Highly Adaptive Lasso
Mark van der Laan
0-HAL-MLE
k-HAL-MLE
Asymptotic Normality
Efficient plug-in HAL
Bootstrapping HAL
Adaptive TMLE
O-A-HAL-C-TMLE
Efficient IPCW
Meta-HAL
Conclusion

Highly Adaptive Lasso

Mark van der Laan

Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics & Statistics
University of California, Berkeley

May 14, 2024, ACIC

Acknowledgements: Rachael Phillips and Lars van der Laan

Zero-Order Spline Highly Adaptive Lasso (HAL)

A maximum likelihood estimator over all, or subset of, cadlag functions with finite (sectional) variation norm.

Key Ingredients

- Any stochastic relation/function we aim to learn from data can be approximated by linear combination (i.e., sum) of spline basis functions $X \to I(X > x_j)$ for knot point x_j.
- The sectional variation norm (i.e., complexity) of the function is the L_1-norm in this representation.
- Optimize empirical performance over all such linear models under fixed L_1-norm that is selected with cross-validation.

---

van der Laan, Mark. "A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso." The International Journal of Biostatistics (2017).

Formal representation of cadlag function as linear combination of zero order splines

- A cadlag function $f \in D^{(0)}([0,1]^d)$ can be represented as

$$f(x) = f(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{(0_s, x_s]} df_s(u),$$

where $f_s(x_s) = f(x(j)I(j \in s) : j = 1, \ldots, d)$ is the section implied by setting coordinates outside s equal to zero, and $x_s = (x(j) : j \in s)$.

- Moreover, we define the sectional variation norm of f as the sum over s of the variation norms of f_s:

$$\| f \|_v^* = \sum_s \| f_s \|_v = | f(0) | + \sum_s \int_{(0_s, 1_s]} | df_s(u) | .$$

- Now, notice that this writes f as an infinite linear combination of $x \rightarrow I(x_s \geq u)$ of zero order splines with knot-point $u \in (0_s, 1_s]$ and coefficient $df_s(u)$, and that the sectional variation norm is the L_1-norm in this representation.

HAL Provides Estimators of Large Variety of Nuisance Parameters Needed in Causal Inference

- Causal Inference requires statistical estimation of nuisance functions.
- In particular, it requires estimation of conditional means; conditional densities or cumulative distribution functions; intensities, conditional hazards. Standard loss functions can be employed for these.
- Moreover, it is often possible to define risk functions that define target functions of interest itself. For example, a variety of risk functions have been proposed for the conditional treatment effect. One can then apply HAL to minimize the empirical risk function.
- In these cases, estimation of the empirical risk function requires itself nuisance parameter estimation, where again HAL could be used.

How to develop an HAL-MLE: parametrize target function in terms of unrestricted function

- Suppose one is interested in a functional parameter $Q(P)$ such as a conditional density.
- Select a loss function or risk function so that $Q(P) = \arg\min_Q R_P(Q)$, where, for example, $R_P(Q) = PL(Q)$ for some loss $L(Q)$.
- Q might be constrained in some ways. Therefore, find a parametrization $Q = Q_f$ in terms of an unconstrained function f. For example, parametrize a conditional density in terms of conditional hazard and represent the latter as $\exp(f)$.
- Now, model f as a linear combination of zero order splines and compute the MLE $\beta_n = \arg\min_{\beta, \|\beta\|_1 < C_n} R_n\left(Q_{\sum_j \beta(j)\phi_j}\right)$, where $R_n(Q)$ is an estimate of the risk $R_{P_0}(Q)$ such as $R_n(Q) = P_n L(Q)$.

Additive models within the cadlag function space to define subspace specific HALs

- Instead of selecting the richest set of basis function such as the ones implied by knot-points $\{X_i(s) : i = 1, \ldots, n, s \subset \{1, \ldots, d\}\}$, one can only model a subset $\mathcal{S}_1$ of all the additive functions $\tilde{f}_s(x_s) = \int I(x_s \geq u) df_s(u)$ by not including all subsets s in $\mathcal{S}_1$.
- If $\mathcal{S}_1$ represents the collection of all subsets we include, then this defines additive models $f(x_s) = f(0) + \sum_{s \in \mathcal{S}_1} \tilde{f}_s(x_s)$. One can then define a corresponding HAL-MLE $f_{n,\mathcal{S}_1}$.
- More generally, we define $D^{(0)}(\mathcal{R}^{0,*})$ as the linear span of $\{\phi_j : j \in \mathcal{R}^{0,*}\}$, and by choosing the richest set $\mathcal{R}^0$, we have $D^{(0)}(\mathcal{R}^0) = D^{(0)}([0,1]^d)$.
- We use $D^{(0)}_M(\mathcal{R}^{0,*})$ when bounding L_1-norm by M.

- Every choice of subset of basis functions then implies an HAL-MLE.
- One could also first use an initial ML-algorithm, such as MARS, to learn the family of subsets, $\mathcal{S}_1$, that appears to be needed, and then compute the resulting HAL-MLE $f_{n,\mathcal{S}_1}$.
- Of course, the screening algorithm is then part of algorithm, which needs to be respected when using cross-validation to select the L_1-norm or select tuning parameter of the initial screening.

Example: HAL of CATE

- An example of a risk function $R_{P_0}(Q)$ that is non-linear in P_0 is the one might use in estimation of CATE $Q_0 = E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)$.
- There are two risk functions of interest defined as $R_{1,P_0}(Q) = E_0 L_{1,P_0}(Q)$ and $R_{2,P_0}(Q) = E_0 L_{2,P_0}(Q)$, where

$$
\begin{aligned}
L_{1,P_0}(Q) &= (Y - E_0(Y \mid W) - (A - E_0(A \mid W))Q(W))^2 \\
L_{2,P_0}(Q) &= (D_{P_0} - Q)^2 \\
D_{P_0}(O) &= \frac{2A-1}{P_0(A \mid W)}(Y - E_0(Y \mid A, W)) \\
&\quad + E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)
\end{aligned}
$$

- Both loss functions are double robust w.r.t. misspecification of the nuisance parameters $P_0(A \mid W)$ and $E_0(Y \mid A, W)$.

