

# Targeted Machine Learning for Real-World Data Science

Mark van der Laan & Rachael Phillips

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6, 2019

# Schedule for Half-Day Tutorial on December 4, 2019

- **9:00A-11:00A:** Overview of Targeted Machine Learning
- **11:00A-12:00P:** TMLE for the Causal Impact of a Single Time-Point Intervention on Survival with Software Exercise in R

# Schedule for Short Course Day 1 on December 5, 2019

- **8:00A-9:30A:** Overview of Targeted Learning
- **9:30A-9:50A:** Break
- **9:50A-11:20P:** Causal Inference and Interventions
- **11:20A-12:40P:** Lunch
- **12:40P-2:10P:** Super (Machine) Learning and Targeted Minimum Loss-Based Estimation
- **2:10P-2:30P:** Break
- **2:30P-4:00P:** Super Learning in the `tlverse` software ecosystem
- **4:00P-4:20P:** Break
- **4:20P-5:00P:** Targeted Maximum Likelihood Estimation of the Average Treatment Effect in the `tlverse` software ecosystem

## Schedule for Short Course Day 2 on December 6, 2019

- **8:00A-9:30A:** Targeted Minimum Loss-Based Estimation of the Effects of Optimal Dynamic and Shift Interventions.
- **9:30A-9:50A:** Break
- **9:50A-11:20P:** Targeted Minimum Loss-Based Estimation of the Treatment Specific Survival Function for Right-Censored Survival Data
- **11:20A-12:40P:** Lunch
- **12:40P-2:10P:** Targeted Minimum Loss-Based Estimation for Longitudinal Data
- **2:10P-2:30P:** Break
- **2:30P- :** Discussion

## Resources

- The latest version of the presentation slides are available here:  
<https://github.com/tlverse/deming2019-workshop/tree/master/slides>.
- The open source and fully-reproducible electronic vignette for the software tutorials can be found here:  
<https://tlverse.org/deming2019-workshop/>.

# Targeted Machine Learning

## Causal Inference for Real-World Data Science

Mark van der Laan

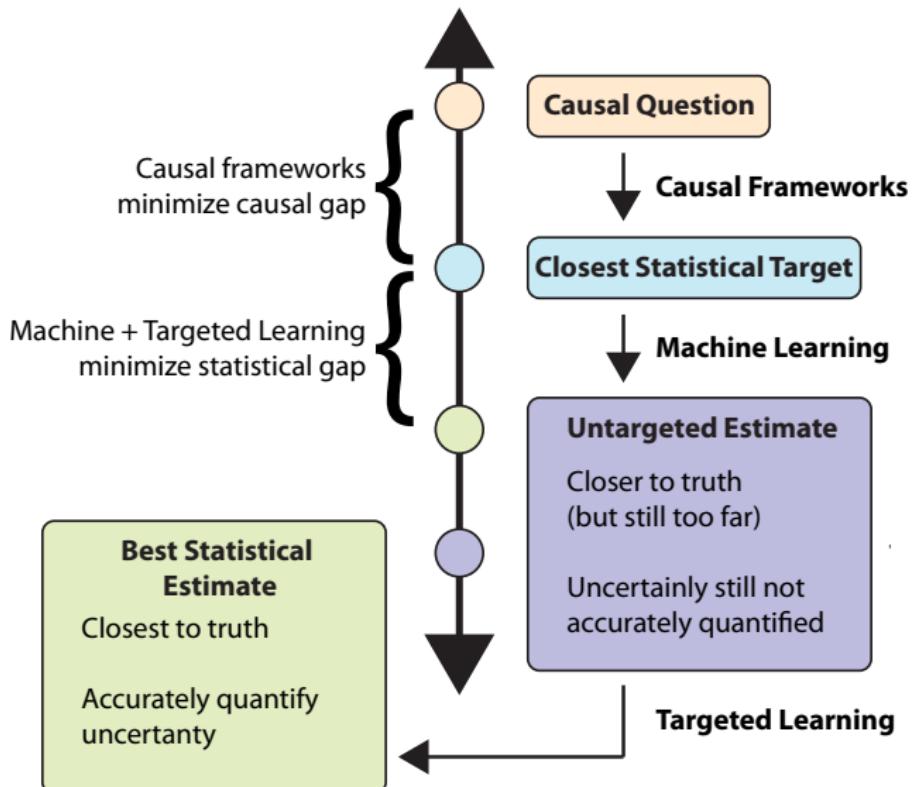
Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**  
December 4-6 2019, Atlantic City NJ

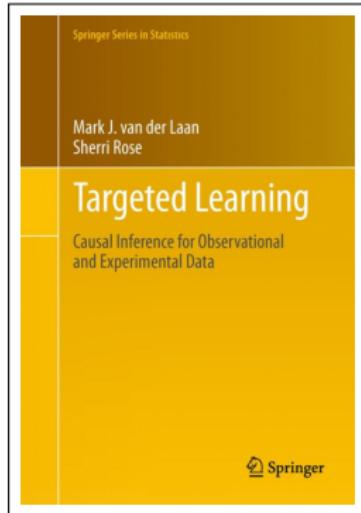
Various slides from Maya Petersen presentation (NIH R01 AI074345)  
and Bill and Melinda Gates Foundation presentation.

# Outline

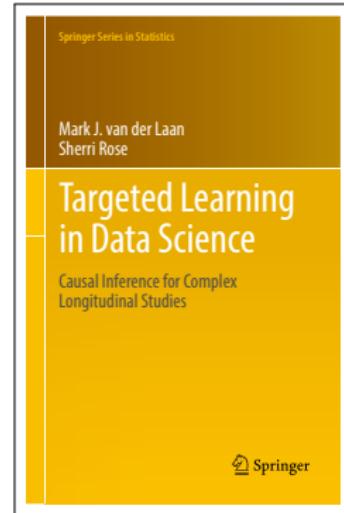
# Targeted Learning fills a much needed gap in machine learning and causal inference



# Targeted Learning is a subfield of statistics



van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.



van der Laan & Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. New York: Springer, 2018.

<https://vanderlaan-lab.org>

# Outline

# Roadmap for Statistical Learning

- ① Describe observed data
- ② Specify statistical model
- ③ Define statistical query (e.g., using causal roadmap)
- ④ Construct estimator
- ⑤ Obtain inference

# Roadmap for Statistical Learning

STEP 1:  
DESCRIBE  
OBSERVED DATA

STEP 2:  
SPECIFY  
STATISTICAL MODEL

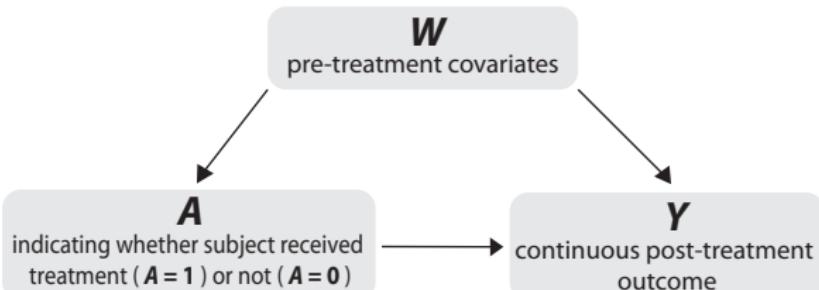
STEP 3:  
DEFINE  
STATISTICAL QUERY

STEP 4:  
CONSTRUCT  
ESTIMATOR

STEP 5:  
OBTAIN INFERENCE

$n = 100$  subjects were sampled independently from each other and from the same population distribution  $P_0$

For each subject, pre-treatment covariates ( $W$ ), treatment ( $A$ ), and outcome ( $Y$ ) vectors were measured



