Nonparametric estimation of the generalized propensity score based on the highly adaptive lasso

Nima Hejazi

Wednesday, 19 May 2021

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

nshejazi
nhejazi
nimahejazi.org
with M. van der Laan, I. Díaz, & D. Benkeser
European Causal Inference Meeting

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_δ 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 $A^\star \sim G^\star(\cdot \mid W)$; static interventions are a special case (degenerate distribution).
- Such an intervention generates a counterfactual RV $Y_{G^\star} := f_Y(A^\star, W, U_Y)$, with distribution P_0^δ.
- We aim to estimate $\psi_{0,\delta} := \mathbb{E}_{P_0^\delta}\{Y_{G^\star}\}$, the counterfactual mean under the post-intervention exposure distribution $G^\star$.

Stochastic interventions for the causal effects of shifts

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

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

- Evaluate outcome under modified *intervention distribution*: $P_\delta(g_0)(A = a \mid W) = g_0(\delta^{-1}(A, W) \mid W)$.
- Díaz and van der Laan (2012) show that $\psi_{0,\delta}$ is identified by a functional of the distribution of O:

$$\psi_{0,\delta} = \int_{\mathcal{W}} \int_{\mathcal{A}} \mathbb{E}_{P_0}\{Y \mid A = \delta(a, w), W = w\} \cdot g_{0,A}(a \mid W = w) \cdot q_{0,W}(w) d\mu(a) d\nu(w)$$

Estimation of the PIE

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

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

where $D^\star(P)$ is the *efficient influence function* (EIF) of Ψ_δ with respect to the model $\mathcal{M}$ at P.

The EIF of Ψ is indexed by two key *nuisance parameters*

$$\overline{Q}_{P,Y}(A, W) := \mathbb{E}_P(Y \mid A, W) \qquad \textit{outcome mechanism}$$

$$g_{P,A}(A, W) := p(A \mid W) \qquad \textit{generalized propensity score}$$

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.

IPW estimators

The IPW estimator $\Psi_\delta(P_n, g_{n,A})$ is a solution to the score equation $P_n U_{g_{n,A}}(\Psi_\delta) = 0$, where $U_{g_A}(O; \Psi_\delta) = \frac{(g_{n,A}(\delta^{-1}(A_i, W_i)|W_i)}{g_{n,A}(A_i|W_i))} Y - \Psi(P)$:

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

- Consistency and convergence rate of IPW relies on those same properties of the 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 $g_A(A \mid W)$ and $g_A(\delta^{-1}(A, W) \mid W)$.
- 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

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

  for $\alpha_{t-1} \leq a < \alpha_t$.
  - 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

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

  - 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 (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 *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 θ that sets the coordinates in *s equal to zero*.

The *variation norm* of θ can be written:

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

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

Variation norm

We can represent the function θ as

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

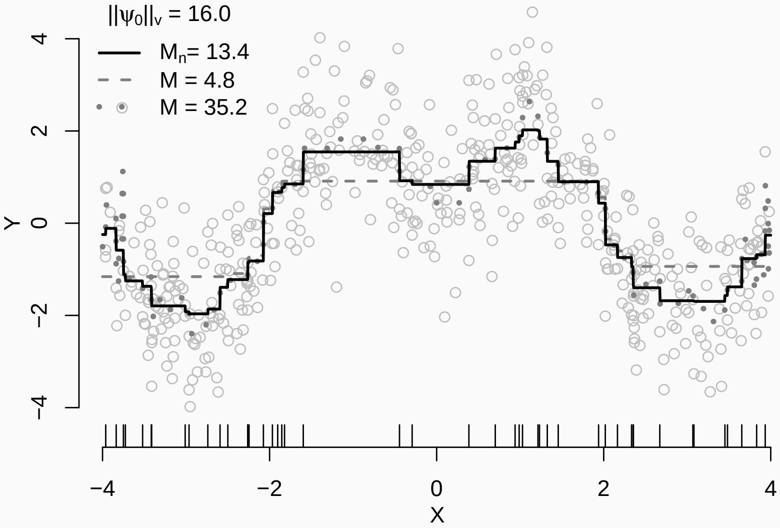
For discrete measures $d\theta_s$ with *support points* $\{u_{s,j} : j\}$ we get a *linear combination* of indicator *basis functions*:

