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

Nima Hejazi

Monday, 30th September 2024

Department of Biostatistics,
T.H. Chan School of Public Health,
Harvard University

nshejazi
nhejazi
nimahejazi.org

Statistics Seminar, University of Edinburgh

joint work with M.J. van der Laan

General setup

Consider as *observed data* unit $O = (L, A, Y) \sim P_0 \in \mathcal{M}$:

- $L \in \mathbb{R}^d$ is a vector of putative confounders;
- $A \in \mathbb{R}$ is a continuous or ordinal exposure; and
- $Y \in \mathbb{R}$ is an outcome of interest.
- L, A, Y follow a fixed temporal ordering for $O_1, \ldots, O_n$.

Let $\mathcal{M}$ be a *non-parametric* or infinite-dimensional model. For any $\mathsf{P} \in \mathcal{M}$, define the *population intervention effect* (PIE):

$$\psi_\delta \equiv \Psi_\delta(\mathsf{P}) \coloneqq \mathbb{E}_{\mathsf{P}}[Y^{A_\delta} - Y] \ ,$$

where A_δ is set by a *modified treatment policy* (MTP).

NPSEM-IE with static interventions

- Use a non-parametric structural equation model (NPSEM-IE) to describe the generating process of O (Pearl 2009), like so

$$\begin{aligned} L &= f_L(U_L) \\ A &= f_A(L, U_A) \\ Y &= f_Y(A, L, U_Y) \ , \end{aligned}$$

  where $U_L \perp\!\!\!\perp U_A \perp\!\!\!\perp U_Y$, that is, "independent errors" (IE).

- *Counterfactual* RVs arise from interventions on the NPSEM.

- A *static* intervention replaces f_A with an a in support $\mathcal{A}$.
  - What specific value $a \in \mathcal{A}$ is of interest *a priori*?
  - Consider all $a \in \mathcal{A}$ for the *causal* dose-response curve, but challenging to identify, estimate (e.g., Kennedy et al. 2017).

Causal effects of stochastic interventions

- A *stochastic* intervention alters A to A_δ by drawing randomly from a *modified* exposure distribution $G_\delta(\cdot \mid L)$.
  - Leads to counterfactual RV as $Y^{A_\delta} \leftarrow f_Y(A_\delta, L, U_Y)$, with the counterfactual $Y^{A_\delta} \sim \mathsf{P}_{0,\delta}$.
  - Static interventions are only a special case of this, in which G_δ is a degenerate distribution that places all mass on $a \in \mathcal{A}$.
- Goal: Estimate counterfactual mean under some modified exposure distribution G_δ, that is, $\psi_{0,\delta} := \mathbb{E}_{\mathsf{P}_{0,\delta}}[Y^{A_\delta}]$.
- Díaz and van der Laan (2012)'s *intervention distribution*: $G_\delta := \mathsf{P}_\delta(g_{0,A})(A = a \mid L) \equiv g_{0,A}(d^{-1}(A, L; \delta) \mid L)$.

Causal effects of modified treatment policies

- Haneuse and Rotnitzky (2013) introduced *modified treatment policies* (MTPs), adopt the intervention schemes like

$$d(a, l; \delta) = \begin{cases} a + \delta, & a + \delta < u(l) \quad \text{(if plausible)} \\ a, & a + \delta \geq u(l) \quad \text{(otherwise)} \end{cases},$$

  and were also studied and extended by Young et al. (2014), Díaz and van der Laan (2018), Díaz et al. (2021), etc.

- $\psi_{0,\delta}$ is identified by a functional of the distribution of O by

$$\psi_{0,\delta} = \Psi(\mathsf{P}_{0,\delta}) := \int_{\mathcal{L}} \int_{\mathcal{A}} \mathbb{E}_{\mathsf{P}_0}\{Y \mid A = d(a, l; \delta), L = l\} \\ g_{0,A}(a \mid L = l) q_{0,L}(l) d\mu(a) d\nu(l) \ .$$

Towards a causal interpretation of the PIE $\psi_{0,\delta}$

Assumption 1: ***Stable Unit Treatment Value (SUTVA)***

- $Y_i^{d(a_i,l_i;\delta)}$ *does not depend on* $d(a_j,l_j;\delta)$ *for* $i=1,\ldots,n$ *and* $j\neq i$*, or lack of interference (Cox 1958).*
- $Y_i^{d(a_i,l_i;\delta)}=Y_i$ *in the event* $A_i=d(a_i,l_i;\delta)$*,* $i=1,\ldots,n$*.*