# Roadmap for Statistical Learning

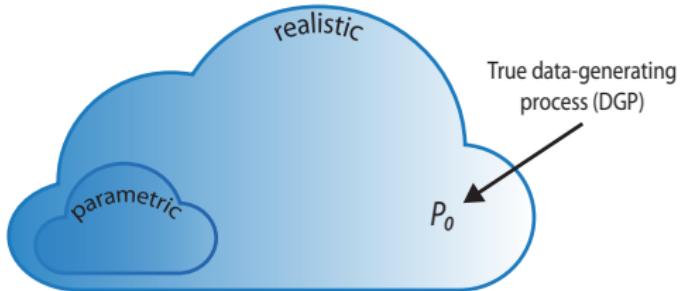
STEP 1:  
DESCRIBE  
OBSERVED DATA

STEP 2:  
SPECIFY  
STATISTICAL MODEL

STEP 3:  
DEFINE  
STATISTICAL QUERY

STEP 4:  
CONSTRUCT  
ESTIMATOR

STEP 5:  
OBTAIN INFERENCE



## Standard Approach

Parametric statistical model

Does not contain  $P_0$ , the DGP  
(i.e., misspecified model)

## Targeted Learning

Realistic semiparametric or  
nonparametric statistical model

Defined to ensure  $P_0$  is  
contained in model

# Roadmap for Statistical Learning

STEP 1:  
DESCRIBE  
OBSERVED DATA

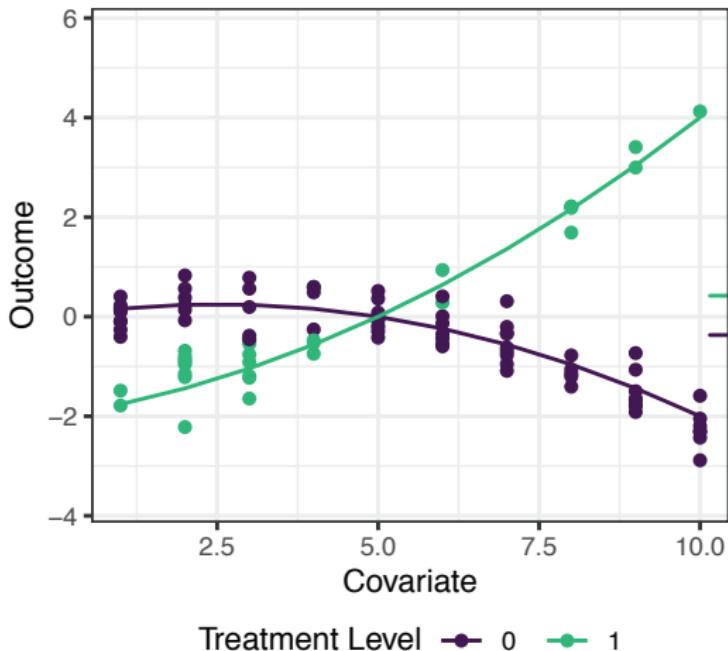
STEP 2:  
SPECIFY  
STATISTICAL MODEL

STEP 3:  
DEFINE  
STATISTICAL QUERY

STEP 4:  
CONSTRUCT  
ESTIMATOR

STEP 5:  
OBTAIN INFERENCE

Example True DGD



# Roadmap for Statistical Learning

STEP 1:  
DESCRIBE  
OBSERVED DATA

STEP 2:  
SPECIFY  
STATISTICAL MODEL

STEP 3:  
DEFINE  
STATISTICAL QUERY

STEP 4:  
CONSTRUCT  
ESTIMATOR

STEP 5:  
OBTAIN INFERENCE

*What is the average difference in outcomes between treatment groups when adjusting for covariates?*

$$\Psi(P_0) = E_0(E_0[Y|A=1, W] - E_0[Y|A=0, W])$$

$\Psi$  is a function that takes as input  $P_0$  and outputs the answer to the question of interest

The **assumption of positivity** is required to estimate of this quantity from the data. That is, it must be possible to observe both levels of treatment for all strata of  $W$ .

Additional assumptions are required to interpret this estimand as causal

# Causal roadmap for obtaining statistical query answering causal question

Step 3 can be carried out using following causal roadmap:

- Define **potential outcomes**  $Y_0, Y_1$  for each subject, representing (counterfactual) outcome we would have seen if subject would have taken treatment 0 and 1, respectively.
- Link desired full-data  $(W, Y_0, Y_1)$  to observed data  $O = (W, A, \mathbf{Y} = \mathbf{Y}_A)$ .
- Define **causal quantity** of interest:  $E(Y_1 - Y_0)$ , called average treatment effect.
- Establish **identification from DGD**: If treatment is independent of potential outcomes, given  $W$ , and positivity holds, then  $E_0(Y_1 - Y_0)$  equals target estimand  $\Psi(P_0)$ .

# Roadmap for Statistical Learning

STEP 1:  
DESCRIBE  
OBSERVED DATA

STEP 2:  
SPECIFY  
STATISTICAL MODEL

STEP 3:  
DEFINE  
STATISTICAL QUERY

STEP 4:  
CONSTRUCT  
ESTIMATOR

STEP 5:  
OBTAIN INFERENCE

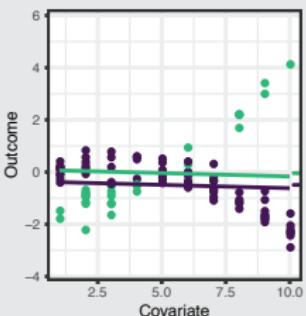
## Standard Approach

Generalized Linear Model (GLM)  
to estimate

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{A} + \beta_2 \mathbf{W} + \epsilon$$

Estimated coefficients  
are biased

Cannot detect heterogeneity  
in treatment effect

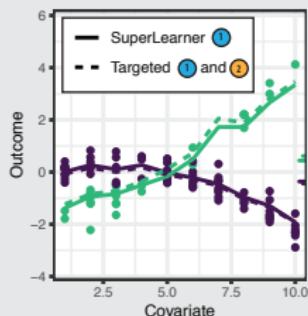


## Targeted Learning

TMLE implements  
a two-step procedure

- 1 initial estimation of  $E_0[Y|A, W]$  with super (machine) learning
- 2 targeting towards optimal bias-variance trade-off for  $\Psi(P_0)$

TMLE estimates are unbiased  
and doubly robust



# Roadmap for Statistical Learning

STEP 1:  
DESCRIBE  
OBSERVED DATA

STEP 2:  
SPECIFY  
STATISTICAL MODEL

STEP 3:  
DEFINE  
STATISTICAL QUERY

STEP 4:  
CONSTRUCT  
ESTIMATOR

STEP 5:  
OBTAIN  
INFERENCE

## Standard Approach

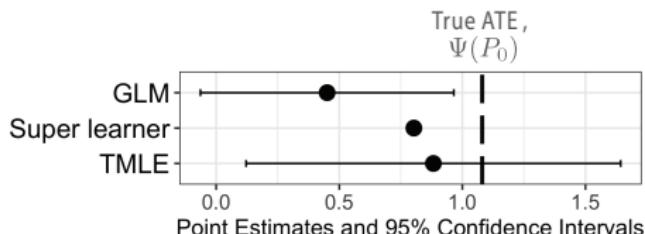
Inference (such as  $p$ -value and confidence interval) assumes parametric model is true

Inference is misleading and erroneous

## Targeted Learning

Targeting (step ②) improves estimate and makes inference possible

Trustworthy inference obtained with efficient influence function



# Roadmap of Statistical Learning Summary

- **Observed data:** Realization of a random variable  $O^n = (O_1, \dots, O_n)$  with a probability distribution (say)  $P_0^n$ , indexed by "sample size"  $n$ .
- **Model stochastic system of observed data realistically:** Statistical model  $\mathcal{M}^n$  is set of possible probability distributions of the data.
- **Define query about stochastic system:** Function  $\Psi$  from model  $\mathcal{M}^n$  to real line, where  $\Psi(P_0^n)$  is the true answer to query about our stochastic system.
- **Estimator:** An a priori-specified algorithm that takes the observed data  $O^n$  and returns an estimate  $\psi_n$  to the *true answer to query*. Benchmarked by a dissimilarity-measure (e.g., MSE) w.r.t true answer to query.
- **Confidence interval for true answer to query:** Establish approximate sampling probability distribution of the estimator (e.g., based on CLT), and corresponding statistical inference.