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

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

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

HAL illustration



Convergence rate of HAL

We have

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

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

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 rate (Bibaut and van der Laan 2019):

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

HAL estimate of $g_{0,A}$

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

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

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

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

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

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

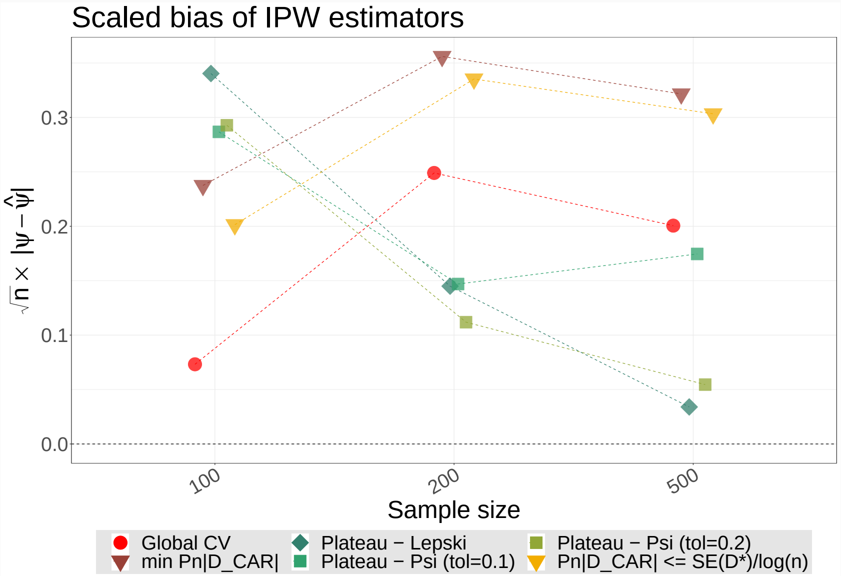
Denote by $g_{n,\lambda} \equiv g_{\beta_{n,\lambda}}$ the HAL estimate of $g_{0,A}$.

Undersmoothing HAL for IPW estimation

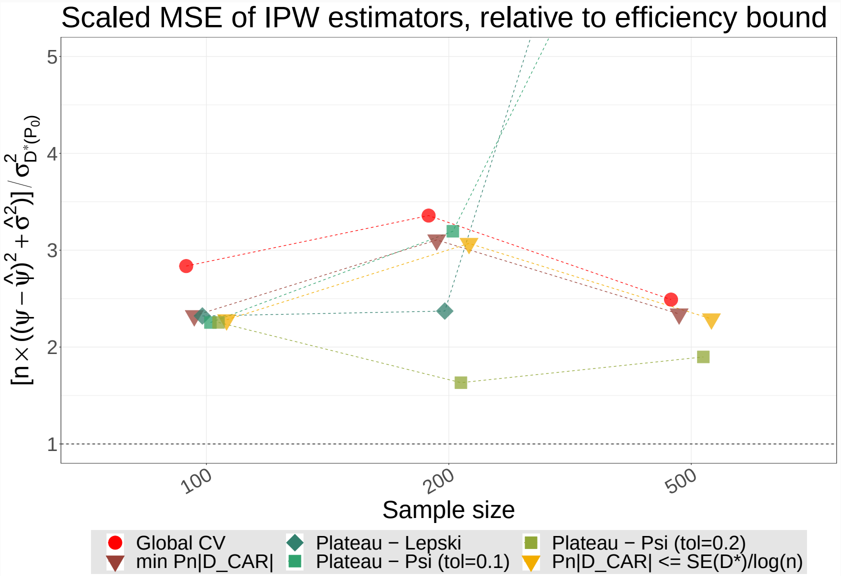
Beyond the cross-validation selector's choice of λ_n, we propose