- Their risk based dissimilarities are $E_0(Q - Q_0)^2 g_1(1 \mid W) g_0(0 \mid W)$ and $E_0(Q - Q_0)^2$, respectively.
- Given the nuisance parameter estimators, the initial rich starting model $D^{(k)}(\mathcal{R}_N)$, the corresponding HAL-MLEs of Q_0 are defined as

$$Q_{n,j} = \arg \min_{Q \in D^{(k)}_{M_n}(\mathcal{R}_N)} P_n L_{j,n}(Q).$$

Finite Sample Performance is Good

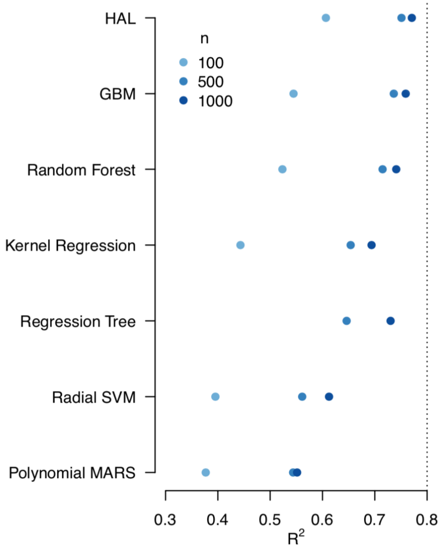
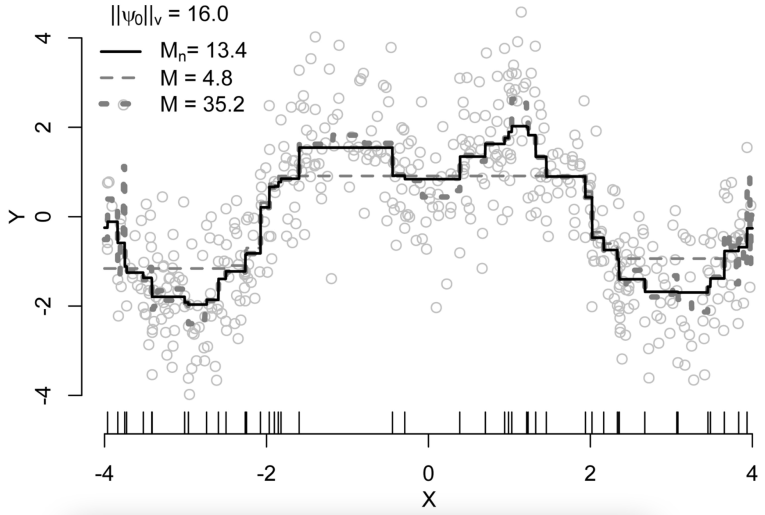


Illustration in Low Dimensions



Rate of convergence of zero order HAL

- Let $Q_n = Q_{f_n}$ be the HAL-MLE.
- Let $d_0(Q, Q_0) = P_0 L(Q) - P_0 L(Q_0)$ be the loss based dissimilarity.
- We have

$$
\begin{aligned}
d_0(Q_n, Q_0) &= P_n\{L(Q_n) - L(Q_0)\} \\
&\quad -(P_n - P_0)\{L(Q_n) - L(Q_0)\} \\
&\leq -(P_n - P_0)\{L(Q_n) - L(Q_0)\}.
\end{aligned}
$$

- Let $\mathcal{F} = \{L(Q_f) - L(Q_0) : f \in D_M^{(0)}([0,1]^d)\}$.
- The (known) covering number for $D_M^{(0)}([0,1]^d)$ implies the same covering number for $\mathcal{F}$.

Using empirical process theory:

- Let $\mathcal{F}(\delta) = \{f \in \mathcal{F} : P_0 f^2 \leq \delta^2\}$.
- $\sup_{f \in \mathcal{F}(\delta)} \mid n^{1/2}(P_n - P_0)f \mid$ can be bounded by the entropy integral $J(\delta, \mathcal{F}(\delta)) = \int_{(0,\delta]} \sqrt{\log N(\epsilon, \mathcal{F}, L^2)} d\epsilon$, which behaves as $\delta^{1/2}$ up till $\log \delta$-factor.
- Using that $P_0(L(Q) - L(Q_0)))^2 \leq C d_0(Q, Q_0)$, we can then apply an iterative proof bounding

$$
\begin{aligned}
d_0(Q_n, Q_0) &\leq n^{-1/2} \sup_{f \in \mathcal{F}(\delta^k)} \mid n^{1/2}(P_n - P_0)f \mid \\
&\sim n^{-1/2} J(\delta^k, \mathcal{F}),
\end{aligned}
$$

starting with $\delta^0 = 1$, $\delta^1 = n^{-1/4}$, and so on.

- The monotone improving (in rate) sequence converges to the fixed point δ^* of equation

$$\delta^2 = n^{-1/2} J(\delta, \mathcal{F}).$$

- This δ^* equals the rate of convergence for $d_0^{1/2}(Q_n, Q_0)$.
- We find $d_0(Q_n, Q_0) = O_P(n^{-2/3}(\log n)^d)$.

Derivation of first order spline representation of cadlag function

- In the presentation of $f \in D^{(0)}_M([0,1]^d)$ there appear integrals $\int_{(0(s),x(s)]} df_s(u)$.
- Assume and write $df_s(u) = df_s/du du$; assume the RN-derivative $f^{(1)}_s = df_s/du \in D^{(0)}_M([0,1]^d)$, and plug-in the zero order spline representation for

$$f^{(1)}_s(u) = f^{(1)}_s(0) + \sum_{s_1 \subset s} \int_{(0(s_1),u(s_1)]} df^{(1)}_{s_1}(u_1).$$

- Apply Fubini's theorem to the double integrals over (u, u_1) to obtain a representation in terms of tensor products of **first** order splines.

