Efficient estimation of modified treatment policy effects based on the generalized propensity score

Nima Hejazi

Wednesday, 17 November 2021

Division of Biostatistics
Department of Population Health Sciences
Weill Cornell Medicine

@nshejazi
O nhejazi
/reference/nimahejazi.org
with M. van der Laan, I. Díaz, & D. Benkeser
Causal Inference Learning Group, Columbia University

Motivating example

The observed data unit is \( O := (W, A, Y) \sim P_0 \in \mathcal{M} \):

- \( W \in \mathbb{R}^d \) is a vector of baseline covariates;
- \( A \in \mathbb{R} \) is a continuous-valued exposure; and
- \( Y \in \mathbb{R} \) is an outcome of interest.

Let \( \mathcal{M} \) be a large semiparametric model and for each \( P \in \mathcal{M} \), define the population intervention effect (PIE) as

\[
\Psi_\delta(P) := \mathbb{E}_P\{Y(A_\delta) - Y\},
\]

where \( A_\delta \) arises from a stochastic intervention.
**NPSEM with static interventions**

- Use a nonparametric structural equation model (NPSEM) to describe the generation of $O$ (Pearl 2009), specifically

  $$W = f_W(U_W); A = f_A(W, U_A); Y = f_Y(A, W, U_Y)$$

- Implies a model for the distribution of counterfactual random variables generated by interventions on the process.

- A *static intervention* replaces $f_A$ with a specific value $a$ in its conditional support $A \mid W$.

- This requires specifying a particular value of the exposure under which to evaluate the outcome *a priori*.

**NPSEM with stochastic interventions**

- *Stochastic interventions* modify the value $A$ would naturally assume by drawing from a modified exposure distribution.

- Consider the post-intervention value $A_\delta \sim G_\delta(\cdot \mid W)$; static interventions are a special case (degenerate distribution).

- Such an intervention generates a counterfactual RV $Y_{G_\delta} := f_Y(A_\delta, W, U_Y)$, with distribution $P_0^\delta$.

- We aim to estimate $\psi_{0,\delta} := \mathbb{E}_{P_0^\delta} \{ Y_{G_\delta} \}$, the counterfactual mean under the post-intervention exposure distribution $G_\delta$. 
Stochastic interventions for the causal effects of shifts

- Díaz and van der Laan (2012; 2018)’s stochastic interventions

\[ \delta(a, w) = \begin{cases} 
    a + \delta, & a + \delta < u(w) \quad \text{(if plausible)} \\
    a, & a + \delta \geq u(w) \quad \text{(otherwise)} 
\end{cases} \]

- Haneuse and Rotnitzky (2013): modified treatment policies

- Evaluate outcome under modified intervention distribution:

\[ P^\delta(g_{0,A})(A = a \mid W) = g_{0,A}(\delta^{-1}(A, W) \mid W). \]

- Díaz and van der Laan (2018) show that \( \psi_{0,\delta} \) is identified by a functional of the distribution of \( O \):

\[ \psi_{0,\delta} = \int_W \int_A E_{P_0}\{Y \mid A = \delta(a, w), W = w\} \cdot 
\quad g_{0,A}(a \mid W = w) \cdot q_{0,W}(w) \, d\mu(a) \, d\nu(w) \]

Estimation of the PIE

An estimator \( \psi_n \) of \( \psi_0 := \Psi(P_0) \) is efficient if and only if

\[ \psi_n - \psi_0 = n^{-1} \sum_{i=1}^{n} D^*(P_0)(O_i) + o_P(n^{-1/2}), \]

where \( D^*(P) \) is the efficient influence function (EIF) of \( \Psi_\delta \) with respect to the model \( M \) at \( P \).

The EIF of \( \Psi \) is indexed by two key nuisance parameters

\[ \overline{Q}_{P,Y}(A, W) := E_P(Y \mid A, W) \quad \text{outcome mechanism} \]
\[ g_{P,A}(A, W) := p(A \mid W) \quad \text{generalized propensity score} \]
Estimation of a counterfactual mean

We’ll rely on empirical process notation throughout:

- \( P_0 f = \mathbb{E}_{P_0} \{ f(O) \} = \int f(o) dP(o) \)
- \( P_n f = \mathbb{E}_{P_n} \{ f(O) \} = n^{-1} \sum_{i=1}^n f(O_i) \)

We can estimate the counterfactual mean \( \Psi_\delta(P) \), using the inverse probability weighted (IPW) estimator

\[
\psi_{\delta,n} = n^{-1} \sum_{i=1}^n \frac{g_{n,A}(\delta^{-1}(A_i,W_i) \mid W_i)}{g_{n,A}(A_i \mid W_i)} Y_i .
\]

Why IPW estimators?

- IPW estimators are the oldest class of causal effect estimators.
- IPW estimators are still very commonly used in practice today.
- Easy to implement and appropriate in many settings, but...
  1. requires a correctly specified estimate of the propensity score;
  2. can be inefficient, never attaining the efficiency bound; and
  3. suffers from an (asymptotic) curse of dimensionality.
The IPW estimator $\Psi_\delta(P_n, g_{n, A})$ is a solution to the score equation

$$D_\text{IPW}(O; \Psi_\delta) = \frac{(g_{n, A}(\delta^{-1}(A_i, W_i) | W_i))}{g_{n, A}(A_i | W_i)} Y_i - \Psi(P):$$

$$\Psi_\delta(P_n, g_{n, A}) = n^{-1} \sum_{i=1}^{n} \frac{g_{n, A}(\delta^{-1}(A_i, W_i) | W_i)}{g_{n, A}(A_i | W_i)} Y_i.$$

- Consistency and convergence rate of IPW relies on those same properties of the generalized propensity score estimator $g_{n, A}$.
- Generally, finite-dimensional (i.e., parametric) models are not flexible enough to consistently estimate $g_{0, A}$.

Nonparametric conditional density estimation

- Our IPW estimator require the generalized propensity score, at both $g_A(A | W)$ and $g_A(\delta^{-1}(A, W) | W)$.
- There is a rich literature on density estimation, we follow the approach first explored in Díaz and van der Laan (2011).
- To build a conditional density estimator, consider

$$g_{n, A, \alpha}(A | W) = \frac{\mathbb{P}(A \in [\alpha_{t-1}, \alpha_t) | W)}{|\alpha_t - \alpha_{t-1}|}.$$

- This is a classification problem, where we estimate the probability that a value of $A$ falls in a bin $[\alpha_{t-1}, \alpha_t]$.
- The choice of the tuning parameter $t$ corresponds roughly to the choice of bandwidth in classical kernel density estimation.
Nonparametric conditional density estimation

- Díaz and van der Laan (2011) propose a reformulation of this classification approach as a set of hazard regressions.

- To effectively employ this proposed reformulation, consider

  \[ P(A \in [\alpha_{t-1}, \alpha_t] \mid W) = P(A \in [\alpha_{t-1}, \alpha_t] \mid A \geq \alpha_{t-1}, W) \times \prod_{j=1}^{t-1} \{1 - P(A \in [\alpha_{j-1}, \alpha_j] \mid A \geq \alpha_{j-1}, W)\} \]

  - Likelihood may be re-expressed as the likelihood of a binary variable in a repeated measures data structure.
  - Specifically, the observation of \( O_i \) is repeated as many times as intervals \([\alpha_{t-1}, \alpha_t]\) are prior to the interval to which \( A_i \) falls, and the indicator variables \( A_i \in [\alpha_{t-1}, \alpha_t) \) are recorded.

Curse of dimensionality

Goal: Construct nuisance parameter estimators that are consistent and converge faster than \( n^{-1/4} \) under minimal assumptions.

Challenging for moderately large \( d \), i.e., curse of dimensionality.

For example, consider kernel regression with bandwidth \( h \) and kernels orthogonal to polynomials in \( W \) of degree \( k \).

- Assume parameter is \( k \) times differentiable.
- Optimal bandwidth \( O(n^{-1/(2k+d)}) \)
- Optimal convergence rate \( O(n^{-k/(2k+d)}) \)
Curse of dimensionality

Broadly, two approaches for handling the curse of dimensionality.

1. Enforce fairly strong smoothness assumptions on the model space (e.g., Hirano et al. 2003).
   - No general guarantee of consistency

2. Ensemble machine learning, e.g., Super Learning (van der Laan et al. 2007).
   - No guarantee of $n^{-1/4}$ convergence rates

An important class of functions

Consider space of cadlag functions with finite variation norm.

Def. cadlag = left-hand continuous with right-hand limits

Variation norm Let $\theta_s(u) = \theta(u_s, 0_{s^c})$ be the section of $\theta$ that sets the coordinates in $s$ equal to zero.