# Outline

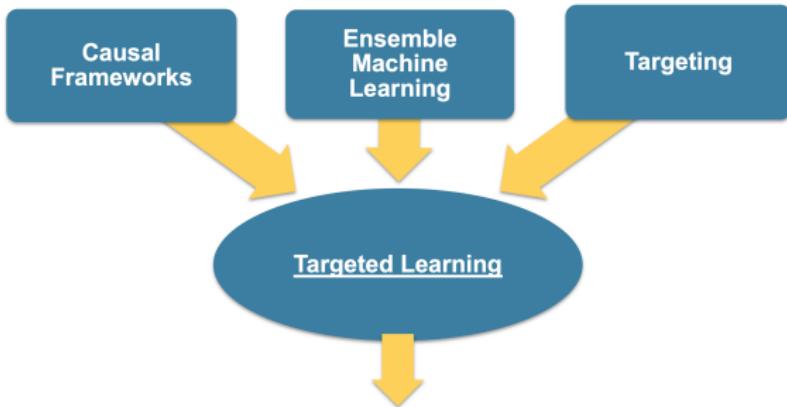
## Example: Nonparametric Estimation of Average Treatment Effect

- Unit (i.i.d.) data  $O \sim P_0$  consists of baseline covariates  $W$ , binary treatment  $A$ , and final binary outcome  $Y$ .
- Statistical model for the data distribution  $P_0$  is nonparametric.
- Statistical target parameter:

$$\Psi(P) = E_P\{P(Y = 1 | A = 1, W) - P(Y = 1 | A = 0, W)\}.$$

- Under causal model, randomization assumption, and positivity assumption,  $\Psi(P) = E(Y_1 - Y_0)$  is the ATE.
- A TMLE will estimate  $P(Y = 1 | A, W)$  with **ensemble machine learning** and a subsequent **Targeting step** using logistic regression with off-set initial fit, and clever covariate  $(2A - 1)/\hat{P}(A|W)$ .

# Targeted Learning



Better (more precise) answers to causal (actionable) questions with  
accurate quantification of uncertainty (signal from noise)

DIA

# Identifying contributing factors for health care spending

© Health Research and Educational Trust

DOI: 10.1111/1475-6773.12848

METHODS ARTICLE

## Robust Machine Learning Variable Importance Analyses of Medical Conditions for Health Care Spending

*Sherri Rose* 

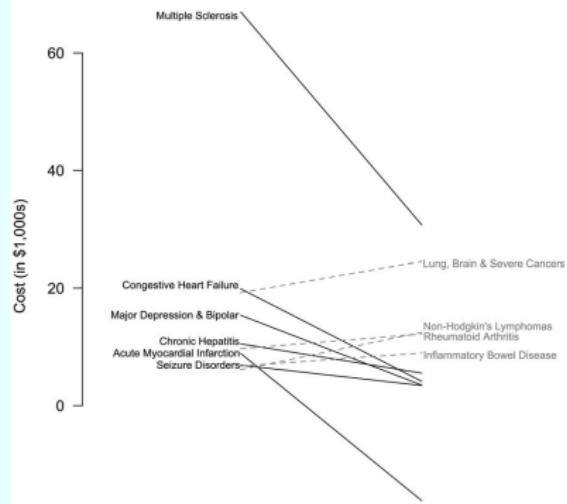
# Identifying contributing factors for health care spending

© Health Research and Educational Trust  
DOI: 10.1111/1475-6773.12848  
METHODS ARTICLE

## Robust Machine Learning Variance Importance Analyses of Medication Conditions for Health Care Spending

Sherri Rose 

Figure 4: Top 10 Largest Targeted Learning Effect Estimates



# Average treatment effect in an observational study

International Archives of Occupational and Environmental Health (2019) 92:629–638  
<https://doi.org/10.1007/s00420-018-1397-1>

ORIGINAL ARTICLE



## An educational intervention to improve knowledge about prevention against occupational asthma and allergies using targeted maximum likelihood estimation

Daloha Rodríguez-Molina<sup>1,2</sup> · Swaantje Barth<sup>1</sup> · Ronald Herrera<sup>1</sup> · Constanze Rossmann<sup>3</sup> · Katja Radon<sup>1</sup> · Veronika Karnowski<sup>4</sup>

Received: 15 March 2018 / Accepted: 13 December 2018 / Published online: 14 January 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

# Average treatment effect in an observational study

International Archives of Occupational and Environmental Health (2019) 92:629–638  
https://doi.org/10.1007/s00420-018-1397-1

## ORIGINAL ARTICLE

### An educational intervention to improve knowledge about occupational asthma and allergies using targeted likelihood estimation

Daloña Rodríguez-Molina<sup>1,2</sup>  · Swaantje Barth<sup>1</sup>  · Ronald Herrera<sup>1</sup>  · Constanze R. Veronika Karnowski<sup>4</sup> 

Received: 15 March 2018 / Accepted: 13 December 2018 / Published online: 14 January 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

**Table 4** Adjusted average treatment effects of the intervention ( $n=116$ ), Bavaria, Germany, 2014

	All six correct measures	At least five correct measures	At least four correct measures
Additive ATE			
Parameter	18.44%	55.53%	29.60%
95% CI	(7.3–29.58%)	(36.96–74.09%)	(12.2–47.0%)
Additive ATT			
Parameter	16.9%	63.07%	62.78%
95% CI	(5.38–28.51%)	(46.02–80.13%)	(41.64–83.93%)
Additive ATC			
Parameter	16.8%	32.28%	18.97%
95% CI	(5.02–28.57%)	(12.84–51.72%)	(1.91–36.02%)

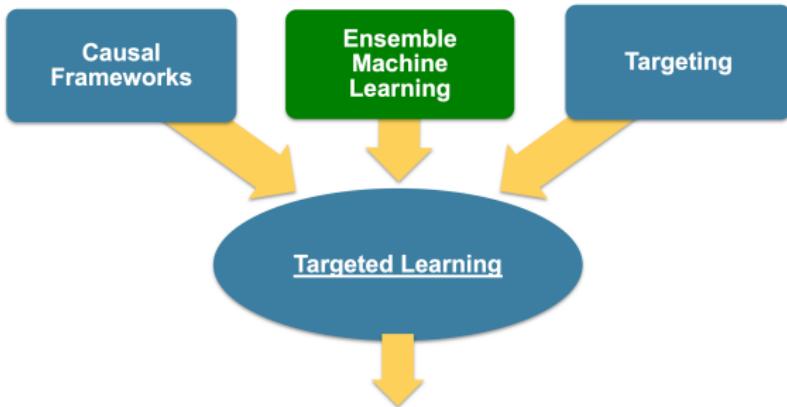
Adjusted for sex, age, education level, smoking status, presence of asthma or rhinoconjunctivitis, riskperception, parental asthma, and knowledge about preventive measures against asthma and allergies before the intervention

The adjusted model using TMLE allowed including both observed data ( $n=47$ ) and missing values ( $n=69$ ) as parameters

ATE average treatment effect, ATT average treatment effect on the treated, CI confidence interval, ATC average treatment effect on the controls, TMLE targeted maximum likelihood estimation

# Outline

# Targeted Learning

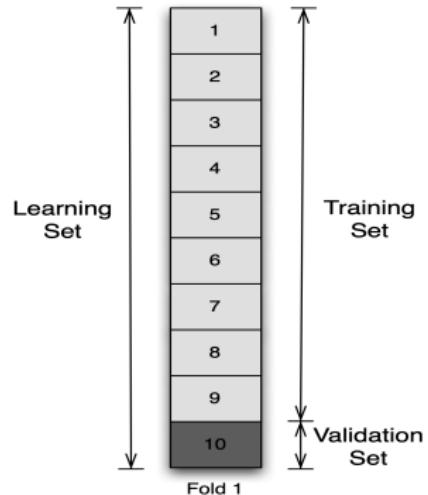


Better (more precise) answers to causal (actionable) questions with  
accurate quantification of uncertainty (signal from noise)