- As an example, let's do it for a univariate function:

$$
\begin{aligned}
f(x) &= f(0) + \int_{(0,x]} df(u) \\
&= f(0) + \int_{(0,x]} f^{(1)}(u)du \\
&= f(0) + \int_{(0,x]} \{f^{(1)}(0) + \int_{(0,u]} df^{(1)}(u_1)\}du \\
&= f(0) + xf^{(1)}(0) + \int_{u_1}\int_u I(u \leq x)I(u_1 \leq u)dudf^{(1)}(u_1) \\
&= f(0) + f^{(1)}(0)x + \int_{u_1} I(x \geq u_1)(x-u_1)df^{(1)}(u_1),
\end{aligned}
$$

which is a linear combination of first order splines $\phi^1_{u_1}(x) = I(x \geq u_1)(x-u_1)$, including $\phi^1_0(x) = x$ implied by knot-point $u_1 = 0$.

- In this manner we obtain the following first order spline representation for a function $f \in D^{(1)}([0,1]^d)$ (a cadlag function satisfying our first orders smoothness assumption):

$$
\begin{aligned}
f(x) = f(0) + \textstyle\sum_{s \subset \{1,\ldots,d\}} \phi_0^1(x(s)) f_s^{(1)}(0) \\
+ \textstyle\sum_{s,s_1 \subset s, |s_1|>0} \bar{\phi}_{s,s_1}(x) \int \phi^1_{u(s_1)}(x(s_1)) f^{(1)}_{s,s_1}(du).
\end{aligned}
$$

- Here $\bar{\phi}_{s,s_1}(x(s/s_1)) = \prod_{l \in s/s_1} x(l)$ and $f^{(1)}_{s,s_1}$ is the s_1-section of $f_s^{(1)} = df_s/du$.

- Notice that finite linear part in this representation of f is just parametric model in $x_1, \ldots, x_d$ and its interactions (e.g., for $d = 2$, $x_1, x_2, x_1 x_2$. The remaining infinite linear part is linear combination in tensor products of first order splines with interior knot-points in $(0(s_1), 1(s_1)]$ while having knots at 0 for components x_j with $j \in s/s_1$.
- The L_1-norm of all coefficients $f^{(1)}_{s,s_1}(du)$ and $f^{(1)}_s(0)$ defines our first order sectional variation norm $\| f \|^*_{v,1}$.
- $D^{(1)}_M([0,1]^d)$ is defined as all functions $f \in D^{(0)}([0,1]^d)$ satisfying this first order smoothness with $\| f \|^*_{v,1} \leq M$.

(fine enough) Finite dimensional first order spline working model

- For each $s \subset \{1, \ldots, d\}$, we can select knot-points $\mathcal{R}^1(s) \equiv \{(0(s/s_1), X_i(s_1)) : i = 1, \ldots, n, s_1 \subset s\} \subset [0(s), 1(s)]$. The corresponding first order splines are:

$$\{\phi^1_{u(s)} : u(s) \in \mathcal{R}^1(s)\}.$$

  For s_1 the empty set this yields $\{0(s) : s \subset \{1, \ldots, d\}\}$, giving the interactions $\phi^1_{0(s)} = \prod_{j \in s} x_j$.

- The total set of N first order splines is thus;

$$\mathcal{R}^1_N \equiv \{\phi^1_{u(s)} : u(s) \in \mathcal{R}^1(s), s \subset \{1, \ldots, d\}\}.$$

- The initial starting model of linear combinations of first order splines is then:

$$D^{(1)}(\mathcal{R}_N^1) = \left\{ \sum_{j \in \mathcal{R}_N^1} \beta(j)\phi_j : \beta \right\}.$$

- This working model $D^{(1)}(\mathcal{R}_N^1) \subset D^{(1)}([0,1]^d)$ represents a close approximation of $D^{(1)}([0,1]^d)$ (providing sup-norm approximations going as fast as $n^{-1/2}$).

First order spline HAL

Highly Adaptive Lasso

Mark van der Laan

0-HAL-MLE

k-HAL-MLE

Asymptotic Normality

Efficient plug-in HAL

Bootstrapping HAL

Adaptive TMLE

O-A-HAL-C-TMLE

Efficient IPCW

Meta-HAL

Conclusion

- We then define

$$\beta_n = \arg \min_{\beta, |\beta|_1 \leq C_n} P_n L(Q_{f_{N,\beta}}),$$

and first order HAL $Q_n = Q_{f_{N,\beta_n}}$.

Rate of convergence for first order HAL

- One can select J zero order splines $\mathcal{R}^0(J)$ so that $D^{(0)}(\mathcal{R}^0(J))$ yields a $O^+(1/J)$ L^2-approximation of $D^{(0)}([0,1]^d)$.
- This explains the dimension free rate of convergence $O^+(n^{-1/3})$ for the zero-order HAL.
- One can select J first order splines, $\mathcal{R}^1(J)$, so that $D^{(1)}(\mathcal{R}^1(J))$ yields a $O^+(1/J^2)$ **supremum norm** approximation of $D^{(1)}([0,1]^d)$.
- The entropy integral $J(\delta, D^{(1)}_M([0,1]^d), \| \cdot \|_\infty) = O^+(\delta^{3/4})$ instead of $O^+(\delta^{1/2})$ for $D^{(0)}_M([0,1]^d)$.
- Our rate of convergence proof yields

$$d_0(Q_n, Q_0) = O^+_P(n^{-4/5}).$$

Higher order spline HAL

- Analogue as above, we can define $D^{(k)}_M([0,1]^d) \subset D^{(k-1)}([0,1]^d)$ as class of k-th order smooth functions with k-th order sectional variation norm bounded by M.
- Analogue we obtain a k-th order spline representation for $f \in D^{(k)}([0,1]^d)$; a finite dimensional linear working model $D^{(k)}(\mathcal{R}^k_N)$ approximating $D^{(k)}([0,1]^d)$ and a corresponding k-th order spline HAL-MLE:
$$f_n = \arg \min_{f \in D^{(k)}_{M_n}(\mathcal{R}^k_N)} P_n L(Q_f),$$
and $Q_n = Q_{f_n}$.
- With J well chosen k-th order spline basis functions we can obtain an $O^+(1/J^{k+1})$ sup-norm approximation of $D^{(k)}([0,1]^d)$.

- As a consequence, we now have $J(\delta, D_M^{(k)}([0,1]^d), \| \cdot \|_\infty) = O^+(\delta^{(2k+1)/(2k+2)})$.
- Our rate of convergence proof now yields:

$$d_0(Q_n^k, Q_0) = O_P^+(n^{-2k^*/(2k^*+1)}),$$

with $k^* = k+1$.

- For example, for $k = 0, 1, 2$, we have the dimension free rates $O_P^+(n^{-1/3})$, $O_P^+(n^{-2/5})$ and $O_P^+(n^{-3/7})$, respectively.

Discrete Super Learner incorporating higher order HAL

- By varying smoothness degree k and starting sets $\mathcal{R}^k(j)$ of basis functions, one can define many HAL-MLEs over $D_C^{(k)}(\mathcal{R}^k(j))$.
- The discrete super learner using this library of (k,j)-specific HAL-MLEs will perform asymptotically exact as well as the oracle choice among all these HAL estimators, thereby achieves the rate of convergence of HAL for the unknown smoothness k_0 and smallest subspace $D^{(k_0)}(\mathcal{R}_{j_0}(d))$ containing true Q_0.
- Thus this cross-validated higher order HAL-MLE is minimax **smoothness adaptive** achieving the minimax smoothness adaptive rate of convergence for univariate function estimation, up till log n-factors.

Asymptotic normality of higher order HAL

- Let $\mathcal{R}_n$ be a set of J_n k-th order splines providing uniform approximation error $O^+(1/J_n^{k+1})$.
- The non-zero coefficients in the HAL-fit does this for us when we choose a fine enough starting model.
- In addition, HAL will select an adaptive selection that works best for Q_0.
- Let $D^{(k)}(\mathcal{R}_n)$ be the linear working model. This set yields the $O(1/J_n^{k+1})$-uniform approximation of Q_0.
- The HAL-MLE $Q_n = \sum_{j\in\mathcal{R}_n} \beta_n(j)\phi_j$ operates as an MLE of the oracle approximation $Q_{0,n} = \sum_{j\in\mathcal{R}_n} \beta_{0,n}(j)\phi_j$ in this working model.

- In particular, if we do the relax-HAL (refitting the selected working model without L_1-penalty), then it is an exact MLE of $Q_{0,n}$.
- One can analyze this parametric MLE Q_n^k w.r.t. $Q_{0,n}$ to establish that $(J_n/n)^{1/2}(Q_n - Q_{0,n})(x) \Rightarrow_d N(0, \sigma_0^2(x))$, while, by our uniform approximation result $||Q_{0,n} - Q_0||_\infty \sim O^+(1/J_n^{k+1})$.

Technical challenge in asymptotic normality proof

- Actually, the asymptotic normality proof runs into an asymptotically linear approximation with influence curve still random through $\mathcal{R}_n$ (a linear combination of the scores of β_j, $j \in \mathcal{R}_n$).
- So, the proof would have worked if Q_n is an MLE of $D^{(k)}(\mathcal{R}_n^{\#})$ for an independent $\mathcal{R}_n^{\#}$.
- Therefore we define this independent set as the result of the same HAL applied to independent $P_n^{\#}$ sample from P_0. We then prove that the actual HAL-MLE Q_n **still** acts as an approximate MLE for this fixed working model $D^{(k)}(\mathcal{R}_n^{\#})$.
- In this manner our pointwise asymptotic normality proof is completed.

Asymptotic normality and uniform rates

- By selecting $J_n \sim n^{-1/(k+2)}$, the pointwise rate equals $n^{-k^*/(2k^*+1)}$ up till $\log n$-factors.
- At cost of another $\log n$-factor this rate is uniform in x.
- Pointwise and uniform confidence intervals follow.
- Beyond inference for Q_0, these results teach us that we have dimension free **uniform** rates of convergence for $\parallel Q_n - Q_0 \parallel_\infty = O^+(n^{-k^*/(2k^*+1)})$.

Coverage Probability of HAL-based Confidence Interval by Delta Method

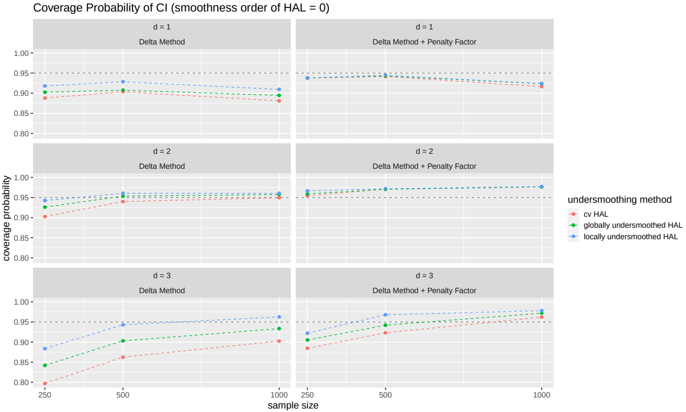


Figure: Coverage Probability of HAL-based CI (smoothness order = 0)

Coverage Probability of HAL-based Confidence Interval by Delta Method

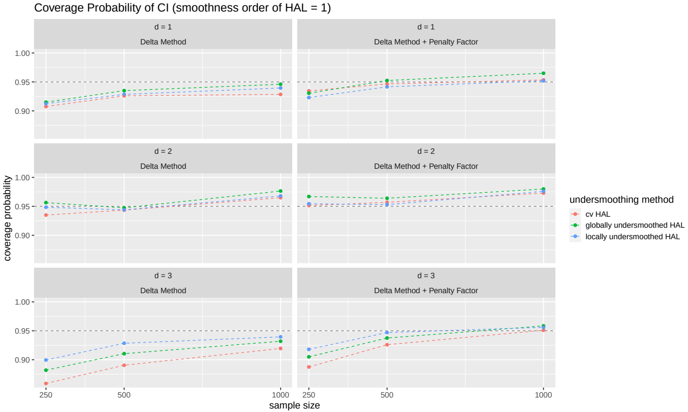


Figure: Coverage Probability of HAL-based CI (smoothness order = 1)

Coverage Probability of HAL-based Confidence Interval by Delta Method: Example of d = 3

