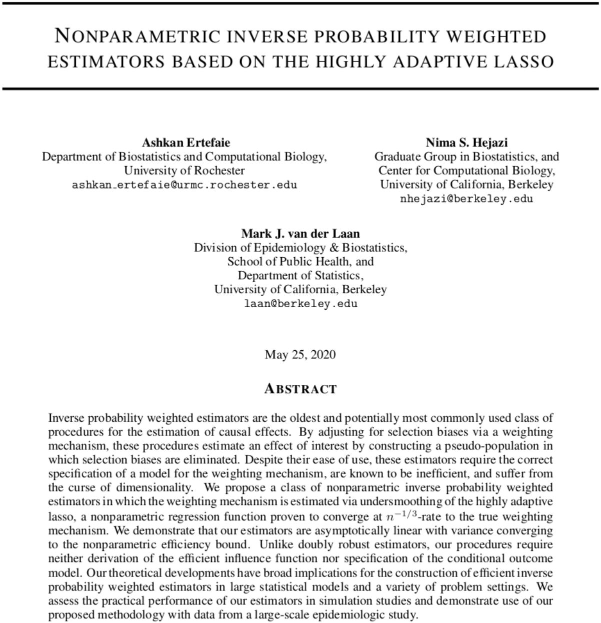
Journal club – undersmoothed HAL

Anders Munch

May 26, 2021

The article



Nonparametric inverse probability weighted estimators based on the highly adaptive lasso

Ashkan Ertefaie
Department of Biostatistics and Computational Biology,
University of Rochester
ashkan_ertefaie@urmc.rochester.edu

Nima S. Hejazi
Graduate Group in Biostatistics, and
Center for Computational Biology,
University of California, Berkeley
nhejazi@berkeley.edu

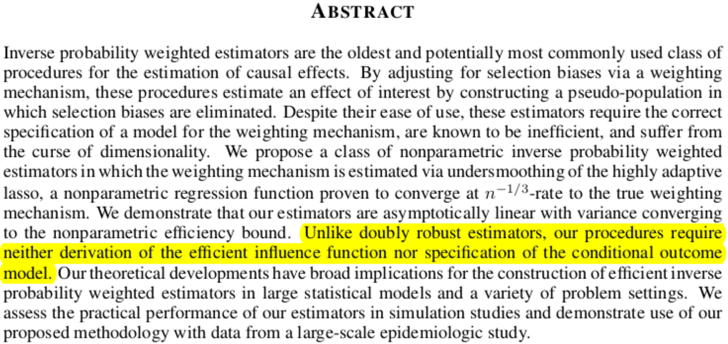
Mark J. van der Laan
Division of Epidemiology & Biostatistics,
School of Public Health, and
Department of Statistics,
University of California, Berkeley
laan@berkeley.edu

May 25, 2020

Abstract

Inverse probability weighted estimators are the oldest and potentially most commonly used class of procedures for the estimation of causal effects. By adjusting for selection biases via a weighting mechanism, these procedures estimate an effect of interest by constructing a pseudo-population in which selection biases are eliminated. Despite their ease of use, these estimators require the correct specification of a model for the weighting mechanism, are known to be inefficient, and suffer from the curse of dimensionality. We propose a class of nonparametric inverse probability weighted estimators in which the weighting mechanism is estimated via undersmoothing of the highly adaptive lasso, a nonparametric regression function proven to converge at $n^{-1/3}$-rate to the true weighting mechanism. We demonstrate that our estimators are asymptotically linear with variance converging to the nonparametric efficiency bound. Unlike doubly robust estimators, our procedures require neither derivation of the efficient influence function nor specification of the conditional outcome model. Our theoretical developments have broad implications for the construction of efficient inverse probability weighted estimators in large statistical models and a variety of problem settings. We assess the practical performance of our estimators in simulation studies and demonstrate use of our proposed methodology with data from a large-scale epidemiologic study.

The article



ABSTRACT

Inverse probability weighted estimators are the oldest and potentially most commonly used class of procedures for the estimation of causal effects. By adjusting for selection biases via a weighting mechanism, these procedures estimate an effect of interest by constructing a pseudo-population in which selection biases are eliminated. Despite their ease of use, these estimators require the correct specification of a model for the weighting mechanism, are known to be inefficient, and suffer from the curse of dimensionality. We propose a class of nonparametric inverse probability weighted estimators in which the weighting mechanism is estimated via undersmoothing of the highly adaptive lasso, a nonparametric regression function proven to converge at $n^{-1/3}$-rate to the true weighting mechanism. We demonstrate that our estimators are asymptotically linear with variance converging to the nonparametric efficiency bound. Unlike doubly robust estimators, our procedures require neither derivation of the efficient influence function nor specification of the conditional outcome model. Our theoretical developments have broad implications for the construction of efficient inverse probability weighted estimators in large statistical models and a variety of problem settings. We assess the practical performance of our estimators in simulation studies and demonstrate use of our proposed methodology with data from a large-scale epidemiologic study.

Setting and notation – the average treatment effect

- We observe n iid. samples from $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the nonparametric model.

Setting and notation – the average treatment effect

- We observe n iid. samples from $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the nonparametric model.
- $W \in \mathcal{W}$ are baseline covariates, $A \in \{0, 1\}$ is treatment indicator, and Y is the outcome of interest.

Setting and notation – the average treatment effect

- We observe n iid. samples from $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the nonparametric model.
- $W \in \mathcal{W}$ are baseline covariates, $A \in \{0, 1\}$ is treatment indicator, and Y is the outcome of interest.
- $G\colon \mathcal{M} \to \mathcal{G}$ with $\mathcal{G} := \{G(P) \,:\, P \in \mathcal{M}\}$ is the functional nuisance parameter $G := G(P)$ denoting the treatment mechanism, i.e., $G(P)(a \mid w) := \mathbb{E}_P(A = a \mid W = w)$.

