Causal Machine Learning Masterclass

Session 3 (14:45 – 16:00) & Session 4 (16:15 – 17:30)

Mark van der Laan Rachael Phillips

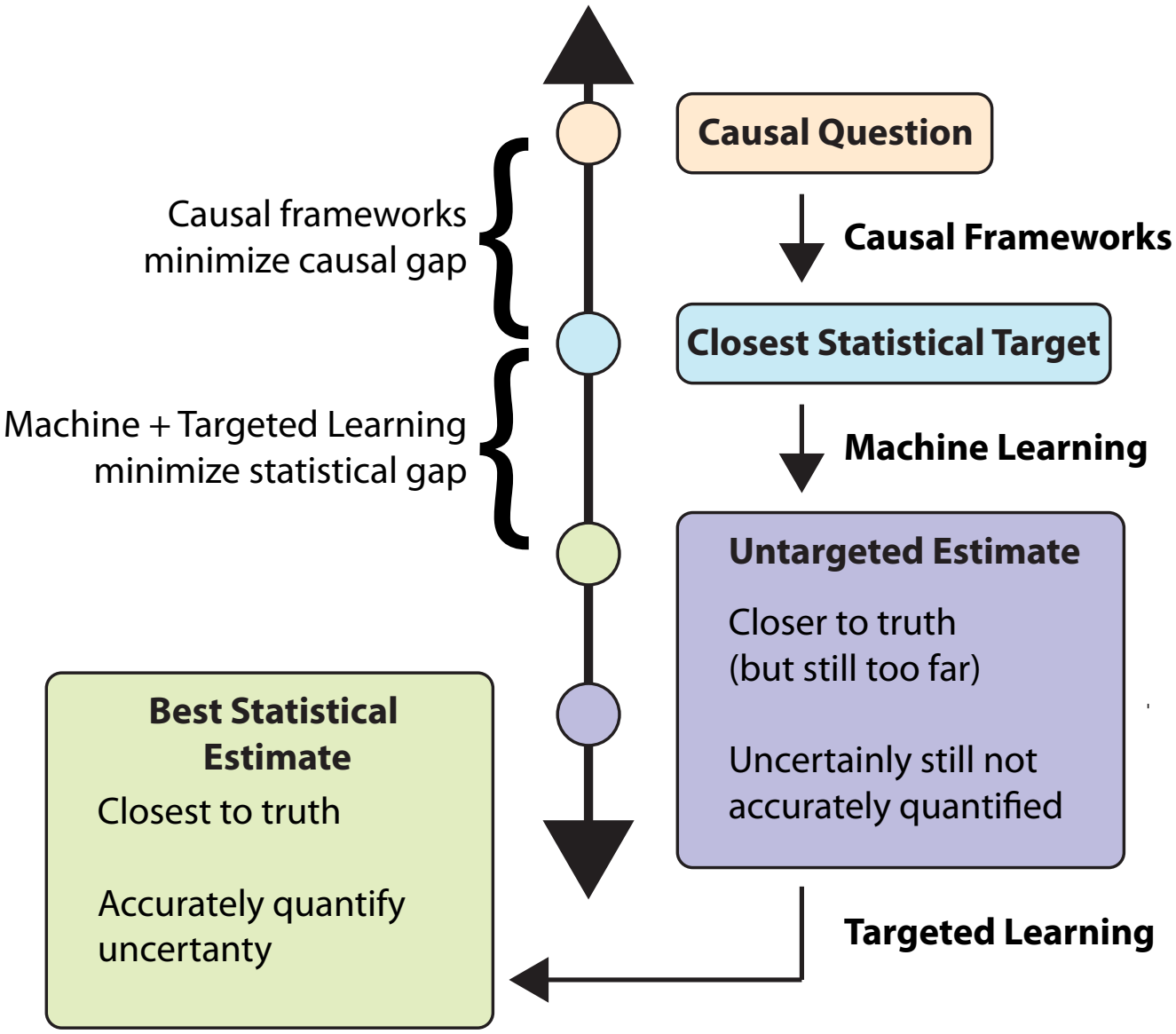
University of California, Berkeley

3 March 2020

Alan Turing Institute in collaboration with the Centre for Statistical Methodology and London School of Hygiene and Tropical Medicine

Outline

Targeted Learning for answering statistical and causal questions with confidence intervals

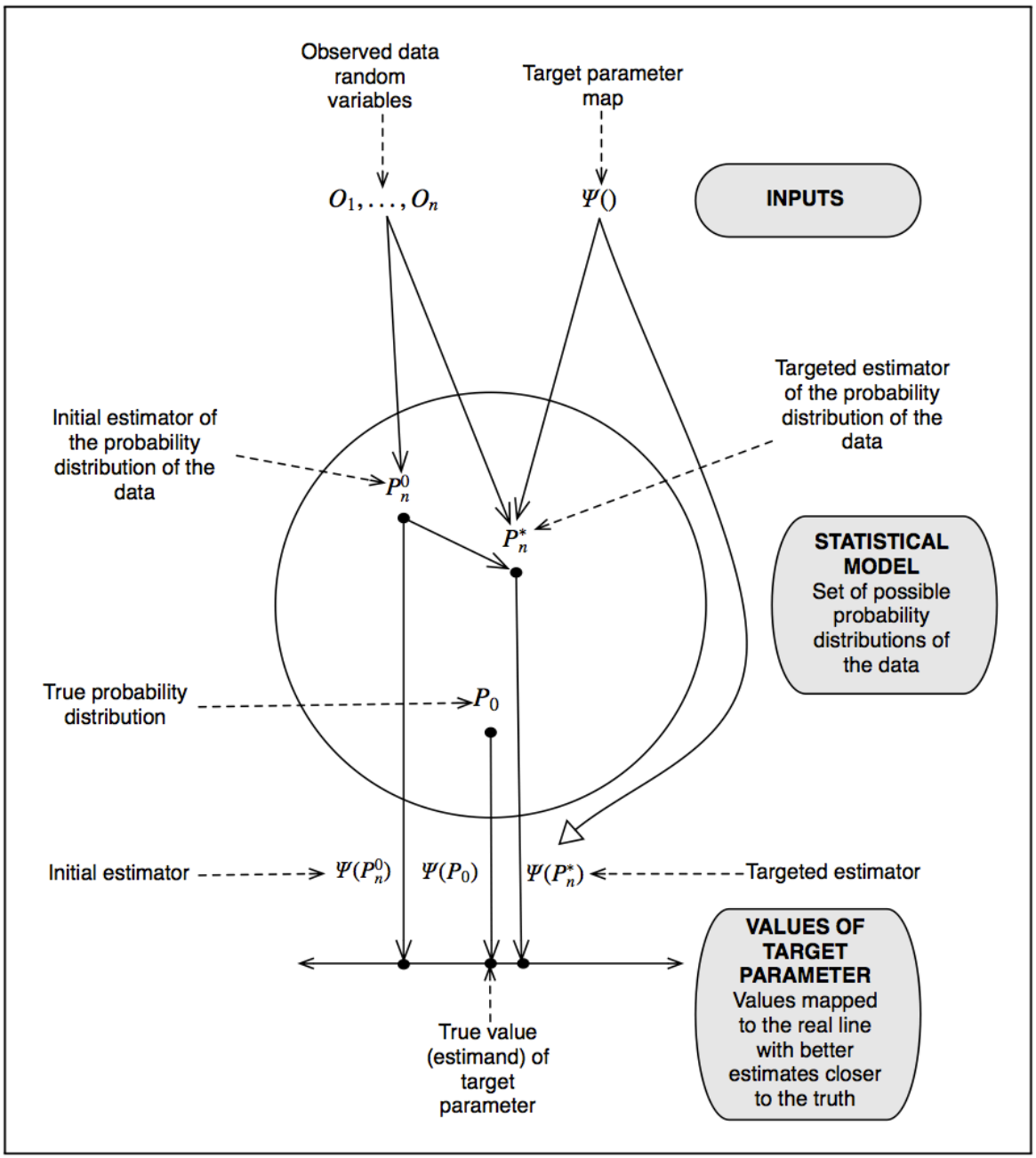


Targeted Update of Machine Learning

- Don't try to do a good job for all questions at once.
- Focus estimation where it matters most for question at hand.