Assumption 2: ***No unmeasured confounding***

$Y_i^{d(a_i,l_i;\delta)}\perp\!\!\!\perp A_i\mid L_i$*, for* $i=1,\ldots,n$*.*

Assumption 3: ***Structural positivity***

$a_i\in\mathcal{A}\implies d(a_i,l_i;\delta)\in\mathcal{A}$ *for all* $l\in\mathcal{L}$*, where* $\mathcal{A}$ *denotes the support of* A *conditional on* $L=l_i$ *for all* $i=1,\ldots n$*.*

Estimation of the PIE $\psi_{0,\delta}$

A RAL *estimator* $\psi_{n,\delta}$ of $\psi_{0,\delta} := \Psi_\delta(\mathrm{P}_0)$ is *efficient* if and only if

$$\psi_{n,\delta} - \psi_{0,\delta} = \frac{1}{n}\sum_{i=1}^{n} D^\star(\mathrm{P}_0)(O_i) + o_{\mathrm{P}}(n^{-1/2}) \ ,$$

where $D^\star(\mathrm{P})$ is the *efficient influence function* (EIF) of $\Psi_\delta(\cdot)$ with respect to the non-parametric model $\mathcal{M}$ at P.

The EIF of $\Psi_\delta(\cdot)$ is indexed by two common nuisance parameters

$$\overline{Q}_{\mathrm{P},Y}(A, L) := \mathbb{E}_{\mathrm{P}}(Y \mid A, L) \qquad \textit{outcome regression}$$

$$g_{\mathrm{P},A}(A, L) := f_{\mathrm{P}}(A \mid L) \qquad \textit{generalized propensity score}$$

IPW Estimation of the PIE $\psi_{0,\delta}$

We can estimate the *counterfactual mean* $\psi_{0,\delta}$, using the inverse probability weighted (IPW) estimator,

$$\psi_{n,\delta}^{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{g_{n,A}(d^{-1}(A_i, L_i; \delta) \mid L_i)}{g_{n,A}(A_i \mid L_i)} \right) Y_i \ ,$$

or could use stabilized IPW instead...

Why bother? Isn't simplicity dead? (Cf., machine learning, AI)

- IPW estimators are the oldest class of causal effect estimators and are still *very commonly used in practice* today.
- Easy to implement and appropriate in many settings, but...
    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.

IPW estimators

The IPW estimator $\psi_{n,\delta}^{\text{IPW}} \equiv \Psi_\delta(\mathsf{P}, g_{n,A})$ is a Z-estimator solving the score equation $\mathbb{E}_{\mathsf{P}} D_{\text{IPW}}(\cdot) = 0$, where D_{IPW} is defined as

$$D_{\text{IPW}}(O; \psi_\delta) := \left(\frac{g_{n,A}(d^{-1}(A_i, L_i; \delta) \mid L_i)}{g_{n,A}(A_i \mid L_i)} \right) Y - \Psi(\mathsf{P}) \ .$$

Some hiccups along the way:

- 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 may not always be flexible enough to consistently estimate $g_{0,A}$.
- Our IPW estimator requires the generalized propensity score (GPS), so we need to estimate a conditional density (hard).

The GPS and conditional density estimation

- Conditional density estimation is a fundamental problem in statistics—with few practical innovations (Ghosh et al. 2024).
- Díaz and van der Laan (2011) proposed

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

  for $[\alpha_{t-1}, \alpha_t)$, $t = 0, \ldots, k$ a discrete partitioning of $\mathcal{A}$.
  - Classification problem: probability of A falling in a given bin $[\alpha_{t-1}, \alpha_t)$ is estimated, then re-mapped to density scale.
  - Classification approach as a series of hazard regressions:

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

  - Others: Rytgaard et al. (2022), Munch et al. (2024).

Curse of dimensionality

Broadly, *two approaches* for handling the *curse of dimensionality*.