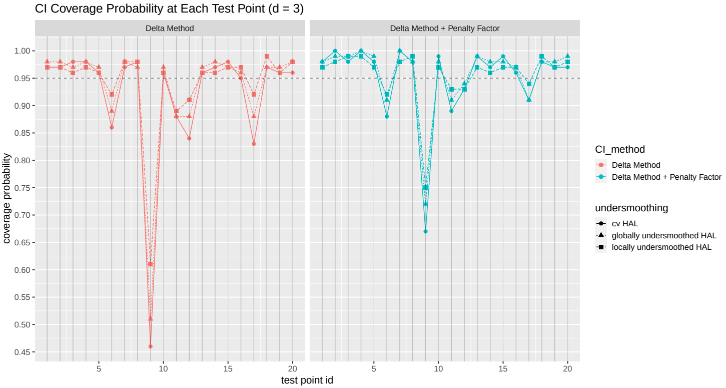


Figure: Pointwise Coverage Probability of HAL-based CI (d = 3; n = 1000; smoothness order = 1)

Data Generating Functions

$$Y = E(Y|X) + \epsilon, \epsilon \sim N(0, 0.25);$$

$$d = 1 : E(Y|X) = \sin\left(\pi X^2 \text{sign}(X)\right),$$

where $X \sim \text{Unif}(-1, 1)$;

$$d = 2 : E(Y|X) = \frac{1}{1 + \exp\left(-(X_1 - X_1 X_2)\right)},$$

where $X_1 \sim 4Z - 2, X_2 \sim Ber(0.5), Z \sim Beta(0.85, 0.85)$;

$$d = 3 : E(Y|X) = \frac{1}{1 + \exp\left(-(-2X_1 \text{sign}(X_1 > 0.5) - X_3 + 2X_2 X_3\right.}$$

where $X_1 \sim Unif(-1, 1), X_2 \sim Ber(0.5), X_3 \sim Unif(-1, 1)$

Efficient HAL-TMLE (vdL, Rubin, 2006)

- Construct initial estimator $\mathbf{P}_n$; determine a least favorable path $\{\mathbf{P}_{n,\epsilon} : \epsilon \in (-\delta, \delta)\} \subset \mathcal{M}$ through $\mathbf{P}_n$ with score $D^*_{\mathbf{P}_n}$ at $\epsilon = 0$.
- Compute MLE $\epsilon_n = \arg\max_\epsilon P_n L(\mathbf{P}_{n,\epsilon})$, where $L(P)$ is a valid loss function so that $\Psi(\arg\min_P P_0 L(P)) = \Psi(P_0)$.
- Let $\mathbf{P}^*_n = \mathbf{P}_{n,\epsilon_n}$. The TMLE is given by $\Psi(\mathbf{P}^*_n)$.
- By construction, $\Psi(\mathbf{P}^*_n) - \Psi(P_0) = (P_n - P_0)D^*_{\mathbf{P}^*_n} + R(\mathbf{P}^*_n, P_0)$, with $R(P, P_0) \equiv \Psi(P) - \Psi(P_0) - (P - P_0)D^*_P$ an exact second order remainder.
- Using the Highly Adaptive Lasso as initial estimator $\mathbf{P}_n$, we are guaranteed that $\Psi(\mathbf{P}^*_n)$ is asymptotically efficient estimator of $\Psi(P_0)$ (van der Laan, 2017).

HAL-TMLE of Target Estimands is Asymptotically Efficient

- TMLE is two-stage method for constructing plug-in efficient estimators $\Psi(P_n^*)$ of a pathwise differentiable target estimand $\Psi(P_0)$.
- The HAL-TMLE (using HAL as initial estimator) is efficient in great generality (vdL, 15).
- The only necessary model assumptions are:
  - The true nuisance parameters have finite sectional variation norm
  - The loss functions of the true nuisance parameters are uniformly bounded, so that oracle inequality applies
  - The strong positivity assumption holds

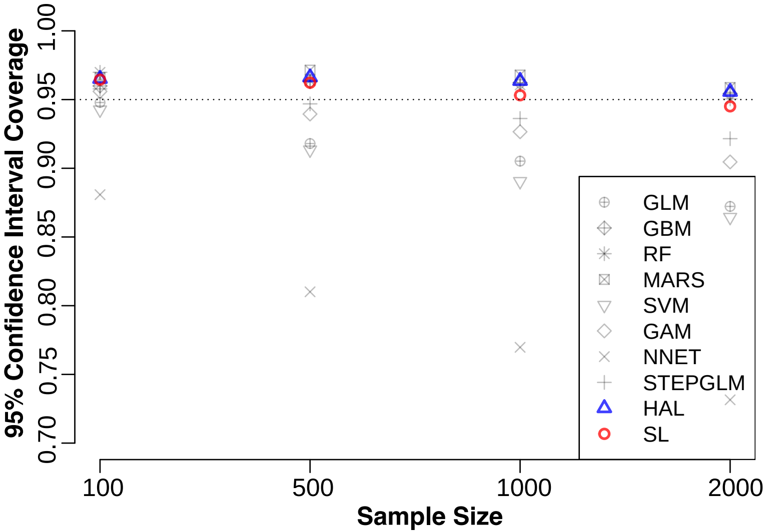
Example: Asymptotic efficiency of (zero-order) HAL-TMLE for treatment-specific mean / ATE

Consider the HAL-TMLE of $EY_1 = EE(Y \mid A = 1, W)$ based on $(W, A, Y) \sim P_0$ in a nonparametric statistical model.

It is asymptotically efficient if

1. $\delta < P_0(A = 1 \mid W)$ for some $\delta > 0$
2. $W \rightarrow E_0(Y \mid A = 1, W)$ and $W \rightarrow P_0(A = 1 \mid W)$ are cadlag
3. $W \rightarrow E_0(Y \mid A = 1, W)$ and $W \rightarrow P_0(A = 1 \mid W)$ have finite sectional variation norm.

Coverage of zero-order HAL-TMLE when randomly selecting data distributions



Undersmoothed HAL-MLE is efficient uniformly over large class of target estimands

- HAL-MLE is efficient for pathwise differentiable target estimands, if L_1-norm is chosen large enough.
- Due to being an MLE, it solves a large class of score equations, in particular, efficient scores corresponding with target estimands.
- By undersmoothing enough, it uniformly solves a class of score equations that approximates all scores with finite variation norm. As a consequence, it is a globally efficient MLE across most pathwise differentiable features.
- This results can be applied to different assumed subspaces (i.e.., additive models of form $D^{(0)}(\mathcal{R}(d))$ of $D^{(0)}([0,1]^d)$.

Nonparametric Bootstrap of HAL-TMLE or HAL-MLE

- Fix M at the cross-validation selector M_n or another selector.
- Draw 10,000 samples of size n from empirical measure P_n. For each bootstrap sample $P_n^{\#}$, recompute the HAL-TMLE(M), say $\mathbf{P}_{n,M}^{\#*}$.
- The HAL on bootstrap sample can be restricted to only include indicator basis functions that were selected by HAL-MLE(M) on original data.
- Use sampling distribution of $\psi_{n,M}^{\#*} = \Psi(\mathbf{P}_{n,M}^{\#*})$, conditional on P_n, to construct 0.95-confidence interval.
- Increase M till plateau in confidence interval for optimal coverage.

Bootstrap works for HAL-TMLE or HAL-MLE

0-HAL-MLE
k-HAL-MLE
Asymptotic Normality
Efficient plug-in HAL
Bootstrapping HAL
Adaptive TMLE
O-A-HAL-C-TMLE
Efficient IPCW
Meta-HAL
Conclusion

- Conditional on the data $(P_n : n \geq 1)$, the bootstrap sampling distribution of ψ_n^* converges to optimal normal limit distribution $N(0, \sigma_0^2)$.
- The approximation error of bootstrap is driven by performance of nonparametric bootstrap for an empirical process indexed by Donsker class (i.e., cadlag functions with sectional variation norm bounded by M).
- This suggests robust finite sample behavior of the nonparametric bootstrap.

Case Study

Compare two confidence intervals for ATE $EY_1 - EY_0$:

1. Wald-type
2. HAL-TMLE bootstrap, using plateau selection of $L1$-norm in HAL.

Simulation

Setting:

- $W \sim N(0, 4^2, -10, 10)$ drawn i.i.d. from a truncated normal distribution, bounded within [-10,10].
- $A \sim Bernoulli(p(W))$ with probability p(W) as a function of W bounded between [0.3, 0.7], given by $p(W) = 0.3 + 0.1Wsin(0.1W) + \varepsilon$, $\varepsilon \sim N(0, 0.05^2)$
- $Y = 3sin(a_1 W) + A + \varepsilon_2$ is a sinusoidal function of W, where $\varepsilon_2 \sim N(0, 1)$.
- a_1 controls the frequency (and true sectional variation norm) of the sinusoidal function.

The value of the parameter of interest, the ATE, is 1.

- The experiment is repeated 500 times, and interval coverages are computed

Simulation for $n = 100$

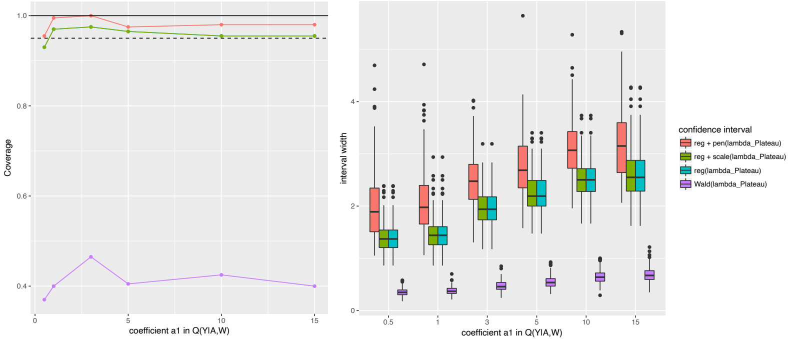


Figure: Coverage (left) and interval width (right) as a function of the a_1 coefficient (i.e., sectional variation norm) of the true data-generating distribution.

Super-efficient estimation for smooth features of target function

- The smoothness adaptive HAL (using cross-validation to select k and additive submodel $D^{(k)}(\mathcal{R}_{j_0})$) will asymptotically act like an HAL for an oracle smoothness k_0 and oracle subspace $D^{(k_0)}(\mathcal{R}_{j_0}(d))$.
- As a consequence, by undersmoothing within the oracle subspace it will be plug-in efficient for smooth target features w.r.t. the oracle statistical model on the data distribution that assumes $Q_0 \in D^{(k_0)}(\mathcal{R}_{j_0}(d))$.
- That is, $\Psi(Q_n)$ is asymptotically super-efficient, and is an example of the Adaptive MLE of (Lars van der Laan et al., 2023).

Finite sample robust TMLE

- The current literature on TMLE has proposed various modifications of TMLE that regularize the TMLE to be better behaved in finite samples when the support is limited (i.e., practical violation of positivity assumption).
- For example, in censored and causal inference literature: adaptive truncation; regularize TMLE step by not fully solving EIC-equation; collaborative TMLE; outcome-adaptive TMLE; super-efficient TMLE (adjusting in PS for outcome regressions $Q_n(1, W)$, $Q_n(0, W)$, beyond a possible baseline model).
- These proposed regularizations have in common that they all concern targeted estimation of the orthogonal nuisance function $g(P_0)$ that is needed in targeting step, while $\Psi(P_0) = \Psi_1(Q_0)$ only depends on certain factors of likelihood.

- Typically, these variations preserve asymptotic efficiency but adapt the targeting step towards the data to carefully trade-off bias reduction with variance gain.
- Finite sample simulations, theoretical results such as collaborative double robustness etc, have shown that these regularizations are crucial for robust finite sample behavior for poorly supported parameters.
- However, we never developed a truly unifying approach (and super-efficiency theory)!
- Fortunately, Lars did: Lars van der Laan et al. (2023), Adaptive debiased machine learning.

Limitations of Efficient Estimators such as TMLE

- Due to lack of nonparametric support, if $\mathcal{M}$ is close to nonparametric, then $D^*_{P_0}(o)$ can have very large values.
- In that case, confidence intervals for standard TMLE will be wide and no significant finding can be obtained.
- However if we would use plug-in HAL $\Psi(\mathbf{P}_n)$ (possibly super-efficient) we might very well still find a significant true result. Somehow, the targeting step can blow up a good initial estimator.
- This is due to an efficient estimator to be regular along all paths through P_0, including paths that appear to be contradicting the data.

