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

nshejazi
nhejazi
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

$$\Psi(P) := \mathbb{E}_P\{\mathbb{E}_P(Y \mid A = 1, W) - \mathbb{E}_P(Y \mid A = 0, W)\} \ .$$

Estimation of the ATE

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 with respect to the model $\mathcal{M}$ at P.

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

$$\begin{aligned} \overline{Q}_P(A, W) &:= \mathbb{E}_P(Y \mid A, W) && \textit{outcome mechanism} \\ g_P(W) &:= \mathbb{E}_P(A \mid W) && \textit{propensity score} \end{aligned}$$

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.

IPW estimators

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,

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

- 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:

$$\begin{aligned}\Psi(P_n, g_n) - \Psi(P_0, g_0) =& P_n U_{g_n}(\Psi) - P_0 U_{g_0}(\Psi) \\ =& (P_n - P_0) U_{g_0}(\Psi) \\ &+ P_0\{U_{g_n}(\Psi) - U_{g_0}(\Psi)\} \\ &+ (P_n - P_0)\{U_{g_n}(\Psi) - U_{g_0}(\Psi)\}.\end{aligned}$$

- 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.

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 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 $O(n^{-k/(2k+d)})$

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:

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

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

Variation norm

We can represent the function θ as

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

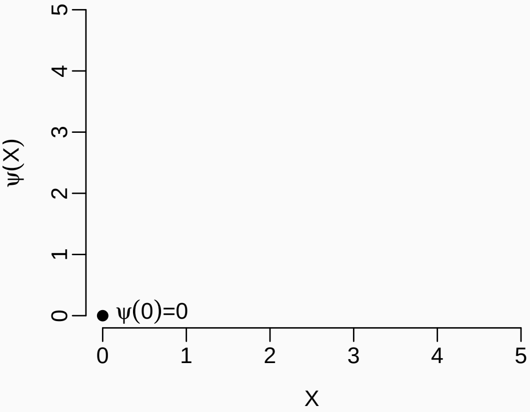
For discrete measures $d\theta_s$ with *support points* $\{u_{s,j} : j\}$ we get a *linear combination* of indicator *basis functions*:

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

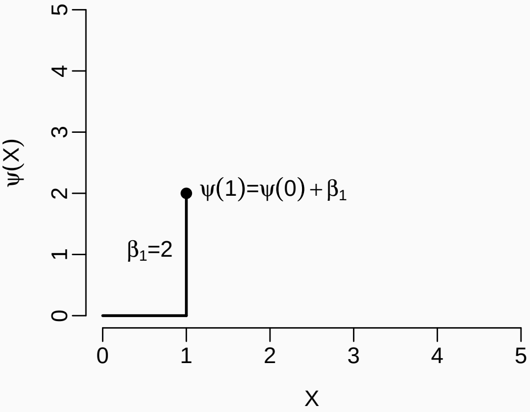
where $\beta_{s,j} = d\theta_s(u_{s,j})$, $\theta_{u_{s,j}}(x) = I(x_s \geq u_{s,j})$, and