1. Less bias (closer to truth).
2. Sampling distribution approximately normal, more accurate quantification of uncertainty.

Targeted Minimum Loss Based Estimation (TMLE)



Targeted Minimum Loss Based Estimation (TMLE)

- Super learning provides an initial estimator $\mathbf{P}_n$ of P_0.
- Determine mathematically the fluctuation strategy (least favorable submodel) $\mathbf{P}_{n,\epsilon}$ of the super-learner fit $\mathbf{P}_n$ with tuning parameter ϵ, so that, for a small fluctuation of ϵ, **the square change in estimated answer $\Psi(\mathbf{P}_{n,\epsilon})$ per increase in log-likelihood** is maximized: i.e., score at ϵ equals canonical gradient/**efficient influence curve** $D^*(\mathbf{P}_{n,\epsilon})$.
- Determine the optimal amount ϵ_n of fluctuation based on the data (e.g., maximum likelihood estimation).
- The resulting update $\mathbf{P}_n^* = \mathbf{P}_{n,\epsilon_n}$ of the initial estimator of stochastic system is the TMLE of P_0 and it implies the TMLE $\Psi(\mathbf{P}_n^*)$ of the answer to query.
- Thanks to TMLE update, TMLE solves optimal score equation $P_n D^*(\mathbf{P}_n^*) \approx 0$, and is asymptotically normally distributed around true answer to query with minimal asymptotic variance.

Three general methods for efficient estimation in literature

Three general methods result in asymptotically efficient estimators, given good initial estimator $\mathbf{P}_n$ of data distribution P_0, using canonical gradient $D^*(P)$ of target estimand as ingredient:

- **One-step estimator:** $\psi_n^1 = \Psi(\mathbf{P}_n) + P_n D^*(\mathbf{P}_n)$.
- **Estimating equation estimator:** Assume estimating function representation $D^*(P) = D^*(\psi, \eta(P))$; let ψ_n solution of $P_n D^*(\psi, \eta(\mathbf{P}_n)) = 0$.
- **TMLE:** $\mathbf{P}_{n,\epsilon}$ least favorable submodel through initial $\mathbf{P}_n$; ϵ_n MLE; $P_n^* = \mathbf{P}_{n,\epsilon_n}$; TMLE is $\Psi(P_n^*)$.
- TMLE is general method that updates initial $\mathbf{P}_n$ into improved fit $\mathbf{P}_n^*$ that solves **user supplied set of equations** $P_n D(\mathbf{P}_n^*) \approx 0$, allowing for various additional statistical properties beyond asymptotic efficiency.

Each one of the methods has a sample splitting analogue removing Donsker class condition.

Outline

Highly Adaptive Lasso (HAL)

Key idea

- Any d-dimensional cadlag function (i.e. right-continuous) can be represented as a possibly infinite linear combination of spline basis functions.
- The variation norm / complexity of a function is the L_1-norm of the vector of coefficients.

Converges to true function at rate $n^{-1/3}(\log n)^{d/2}$.

Representation of cadlag function as linear combination of indicators

For a cadlag function $\psi : [0, \tau] \subset \mathbb{R}^d \to \mathbb{R}$ with finite variation norm (and thus generates a signed measure), we have

$$\psi(x) \quad = \quad \sum_{s \subset \{1,\ldots,d\}} \int I(x_s \geq u_s) d\psi_s(u_s),$$

where $\psi_s(u) = \psi(u_s, 0_{s^c})$ is the section of ψ that sets the coordinates in s equal to zero. Here $x_s = (x(j) : j \in s)$ and the sum is over all subsets of $\{1, \ldots, d\}$.

Variation norm

The variation norm of ψ can be defined as:

$$\| \psi \|_v = \sum_{s \subset \{1,\ldots,d\}} \int | d\psi_s(u_s | .$$

Representation for discrete cadlag functions

For discrete measures $d\psi_s$ with support points $\{u_{s,j} : j\}$ one obtains the following linear combination of indicator basis functions:

$$\psi(x) = \sum_{s \subset \{1,\ldots,d\}} \sum_{j} \beta_{s,j} \phi_{u_{s,j}}(x),$$

where $\beta_{s,j} = d\psi_s(u_{s,j})$, and

$$\begin{aligned} \| \psi \|_v &= \sum_{s \subset \{1,\ldots,d\}} \sum_{j} | \beta_{s,j} | \\ &\equiv \| \beta \|_1 . \end{aligned}$$

Highly Adaptive Lasso

Consider a loss function $L(\psi)$ such as $L(\psi)(X, Y) = (Y - \psi(X))^2$, let $\psi_0 = \arg\min_\psi P_0 L(\psi)$ and let

$$d_0(\psi, \psi_0) = P_0 L(\psi) - P_0 L(\psi_0)$$

be the loss-based dissimilarity. Consider the constrained MLE:

$$\psi_{n,M} = \arg \min_{\psi, \|\psi\|_v < M} P_n L(\psi).$$

Highly Adaptive Lasso

Given that this MLE is attained at a discrete measure $d\psi_{n,M}$, this MLE is given by $\psi_{n,M} = \sum_{s \subset \{1,\ldots,d\}} \beta_{n,M,s,j} \phi_{u_{s,j}}$, where

$$\beta_{n,M} = \arg \min_{\beta, \|\beta\|_1 < M} \frac{1}{n} \sum_{i-1}^{n} (Y_i - \sum_{s,j} \beta_{s,j} \phi_{u_{s,j}}(X_i))^2.$$

In other words, $\beta_{n,M}$ is computed with the Lasso.

Cross-validation to select variation norm

As in the Lasso, we select M with cross-validation. Let M_n be the cross-validation selector and

$$\psi_n = \psi_{n,M_n}.$$

We refer to ψ_n as the Highly Adaptive Lasso estimator (HAL-E).

Guaranteed rate of convergence $n^{-1/3}$

We have

$$d_0(\psi_{n,M}, \psi_{0,M}) = o_P(n^{-2/3}(\log n)^d).$$

Thus, if we select $M > \| \psi_0 \|_v$, then

$$d_0(\psi_{n,M}, \psi_0) = o_P(n^{-2/3}(\log n)^d).$$

Due to oracle inequality for the cross-validation selector M_n, as long as $\| \psi_0 \|_v < \infty$, we have

$$d_0(\psi_n = \psi_{n,M_n}, \psi_0) = o_P(n^{-2/3}(\log n)^d).$$

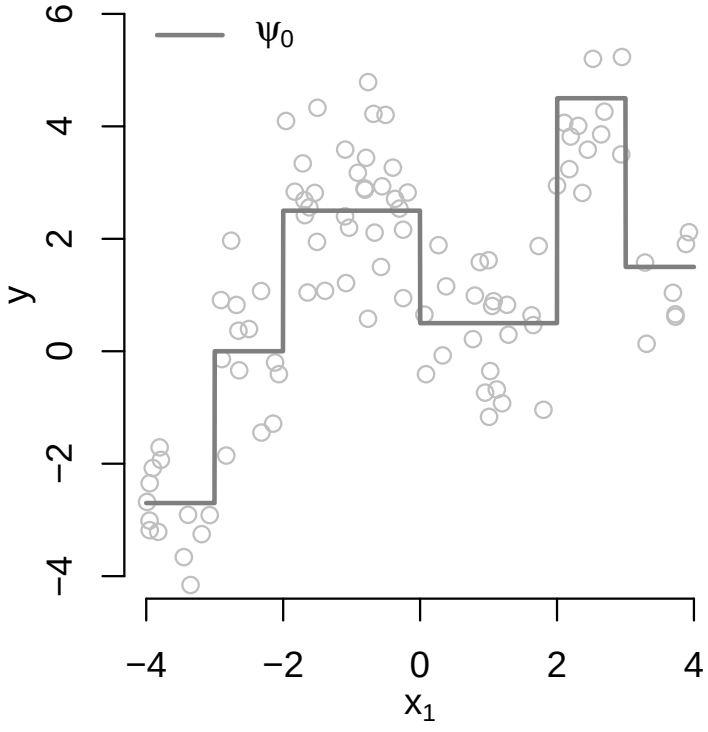
Super Learner should use HAL to guarantee this minimal rate of convergence: HAL-SL

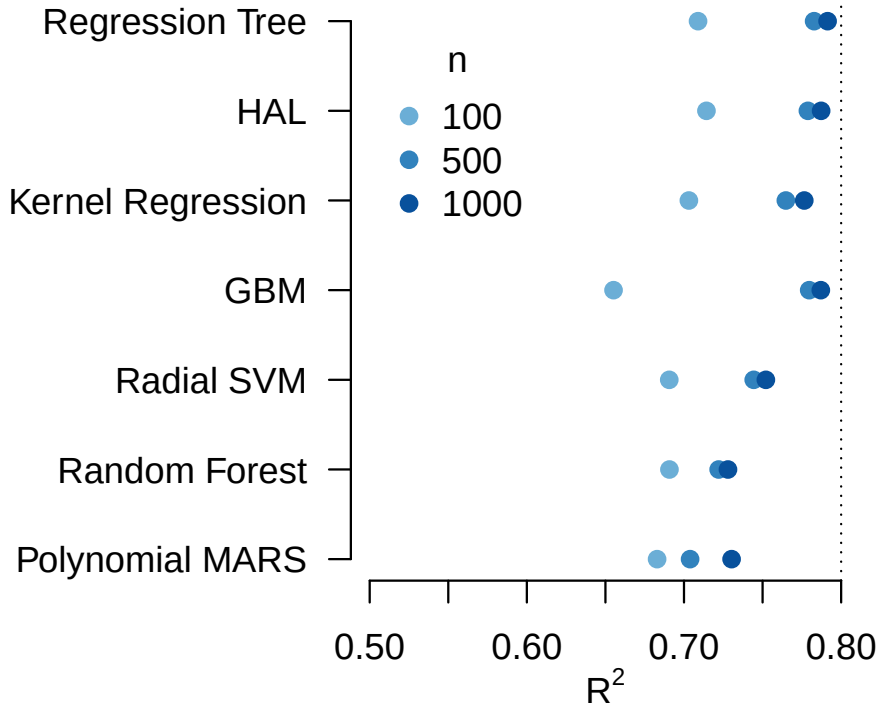
HAL Performance

For each simulations 20 data sets of sample size n were drawn from P_0. Each data generating mechanism was chosen such that the optimal $R^2 = 0.8$.

HAL was evaluated against competitor algorithms based on R^2 calculated on an independent evaluation data set of size 10,000 averaged across the 20 data sets.

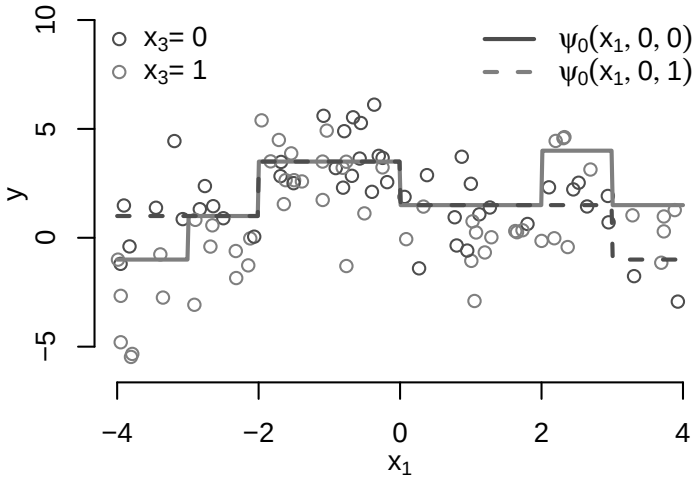
Jump functions, $d = 1$

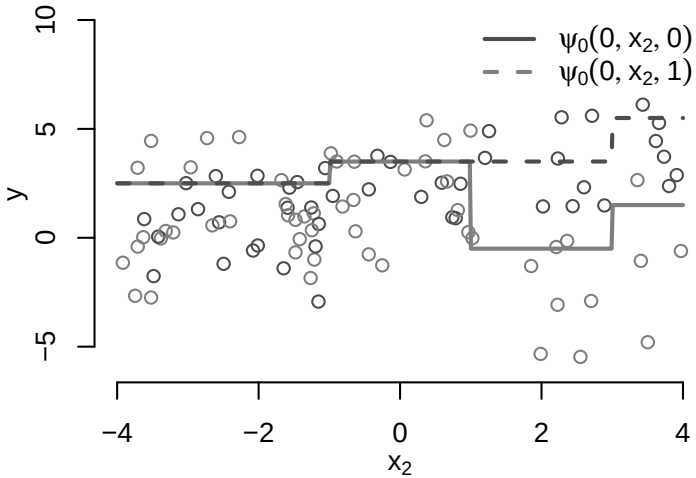


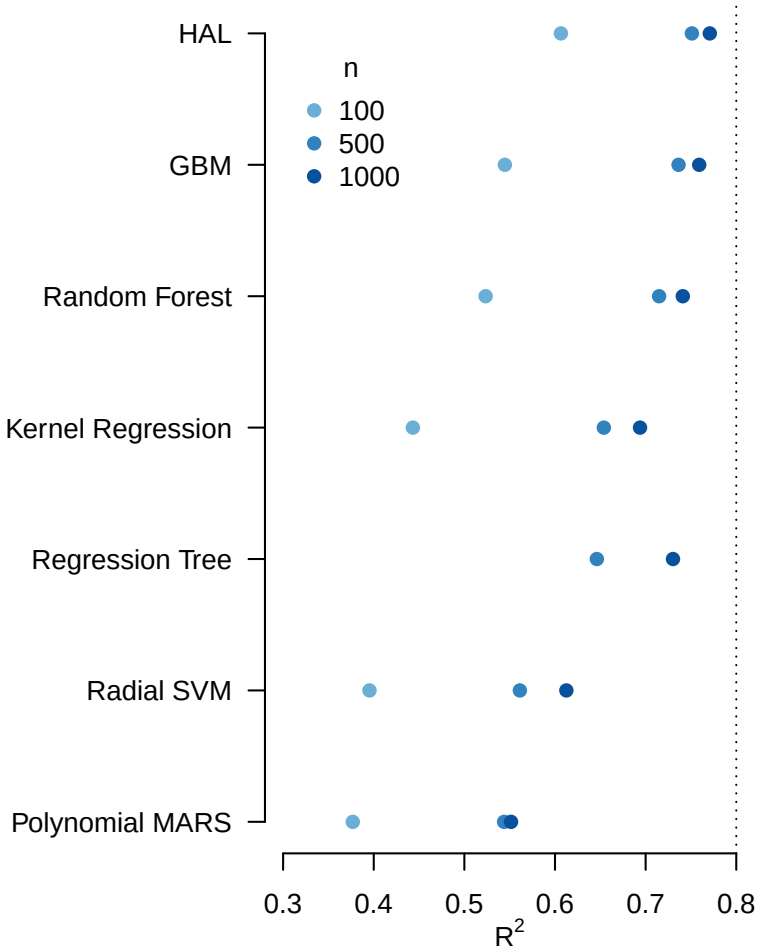


Jump functions, $d = 3$

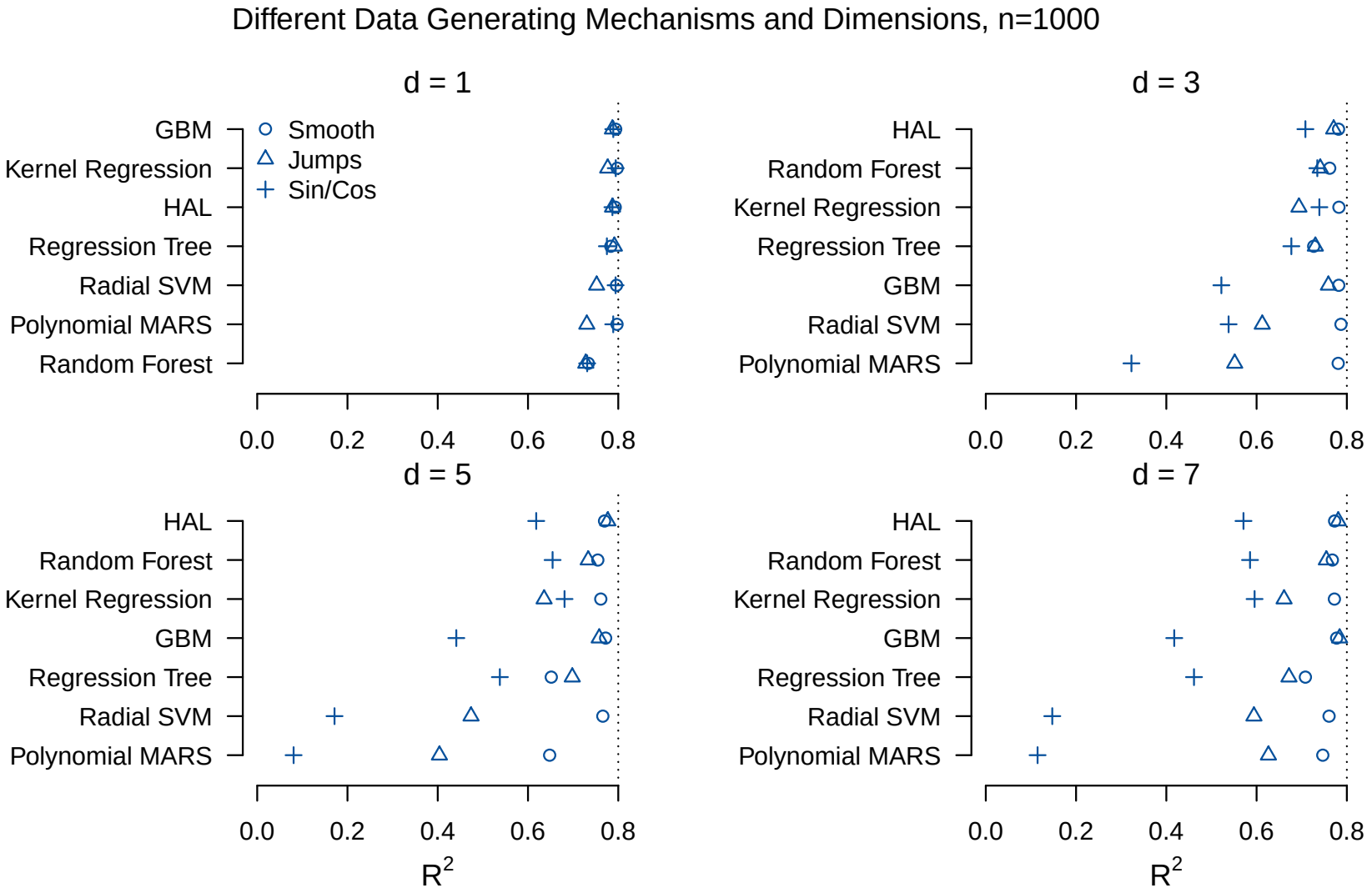
$$\psi_0(x) = -2x_3\, I(x_1 < -3) + 2.5\, I(x_1 > -2) - 2\, I(x_1 > 0) + 2.5x_3\, I(x_1 > 2)$$
$$-2.5\, I(x_1 > 3) + I(x_2 > -1) - 4x_3\, I(x_2 > 1) + 2\, I(x_2 > 3)$$







Overall Performance



Objective simulation with HAL-TMLE of ATE

We repeatedly sampled random data generating mechanisms and simulated samples of size $n \in \{100, 500, 1000, 2000\}$ for a total of 25,000 different data generating mechanisms of (W, A, Y). We computed TMLEs of the