An efficient estimator cannot adapt to structure in the true data distribution

- For example, suppose $O = (W, A, Y)$ and $\Psi(P) = E_P E_P(Y \mid A = 1, W)$. A plug-in HAL might end up fitting an additive model for the regression $E(Y \mid A, W)$ that contains the true regression.
- An efficient estimator of $\Psi(P_0)$ for this additive model would be well supported and have reasonable variance.
- But an efficient estimator for a nonparametric model protects itself against any kind of fluctuation of the data distribution, and as a consequence carries out a bias reduction that (**asymptotically**) holds up uniform in P in a ball around P_0.
- In other words, it needs to remain asympotically unbiased under fluctuations adding a 100 way interaction.

Lets reset benchmarks for our estimator

- **Super-efficient estimator:** Suppose that we require that our estimator of $\Psi(P_0)$ is asymptotically linear at any $P_0 \in \mathcal{M}$, and regular along paths through P_0 that stay in an oracle model $\mathcal{M}_0 \subset \mathcal{M}$, approximated by a data adaptive submodel $\mathcal{M}_n$ satisfying $P_0 \in \mathcal{M}_0$.
- Then, we can construct $\mathcal{M}_0$-super-efficient estimators that still provide asymptotically valid confidence interval and are still robust under perturbations of P_0 that stay in the model $\mathcal{M}_0$.
- **Regularized efficient estimator:** If our data adaptive model $\mathcal{M}_n$ approximates $\mathcal{M}_0 = \mathcal{M}$, let the estimator behave as an efficient estimator under model $\mathcal{M}_n$ that approximates the a-priori specified model $\mathcal{M}$ as sample size converges to infinity but in a way that carefully balances finite sample bias and variance.

Adaptive TMLE

- Data adaptively learn a model $\mathcal{M}_n \subset \mathcal{M}$. Make sure that $d(P_0, \mathcal{M}_n) = o_P(n^{-1/4})$ (e.g., use HAL, or cross-validation selection among a large family of submodels).
- Define the projection parameter $\Psi_{\mathcal{M}_n} : \mathcal{M} \rightarrow \mathbb{R}$ defined by $\Psi_{\mathcal{M}_n}(P_0) = \Psi(\Pi_{\mathcal{M}_n} P_0)$, where $P_{0,n} \equiv \Pi_{\mathcal{M}_n}(P_0) = \arg\min_{P \in \mathcal{M}_n} P_0 L(P)$ is the log-likelihood (or loss based) projection of P_0 onto $\mathcal{M}_n$.
- Construct a TMLE $\Psi_{\mathcal{M}_n}(P_n^*)$ of $\Psi_{\mathcal{M}_n}(P_0)$ based on canonical gradient $D^*_{\mathcal{M}_n,P}$ of this projection parameter. Generally speaking, for log-likelihood loss functions $\Psi_{\mathcal{M}_n}(P_n^*) = \Psi(P_n^*)$ with $P_n^* \in \mathcal{M}_n$.
- Provide confidence intervals for $\Psi_{\mathcal{M}_n}(P_0)$ as usual based on $D_{\mathcal{M}_n,P_0}$.
- Possibly cross-fit, like CV-TMLE.

Why does A-TMLE work: 1) standard TMLE analysis

- From TMLE analysis we will have that $\Psi_{\mathcal{M}_n}(P_n^*) - \Psi_{\mathcal{M}_n}(P_0)$ behaves as $(P_n - P_0)D^*_{\mathcal{M}_n,P_0}$. Therefore, under an asymptotic stability condition on $\mathcal{M}_n$ so that $D^*_{\mathcal{M}_n,P_0} \rightarrow_p D^*_{\mathcal{M}_0,P_0}$ we have that it behaves as $P_n D^*_{\mathcal{M}_0,P_0}$ and is thus asymptotically normal with mean zero and variance $\sigma^2_{\mathcal{M}_0}(P_0) = P_0\{D^*_{\mathcal{M}_0,P_0}\}^2$.
- Cross-fitting weakens this need for asymptotic stability. Moreover, one might still have asymptotic normality by standardizing by a variance estimator.
- $D^*_{\mathcal{M}_0,P_0}$ equals the efficient influence curve of $\Psi : \mathcal{M}_0 \rightarrow \mathbb{R}$: i.e. we achieve the efficiency we would achieve with TMLE if we would a priori know that $P_0 \in \mathcal{M}_0$.

Why does A-TMLE work: 2) data adaptive model bias negligible

- Let $R_{\mathcal{M}_0}(P,P_0)=\Psi_{\mathcal{M}_0}(P)-\Psi_{\mathcal{M}_0}(P_0)+P_0D^*_{\mathcal{M}_0,P}$.
- We have

$$\begin{aligned}&\Psi_{\mathcal{M}_n}(P_0)-\Psi(P_0)=\\&(P_{0,n}-P_0)\{D^*_{\mathcal{M}_0,P_{0,n}}-\Pi_n(D^*_{\mathcal{M}_0,P_{0,n}}\mid T_{\mathcal{M}_n}(P_{0,n}))\}\\&\qquad\qquad+R_{\mathcal{M}_0}(P_{0,n},P_0),\end{aligned}$$

where the projection Π_n projects $D^*_{\mathcal{M}_0,P_{0,n}}$ onto tangent space of $\mathcal{M}_n$ at $P_{0,n}$.

- This is a very nice second order remainder (i.e., $o_P(n^{-1/2})$).
- Therefore, our adaptive TMLE is asymptotically linear estimator of $\Psi(P_0)$ with (super-efficient) influence curve $D^*_{\mathcal{M}_0,P_0}$.
- Since it operates as an efficient estimator of $\Psi_{\mathcal{M}_n}(P_0)$ it will also be regular along any path through P_0 that stays in the limit oracle model $\mathcal{M}_0$.