1. EIF-based: choose λ_n s.t.
$$\lambda_n = \arg\min_{\lambda} |P_n D^\star(g_{n,A,\lambda}, \overline{Q}_{n,Y})|,$$
where $\overline{Q}_{n,Y}$ is an estimate of $\overline{Q}_{0,Y}(A, W)$.
2. Plateau-based: choose λ_n as the first in $\lambda_1, \ldots, \lambda_K$ s.t.
$$|\psi_{n,\lambda_{j+1}} - \psi_{n,\lambda_j}|_{j=1}^{K-1} \leq Z_{(1-\alpha/2)}[\sigma_{n,\lambda_{j+1}} - \sigma_{n,\lambda_j}]_{j=1}^{K-1},$$
where σ_{n,λ_j} is a variance estimate at λ_j.
3. Plateau-based: choose λ_n as the first in $\lambda_1, \ldots, \lambda_K$ s.t.
$$\left[\frac{|\psi_{n,\lambda_{j+1}} - \psi_{n,\lambda_j}|}{\max_j |\psi_{n,\lambda_{j+1}} - \psi_{n,\lambda_j}|}\right]_{j=1}^{K-1} \leq \kappa$$
for an arbitrary tolerance level κ.

Simulation: Bias



Simulation: MSE



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.

Thank you!

https://nimahejazi.org

https://twitter.com/nshejazi

https://github.com/nhejazi

Appendix

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* $A_i = \delta(a_i, w_i)$, $i = 1, \ldots, n$

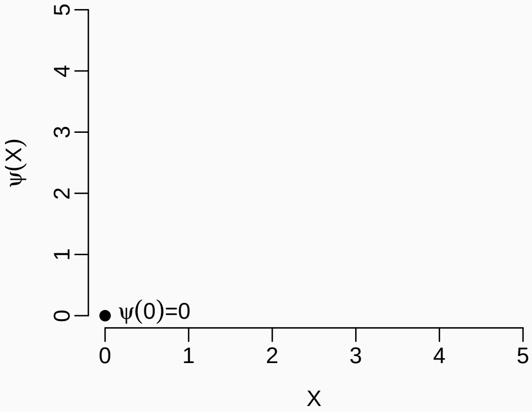
Assumption 2: ***Ignorability***

$A_i \perp\!\!\!\perp Y_i^{\delta(a_i,w_i)} \mid W_i$, *for* $i = 1, \ldots, n$

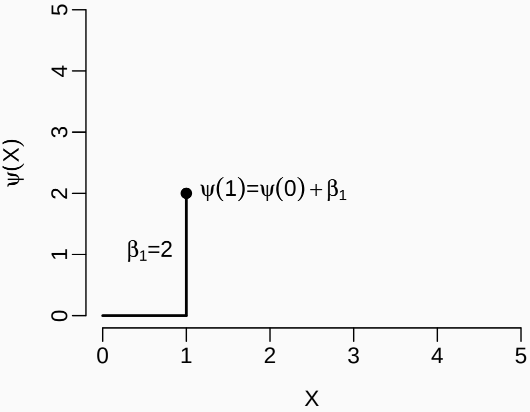
Assumption 3: ***Positivity***

$a_i \in \mathcal{A} \implies \delta(a_i, w_i) \in \mathcal{A}$ *for all* $w \in \mathcal{W}$, *where* $\mathcal{A}$ *denotes the support of* A *conditional on* $W = w_i$ *for all* $i = 1, \ldots n$