The variation norm of $\theta$ can be written:

$$|\theta|_v = \sum_{s \subseteq \{1, \ldots, d\}} \int |d\theta_s(u_s)|,$$

where $x_s = (x(j) : j \in s)$ and the sum is over all subsets.
Variation norm

We can represent the function $\theta$ as

$$\theta(x) = \theta(0) + \sum_{s \subset \{1, \ldots, d\}} \int \mathbb{I}(x_s \geq u_s) d\theta_s(u_s),$$

For discrete measures $d\theta_s$ with support points $\{u_{s,j} : j\}$ we get a linear combination of indicator basis functions:

$$\theta(x) = \theta(0) + \sum_{s \subset \{1, \ldots, d\}} \sum_j \beta_{s,j} \theta_{u_{s,j}}(x),$$

where $\beta_{s,j} = d\theta_s(u_{s,j})$, $\theta_{u_{s,j}}(x) = \mathbb{I}(x_s \geq u_{s,j})$, and

$$|\theta|_v = \theta(0) + \sum_{s \subset \{1, \ldots, d\}} \sum_j |\beta_{s,j}|.$$
Convergence rate of HAL

We have, for \( \alpha(d) = \frac{1}{d+1} \),

\[
|\theta_{n,M} - \theta_{0,M}|_{P_0} = o_P(n^{-\left(\frac{1}{4} + \alpha(d)/8\right)}).
\]

Thus, if we select \( M > |\theta_0|_v \), then

\[
|\theta_{n,M} - \theta_0|_{P_0} = o_P(n^{-\left(\frac{1}{4} + \alpha(d)/8\right)}).
\]

Due to oracle inequality for the cross-validation selector \( M_n \),

\[
|\theta_{n,M_n} - \theta_0|_{P_0} = o_P(n^{-\left(\frac{1}{4} + \alpha(d)/8\right)}).
\]

Improved convergence rate (Bibaut and van der Laan 2019):

\[
|\theta_{n,M_n} - \theta_0|_{P_0} = o_P(n^{-\frac{1}{3} \log(n)^{d/2}}).
\]

HAL estimate of \( g_{0,A} \)

If the nuisance functional \( g_{0,A} \) is cadlag with finite sectional variation norm, \( \logit g \) can be expressed (Gill et al. 1995):

\[
\logit g_\beta = \beta_0 + \sum_{s \subset \{1, \ldots, d\}} \sum_{i=1}^n \beta_{s,i} \phi_{s,i},
\]

where \( \phi_{s,i} \) is an indicator basis function.

The loss-based HAL estimator \( \beta_n \) may then be defined as

\[
\beta_{n,\lambda} = \arg \min_{\beta:|\beta_0| + \sum_{s \subset \{1, \ldots, d\}} \sum_{i=1}^n |\beta_{s,i}| < \lambda} P_n \mathcal{L}(\logit g_\beta),
\]

where \( \mathcal{L}(\cdot) \) is an appropriate loss function.

Denote by \( g_{n,\lambda} \equiv g_{\beta_{n,\lambda}} \) the HAL estimate of \( g_{0,A} \).
Targeted selection of $\lambda_n$ for IPW estimation

1. CV-based: choose $\lambda_n$ as cross-validated empirical minimizer of negative log-density loss (Dudoit and van der Laan 2005):

$$L(\cdot) = -\log(g_{n,A,\lambda}(A | W)).$$

n.b., “targeted” but incorrect tradeoff ($g_{n,A,\lambda}$ instead of $\psi_{n,\delta}$).

2. EIF-based: choose $\lambda_n$ to solve the EIF estimating equation:

$$\lambda_n = \arg \min_{\lambda} |P_n D_{\text{CAR}}(g_{n,A,\lambda}, \bar{Q}_n,Y)|,$$

where $\bar{Q}_n,Y$ is an estimate of $\bar{Q}_0,Y$ and $D^* = D_{\text{IPW}} - D_{\text{CAR}}$.

Agnostic selection of $\lambda_n$ for IPW estimation

What if we dispensed with criteria based on $\psi_{n,\delta}$ altogether?

1. Plateau-based: choose $\lambda_n$ as the first in $\lambda_1, \ldots, \lambda_K$ s.t.

$$|\psi_{n,\lambda_{j+1}} - \psi_{n,\lambda_j}|^{K-1} \leq Z(1-\alpha/2)[\sigma_{n,\lambda_{j+1}} - \sigma_{n,\lambda_j}]^{K-1},$$

where $\sigma_{n,\lambda_j}$ is a variance estimate at $\lambda_j$.