Outcome-regression weighted LASSO (OAL)

Shortreed & Artefaie (2017) proposed outcome-regression weighted Lasso (OAL) for propensity score (PS) estimation:

- Fit unpenalized linear model for $\mathbb{E}(Y|A,W)$:

$$(\hat{\alpha}, \hat{\eta}) = \arg\min_{\alpha,\eta} l_n(\alpha; Y, A, W).$$

  where η is the coefficient for A, and α is the coefficient for W.

- Denote the coefficient for variable W_j in the outcome regression with $\hat{\alpha}_j$.
- Fit PS with Lasso using regularization term

$$\lambda \sum_j ||\alpha_j||^{-\gamma}||\beta_j|| \text{ instead of usual } \lambda \textstyle\sum_j \mid \beta_j \mid.$$

HAL-based OAL for PS Estimation

The theoretical property of OAL relies on the ***correct parametric formula***, which is often unknown in practice.

We extend OAL to outcome-regression weighted HAL (OHAL):

1. Compute the outcome regression using Lasso of form $\sum_j \alpha_n(j)\phi_j(W, A)$.
2. Use as basis functions for the PS $\{\phi_j(1, W), \phi_j(0, W) : j\}$.
3. Both of these two basis functions will be associated with same weight $\alpha_n(j)$.
4. Compute the propensity score using a Lasso logistic regression using the above basis functions. The L_1-constraint for β_j's, the coefficient for ϕ_j, is defined as the weighted L_1-norm above.
5. Or simply define $\parallel \beta \parallel_1 = \sum_{j,\alpha_n(j)\neq 0} \mid \beta(j) \mid$.

C-TMLE to select L_1-norm

- Tune the L_1-norm with C-TMLE: i.e. optimize in λ increase in likelihood of TMLE-step using $g_{n,\lambda}$.

OHAL Performance based on Kang & Shafer (2007) Simulation

- Pre-treatment covariates $(Z_{i1}, \cdots, Z_{i4})$ are generated from uncorrelated standard normal distributions.
- Treatment indicator is then generated from a Bernoulli distribution with:

$$P(A_i = 1|Z_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4})$$

- Potential outcomes are generated by:

$$Y_i^{(a)} = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4} + \epsilon$$
$$\epsilon \sim N(0,1)$$

- Thus, the value of the estimand, the ATE, is 0.

Simulation to Assess OHAL Performance (Continued)

- Only transformed covariates W are observed:

$$
\begin{aligned}
W_{i1} &= \exp(Z_{i1}/2) \\
W_{i2} &= z_{i2}/(1+\exp(Z_{i1})+10 \\
W_{i3} &= (Z_{i1}Z_{i3}/25+0.6)^3 \\
W_{i4} &= (Z_2+Z_4+20)^2.
\end{aligned}
$$

- In our experiment, we also included an instrumental variable W_{i5}, and the PS was modified to:

$$PS = \text{expit}\left(\frac{-Z_1+0.5Z_2-0.25Z_3-0.1Z_4}{2}+W_5\right)$$

OHAL Simulation Results

We use main-term linear regression to create biased initial estimator $\bar{Q}_n^0$ for TMLE/C-TMLE.

	TMLE-HAL	CTMLE-HAL	CTMLE-OHAL	Oracle
N=500	7.06	6.34	**2.94**	3.64
N=1000	4.42	3.55	**1.40**	1.70
N=2000	2.94	1.85	**0.83**	0.87

Table: MSE for each estimator across 200 replications with different sample size.

Inverse Probability of Treatment/Censoring Weighted Estimators using HAL (Ertefaije, Hejazi, vdL, 2023)

- Let $O = \Phi(C, X)$, $P_X \in \mathcal{M}^F$, $G_{C|X}$ satisfying CAR, and let $\Psi^F(P_X)$ be the target.
- One can then construct IPCW-estimators $\psi_{n,IPCW}$ as solutions of $P_n U_{G_n}(D^F(\psi)) = 0$ in ψ with $U_G(D^F)(O)$ an inverse probability weighted full-data estimating function D^F satisfying $E_G(U_G(D^F)(O) \mid X) = D^F(X)$.
- By using undersmoothed HAL as estimator of the censoring mechanism G, the HAL-IPCW estimator $\psi_{n,IPCW}$ is asymptotically linear with influence curve the IPCW-function $U_G(D^F(\psi))$ minus its projection on the tangent space T_{CAR}.

- In particular, for nonparametric full-data models, the IPCW-estimator using undersmoothed HAL is asymptotically efficient.
- Similarly, if one uses HAL to estimate G for double robust estimators, then inference and efficiency is preserved under misspecified estimators of the full data distribution P_X.

Meta-HAL Super Learning (Wang, Zhang, vdL, 2024)

- Given a library of candidate machine learning algorithms, one can construct a cross-fitted data set in which the covariates are the cross-fitted predictions.
- One can then use HAL at this Meta-level.
- This is the same as using HAL to do super-learning instead of convex linear regression.
- One can include this Meta-HAL SL as candidates in a discrete super-learner to tune the meta-level HAL optimally.
- Meta-HAL satisfies the same properties at HAL, but at the Meta-level: rates of convergence; plug-in efficiency; asymptotic normality.
- This provides a powerful approach for targeting learning incorporating multimodal data sources, utilizing algorithms

Concluding Remarks

- HAL is the first general nonparametric MLE, with formal theoretical properties analogue to parametric MLE.
- It represents a large class of HAL estimators by restricting function space, a priori, or data adaptively (screening).
- Its sparse representation makes it an excellent algorithm for interpretable machine learning.
- Fast dimension free rate of convergence, even uniform convergence, and has pointwise normal limit distribution.
- Can be made adaptive to unknown smoothness and support.
- HAL-TMLE is guaranteed asymptotically efficient for smooth features.

- Smoothness and subspace adaptive HAL super-efficient for smooth features, under conditions of the A-MLE/A-TMLE of Lars vdL et al. (2023).
- It yields delta-method based Wald type confidence intervals, and it allows for nonparametric bootstrap to improve finite sample coverage.
- It has many applications to further enhance TMLE beyond using HAL in discrete SL: e.g., outcome adaptive HAL-TMLE, higher order TMLE, Collaborative TMLE, etc.