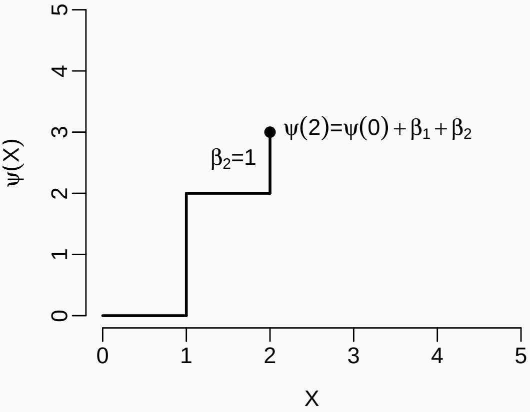
HAL illustration



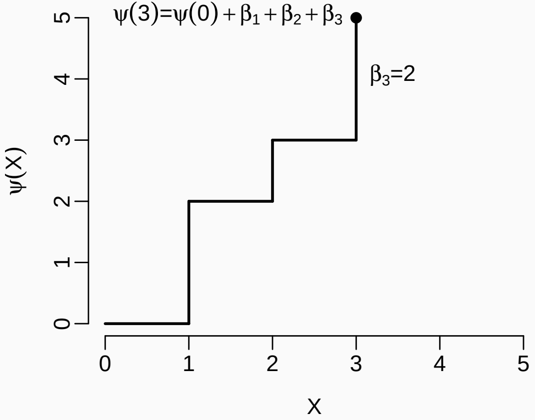
HAL illustration



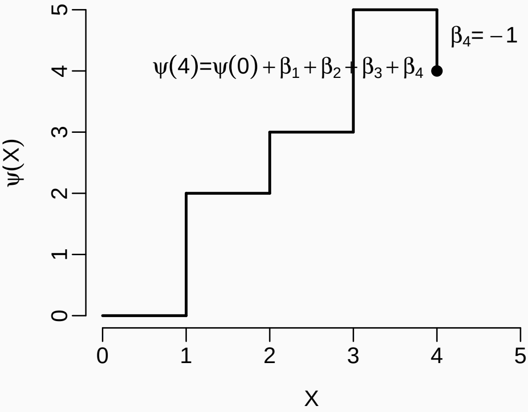
HAL illustration



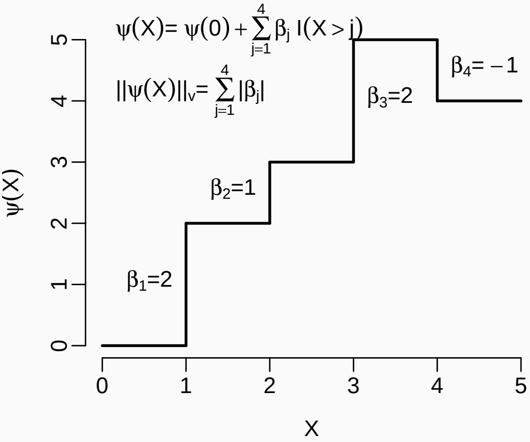
HAL illustration



HAL illustration



HAL illustration



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:

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

- 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, accomodating MTP-type shifts $\delta(A, W)$ proposed after their earlier work.
- To protect against positivity violations, considers a specific shifting mechanism:

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

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

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Benkeser, D., Carone, M., van der Laan, M. J., and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.

Benkeser, D. and van der Laan, M. J. (2016). The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE.

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*.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169:571–584.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Coyle, J. R., Hejazi, N. S., and van der Laan, M. J. (2019). *hal9001: The scalable highly adaptive lasso*. R package version 0.2.5.

Coyle, J. R., Hejazi, N. S., and van der Laan, M. J. (2020). hal9001: The scalable highly adaptive lasso. R package version 0.2.7.

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 Science & Business Media.

Gill, R. D., van der Laan, M. J., and Wellner, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 545–597.

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

Hejazi, N. S., Coyle, J. R., and van der Laan, M. J. (2020). hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, pages 561–570.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. CRC Boca Raton, FL.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Klaassen, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562.

Mukherjee, R., Newey, W. K., and Robins, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*.

Owen, A. B. (2005). Multidimensional variation for quasi-monte carlo. In *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pages 49–74. World Scientific.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

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

Qiu, H., Luedtke, A., and Carone, M. (2020). Universal sieve-based strategies for efficient estimation using machine learning tools. *arXiv preprint arXiv:2003.01856*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, J., Li, L., Tchetgen Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen Tchetgen, E., and van der Vaart, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68(1):129–137.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

van der Laan, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2).

van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics*, 10(1):29–57.

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

van der Laan, M. J. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2).

van der Laan, M. J., Benkeser, D., and Cai, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *arXiv preprint arXiv:1908.05607*.

van der Laan, M. J. and Bibaut, A. F. (2017). Uniform consistency of the highly adaptive lasso estimator of infinite-dimensional parameters. *arXiv preprint arXiv:1709.06256*.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Vansteelandt, S., Carpenter, J., and Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology*.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Vermeulen, K. and Vansteelandt, S. (2016). Data-adaptive bias-reduced doubly robust estimation. *The international journal of biostatistics*, 12(1):253–282.

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.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer.