Nonparametric inverse probability weighted estimators based on the highly adaptive lasso

Nima Hejazi

Wednesday, 17 March 2021

Division of Biostatistics, and
Center for Computational Biology,
University of California, Berkeley

Twitter: nshejazi
Instagram: nhejazi
Website: nimahejazi.org

with A. Ertefaie & M. van der Laan
Causal inference seminar, UC Berkeley
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 covariates;
- \( A \in \{0, 1\} \) is a binary treatment; 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 average treatment effect (ATE) as

\[
\Psi(P) := \mathbb{E}_P\{\mathbb{E}_P(Y \mid A = 1, W) - \mathbb{E}_P(Y \mid A = 0, W)\}.
\]
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 $\Psi$ with respect to the model $\mathcal{M}$ at $P$.

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

$$\overline{Q}_P(A, W) := \mathbb{E}_P(Y \mid A, W) \quad \text{outcome mechanism}$$

$$g_P(W) := \mathbb{E}_P(A \mid W) \quad \text{propensity score}$$
Estimation of a counterfactual mean

We’ll rely on *empirical process notation* throughout:

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

Consider estimating the *counterfactual mean in the treatment arm*:

$$\Psi(P) = \mathbb{E}_P \{ \mathbb{E}_P(Y \mid A = 1, W) \},$$

using the inverse probability weighted (IPW) estimator

$$\psi_n = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{g_n(1 \mid W_i)}.$$
IPW estimators

- IPW estimators are the oldest class of causal effect estimators, and they are still very commonly used in practice today.
- IPW is easy to implement and appropriate across a variety of settings, but IPW estimators have several disadvantages:
  1. require a correctly specified estimate of the propensity score;
  2. can be inefficient, never attaining the efficiency bound; and
  3. suffer from an (asymptotic) curse of dimensionality.
An IPW estimator $\Psi(P_n, g_n)$ is a solution to the score equation $P_n U_{g_n}(\Psi) = 0$, where $U_{g}(O; \Psi) = \frac{AY}{g(1|W)} - \Psi(P)$. That is,

$$\Psi(P_n, g_n) = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i}{g_n(1 | W_i)}.$$ 

- Consistency and convergence rate of IPW relies on those same properties of the propensity score estimator $g_n$.
- Generally, finite-dimensional (i.e., parametric) models are not flexible enough to consistently estimate $g_0$. 

Data-adaptive estimators

Data-adaptive regression can improve consistency of \( g_n \) for \( g_0 \) but establishing asymptotic linearity is challenging:

\[
\Psi(P_n, g_n) - \Psi(P_0, g_0) = P_n U_{g_n}(\Psi) - P_0 U_{g_0}(\Psi) = (P_n - P_0) U_{g_0}(\Psi) + P_0 \{ U_{g_n}(\Psi) - U_{g_0}(\Psi) \} + (P_n - P_0) \{ U_{g_n}(\Psi) - U_{g_0}(\Psi) \}.
\]

- Using only standard empirical process theory and the assumption of consistency, the blue term is \( o_p(n^{-1/2}) \).
- Asymptotic linearity of our IPW estimator relies on the asymptotic linearity of the red term.
Goal: Construct nuisance parameter estimators that are consistent and converge faster than $n^{-1/4}$ under minimal assumptions.

Challenging for moderately large $d$ due to the 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 $O(n^{-k/(2k+d)})$
Broadly, *two approaches* for handling the *curse of dimensionality*.

[1] Enforce (strong) *smoothness assumptions* on model space.
   - No guarantee of *consistency*

[2] Ensemble machine learning, e.g., *Super Learning*
   - No guarantee of *quarter rates*
An important class of functions

Consider space of \textit{cadlag} functions with \textit{finite variation norm}.

**Def.** cadlag = \textit{left-hand continuous with right-hand limits}

**Variation norm** Let $\theta_s(u) = \theta(u_s, 0_{sc})$ be the \textit{section} of $\theta$ that sets the coordinates in $s$ \textit{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

\[
\theta(x) = \theta(0) + \sum_{s \subset \{1, \ldots, d\}} \int l(x_s \geq u_s) d\theta_s(u_s),
\]

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 \( \beta_{s,j} = d\theta_s(u_{s,j}) \), \( \theta_{u_{s,j}}(x) = l(x_s \geq u_{s,j}) \), and

\[
|\theta|_V = \theta(0) + \sum_{s \subset \{1, \ldots, d\}} \sum_j |\beta_{s,j}|.
\]
Illustration

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

\[ \beta_1 = 2 \]
Illustration

\[ \psi(X) \]

0 1 2 3 4 5

\( \beta_2 = 1 \)

\[ \psi(2) = \psi(0) + \beta_1 + \beta_2 \]
\[ \psi(3) = \psi(0) + \beta_1 + \beta_2 + \beta_3 \]

\[ \beta_3 = 2 \]
\[ \psi(4) = \psi(0) + \beta_1 + \beta_2 + \beta_3 + \beta_4 \]

\[ \beta_4 = -1 \]
Illustration

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

\[ ||\psi(X)||_v = \sum_{j=1}^{4} |\beta_j| \]

- \( \beta_1 = 2 \)
- \( \beta_2 = 1 \)
- \( \beta_3 = 2 \)
- \( \beta_4 = -1 \)
Convergence rate of HAL

We have

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

where $\alpha(d) = 1/(d+1)$.