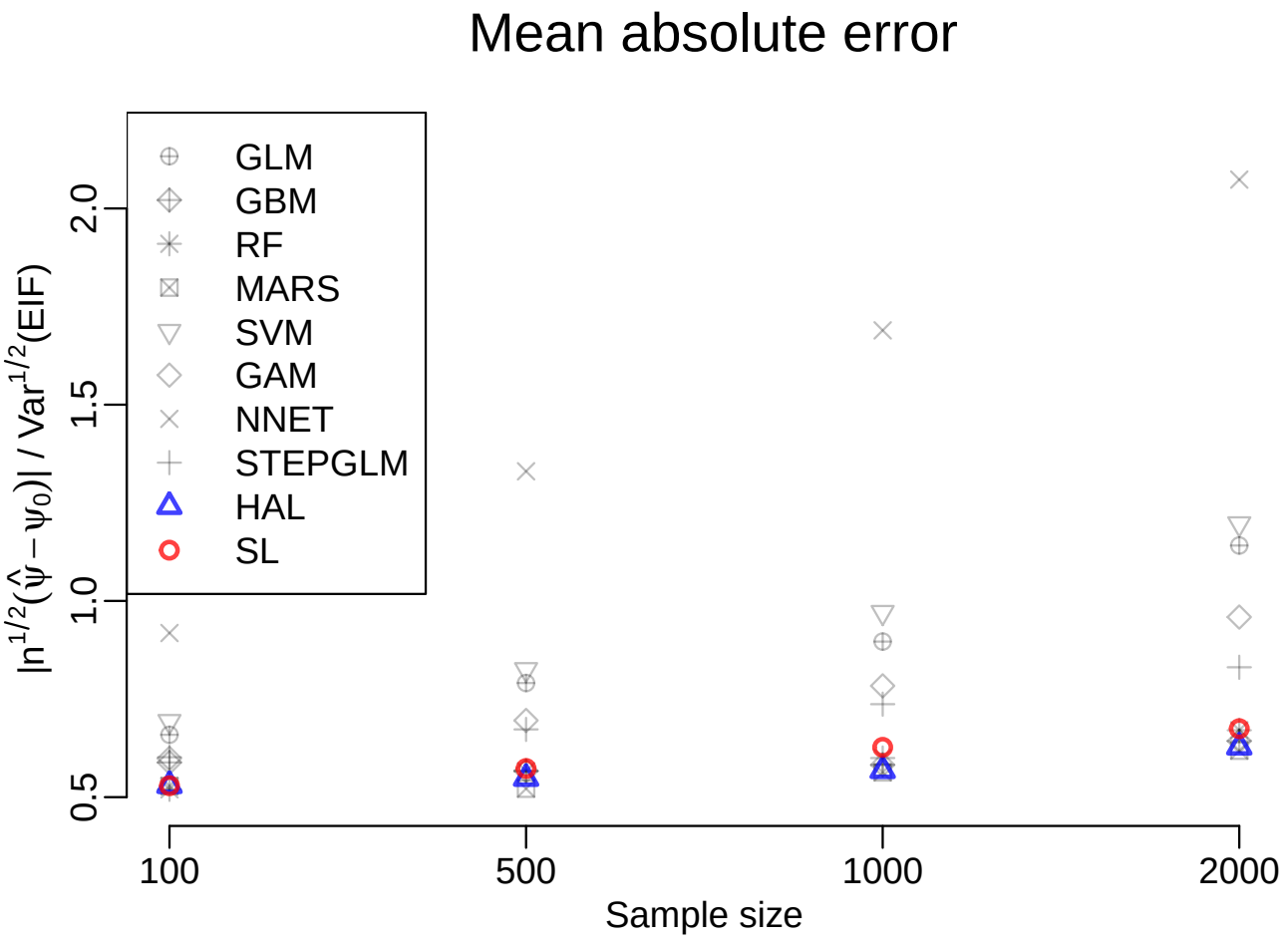
ATE based on different estimators of $E_0(Y \mid A, W)$ and $P_0(A = 1 \mid W)$:

- GLM, Bayes GLM, stepwise GLM (AIC), stepwise GLM (p-value), stepwise GLM with two-way interactions, intercept-only GLM, GAM, GBM*, random forest*, linear SVM*, neural nets*, regression trees*, HAL
- Super Learner (based on these algorithms)
- * = tuning parameters selected via cross-validation

Estimators compared on their absolute error (relative to best achievable SE) and coverage probability of 95% oracle confidence intervals.

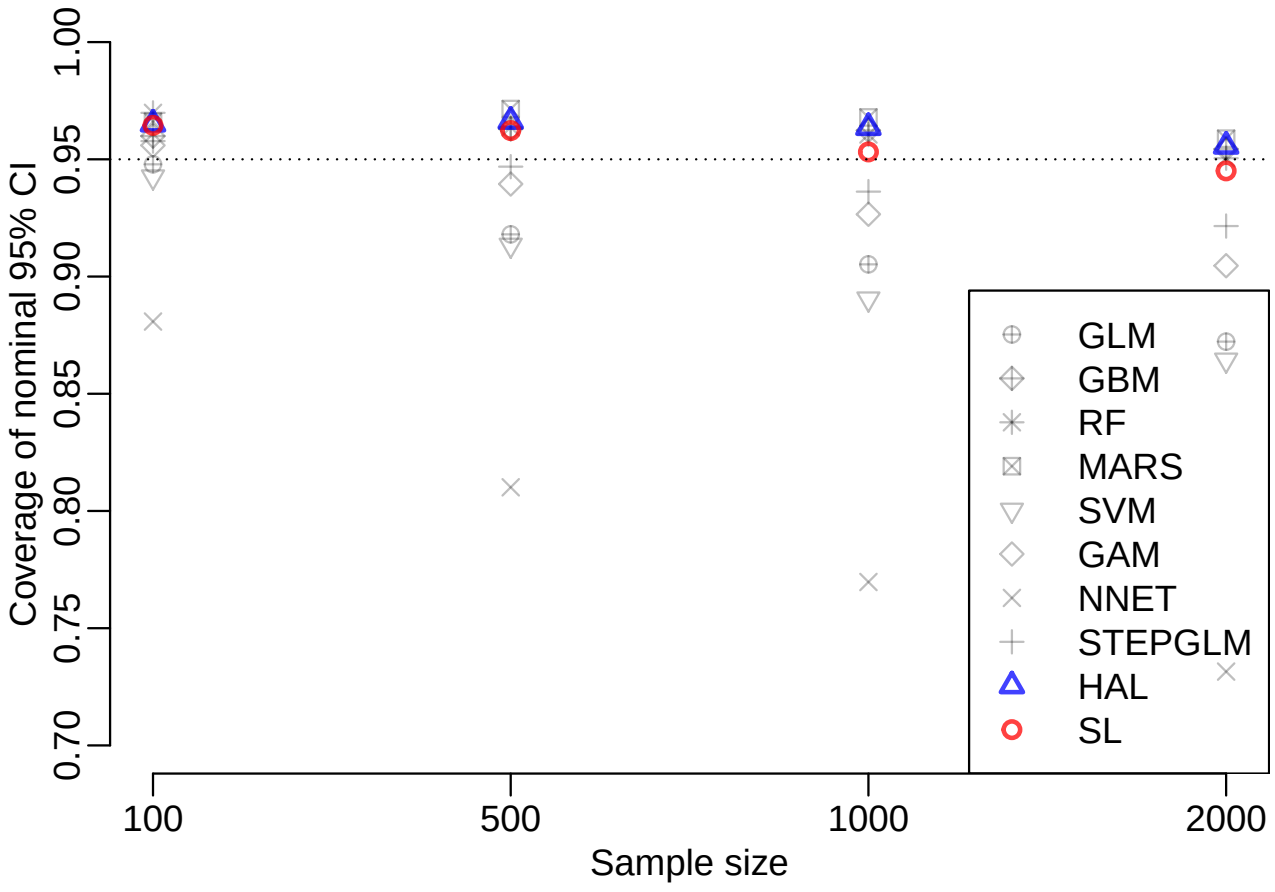
Results – absolute error by sample size

HAL-TMLE exhibited excellent accuracy relative to competitors.



Results – coverage by sample size

HAL-TMLE achieves approximate Normality in reasonable sample sizes.



Outline

General longitudinal data structure

We observe n i.i.d. copies of a longitudinal data structure

$$O = (L(0), A(0), \ldots, L(K), A(K), Y = L(K+1)),$$

where

- $A(t)$ denotes a discrete valued **intervention node** whose effect we desire to evaluate
- $L(t)$ is an **intermediate covariate/outcome** realized in between intervention nodes $A(t-1)$ and $A(t)$, $t = 0, \ldots, K$
- Y is the **final outcome** of interest

Survival outcome example

$$
\begin{aligned}
A(t) &= (A_1(t), A_2(t)) \\
A_1(t) &= \text{Indicator of being treated at time } t \\
A_2(t) &= \text{Indicator of being right-censored at time } t \\
Y(t) &= \text{Indicator of observing a failure by time } t \\
L(t) &= \text{Vector of time-dependent measurements} \\
Y(t) &\subset L(t) \text{and } Y = Y(K+1).
\end{aligned}
$$

Outline

Outline

Likelihood and statistical model

The probability distribution P_0 of O can be factorized according to the time-ordering as

$$\begin{aligned} p_0(O) &= \prod_{t=0}^{K+1} p_0(L(t) \mid Pa(L(t))) \prod_{t=0}^{K} p_0(A(t) \mid Pa(A(t))) \\ &\equiv \prod_{t=0}^{K+1} q_{0,L(t)}(O) \prod_{t=0}^{K} g_{0,A(t)}(O) \\ &\equiv q_0 g_0, \end{aligned}$$

where $Pa(L(t)) \equiv (\bar{L}(t-1), \bar{A}(t-1))$ and $Pa(A(t)) \equiv (\bar{L}(t), \bar{A}(t-1))$ denote the parents of $L(t)$ and $A(t)$, respectively.

The g_0-factor represents the intervention mechanism.

Statistical Model: We make no assumptions on q_0, but could make assumptions on g_0.

Target estimand

- $p_0^{g^*} = q_0(o)g^*(o)$ is the G-computation formula for the post-intervention distribution of O under the stochastic intervention $g^* = \prod_{t=0}^{K} g^*_{A(t)}(O)$.

- Target estimand $\Psi(P) = E_{P_{g^*}} Y$, i.e., mean outcome under P_{g^*}.

Sequentially integration out $L(k+1)$ and $A(k)$, going backwards

- Let $\bar{Q}_{L(K+1)} = E_P(Y \mid \bar{A}(K), \bar{L}(K))$.
- Let $\bar{Q}_{A(K)} = E_{g_{A^*(K)}} \bar{Q}^{g^*}$.
- Let $\bar{Q}_{L(K)} = E_{Q_{L(K)}} \bar{Q}_{A(K)}$.
- Let $barQ_{A(K-1)} = E_{g_{A^*(K-1)}} \bar{Q}_{L(K)}$.

Iterate till all variables are integrated out

- Set $k = K - 1$. We just evaluated $\bar{Q}_{A(k)}$.
- Let $\bar{Q}_{L(k)} = E_{Q_{L(k)}} \bar{Q}_{A(k)}$.
- Let $\bar{Q}_{A(k-1)} = E_{g_{A^*(k-1)}} \bar{Q}_{L(k)}$.
- Let $k = k - 1$ and repeat above 2 steps, and iterate till we obtain $\bar{Q}_{A(0)}(L(0))$.

Sequential regression representation of target parameter

- Let $\bar{Q}_{L(0)} = E_{Q_{L(0)}} \bar{Q}_{A(0)}$, marginal expectation over $L(0)$.
- Then $\Psi_{g^*}(P) = \Psi_{g^*}(Q) = \bar{Q}_{L(0)}$.
- Note that $\Psi_{g^*}(P)$ depends on P through $\bar{Q} = (\bar{Q}_{L(0)}, \ldots, \bar{Q}_{L(K+1)})$.
- All the conditional regressions $\bar{Q}_{A(k)}$ are w.r.t. known stochastic intervention, and do thus not represent parameter of P.
- If various intervention are considered, then we would use notation $\bar{Q}^{g^*}_{L(k)}$, $\bar{Q}^{g^*}_{A(k)}$ and $\bar{Q}^{g^*}$ for the above defined parameters.

Sequential super-learning

- Note that each $\bar{Q}_{L(k)}$ is a regression of previous $\bar{Q}_{A(k)}$ (outcome) on past $\bar{A}(k-1), \bar{L}(k-1)$, and similarly, for $\bar{Q}_{A(k)}$.
- Therefore, we could estimate any $\bar{Q}_{L(k)}$ by sequentially fitting a regression (e.g., using super-learning) where the previously fitted regression curve represents the outcome in this regression.
- The evaluations $\bar{Q}_{A(k)}$ are known and are thus not involving estimation.
- In particular, sequential regression estimation can be used to estimate $\bar{Q}_{L(0)} = \Psi_{g^*}(P)$.

Utilize known degeneracies in data distribution

- The conditional expectation in $\bar{Q}_{L(k)}$ is known for any history $\bar{L}(k-1), \bar{A}(k-1)$ for which $\tilde{T} \leq k-1$.
- So estimation of $\bar{Q}_{L(k)}$ focusses on estimation conditional on $\tilde{T} > k-1$.
- If other degeneracies are present (e.g., $L(k) = L(k-1)$ for history $\bar{L}(k-1), \bar{A}(k-1)$), then these should be respected as well.

A real-world CER study comparing different rules for treatment intensification for diabetes

- Data extracted from diabetes registries of 7 HMO research network sites:
  - Kaiser Permanente
  - Group Health Cooperative
  - HealthPartners
- Enrollment period: Jan 1^{st} 2001 to Jun 30^{th} 2009

Enrollment criteria:

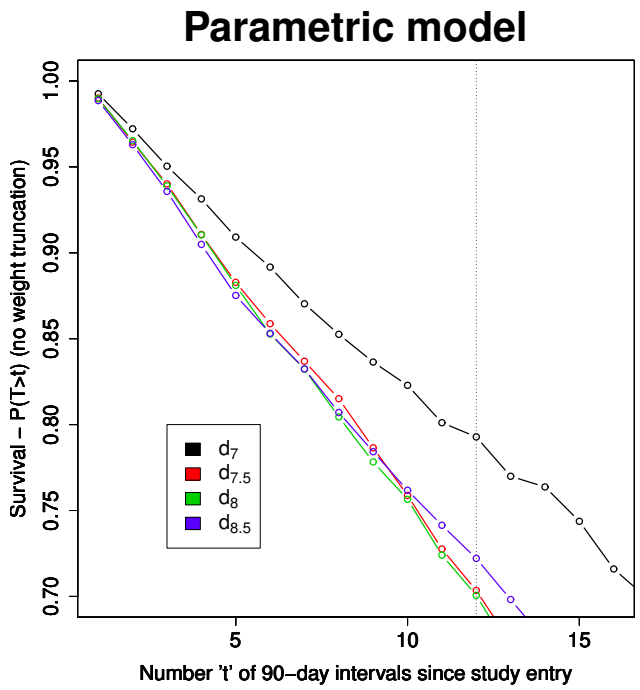
- past A1c$<$ 7% (glucose level) while on 2+ oral agents or basal insulin
- 7% $\leq$ latest A1c $\leq$ 8.5% (study entry when glycemia was no longer reined in)

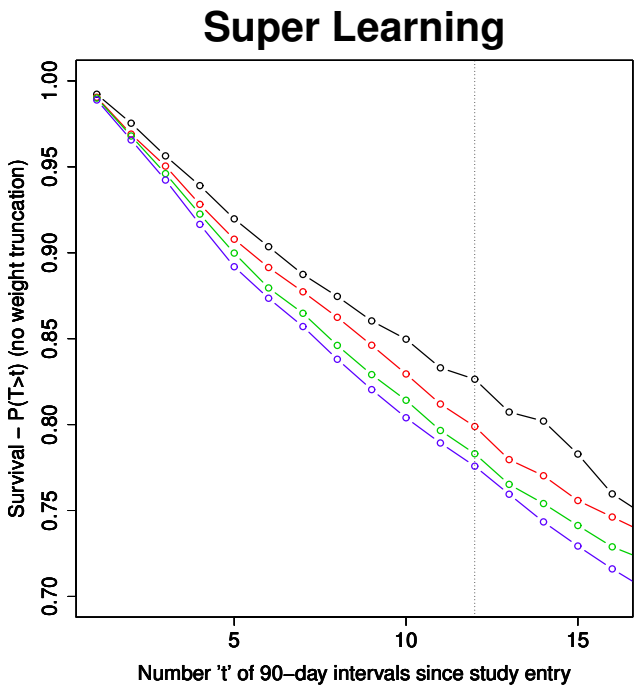
Longitudinal data

- Follow-up til the earliest of Jun 30^{th} 2010, death, health plan disenrollment, or the failure date
- Failure defined as onset/progression of albuminuria (a microvascular complication)
- Treatment is the indicator being on "treatment intensification" (TI)
- $n \approx 51,000$ with a median follow-up of 2.5 years

Back to the TI study...