1. Enlist *smoothness or sparsity assumptions* on the nuisance parameter space (i.e., for the GPS $g_{0,A}(A \mid L)$ in our case).
   - No *general* guarantee of achieving *consistency*.
   - Rates tied to specific learning algorithms and function classes, e.g., functions being Hölder-smooth, s-sparse.
   - Hirano et al. (2003) took a series regression approach but required 7-times differentiability of $g_{0,A}(A \mid L)$ wrt d.
2. Cross-validation with machine learning or ensemble machine learning (e.g., van der Laan et al. (2007)'s *super learner*), but no *general* guarantee of $n^{-1/4}$ *convergence rates*.

A class of functions

Consider space of *càdlàg* functions with *finite variation norm*.

Def. càdlàg: *left-hand continuous* with *right-hand limits* (RCLL)

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

The *variation norm* of θ can be written:

$$|\theta|_v = \sum_{s \subset \{1,\dots,d\}} \int \mid d\theta_s(u_s) \mid ,$$

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

Variation norm

We can represent the function θ as

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

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

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

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

$$|\theta|_v = \theta(0) + \sum_{s \subset \{1,\dots,d\}} \sum_j |\beta_{s,j}| \ .$$

HAL estimator of GPS $g_{0,A}$

If the nuisance functional $g_{0,A}$ is cadlag with a finite sectional variation norm, $\operatorname{logit} g$ can be expressed (van der Laan 2015):

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

where ϕ_s are indicator basis functions.

The loss-based HAL estimator β_n may then be defined as

$$\beta_{n,\lambda} = \underset{\beta : |\beta_0| + \sum_{s \subset \{1,\ldots,d\}} \sum_{i=1}^{n} |\beta_{s,i}| < \lambda}{\arg\min} \mathbb{E}_{\mathsf{P}_n} \mathcal{L}(\operatorname{logit} g_\beta),$$

where $\mathcal{L}(\cdot)$ is an appropriately selected loss function.

Denote by $g_{n,\lambda} \equiv g_{\beta_{n,\lambda}}$ the HAL estimator of the GPS $g_{0,A}$.

Targeted selection of λ_n for IPW estimation

1. CV: select λ_n as cross-validated empirical risk minimizer of negative log-density loss (Dudoit and van der Laan 2005):

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

   n.b., incorrect tradeoff optimizing $g_{n,A,\lambda}$ instead of $\psi_{n,\delta}$.

2. EIF[1]: select λ_n to minimize mean of EIF estimating equation:

   $$\lambda_n = \arg\min_{\lambda} |\mathbb{E}_{\mathsf{P}_n} D_{\mathsf{CAR}}(g_{n,A,\lambda}, \overline{Q}_{n,Y})|,$$

   where $\overline{Q}_{n,Y}$ is an estimate of $\overline{Q}_{0,Y}$ and $D^{\star} = D_{\mathsf{IPW}} - D_{\mathsf{CAR}}$ by AIPW representation (Robins and Rotnitzky 1992; 1995).

[1] Used for nonparametric IPW (when $A \in \{0, 1\}$) by Ertefaie et al. (2022).

Agnostic selection of λ_n for IPW estimation

What if we dispensed with criteria based on the EIF altogether?