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

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

Due to oracle inequality for the cross-validation selector $M_n$,

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

Improved rate (Bibaut and van der Laan 2019):

$$|\theta_{n,M_n} - \theta_0|_{P_0} = o_P\left(n^{-1/3} \log(n)^{d/2}\right).$$
HAL estimate of $g_0$

Under the assumption that our nuisance functional parameter $g$ is a cadlag function with finite sectional variation norm, logit $g$ may be approximated as (Gill et al. 1995):

$$\text{logit } g_\beta = \beta_0 + \sum_{s \subseteq \{1, \ldots, d\}} \sum_{i=1}^n \beta_{s,i} \phi_{s,i},$$

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

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

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

where $L(\cdot)$ is an appropriate loss function.

Denote by $g_{n,\lambda} \equiv g_{\beta_{n,\lambda}}$ the HAL estimate of $g_0$. 
The efficient influence function expansion is of the form

$$\psi(P_n, g_n) - \psi(P_0, g_0) = P_n\{U_{g_0}(\psi) - D_{\text{CAR}}(P_0)\} + o_p(n^{-1/2}).$$

In particular, the EIF may be expressed

$$D^*(P_0) := U_{g_0}(\psi) - D_{\text{CAR}}(P_0)$$

$$= \left[ \frac{AY}{g(1 \mid W)} - \psi(P, g) \right] - \left[ \frac{Q(1, W)}{g(1 \mid W)} \right] \{A - g(1 \mid W)\}.$$

The term $D_{\text{CAR}}(g_n, Q_0)$ is key to both efficiency and asymptotic linearity. When the HAL estimator $g_n$ is properly \textit{undersmoothed}

$$P_n D_{\text{CAR}}(g_n, Q_0) = o_p(n^{-1/2}).$$
The score function of the HAL fit is

\[ S_h(g) = \Phi(A - g_n, \lambda_n) \]

where \( \Phi \) is a vector consisting of indicator basis functions \( \phi_s \). As we undersmooth, the dimension of \( \Phi \) increases, and thus, we start solving more and more equations.

Recall, \( D_{CAR} = f(W)(A - g_n, \lambda_n) \) where \( f(W) = Q(1, W)/g(A | W) \). The \( f \) function can be approximated with \( \sum_j \alpha_i \phi_j \).

If we undersmooth enough then we would also solve

\[ P_n D_{CAR}(g_n, Q_n) = o_P(n^{-1/2}). \]
We propose two criteria.

1. \( D_{\text{CAR}} \) based:

\[
\lambda_n = \arg \min_{\lambda} \left| P_n D_{\text{CAR}}(g_n, \lambda, Q_n) \right|
\]

where \( Q_n \) is a HAL estimate of \( Q_0(1, W) \).

2. Score based:

\[
\lambda_n = \arg \min_{\lambda} \left[ \sum_{(s,j) \in \mathcal{J}_n} \frac{1}{\| \beta_{n,\lambda} \|_{L_1}} \left| P_n \tilde{S}_{s,j}(\phi, g_n, \lambda) \right| \right]
\]

in which \( \tilde{S}_{s,j}(\phi, g_n, \lambda) = \phi_{s,j}(W)\{A - g_n, \lambda(1 \mid W)\}\{g_n, \lambda(1 \mid W)\}^{-1} \).
Figure 1: Circle: parametric; Triangle: NP with cross-validated $\lambda$ selector; “▽”: $D_{\text{CAR}}$-based $\lambda$ selector; “⋄”: score-based $\lambda$ selector.
1. Unlike standard IPW estimators, our estimators avoid the asymptotic curse of dimensionality, and are asymptotically efficient;

2. in contrast to targeted IPW estimators, our estimators do not suffer from irregularity issues; and

3. in contrast with typical doubly robust estimators, our estimators rely on a single nuisance parameter and may be formulated without the form of the EIF.
Thank you!

https://nimahejazi.org

https://twitter.com/nshejazi

https://github.com/nhejazi

Appendix
Let $Q_0(1) = \mathbb{E}(Y \mid A = 1, \mathcal{W})$. Then,

$$
P_0 \left\{ U_{G_n}(\Psi) - U_{G_0}(\Psi) \right\}
= P_0 \left\{ G_0 Q_0(1) \left( \frac{G_0 - G_n}{G_n G_0} \right) \right\}
= P_0 \left\{ Q_0(1) \left( \frac{G_0 - G_n}{G_0} \right) \right\} + P_0 \left\{ \frac{Q_0(1)}{G_n} \left( G_0 - G_n \right)^2 \right\}
= P_0 \left\{ Q_0(1) \left( \frac{G_0 - G_n}{G_0} \right) \right\} + o_p(n^{-1/2})
= -(P_n - P_0) \{ D_{\text{CAR}}(P_0) \} - P_n \{ D_{\text{CAR}}(Q_0, G_0, G_n) \} + o_p(n^{-1/2}),
$$

where $D_{\text{CAR}}(Q_0, G_0, G_n) = Q_0(1) \left( \frac{A - G_n}{G_0} \right)$ and $D_{\text{CAR}}(P_0) = Q_0(1) \left( \frac{A - G_0}{G_0} \right)$. 
References