Impact of machine learning on inference with IPW 1:

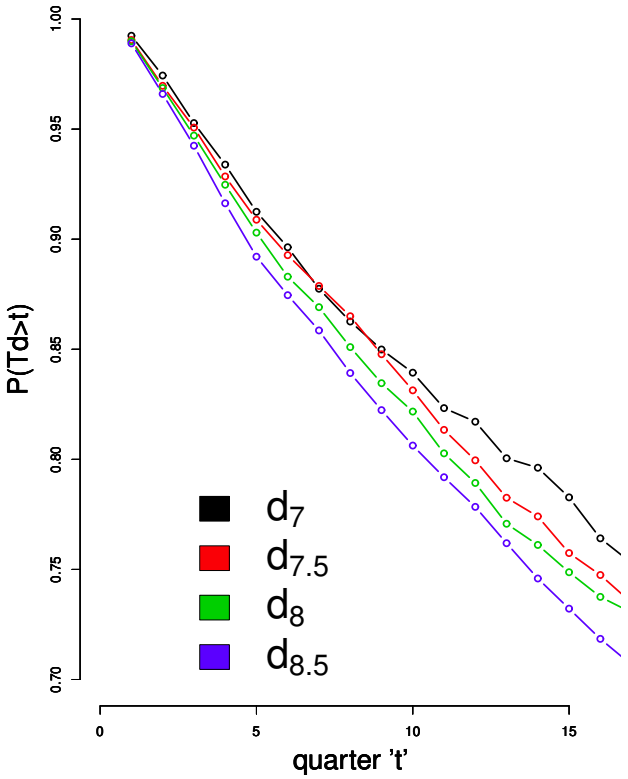




No/weak evidence of protective effect

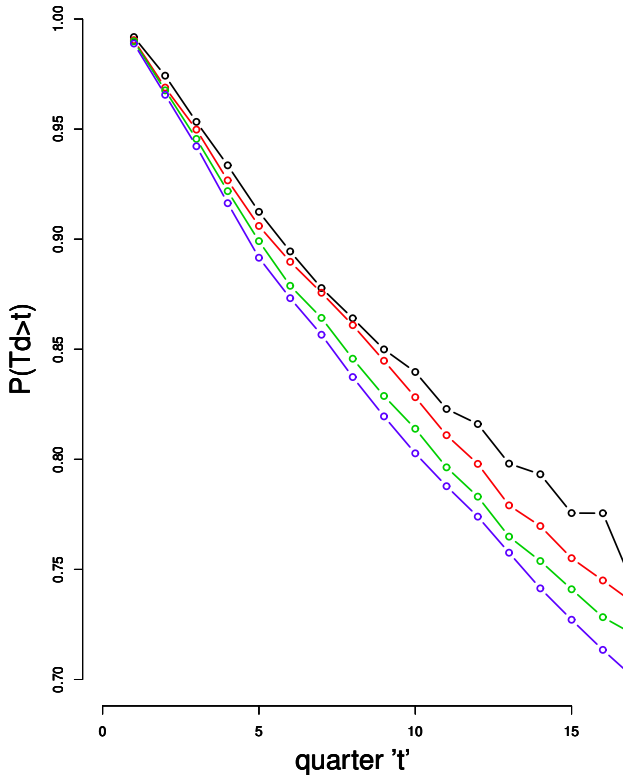
Strong significant evidence

Practical performance



IPW estimator 3 + SL

(hazard-based)



TMLE + SL

$$1.07 \leq \frac{\sigma_{IPW3}}{\sigma_{TMLE}} \leq 1.11$$

Analysis of Tshepo Study (Gruttolas et al.): RCT with time-to-event outcome

TMLE of causal effect of point treatment on survival needs to account for potential bias due to informative censoring by **time-dependent covariates** CD4 and viral load that have an effect on both time to drop-out and the time to event of interest.

We will directly compare results using this TMLE that only incorporates the baseline covariates to the TMLE that accounts for time dependent confounding in the form of informative censoring due to the time-dependent covariates. Moreover, we will compare these results to results based on an **IPCW estimator** and a locally efficient double robust **estimating equation based estimator** (A-IPCW).

Analysis of Tshepo Study (Gruttolas et al.): RCT with time-to-event outcome

For the analysis performed here we evaluate the **effect modification by gender** on the two cART treatments for Time to death censored by treatment modification or end of study (DEATH).
We estimate the difference in additive risk by gender at 36 months after randomization to cART therapy. We will estimate this effect modification parameter using the six estimators.

Results of Tshepo Study (Gruttolas et al.)

Risk Difference @ 36 Months

	Time Dependent			Baseline		
	TMLE	A-IPCW	IPCW	TMLE	A-IPCW	IPCW
Est	6.3%	6.5%	5.2%	5.1%	5.1%	5.2%
SE	2.3%	2.3%	12.5%	2.4%	2.4%	12.5%
p-value	0.005	0.004	0.680	0.029	0.030	0.680

Table: Gender Effect Modification on Death

The TMLE results indicate that gender does in fact modify the effect of the drug treatment EFV/NVP and the difference in the effect between males and females at 36 months is estimated at 6.3 percent.

Outline

Optimal dynamic treatment

- Let observed data be n i.i.d. copies of $O = (W, A, Y)$.
- Let Y_d be counterfactual outcome/reward for unit if it would have received treatment $A = d(W)$.
- Let $E_0 Y_d$ be population mean reward, $\mathcal{D}$ a class of treatment allocation rules. Then optimal rule among this class is defined as

$$d_0 = \arg\min_{d \in \mathcal{D}} E_0 Y_d.$$

- $E_0 Y_{d_0}$ is optimal reward, achieved under optimal rule.

Form of optimal rule

- Let Y_a be a treatment specific potential outcomes, one for each level of treatment, and $Y = Y_A$.
- Assume A is independent of Y_0, Y_1, given W.
- Then, $d_0(W) = \arg\min_a E_0(Y \mid A = a, W)$.
- If treatment is binary, then it is the indicator of $B_0(W) \equiv E(Y \mid A = 1, W) - E(Y \mid A = 0, W)$ being positive.

Super learning of optimal rule

- We can construct a super-learner of the conditional mean of outcome, $E(Y \mid A, W)$, which implies an estimator of the optimal rule.
- We can also construct a super-learner of the optimal rule directly, using as criterion/loss EY_d (equivalent with a weighted classification problem). The library for the super-learner of d_0 can utilize both classification algorithms and prediction algorithms in the machine learning literature.
- The latter appears preferable if the rule itself is the goal.

TMLE of mean outcome under optimal rule

- We have a TMLE of counterfactual mean outcome EY_d.
- This TMLE uses in targeting step the clever covariate $I(A = d(W))/\hat{P}(A|W)$ to update initial outcome regression.
- We can apply this TMLE to $d = d_n$ being the estimate of the optimal rule.
- Better to use CV-TMLE: let $d_{n,v}$ be estimate on v-th training sample; run TMLE of $EY_{d_{n,v}}$ with initial estimator from training but targeting on validation sample, and average the V resulting TMLEs: $1/V \sum_v \psi^*_{n,v}$.