1. Plateau-based[2]: Choose λ_n as the first in $\lambda_1, \ldots, \lambda_K$ s.t.

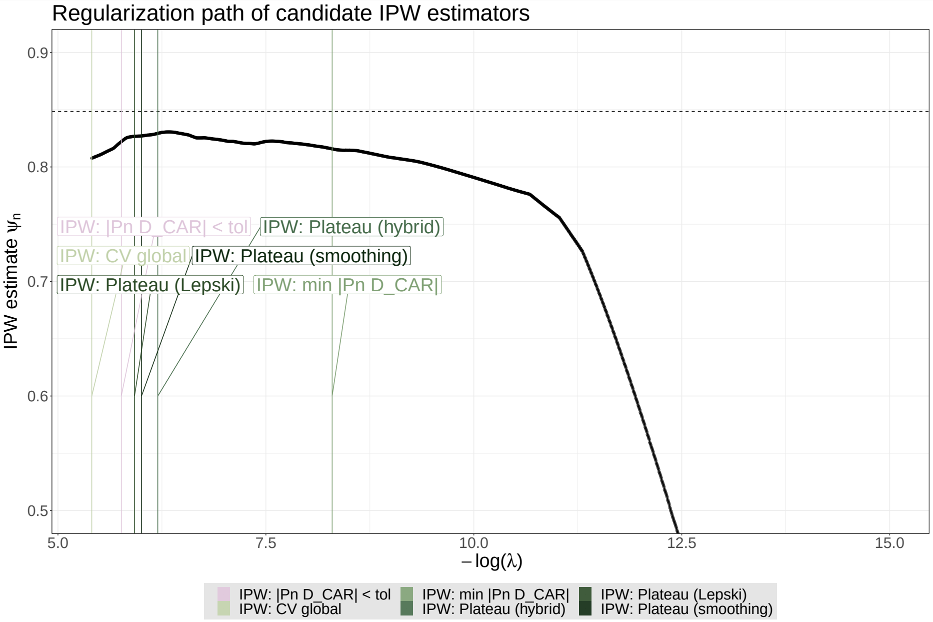
$$|\psi_{n,\delta,\lambda_{j+1}} - \psi_{n,\delta,\lambda_j}| \leq \frac{Z_{(1-\alpha/2)}}{\log n} |\sigma_{n,\lambda_{j+1}} - \sigma_{n,\lambda_j}| \quad \text{for } j = \{1, \ldots, K-1\}$$

   where σ_{n,λ_j} is a standard error estimate for $\psi_{n,\delta,\lambda_j}$ at λ_j.

2. Smoothing-based: Choose λ_n by trimming $\lambda_1, \ldots, \lambda_K$ to run from λ_{CV} to a multiple $C\lambda_{\text{CV}}$, and then finding an inflection point in the estimator's trajectory $\{\psi_{n,\lambda_{\text{CV}}}, \ldots, \psi_{n,C\lambda_{\text{CV}}}\}$.

[2] Inspired by ideas proposed by Lepskii (1993), Lepskii and Spokoiny (1997).

Proof by picture: Smoothing-based selection



Some other efficient estimators

- The one-step bias-corrected estimator:

$$\psi_n^+ = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_{n,Y}(d(A_i, L_i), L_i) + D_n^\star(O_i) \ .$$

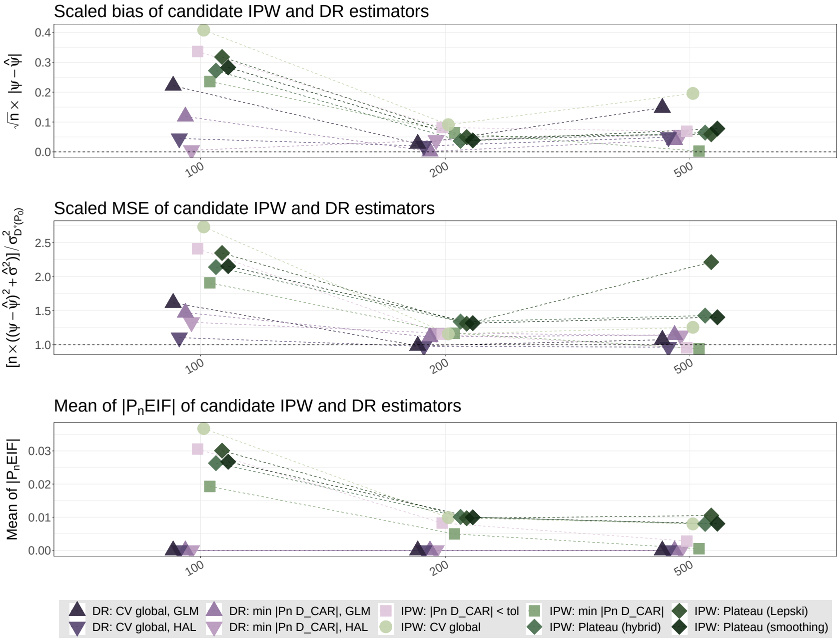
- A TML estimator updates initial estimates of $\overline{Q}_{n,Y}$ by tilting:

$$\psi_n^\star = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_{n,Y}^\star(d(A_i, L_i), L_i) \ ,$$

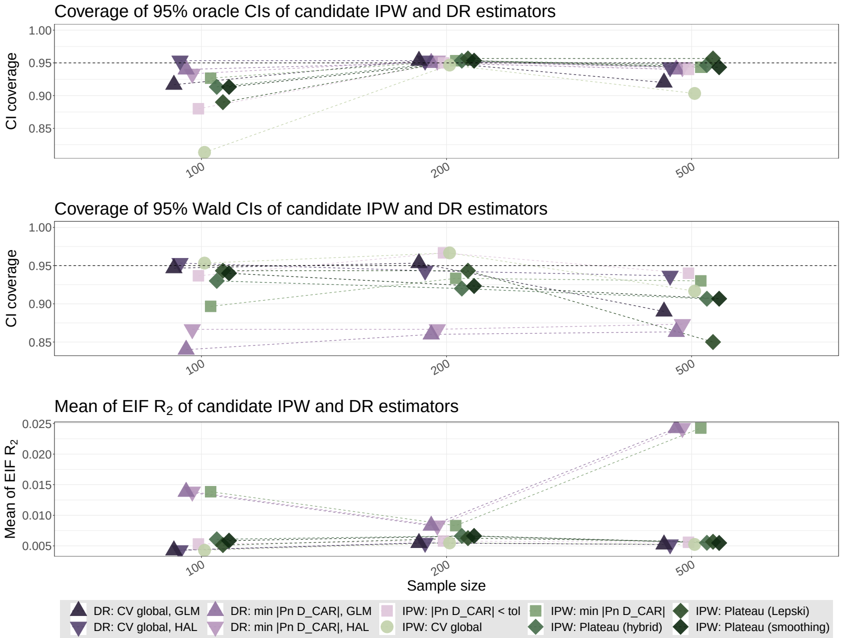
  where $\overline{Q}_{n,Y}^\star$ is arrived at by updating its initial counterpart: $\text{logit}(\overline{Q}_{n,Y}^\star(A_i, L_i)) = \text{logit}(\overline{Q}_{n,Y}(A_i, L_i)) + \epsilon \frac{g_{n,A}(d^{-1}(A_i, L_i; \delta) | L_i)}{g_{n,A}(A_i | L_i)}$.

- Both are doubly robust (DR), allowing flexible methods to estimate the nuisance parameters $g_{n,A}$ and $\overline{Q}_{n,Y}$.

Simulation evidence: A first look



Simulation evidence: A bit deeper



Augmenting TMLE via undersmoothing

- Consider again the TML estimator (with a twist):

$$\psi_n^\star = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_{n,Y}^{\star,\lambda}(d(A_i, L_i), L_i) \ ,$$

  but adjust the update of $\overline{Q}_{n,Y}^\star$ to now involve $g_{n,A,\lambda}$, that is, $\text{logit}(\overline{Q}_{n,Y}^{\star,\lambda}(A_i, L_i)) = \text{logit}(\overline{Q}_{n,Y}(A_i, L_i)) + \epsilon \frac{g_{n,A,\lambda}(d^{-1}(A_i, L_i; \delta) \mid L_i)}{g_{n,A,\lambda}(A_i \mid L_i)}$.
- Update to $\overline{Q}_{n,Y}^{\star,\lambda}$ involves ϵ_n and selection of λ_n via selectors.
- What might be gained from this augmentation?
  - If $\overline{Q}_{n,Y}$ consistent, asymptotic efficiency by bias correction.
  - If $\overline{Q}_{n,Y}$ and $g_{n,A}$ both consistent, then influences higher-order behavior (e.g., second-order bias).

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 popular DR estimators, these IPW estimators can be formulated without the form of the EIF.
4. Analogous ideas (as for IPW) can improve DR estimators too.
5. 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

https://arxiv.org/abs/2205.05777

Appendix

Literature: Haneuse and Rotnitzky (2013)

- *Proposal:* Re-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:

$$g_{0,\delta}(a \mid l) = \sum_{j=1}^{J(s)} S_{\delta,j}\{h_j(a,l), s\} g_0\{h_j(a,l) \mid s\} h_j'(a,l)$$

- Such intervention policies account for the natural value of the intervention A directly yet are interpretable as the imposition of an altered intervention mechanism.
- Only requires that the MTP $d(A, L; \delta)$ have an "amenable" form, by way of a *j*-sectional inverse $h_j(a,l)$ existing.

Literature: Young et al. (2014)

- Establishes equivalence between G-formula when proposed intervention depends on natural value of A vs. when 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 $\mathbb{E}Y^{d(A,L;\delta)}$ or $\mathbb{E}Y^{d(L;\delta)}$.
- The authors also consider some limits on implementing MTPs $d(A, L; \delta)$, and address working in a longitudinal setting.

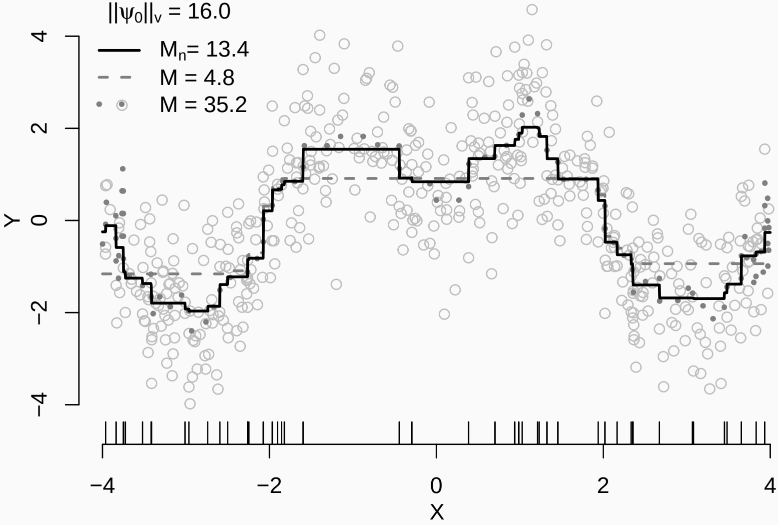
Literature: Díaz and van der Laan (2018)

- Builds on the proposal of Haneuse and Rotnitzky (2013) to accommodate MTPs $d(A, L; \delta)$, proposed after Díaz and van der Laan (2012)'s work with interventional distributions.
- To protect against *structural* positivity violations (Hernán and Robins 2024), considers an MTP mechanism that can avoid these via the guardrail encoded in $u(l)$:

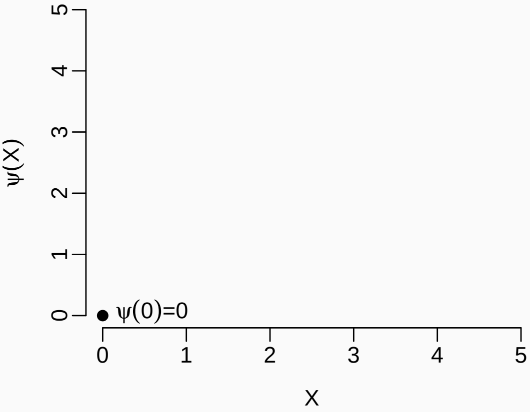
$$d(a, l; \delta) = \begin{cases} a + \delta, & a + \delta < u(l) \\ a, & \text{otherwise} \end{cases}$$

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

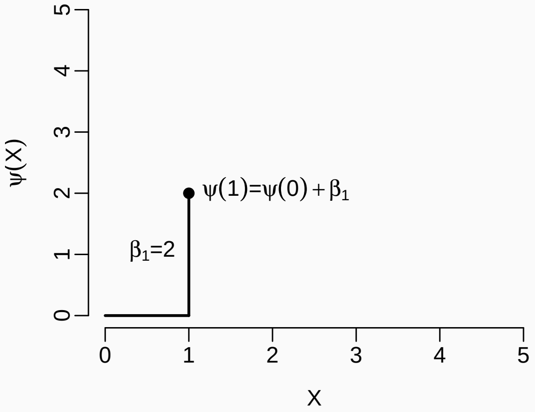
Highly Adaptive Lasso (HAL) illustration