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

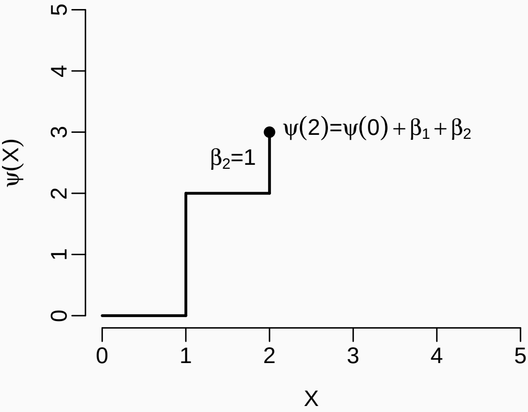
Illustration



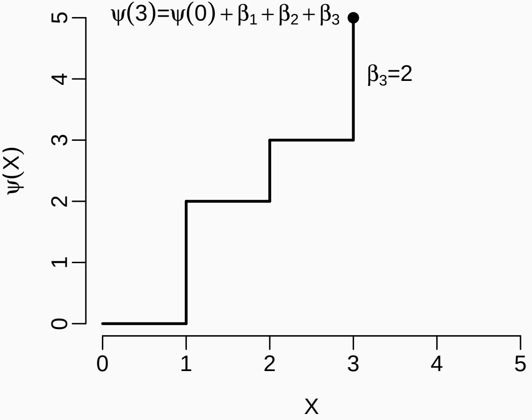
Illustration



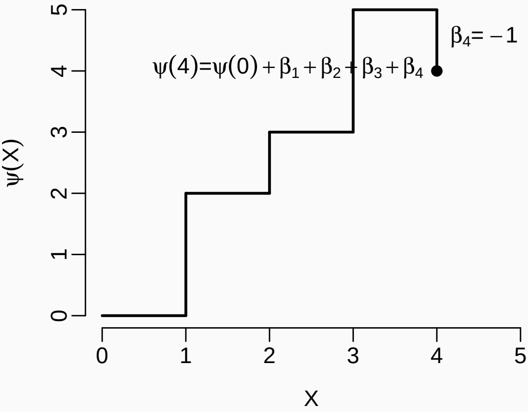
Illustration



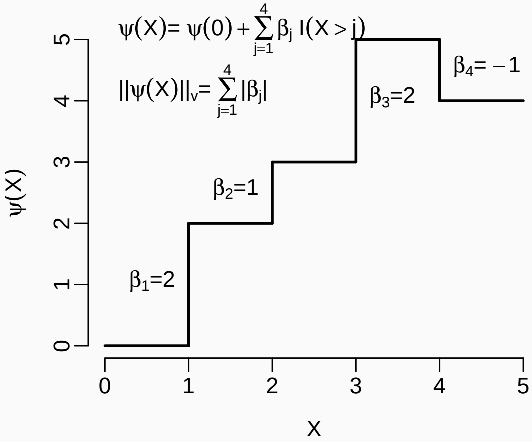
Illustration



Illustration



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

Under the assumption that our nuisance functional parameter g is a cadlag function with finite sectional variation norm, $\operatorname{logit} g$ may be approximated as (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.

The proposal

The efficient influence function expansion is of the form

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

In particular, the EIF may be expressed

$$\begin{aligned} D^\star(P_0) &:= U_{g_0}(\Psi) - D_{\mathrm{CAR}}(P_0) \\ &= \left[\frac{AY}{g(1 \mid W)} - \Psi(P, g)\right] - \left[\frac{\overline{Q}(1, w)}{g(1 \mid W)}\{A - g(1 \mid W)\}\right] \ . \end{aligned}$$

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

$$P_n D_{\mathrm{CAR}}(g_n, Q_0) = o_p(n^{-1/2}).$$

The score function

The score function of the HAL fit is

$$S_h(g) = \Phi(A - g_{n,\lambda_n})$$

where Φ is a vector consisting of indicator basis functions ϕ_s. As we undersmooth, the dimension of Φ increases, and thus, we start solving more and more equations.

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

If we undersmooth enough then we would also solve $P_n D_{CAR}(g_n, Q_n) = o_P(n^{-1/2})$.

Undersmoothing in practice

We propose two criteria.

1. D_{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 $\tilde{S}_{s,j}(\phi, g_{n,\lambda}) = \phi_{s,j}(W)\{A - g_{n,\lambda}(1 \mid W)\}\{g_{n,\lambda}(1 \mid W)\}^{-1}$.

Simulation

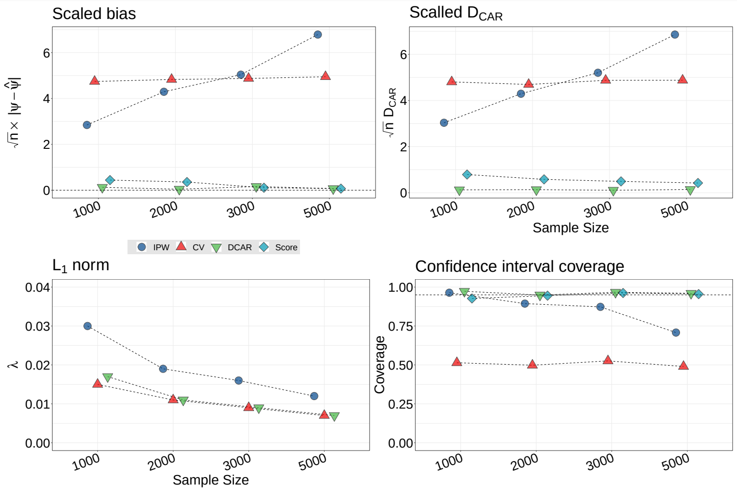


Figure 1: Circle: parametric; Triangle: NP with cross-validated λ selector; "$\bigtriangledown$": D_{CAR}-based λ selector; "$\diamond$": score-based λ selector.

The Big Picture

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

https://arxiv.org/abs/2005.11303

Appendix

The red term

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

$$
\begin{aligned}
& P_0 \left\{ U_{G_n}(\Psi) - U_{G_0}(\Psi) \right\} \\
&\quad = P_0 \left\{ G_0 Q_0(1) \left(\frac{G_0 - G_n}{G_n G_0} \right) \right\} \\
&\quad = P_0 \left\{ Q_0(1) \left(\frac{G_0 - G_n}{G_0} \right) \right\} + {\color{blue} P_0 \left\{ \frac{Q_0(1)}{G_n} \left(G_0 - G_n \right)^2 \right\}} \\
&\quad = P_0 \left\{ Q_0(1) \left(\frac{G_0 - G_n}{G_0} \right) \right\} + o_p(n^{-1/2}) \\
&\quad = -(P_n - P_0) \left\{ D_{\text{CAR}}(P_0) \right\} - {\color{red} P_n \left\{ D_{\text{CAR}}(Q_0, G_0, G_n) \right\}} + o_p(n^{-1/2}),
\end{aligned}
$$

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

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.

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.

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.

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.

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