DIA

# Super Learning: Ensemble Machine Learning

- “Competition” of algorithms
  - Parametric models
  - Data-adaptive (ex. Random forest, Neural nets)
- Best “team” wins
  - Convex combination of algorithms
- Performance judged on independent data
  - V-fold cross validation (Internal data splits)
- Customizable optimality criterion
  - Standard loss function
  - Minimize false negatives with bounded false positives
  - Respect resource constraints



Van der Laan, Polley, 2007

DIA

# V-fold Cross Validation

1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9
10	10	10	10	10	10	10	10	10	10

Fold 1      Fold 2      Fold 3      Fold 4      Fold 5      Fold 6      Fold 7      Fold 8      Fold 9      Fold 10

DIA

# Improving upon the current standard of predictive analytics in the ICU

## THE LANCET Respiratory Medicine

Volume 3, Issue 1, January 2015, Pages 42-52



### Articles

Mortality prediction in intensive  
care units with the Super ICU  
Learner Algorithm (SICULA): a  
population-based study

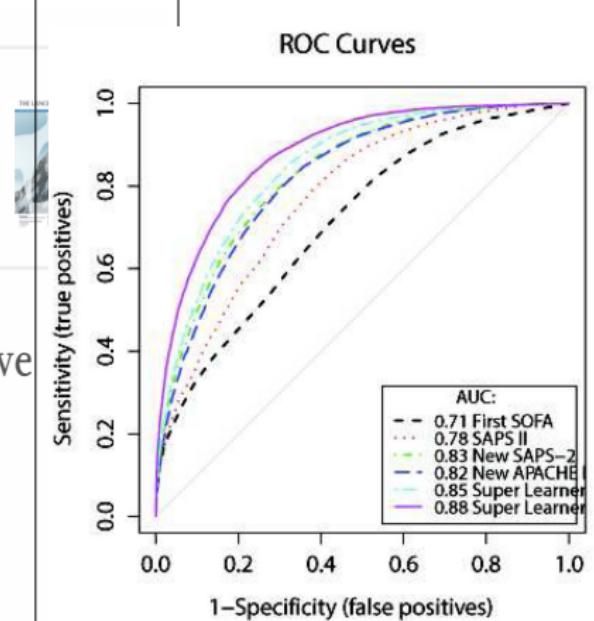
# Improving upon the current standard of predictive analytics in the ICU

## THE LANCET Respiratory Medicine

Volume 3, Issue 1, January 2015, Pages 42-52

### Articles

Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study



## Cross-validation is optimal for selection among estimators

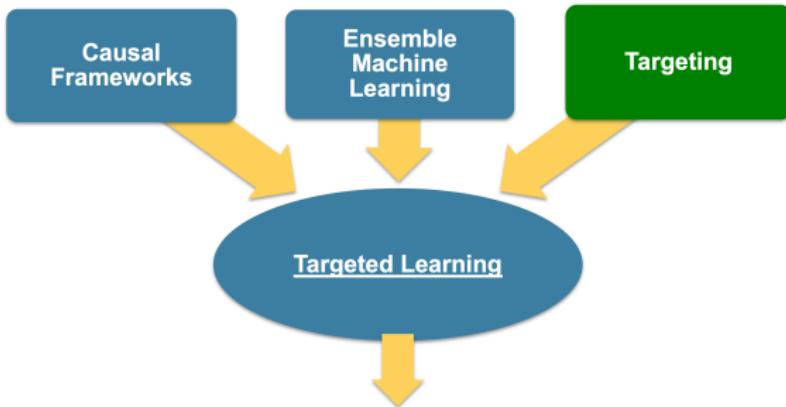
- We established an oracle inequality for the cross-validation selector among a collection of candidate estimators (e.g, van der Laan, Dudoit, 03, van der Vaart et al 06).
- Oracle selector chooses the estimator closest to the true function w.r.t. loss-based dissimilarity.
- It establishes that the loss-based dissimilarity with truth of the cross-validated selected estimator divided by the loss-based dissimilarity of the oracle selected estimator converges to 1, even as the number of candidate estimators converges to infinity as a polynomial in sample size.
- Only condition is that loss-function is uniformly bounded.