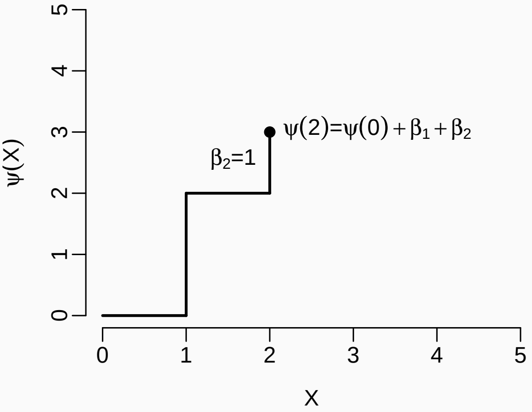
HAL illustration



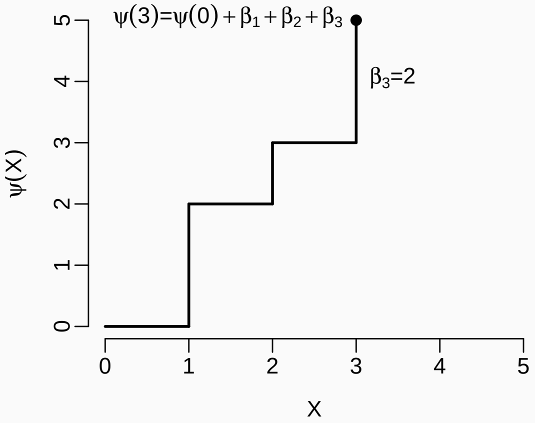
HAL illustration



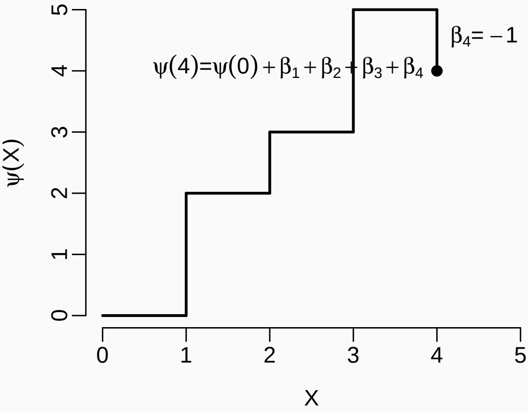
HAL illustration



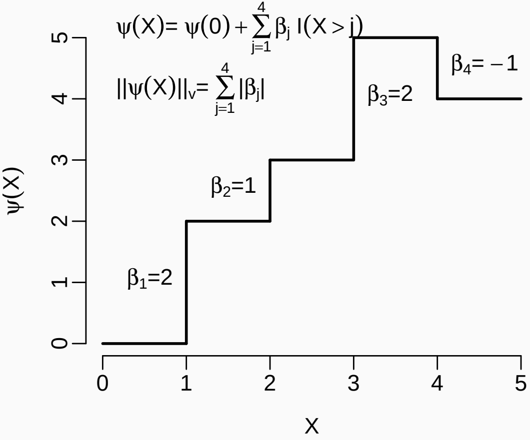
HAL illustration



HAL illustration



HAL illustration



Convergence rate of HAL

We have, for $\alpha(d) = 1/(d+1)$,

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

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

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

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

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

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

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

References

Bibaut, A. F. and van der Laan, M. J. (2019). Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*.

Cox, D. R. (1958). Planning of experiments.

Díaz, I. and van der Laan, M. J. (2011). Super learner based conditional density estimation with application to marginal structural models. *The international journal of biostatistics*, 7(1):1–20.

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

Díaz, I. and van der Laan, M. J. (2018). Stochastic treatment regimes. In *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 167–180. Springer.

Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2021). Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 118(542):846–857.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

Ertefaie, A., Hejazi, N. S., and van der Laan, M. J. (2022). Nonparametric inverse-probability-weighted estimators based on the highly adaptive lasso. *Biometrics*, (in press).

Ghosh, T., Yu, M., and Zhao, J. (2024). A selective review on conditional density estimation. *Statistics and Its Interface*, 17(3):549–564.

Haneuse, S. and Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277.

Hernán, M. A. and Robins, J. M. (2024?). *Causal Inference: What If*. CRC Press.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245.

Lepskii, O. V. (1993). Asymptotically minimax adaptive estimation II — Schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448.

Lepskii, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546.

Munch, A., Gerds, T. A., van der Laan, M. J., and Rytgaard, H. C. W. (2024). Estimating conditional hazard functions and densities with the highly-adaptive lasso. *arXiv preprint arXiv:2404.11083*.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, pages 297–331. Springer.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Rytgaard, H. C., Gerds, T. A., and van der Laan, M. J. (2022). Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. *The Annals of Statistics*, 50(5):2469–2491.

van der Laan, M. J. (2015). A generally efficient targeted minimum loss based estimator.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Young, J. G., Hernán, M. A., and Robins, J. M. (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic methods*, 3(1):1–19.