2. Plateau-based: choose $\lambda_n$ as the first in $\lambda_1, \ldots, \lambda_K$ s.t.

$$\left[ \frac{|\psi_{n,\lambda_{j+1}} - \psi_{n,\lambda_j}|}{\max_j |\psi_{n,\lambda_{j+1}} - \psi_{n,\lambda_j}|} \right]^{K-1} \leq \tau$$

for an arbitrary tolerance level $\tau$. 

The big picture

1. Unlike classical IPW estimators, ours avoid the asymptotic curse of dimensionality and are asymptotically efficient;

2. Our approach leverages flexible conditional density estimation for initial generalized propensity score estimates; and

3. In contrast with doubly robust estimators, our estimators can be formulated without the form of the EIF.

4. Check out the R packages that make this possible
   - hal9001: https://github.com/tlverse/hal9001
   - haldensify: https://github.com/nhejazi/haldensify

Thank you!

https://nimahejazi.org
https://twitter.com/nshejazi
https://github.com/nhejazi
Manuscript coming soon — stay tuned!
From the causal to the statistical target parameter

**Assumption 1: Stable Unit Treatment Value (SUTVA)**
- \( Y_i^\delta(a_i, w_i) \) does not depend on \( \delta(a_j, w_j) \) for \( i = 1, \ldots, n \) and \( j \neq i \), or lack of interference (Rubin 1978; 1980)
- \( Y_i^\delta(a_i, w_i) = Y_i \) in the event \( A_i = \delta(a_i, w_i) \), \( i = 1, \ldots, n \)

**Assumption 2: Ignorability**
\( A_i \perp Y_i^\delta(a_i, w_i) \mid W_i, \text{ for } i = 1, \ldots, n \)

**Assumption 3: Positivity**
\( a_i \in \mathcal{A} \implies \delta(a_i, w_i) \in \mathcal{A} \text{ for all } w \in \mathcal{W}, \text{ where } \mathcal{A} \text{ denotes the support of } A \text{ conditional on } W = w_i \text{ for all } i = 1, \ldots, n \)
\[ \psi(0) = 0 \]

\[ \psi(1) = \psi(0) + \beta_1 \]

\[ \beta_1 = 2 \]
\[\psi(2) = \psi(0) + \beta_1 + \beta_2\]

\[\psi(3) = \psi(0) + \beta_1 + \beta_2 + \beta_3\]
\( \psi(4) = \psi(0) + \beta_1 + \beta_2 + \beta_3 + \beta_4 \)

\( \beta_4 = -1 \)

\( \psi(X) = \psi(0) + \sum_{j=1}^{4} \beta_j I(X > j) \)

\( \| \psi(X) \|_v = \sum_{j=1}^{4} |\beta_j| \)
**Literature: Haneuse and Rotnitzky (2013)**

- **Proposal:** Characterization of stochastic interventions as *modified treatment policies* (MTPs).

- **Assumption of piecewise smooth invertibility** allows for the intervention distribution of any MTP to be recovered:
  
  \[
  g_{0,\delta}(a \mid w) = \sum_{j=1}^{J(w)} I_{\delta_j} \{ h_j(a, w) \mid \delta \} g_0 \{ h_j(a, w) \mid \delta \} h_j'(a, w)
  \]

- Such intervention policies account for the natural value of the intervention \( A \) directly yet are interpretable as the imposition of an altered intervention mechanism.

- Identification conditions for assessing the parameter of interest under such interventions appear technically complex (at first).

---

**Literature: Young et al. (2014)**

- Establishes equivalence between g-formula when proposed intervention depends on natural value and when it does not.

- This equivalence leads to a sufficient positivity condition for estimating the counterfactual mean under MTPs via the same statistical functional studied in Díaz and van der Laan (2012).

- Extends earlier identification results, providing a way to use the same statistical functional to assess \( \mathbb{E} Y_{\delta(A, W)} \) or \( \mathbb{E} Y_{\delta(W)} \).

- The authors also consider limits on implementing shifts \( \delta(A, W) \), and address working in a longitudinal setting.
Literature: Díaz and van der Laan (2018)

- Builds on the original proposal, accommodating MTP-type shifts $\delta(A, W)$ proposed after their earlier work.

- To protect against positivity violations, considers a specific shifting mechanism:

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

- Proposes an improved TMLE algorithm, with a single auxiliary covariate for constructing the TML estimator.

References