Inference for mean outcome under optimal rule, or estimate of optimal rule

- The TMLE and CV-TMLE are efficient estimators of mean outcome EY_{d_0} under optimal rule, under assumption $EY_{d_n} - EY_{d_0} = o_P(n^{-1/2})$.
- The latter will hold under reasonable conditions: trivially, if $P(B_0(W) < \delta) = 0$ for some $\delta > 0$.
- In general, assuming rate of convergence of estimator of $B_0(W)$, and assuming density of $B_0(W)$ vanishes at zero.
- One can also interpret the inference as inference for EY_{d_n} or $1/V \sum_v EY_{d_{n,v}}$, mean outcome under estimate of optimal rule. In that case, we do not need this extra condition.
- The variance of the TMLE and CV-TMLE can be estimated as if the rule d_n (or $d_{n,v}$) was a priori given.

Outline

Trauma Medicine – Data-Rich but Chaotic Environment

- Doctors have access to a range of indicators about the state of a patient's health.
  - Vital signs, lab results, demographic information (16 covariates)

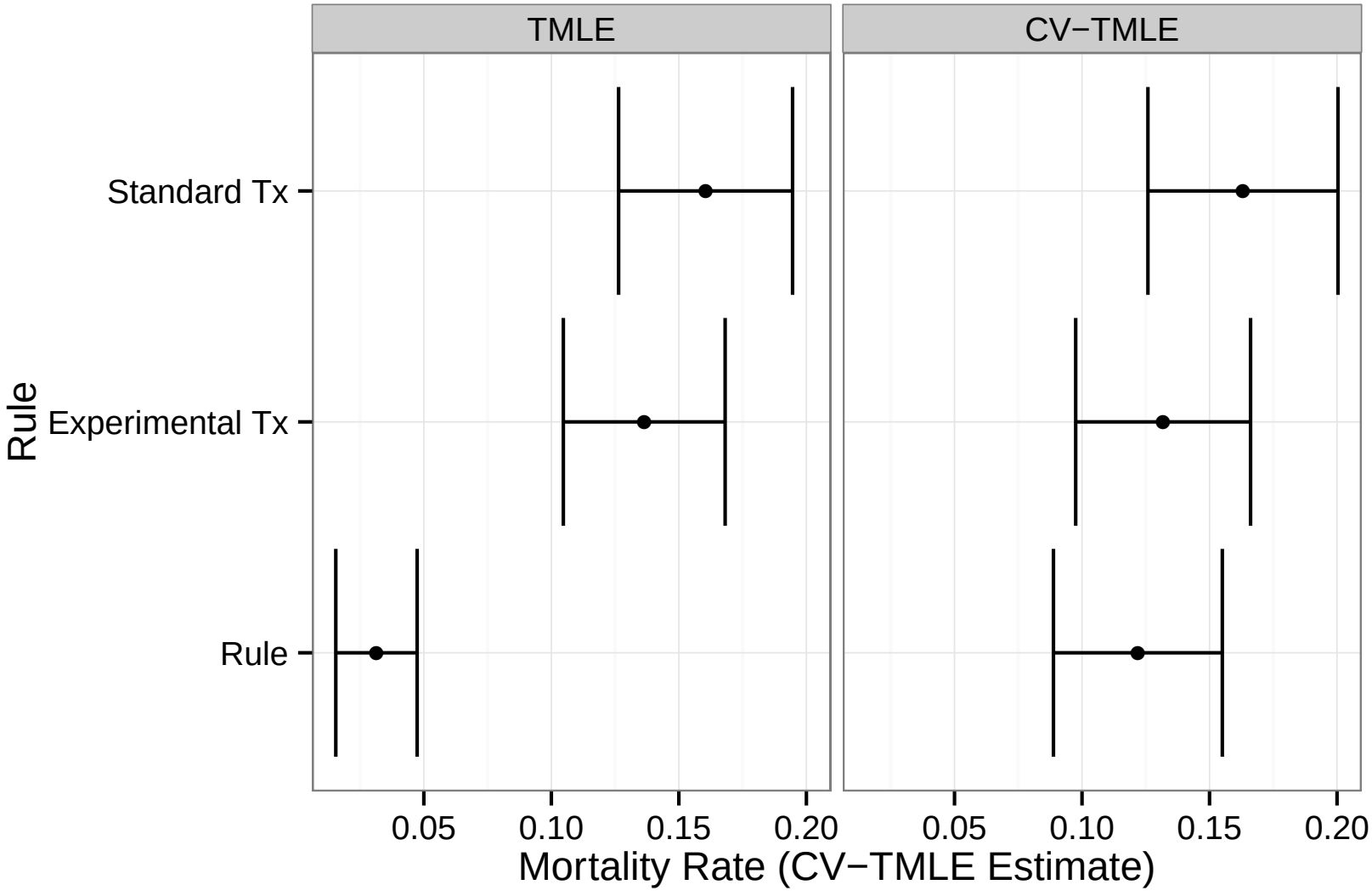
Trauma Medicine – Data-Rich but Chaotic Environment

- Doctors have access to a range of indicators about the state of a patient's health.
  - Vital signs, lab results, demographic information (16 covariates)
- We applied our optimal treatment methodology to estimate a treatment rule for blood transfusion protocols for victims of traumatic injury.
  - Rule decides between "standard" and "experimental" blood protocols.

Trauma Medicine – Data-Rich but Chaotic Environment

- Doctors have access to a range of indicators about the state of a patient's health.
  - Vital signs, lab results, demographic information (16 covariates)
- We applied our optimal treatment methodology to estimate a treatment rule for blood transfusion protocols for victims of traumatic injury.
  - Rule decides between "standard" and "experimental" blood protocols.
- Objective is to minimize 24 hour all-cause mortality rate.

Results


TMLE
CV−TMLE
Standard Tx
Experimental Tx
Rule
Rule
0.05
0.10
0.15
0.20
0.05
0.10
0.15
0.20
Mortality Rate (CV−TMLE Estimate)

Some references for estimation of optimal rule

- A.R. Luedtke and M.J. van der Laan (2016), Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy, Annals of Statistics, vol. 44, no. 2, pp. 713-742, 2016.
- A. Luedtke and M. van der Laan. Optimal Individualized Treatments in Resource-Limited Settings. Int J Biostat, 12(1):283303, 05 2016.
- A.R. Luedtke and M.J. van der Laan (2016), Super-learning of an optimal dynamic treatment rule, International Journal of Biostatistics, vol. 12, no. 1, pp. 305-332, 2016. Also available at http://biostats.bepress.com/ucbbiostat/paper326/
- M.J. van der Laan and A.R. Luedtke (2015), Targeted learning of the mean outcome under an optimal dynamic treatment rule, Journal of Causal Inference, vol. 3, no. 1, pp. 61-95, 2015. PMCID:PMC4517487. Also available at http://biostats.bepress.com/ucbbiostat/paper329/

Outline

Concluding Remarks

- Targeted Learning learns unbiased and reproducible answers to actionable questions (potentially causal) with confidence, which result in improved policy, treatments, evaluations, etc.
- It integrates **causal inference**, **machine learning**, **statistical theory**.
- **Targeted Learning** *optimally estimates* the (potentially causal) impact of an intervention on an outcome for complex real-world data.
- The estimate is accompanied with accurate quantification of uncertainty such as **confidence interval and p-value**.
- We have developed an ongoing targeted learning software environment `tlverse` with growing number of tools and tutorials, such as *The Hitchhiker's Guide to the `tlverse`: A Targeted Learning Practitioner's Handbook*.