Setting and notation – the average treatment effect

- We observe n iid. samples from $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the nonparametric model.
- $W \in \mathcal{W}$ are baseline covariates, $A \in \{0, 1\}$ is treatment indicator, and Y is the outcome of interest.
- $G\colon \mathcal{M} \to \mathcal{G}$ with $\mathcal{G} := \{G(P) : P \in \mathcal{M}\}$ is the functional nuisance parameter $G := G(P)$ denoting the treatment mechanism, i.e., $G(P)(a \mid w) := \mathbb{E}_P(A = a \mid W = w)$.
- We let Y^a denote the potential outcome under the intervention $\mathrm{do}(A = a)$, and the full data unit as $X = (W, Y^0, Y^1) \sim P_X \in \mathcal{M}^F$.

Setting and notation – the average treatment effect

- We observe n iid. samples from $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the nonparametric model.
- $W \in \mathcal{W}$ are baseline covariates, $A \in \{0, 1\}$ is treatment indicator, and Y is the outcome of interest.
- $G \colon \mathcal{M} \to \mathcal{G}$ with $\mathcal{G} := \{G(P) \, : \, P \in \mathcal{M}\}$ is the functional nuisance parameter $G := G(P)$ denoting the treatment mechanism, i.e., $G(P)(a \mid w) := \mathbb{E}_P(A = a \mid W = w)$.
- We let Y^a denote the potential outcome under the intervention $\mathrm{do}(A = a)$, and the full data unit as $X = (W, Y^0, Y^1) \sim P_X \in \mathcal{M}^F$.
- The target parameter is $\Psi^F \colon \mathcal{M}^F \to \mathbb{R}$, $\Psi^F(P_X) := \mathbb{E}_{P_X}(Y^1)$, i.e., the mean counterfactual outcome under treatment.

Setting and notation – the average treatment effect

- We observe n iid. samples from $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the nonparametric model.
- $W \in \mathcal{W}$ are baseline covariates, $A \in \{0, 1\}$ is treatment indicator, and Y is the outcome of interest.
- $G\colon \mathcal{M} \to \mathcal{G}$ with $\mathcal{G} := \{G(P) : P \in \mathcal{M}\}$ is the functional nuisance parameter $G := G(P)$ denoting the treatment mechanism, i.e., $G(P)(a \mid w) := \mathbb{E}_P(A = a \mid W = w)$.
- We let Y^a denote the potential outcome under the intervention $\mathrm{do}(A = a)$, and the full data unit as $X = (W, Y^0, Y^1) \sim P_X \in \mathcal{M}^F$.
- The target parameter is $\Psi^F\colon \mathcal{M}^F \to \mathbb{R}$, $\Psi^F(P_X) := \mathbb{E}_{P_X}(Y^1)$, i.e., the mean counterfactual outcome under treatment.
- Under standard identification assumptions

$$\Psi^F(P_X) = \Psi(P) := \mathbb{E}_P[\mathbb{E}_P(Y \mid A = 1, W)],$$

where $\Psi\colon \mathcal{M} \to \mathbb{R}$.

Score functions and canonical gradient

Score function using inverse probability weights:

$$U_G(O; \Psi) := \frac{AY}{G(1 \mid W)} - \Psi(P).$$

Score functions and canonical gradient

Score function using inverse probability weights:

$$U_G(O;\Psi) := \frac{AY}{G(1 \mid W)} - \Psi(P).$$

Score function based on the canonical gradient

$$D^{\star}(O;P) := U_G(O;\Psi) - D_{\text{CAR}}(P),$$

where $D_{\text{CAR}}(P) = \Pi(U_G(\Psi) \mid T_{\text{CAR}})$ is the projection onto the nuisance tangent space T_{CAR}.

Score functions and canonical gradient

Score function using inverse probability weights:

$$U_G(O; \Psi) := \frac{AY}{G(1 \mid W)} - \Psi(P).$$

Score function based on the canonical gradient

$$D^\star(O; P) := U_G(O; \Psi) - D_{\mathrm{CAR}}(P),$$

where $D_{\mathrm{CAR}}(P) = \Pi(U_G(\Psi) \mid T_{\mathrm{CAR}})$ is the projection onto the nuisance tangent space T_{CAR}. The projection is given as

$$D_{\mathrm{CAR}}(P) = \frac{A - G(A \mid W)}{G(A \mid W)} Q(1, W),$$

where $Q(1, W) := \mathbb{E}_P(Y \mid A = 1, W)$ is the conditional mean outcome [Robins et al., 1994, Van der Laan et al., 2003].

Estimating the ATE – solve a score function

Estimating the ATE – solve a score function

Solve $P_n[U_{G_n}] = 0$: The inverse probability weighted estimator

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

Estimating the ATE – solve a score function

Solve $P_n[U_{G_n}] = 0$: The inverse probability weighted estimator

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

Solve $P_n[D^\star_{G_n,Q_n}] = 0$: The augmented IPW estimator

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

Estimating the ATE – solve a score function

Solve $P_n[U_{G_n}] = 0$: The inverse probability weighted estimator

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

Solve $P_n[D^{\star}_{G_n,Q_n}] = 0$: The augmented IPW estimator

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

✓ Vanishing first order bias – rate $n^{-1/4}$ sufficient for G_n and Q_n

Estimating the ATE – solve a score function

Solve $P_n[U_{G_n}] = 0$: The inverse probability weighted estimator

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

Solve $P_n[D^{\star}_{G_n,Q_n}] = 0$: The augmented IPW estimator

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

✓ Vanishing first order bias – rate $n^{-1/4}$ sufficient for G_n and Q_n

× Two nuisance parameters to estimate

Estimating the ATE – solve a score function

Solve $P_n[U_{G_n}] = 0$: The inverse probability weighted estimator

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

Solve $P_n[D^\star_{G_n,Q_n}] = 0$: The augmented IPW estimator

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

- ✓ Vanishing first order bias – rate $n^{-1/4}$ sufficient for G_n and Q_n
- × Two nuisance parameters to estimate
- × Need to find the EIF

The Highly Adaptive Lasso (HAL) – nuisance estimator

The Highly Adaptive Lasso (HAL) – nuisance estimator

$\mathbb{D}[0,\tau]$ is the Banach space of real-valued càdlàg functions on a cube $[0,\tau] \in \mathbb{R}^d$. For a function $f \in \mathbb{D}[0,\tau]$ and a subset $s \subset \{1,\ldots,d\}$ define

$$f_s \colon [0_s, \tau_s] \to \mathbb{R}, \quad f_s(u_s) := f(u_s, 0_{-s}),$$

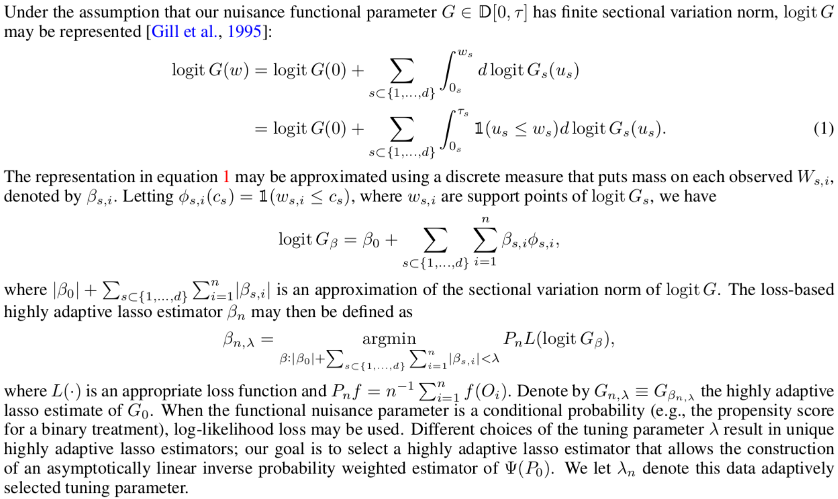
where $u_s = (u_j \ : \ j \in s)$ and u_{-s} is the complement of u_s.

The sectional variation norm of a function $f \in \mathbb{D}[0,\tau]$ is

$$\|f\|_\nu^\star := |f(0)| + \sum_{s \subset \{1,\ldots,d\}} \int_{0_s}^{\tau_s} |\,\mathrm{d}f_s(u_s)|,$$

where the sum is over all subset of the coordinates $\{1,\ldots,d\}$.

The Highly Adaptive Lasso (HAL) – nuisance estimator



Under the assumption that our nuisance functional parameter $G \in \mathbb{D}[0, \tau]$ has finite sectional variation norm, $\operatorname{logit} G$ may be represented [Gill et al., 1995]:

$$\begin{aligned}
\operatorname{logit} G(w) &= \operatorname{logit} G(0) + \sum_{s \subset \{1,\dots,d\}} \int_{0_s}^{w_s} d \operatorname{logit} G_s(u_s) \\
&= \operatorname{logit} G(0) + \sum_{s \subset \{1,\dots,d\}} \int_{0_s}^{\tau_s} \mathbb{1}(u_s \leq w_s) d \operatorname{logit} G_s(u_s). \qquad (1)
\end{aligned}$$

The representation in equation 1 may be approximated using a discrete measure that puts mass on each observed $W_{s,i}$, denoted by $\beta_{s,i}$. Letting $\phi_{s,i}(c_s) = \mathbb{1}(w_{s,i} \leq c_s)$, where $w_{s,i}$ are support points of $\operatorname{logit} G_s$, we have

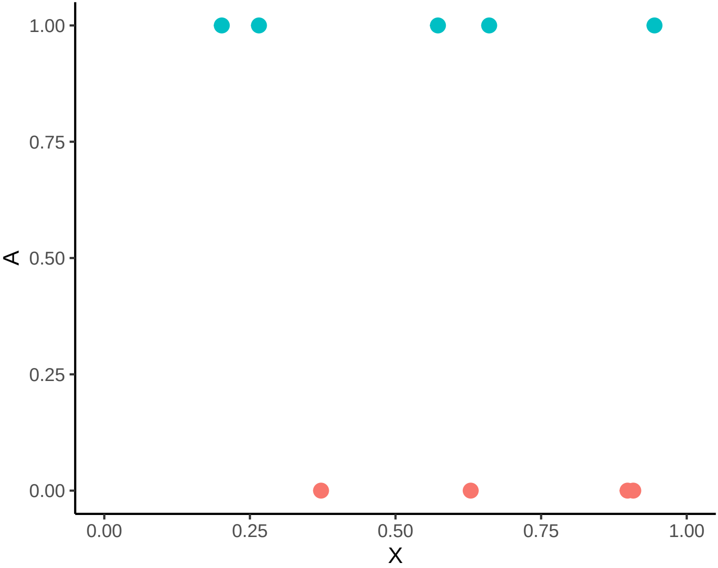
$$\operatorname{logit} G_\beta = \beta_0 + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} \beta_{s,i} \phi_{s,i},$$

where $|\beta_0| + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} |\beta_{s,i}|$ is an approximation of the sectional variation norm of $\operatorname{logit} G$. The loss-based highly adaptive lasso estimator β_n may then be defined as

$$\beta_{n,\lambda} = \operatorname*{argmin}_{\beta : |\beta_0| + \sum_{s \subset \{1,\dots,d\}} \sum_{i=1}^{n} |\beta_{s,i}| < \lambda} P_n L(\operatorname{logit} G_\beta),$$

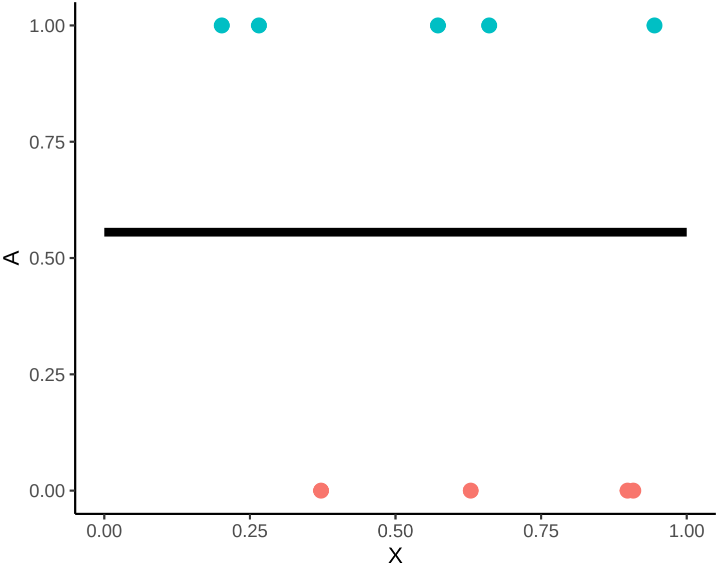
where $L(\cdot)$ is an appropriate loss function and $P_n f = n^{-1} \sum_{i=1}^{n} f(O_i)$. Denote by $G_{n,\lambda} \equiv G_{\beta_{n,\lambda}}$ the highly adaptive lasso estimate of G_0. When the functional nuisance parameter is a conditional probability (e.g., the propensity score for a binary treatment), log-likelihood loss may be used. Different choices of the tuning parameter λ result in unique highly adaptive lasso estimators; our goal is to select a highly adaptive lasso estimator that allows the construction of an asymptotically linear inverse probability weighted estimator of $\Psi(P_0)$. We let λ_n denote this data adaptively selected tuning parameter.

The smoothing hyperparameter



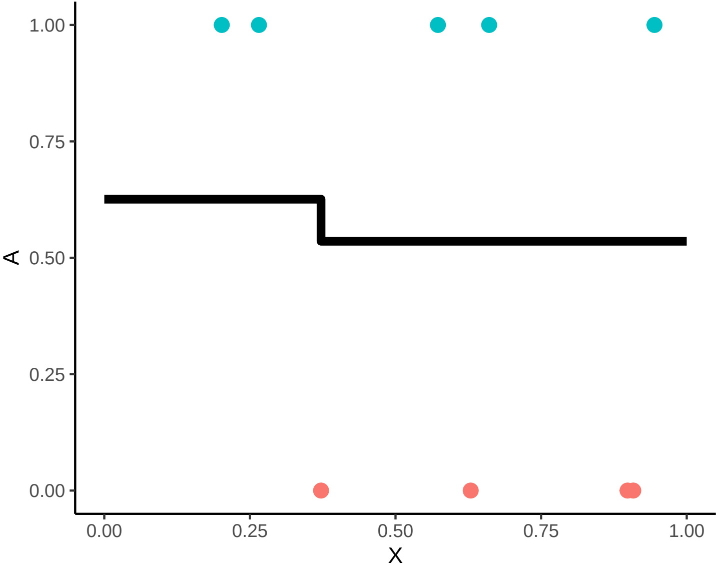
The smoothing hyperparameter

small λ large



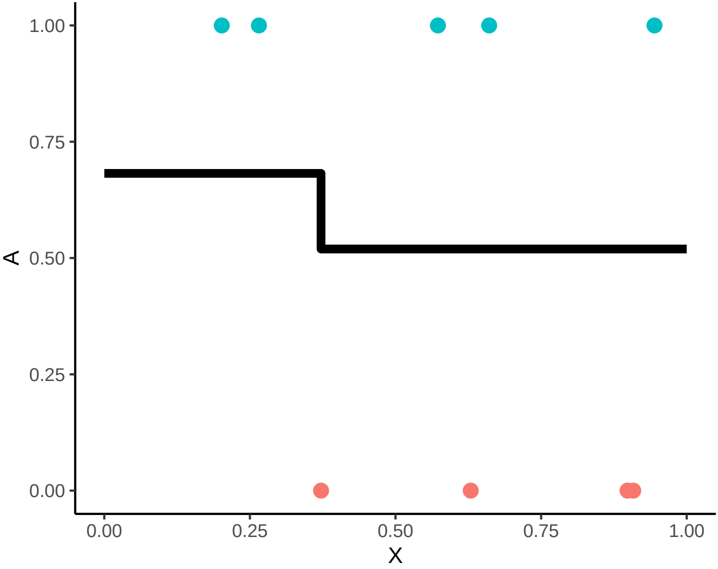
The smoothing hyperparameter

small λ large



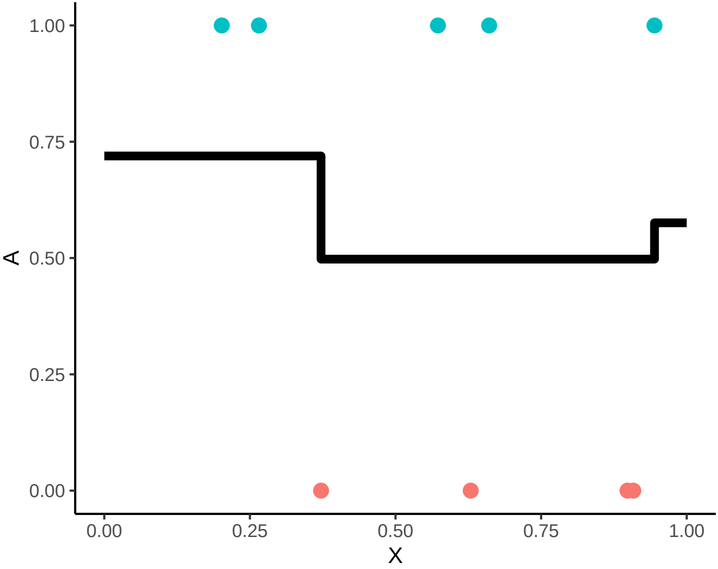
The smoothing hyperparameter

small λ large



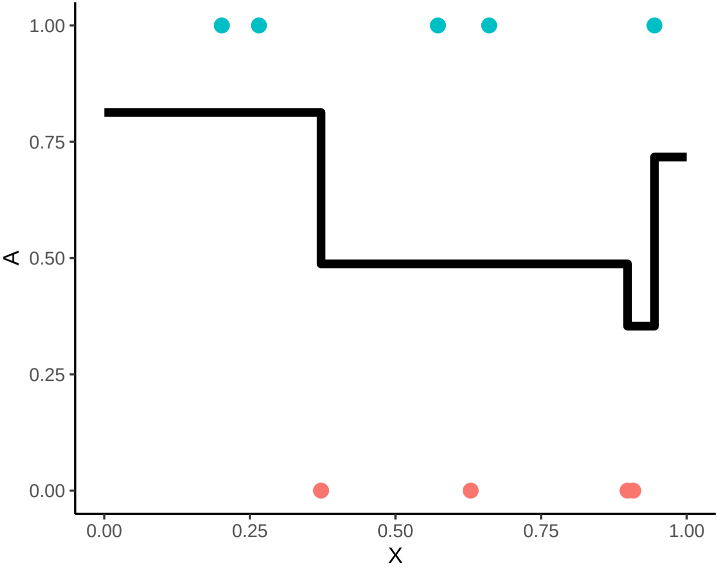
The smoothing hyperparameter

small λ large



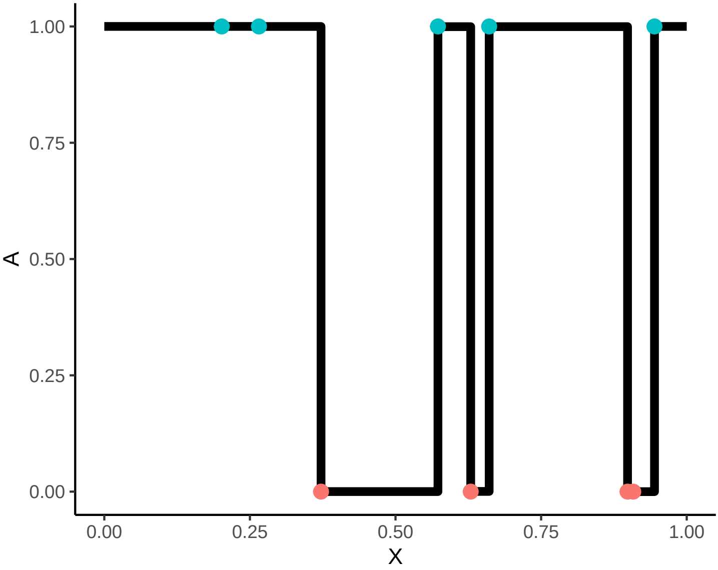
The smoothing hyperparameter

small λ large



The smoothing hyperparameter

small λ large



Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3} \log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3}\log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].
- Belongs to a Donsker function class.

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3}\log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].
- Belongs to a Donsker function class.
- We can use cross-validation to select the penalization/smoothing parameter λ.

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3} \log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].
- Belongs to a Donsker function class.
- We can use cross-validation to select the penalization/smoothing parameter λ.

Undersmoothed HAL

Using CV we find the choice of λ which gives the optimal bias-vaiance trade-off with respect to the *nuisance parameter*.

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3}\log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].
- Belongs to a Donsker function class.
- We can use cross-validation to select the penalization/smoothing parameter λ.

Undersmoothed HAL

Using CV we find the choice of λ which gives the optimal bias-vaiance trade-off with respect to the *nuisance parameter*. We want instead to pick the hyperparamter λ to get the correct bias-vaiance trade-off with respect to the *target parameter*

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3}\log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].
- Belongs to a Donsker function class.
- We can use cross-validation to select the penalization/smoothing parameter λ.

Undersmoothed HAL

Using CV we find the choice of λ which gives the optimal bias-vaiance trade-off with respect to the *nuisance parameter*. We want instead to pick the hyperparamter λ to get the correct bias-vaiance trade-off with respect to the *target parameter* → undersmooth the HAL estimator.

Properties of the HAL estimator

- Only assumption is that the target function is càdlàg with finite sectional variation norm.
- Converges at rate faster than $n^{-1/4}$ regardless of the dimension d of the covariate space; specifically, at rate $n^{-1/3} \log(n)^{d/2}$ [van der Laan and Bibaut, 2017, van der Laan, 2017].
- Belongs to a Donsker function class.
- We can use cross-validation to select the penalization/smoothing parameter λ.

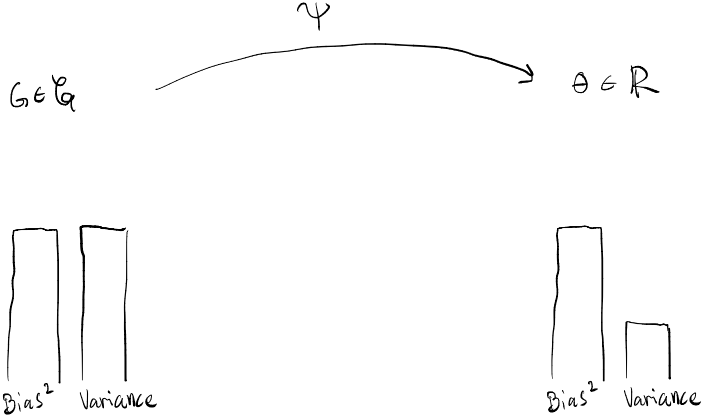
Undersmoothed HAL

Using CV we find the choice of λ which gives the optimal bias-vaiance trade-off with respect to the *nuisance parameter*. We want instead to pick the hyperparamter λ to get the correct bias-vaiance trade-off with respect to the *target parameter* $\rightarrow$ undersmooth the HAL estimator.

Old-school knowledge that undersmoothing is needed in other similar settings (density estimation) [Laurent et al., 1996, Goldstein and Khasminskii, 1996, Bickel et al., 2003, Goldstein and Messer, 1992].

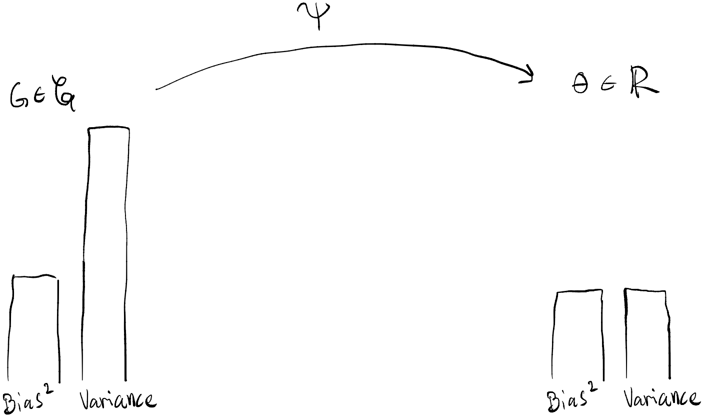
Undersmoothing

Optimizing the nuisance estimator



Undersmoothing

Undersmoothing the nuisance estimator



Undersmoothing in theory

Undersmoothing in theory

Theorem (Lemma 1 and Theorem 1 of the article)

Let G_{n,λ_n} be a HAL estimator of G_0 with λ_n chosen to satisfy

$$\min_{(s,j)\in\mathcal{J}_n} \left\| P_n \left[\frac{\partial}{\partial\varepsilon} L(\operatorname{logit} G_{n,\lambda_n} + \varepsilon\varphi_{s,j}) \right] \right\| = \mathcal{O}_P(n^{-\frac{1}{2}}), \qquad (1)$$

where $L(\cdot)$ is the log-likelihood loss and $\mathcal{J}_n$ is a set of indices for the basis functions such that $\beta_{n,j,s} \neq 0$. Then the (IPW) estimator

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

is asymptotically efficient.

Undersmoothing in theory

Theorem (Lemma 1 and Theorem 1 of the article)

Let G_{n,λ_n} be a HAL estimator of G_0 with λ_n chosen to satisfy

$$\min_{(s,j)\in\mathcal{J}_n}\left\|P_n\left[\frac{\partial}{\partial\varepsilon}L(\operatorname{logit}G_{n,\lambda_n}+\varepsilon\varphi_{s,j})\right]\right\|=\mathcal{O}_P(n^{-\frac{1}{2}}),\qquad(1)$$

where $L(\cdot)$ is the log-likelihood loss and $\mathcal{J}_n$ is a set of indices for the basis functions such that $\beta_{n,j,s}\neq 0$. Then the (IPW) estimator

$$\Psi(P_n,G_{n,\lambda_n})=\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{G_{n,\lambda_n}(A_i\mid W_i)}$$

is asymptotically efficient.

Sketch of proof:

Use empirical process theory and convergence rates of HAL to write

$$\Psi(P_n,G_{n,\lambda_n})-\Psi(P_0,G_0)=P_n[D^\star]-P_n[D_{\mathrm{CAR}}(Q_0,G_0,G_{n,\lambda_n})]+\mathcal{O}_P(n^{-\frac{1}{2}})$$

Undersmoothing in theory

Theorem (Lemma 1 and Theorem 1 of the article)

Let G_{n,λ_n} be a HAL estimator of G_0 with λ_n chosen to satisfy

$$\min_{(s,j)\in\mathcal{J}_n} \left\| P_n \left[\frac{\partial}{\partial\varepsilon} L(\operatorname{logit} G_{n,\lambda_n} + \varepsilon\varphi_{s,j}) \right] \right\| = o_P(n^{-\frac{1}{2}}), \qquad (1)$$

where $L(\cdot)$ is the log-likelihood loss and $\mathcal{J}_n$ is a set of indices for the basis functions such that $\beta_{n,j,s} \neq 0$. Then the (IPW) estimator

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

is asymptotically efficient.

Sketch of proof:

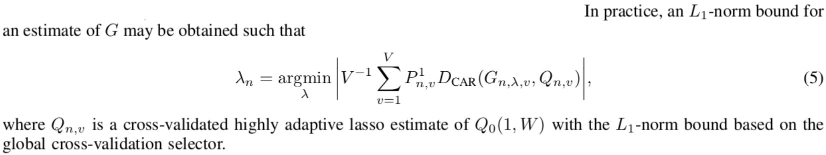
Use empirical process theory and convergence rates of HAL to write

$$\Psi(P_n, G_{n,\lambda_n}) - \Psi(P_0, G_0) = P_n[D^\star] - P_n[D_{\text{CAR}}(Q_0, G_0, G_{n,\lambda_n})] + o_P(n^{-\frac{1}{2}})$$

Lemma 1 states that (1) implies $P_n[D_{\text{CAR}}(Q_0, G_0, G_{n,\lambda_n})] = o_P(n^{-\frac{1}{2}})$ □

Undersmoothing in practice

Undersmoothing in practice

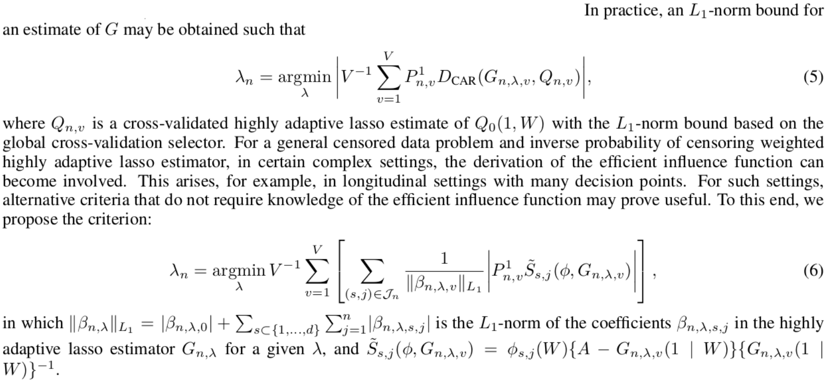


In practice, an L_1-norm bound for an estimate of G may be obtained such that

$$\lambda_n = \underset{\lambda}{\operatorname{argmin}} \left| V^{-1} \sum_{v=1}^{V} P^1_{n,v} D_{\text{CAR}}(G_{n,\lambda,v}, Q_{n,v}) \right|, \tag{5}$$

where $Q_{n,v}$ is a cross-validated highly adaptive lasso estimate of $Q_0(1, W)$ with the L_1-norm bound based on the global cross-validation selector.

Undersmoothing in practice



In practice, an L_1-norm bound for an estimate of G may be obtained such that

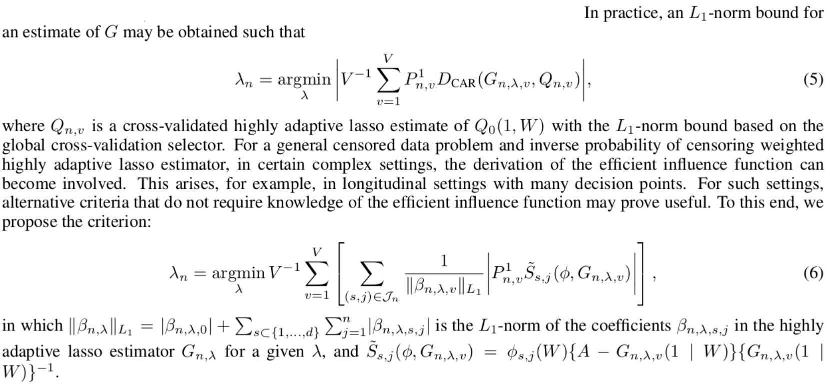
$$\lambda_n = \operatorname*{argmin}_{\lambda} \left| V^{-1} \sum_{v=1}^{V} P^1_{n,v} D_{\mathrm{CAR}}(G_{n,\lambda,v}, Q_{n,v}) \right|, \tag{5}$$

where $Q_{n,v}$ is a cross-validated highly adaptive lasso estimate of $Q_0(1, W)$ with the L_1-norm bound based on the global cross-validation selector. For a general censored data problem and inverse probability of censoring weighted highly adaptive lasso estimator, in certain complex settings, the derivation of the efficient influence function can become involved. This arises, for example, in longitudinal settings with many decision points. For such settings, alternative criteria that do not require knowledge of the efficient influence function may prove useful. To this end, we propose the criterion:

$$\lambda_n = \operatorname*{argmin}_{\lambda} V^{-1} \sum_{v=1}^{V} \left[\sum_{(s,j)\in\mathcal{J}_n} \frac{1}{\|\beta_{n,\lambda,v}\|_{L_1}} \left| P^1_{n,v} \tilde{S}_{s,j}(\phi, G_{n,\lambda,v}) \right| \right], \tag{6}$$

in which $\|\beta_{n,\lambda}\|_{L_1} = |\beta_{n,\lambda,0}| + \sum_{s\subset\{1,\ldots,d\}} \sum_{j=1}^{n} |\beta_{n,\lambda,s,j}|$ is the L_1-norm of the coefficients $\beta_{n,\lambda,s,j}$ in the highly adaptive lasso estimator $G_{n,\lambda}$ for a given λ, and $\tilde{S}_{s,j}(\phi, G_{n,\lambda,v}) = \phi_{s,j}(W)\{A - G_{n,\lambda,v}(1 \mid W)\}\{G_{n,\lambda,v}(1 \mid W)\}^{-1}$.

Undersmoothing in practice



In practice, an L_1-norm bound for an estimate of G may be obtained such that

$$\lambda_n = \operatorname*{argmin}_{\lambda} \left| V^{-1} \sum_{v=1}^{V} P_{n,v}^{1} D_{\mathrm{CAR}}(G_{n,\lambda,v}, Q_{n,v}) \right|, \tag{5}$$

where $Q_{n,v}$ is a cross-validated highly adaptive lasso estimate of $Q_0(1, W)$ with the L_1-norm bound based on the global cross-validation selector. For a general censored data problem and inverse probability of censoring weighted highly adaptive lasso estimator, in certain complex settings, the derivation of the efficient influence function can become involved. This arises, for example, in longitudinal settings with many decision points. For such settings, alternative criteria that do not require knowledge of the efficient influence function may prove useful. To this end, we propose the criterion:

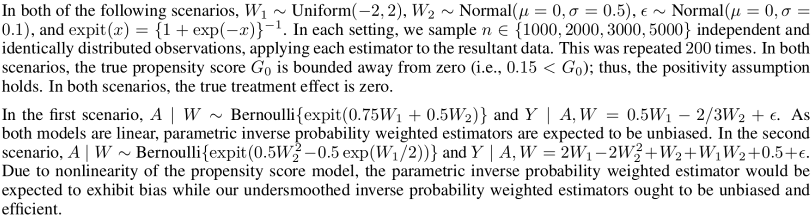
$$\lambda_n = \operatorname*{argmin}_{\lambda} V^{-1} \sum_{v=1}^{V} \left[\sum_{(s,j)\in\mathcal{J}_n} \frac{1}{\|\beta_{n,\lambda,v}\|_{L_1}} \left| P_{n,v}^{1} \tilde{S}_{s,j}(\phi, G_{n,\lambda,v}) \right| \right], \tag{6}$$

in which $\|\beta_{n,\lambda}\|_{L_1} = |\beta_{n,\lambda,0}| + \sum_{s\subset\{1,\ldots,d\}} \sum_{j=1}^{n} |\beta_{n,\lambda,s,j}|$ is the L_1-norm of the coefficients $\beta_{n,\lambda,s,j}$ in the highly adaptive lasso estimator $G_{n,\lambda}$ for a given λ, and $\tilde{S}_{s,j}(\phi, G_{n,\lambda,v}) = \phi_{s,j}(W)\{A - G_{n,\lambda,v}(1 \mid W)\}\{G_{n,\lambda,v}(1 \mid W)\}^{-1}$.

But no theoretical results about this in the article

No proof that this achieves the theoretical undersmoothing rate. . . ?

Numerical studies:



In both of the following scenarios, $W_1 \sim \text{Uniform}(-2, 2)$, $W_2 \sim \text{Normal}(\mu = 0, \sigma = 0.5)$, $\epsilon \sim \text{Normal}(\mu = 0, \sigma = 0.1)$, and $\text{expit}(x) = \{1 + \exp(-x)\}^{-1}$. In each setting, we sample $n \in \{1000, 2000, 3000, 5000\}$ independent and identically distributed observations, applying each estimator to the resultant data. This was repeated 200 times. In both scenarios, the true propensity score G_0 is bounded away from zero (i.e., $0.15 < G_0$); thus, the positivity assumption holds. In both scenarios, the true treatment effect is zero.

In the first scenario, $A \mid W \sim \text{Bernoulli}\{\text{expit}(0.75W_1 + 0.5W_2)\}$ and $Y \mid A, W = 0.5W_1 - 2/3W_2 + \epsilon$. As both models are linear, parametric inverse probability weighted estimators are expected to be unbiased. In the second scenario, $A \mid W \sim \text{Bernoulli}\{\text{expit}(0.5W_2^2 - 0.5\exp(W_1/2))\}$ and $Y \mid A, W = 2W_1 - 2W_2^2 + W_2 + W_1W_2 + 0.5 + \epsilon$. Due to nonlinearity of the propensity score model, the parametric inverse probability weighted estimator would be expected to exhibit bias while our undersmoothed inverse probability weighted estimators ought to be unbiased and efficient.

Scenario 1 Correctly specified parametric model

Scenario 2 Mis-specified parametric model

Scenario 1: Correctly specified parametric model

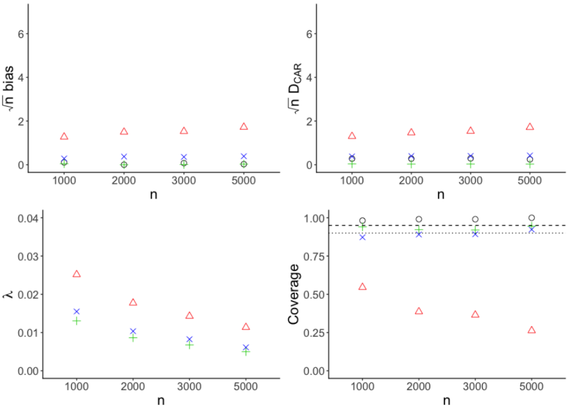


Figure 1: Comparative performance of inverse probability weighting variants in scenario 1. Circle: parametric; Triangle: nonparametric with cross-validated λ selector; "+": D_{CAR}-based λ selector; "x": score-based λ selector.

Scenario 1: Correctly specified parametric model

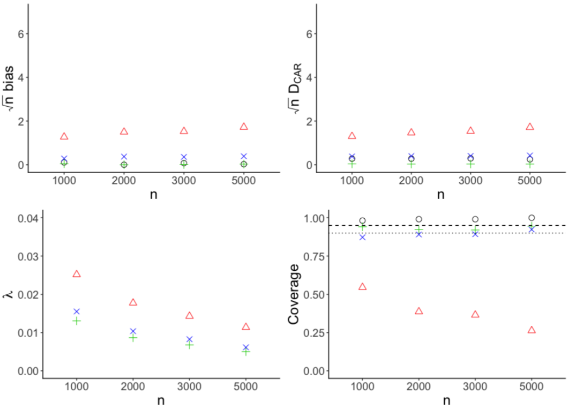


Figure 1: Comparative performance of inverse probability weighting variants in scenario 1. Circle: parametric; Triangle: nonparametric with cross-validated λ selector; "+": D_{CAR}-based λ selector; "x": score-based λ selector.

Coverage... how?

No variance estimator?

Scenario 2: Mis-specified parametric model

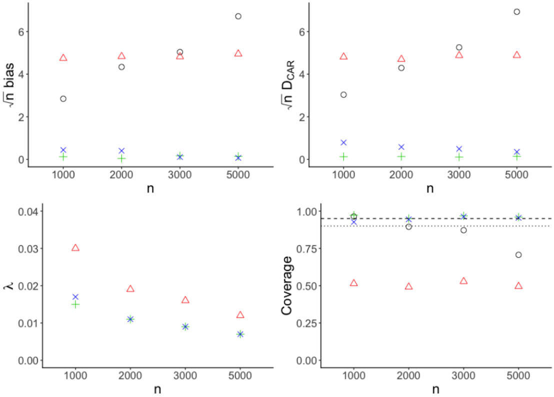


Figure 2: Performance of inverse probability weighting estimators in scenario 2. Circle: parametric; Triangle: non-parametric with cross-validated λ selector; "+": D_{CAR}-based λ selector; "x": score-based λ selector.

Perspective, thoughts, summary, and discussion

Perspective

- Nice to not need to find the EIF. Probably not so important for the ATE but potentially for more complex problems.
- Spend computational energy on optimizing the right bias-variance trade-off.
- Could be nice to generalize to other nuisance estimators. These might not achieve $n^{-1/4}$ convergence in high-dimensions, so undersmoothing could be needed even when using the EIF.

Perspective, thoughts, summary, and discussion

Perspective

- Nice to not need to find the EIF. Probably not so important for the ATE but potentially for more complex problems.
- Spend computational energy on optimizing the right bias-variance trade-off.
- Could be nice to generalize to other nuisance estimators. These might not achieve $n^{-1/4}$ convergence in high-dimensions, so undersmoothing could be needed even when using the EIF.

Thoughts

- No theoretical result for how to do undersmoothing in practice.
- Variance estimator???

Perspective, thoughts, summary, and discussion

Perspective

- Nice to not need to find the EIF. Probably not so important for the ATE but potentially for more complex problems.
- Spend computational energy on optimizing the right bias-variance trade-off.
- Could be nice to generalize to other nuisance estimators. These might not achieve $n^{-1/4}$ convergence in high-dimensions, so undersmoothing could be needed even when using the EIF.

Thoughts

- No theoretical result for how to do undersmoothing in practice.
- Variance estimator???

Questions and comments?

P. J. Bickel, Y. Ritov, et al. Nonparametric estimators which can be "plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003.

L. Goldstein and R. Khasminskii. On efficient estimation of smooth functionals. *Theory of Probability & Its Applications*, 40(1):151–156, 1996.

L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation. *The annals of statistics*, pages 1306–1328, 1992.

B. Laurent et al. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

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

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

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

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