Reponse and Vaccine Efficacy

- *Goal:* Evaluate the immune response variables among vaccine recipients as predictors of HIV-1 infection.
- Cox proportional hazard models that account for case-control sampling design and adjust for the baseline covariates.
- $\lambda\left(t; Z = z\right) = \lambda_0(t) \exp\left(\beta^T z\right), \quad t \geq 0.$
  - Semiparametric overall.
  - nonparametric in λ_0, parametric in β.
- Corrections for multiple testing performed, with q-values below 0.20 considered significant.

- Principal components analysis (PCA) was used to discover unique immune response profiles among vaccine recipients.
- First and second principal components were associated with HIV-1 infection using Cox proportional hazards regression models that account for the sampling design and baseline covariates.
- Logistic regression models with lasso penalty and weights to account for case-control sampling were used to identify the baseline covariates and immune response variables that best predict HIV-1 infection.

Motivation: Science Before Statistics

- Cox model: assumption of proportional hazards.
- Such models are a matter of convenience: does $\hat{\beta}$ answer our scientific questions?
  - Perhaps not.
- Is consideration being given to whether the data could have been generated by a process that is consistent with the assumptions of the Cox model?
  - Perhaps not.

Interlude: Causal Inference

1. Motivation: “We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.” –Aristotle
2. Our question of interest concerns the manner in which changes in a given immune response profile affect risk of HIV-1 infection.
   - This is a question of causality.
   - How does *intervening* on immune response profile cause changes in the risk of HIV-1 infection.
3. But how do we go about thinking about intervening on continuous quantities (e.g., immune response profile measures)?
4. Classical causal parameters (e.g., ATE) are not well suited for answering these sorts of questions.

Causal Inference and Vaccine Efficacy

- Consider observing n individuals in a data structure of the form specified above.
- To formalize, consider $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where we make no assumptions on the statistical model containing P_0.
- For the treatment A, we would normally be limited to thinking about counterfactual means (i.e., $\mathbb{E}Y_a$ for $A = a$) or similar quantities.
- This requires specifying a particular value of the treatment (i.e., $A = a$) under which to evaluate the outcome.

Stochastic Treatment Regimes

- Rather than a deterministic intervention, consider a shift of the treatment (i.e., instead of $A = a$, consider $A = a + \delta$).
- This is a far more flexible approach. We need not specify a given value of the treatment but rather a shift (δ) of the treatment.
- In this setting, the effect of the intervention appears as $\mathbb{E}Y_{a+\delta} - \mathbb{E}Y_a$, where $A = a$ is simply the observed value of treatment.
- To compare with the linear model, the shift δ may be thought of as analogous to shifts in the slope of the regression line.

Problems with Stochastic Interventions

- Even though we employ a more flexible type of intervention, the common assumptions (and problems!) of causal inference still arise.
  - Randomization: $Y_{d(a,w)} \perp\!\!\!\perp A \mid W$
  - Positivity: $0 < P(A \mid W) < 1$ everywhere. The propensity score is bounded in $(0, 1)$.
- To protect against positivity violations, a clever shifting mechanism: $d(a, w) = a + \delta$, if $a + \delta < u(w)$ and $d(a, w) = a$ otherwise.
- The shift $d(A, W)$ is now a function of the observed data, and the shift intervention $(a + \delta)$ is only applied when there is support in the observed data.

Parameters for Treatment Shifting

- Let's consider a simple statistical target parameter:

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

- Assume *piecewise smooth invertibility* of $d(a, w)$ in order to obtain a pathwise differentiability of the parameter.
- This makes semiparametric-efficient estimation in the nonparametric model possible when relying on stochastic interventions.
- The parameter now corresponds to our scientific question of interest: How does shifting immune response by an amount δ affect the risk of HIV-1 infection?

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Semiparametric-Efficient Estimation

- Recall that our statistical parameter of interest is

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

- For which the efficient influence function (EIF) is

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

- The auxiliary covariate introduced (i.e., $H(a, w)$) may be expressed

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

The auxiliary covariate simplifies when the treatment is in the limits (conditional on W) — i.e., for $A_i \in (u(w) - \delta, u(w))$, then we have $H(a, w) = \frac{g_0(a-\delta|w)}{g_0(a|w)} + 1$.

Target Minimum Loss-Based Estimation

- TMLEs provide semiparametric-efficient estimation and robust inference in nonparametric models.
- **Asymptotic linearity:**

$$\Psi(P_n^*) - \Psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

- **Limiting distribution:**

$$\sqrt{n}(\Psi_n - \Psi) \to N(0, Var(D(P_0)))$$

- **Statistical inference:**

$$\Psi_n \pm z_\alpha \cdot \frac{\sigma_n}{\sqrt{n}}$$

Under the additional condition that the remainder term $R(\hat{P}^*, P_0)$ decays as $o_P\left(\frac{1}{\sqrt{n}}\right)$, we have that $\Psi_n - \Psi_0 = (P_n - P_0) \cdot D(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right)$, which, by a central limit theorem, establishes a Gaussian limiting distribution for the estimator, with variance $V(D(P_0))$, the variance of the efficient influence function when Ψ admits an asymptotically linear representation.

The above implies that Ψ_n is a $\sqrt{n}$-consistent estimator of Ψ, that it is asymptotically normal (as given above), and that it is locally efficient. This allows us to build Wald-type confidence intervals, where σ_n^2 is an estimator of $V(D(P_0))$. The estimator σ_n^2 may be obtained using the bootstrap or computed directly via $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} D^2(\bar{Q}_n^*, g_n)(O_i)$

Statistical Inference for TMLEs

- Asymptotic distribution of TML estimators has been studied thoroughly:
  $\psi_n - \psi_0 = (P_n - P_0) \cdot D(P_0) + R(\hat{P}^*, P_0)$, giving
  $\psi_n - \psi_0 = (P_n - P_0) \cdot D(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right)$.
- Have a *Gaussian limiting distribution* $\sqrt{n}(\psi_n - \psi) \rightarrow N(0, V(D(P_0)))$ when ψ exhibits asymptotically linearity.
- **Statistical inference** using Wald-type confidence intervals: $\Psi_n \pm z_\alpha \cdot \frac{\sigma_n}{\sqrt{n}}$, where σ_n^2 is an estimator of $V(D(P_0))$.
- Bootstrap for σ_n^2 or compute directly via $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} D^2(\bar{Q}_n^*, g_n)(O_i)$.

1. If $D(\bar{Q}_n^*, g_n)$ converges to $D(P_0)$ in $L_2(P_0)$ norm.
2. The size of the class of functions $\bar{Q}_n^*$ and g_n is bounded (technically, $\exists \mathcal{F}$ st $D(\bar{Q}_n^*, g_n) \in \mathcal{F}$ whp, where $\mathcal{F}$ is a Donsker class)

Complication: Multi-Stage Sampling

- In the 505 HIV-1 trial, all infected individuals are matched to pairs using a complex mechanism.
- Using our observed data structure $O = (W, A, Y)$, let us introduce $V = (W, Y)$, where V is the set of variables used to define the sampling mechanism.
- Thus, the observed data structure is now represented $O = (W, \Delta A, Y)$ wrt to the full data structure.
  - In the above, let $\Delta = f(V)$ be binary st $\Delta \in \{0, 1\}$.
  - Further, let $\Pi_0(V) = P(\Delta = 1 \mid V)$ and $\Pi_n(V)$ be an estimator of $\Pi_0(V)$.
- In this way, our approach accounts for multi-stage sampling (e.g., matched or case-control designs).

Multi-Stage Sampling with TMLEs

- Rose & van der Laan (2011) introduce an IPCW-TMLE to be used when the data structure takes the form $O = (V, \Delta, \Delta X)$, for multi-stage sampling designs.
- How? Use an IPC-weighted loss function:

$$\mathcal{L}(P_X)(O) = \frac{\Delta}{\Pi_n(V)} \mathcal{L}^F(P_X)(X)$$

- The IPCW-TMLE solves the full-data efficient influence function (EIF) equation:

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\Pi_n(V_i)} D^F(P^*_{X,n})(X_i).$$

Note that the IPCW-TMLE is not fully semiparametric-efficient, unless the estimator of $\Pi_n(V)$ is nonparametric, which requires that V be discrete. Of course, violation of these assumptions can be highly problematic.

Efficiency Under Multi-Stage Sampling

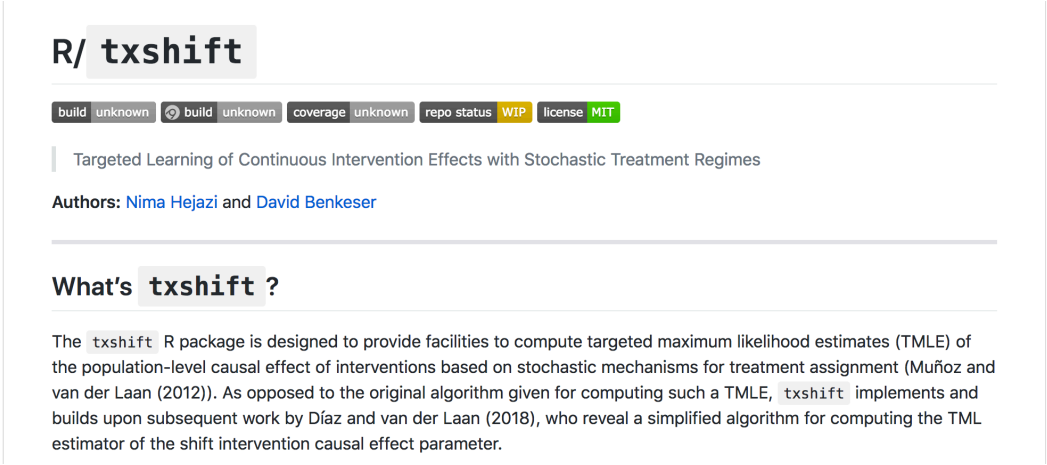
- When working in a nonparametric model, it is necessary to use a nonparametric estimator of the missingness mechanism to obtain full efficiency.
- In many practical settings, this further complicates the efficient influence function estimating equation

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

Putting It Together: Multiple Robustness

- We now have a semiparametric-efficient and robust procedure for assessing the effect of the intervention $d(a, w) = a + \delta$ even in the presence of multi-stage sampling.
- Due to the nature of the IPCW-TMLE, we have a form of multiple double robustness — in terms of combinations of (g, Q) and $(\Pi, \mathbb{E}_0(D^F(P^F) \mid V))$.
- This allows us to assess how simple (additive) shifts of immune response variables affect the risk of HIV-1 infection.

Software package: R/txshift



R/ txshift

build unknown | build unknown | coverage unknown | repo status WIP | license MIT

> Targeted Learning of Continuous Intervention Effects with Stochastic Treatment Regimes

Authors: Nima Hejazi and David Benkeser

What's txshift?

The `txshift` R package is designed to provide facilities to compute targeted maximum likelihood estimates (TMLE) of the population-level causal effect of interventions based on stochastic mechanisms for treatment assignment (Muñoz and van der Laan (2012)). As opposed to the original algorithm given for computing such a TMLE, `txshift` implements and builds upon subsequent work by Díaz and van der Laan (2018), who reveal a simplified algorithm for computing the TML estimator of the shift intervention causal effect parameter.

Figure: https://github.com/nhejazi/txshift

- Variable importance for continuous interventions.
- Take it for a test drive! Coming soon . . .

- Contribute on GitHub: https://github.com/nhejazi/txshift.
- Reach out to us with questions and any feature requests.

Software Ecosystem: The tlverse!

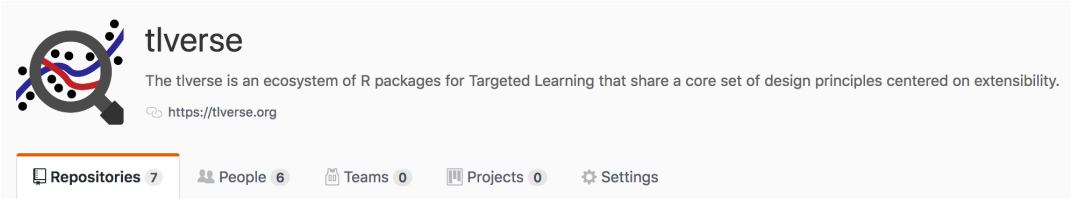


Figure: `https://github.com/tlverse`

- This is a new framework for Targeted Learning with a focus on extensibility.
- "txshift" will be the first of many connector packages — collaboration with Jeremy Coyle and others.

- Contribute on GitHub: `https://github.com/tlverse`.
- Reach out to us with questions and any feature requests.

Future Work

- Exploration of different forms of stochastic treatment shifts — EH Kennedy provides a shift in propensity score space in a recent JASA manuscript currently in press (collaboration in progress).
- Further refinement of the available software, explore how to provide a more efficient and extensible system, including stronger integration with the tlverse.
- Refinements of statistical theory so as to better work with quantities common in survival analysis: hazards? survival?
- Assessment of efficacy trials other than the HVTN 505 HIV-1 vaccine trial — perhaps further scientific findings?

Review: Summary

- The evaluation of vaccine efficacy is a high-impact scientific problem that leads to numerous statistical challenges.
- Stochastic interventions provide a flexible framework through which these statistical problems may be viewed from the perspective of causal inference.
- Standard targeted minimum loss-based estimation may be augmented to handle multi-stage sampling designs, like those common in efficacy trials.
- Statistical software is now readily available for deploying these types of techniques in a number of settings. We apply these methods in efficacy trials.

It's always good to include a summary.

References I

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.

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.

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

References II

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

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.

Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Mark J van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media, 2018.

Acknowledgments

Collaborators:

David C. Benkeser	Emory University
Mark J. van der Laan	University of California, Berkeley
Peter B. Gilbert	Fred Hutchinson Cancer Research Center
Holly E. Janes	Fred Hutchinson Cancer Research Center

Funding source:
UC Berkeley NIH BD2K Biomedical Big Data Training Program: T32-LM012417-02

Thank you.

Slides: goo.gl/LAoDUJ

Notes: goo.gl/Vq6v5o

`stat.berkeley.edu/~nhejazi`

`nimahejazi.org`

`twitter/@nshejazi`

`github/nhejazi`