## Highly Adaptive Lasso (HAL)

- This is a machine learning algorithm that estimates functionals (e.g outcome regression and propensity score) by approximating them with linear model in **tensor product of spline basis functions** and constraining the  $L_1$ -norm of the coefficients.
- Can be computed with **Lasso**-software implementations.
- Guaranteed to converge to truth at rate  $n^{-1/3}$  (up till  $\log n$ -factors) in sample size  $n$ .
- When used in super-learner library, TMLE (targeted learning) is guaranteed **consistent, (double robust) asymptotically normal and efficient**: one only needs to assume *strong positivity assumption*.

# Outline

# Targeted Learning



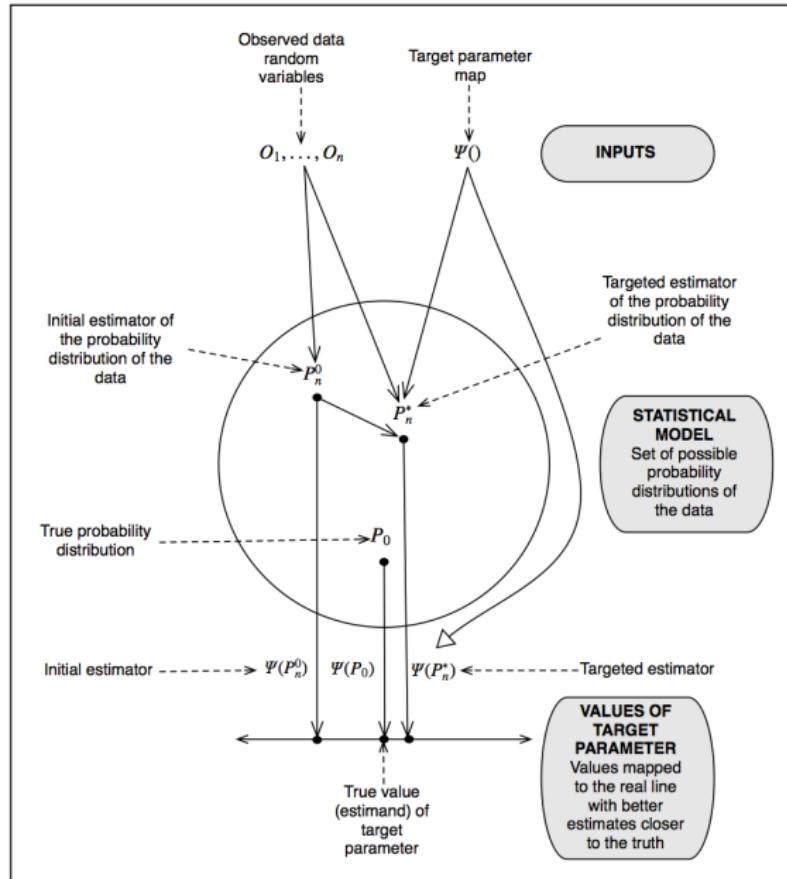
Better (more precise) answers to causal (actionable) questions with  
accurate quantification of uncertainty (signal from noise)

DIA

## 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.
- ➊ Less bias (closer to truth).
  - ➋ 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 stochastic system  $P_0$ .
- Determine mathematically the fluctuation strategy (least favorable submodel)  $\mathbf{P}_{n,\epsilon}$  of the super-learner fit  $\mathbf{P}_n$  with tuning parameter  $\epsilon$  **so that a small change in  $\epsilon$  corresponds with a maximal small change** in estimated answer  $\Psi(\mathbf{P}_{n,\epsilon})$  to query  $\Psi(P_0)$ : i.e., score equals canonical gradient/**efficient influence curve**  $D^*(\mathbf{P}_n)$ .
- Determine the optimal amount  $\epsilon_n$  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  $\psi_n$  solution of  $P_n D^*(\psi, \eta(\mathbf{P}_n)) = 0$ .
- **TMLE:**  $\mathbf{P}_{n,\epsilon}$  least favorable submodel through initial  $\mathbf{P}_n$ ;  $\epsilon_n$  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.

## 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  $(W, A, Y)$ .

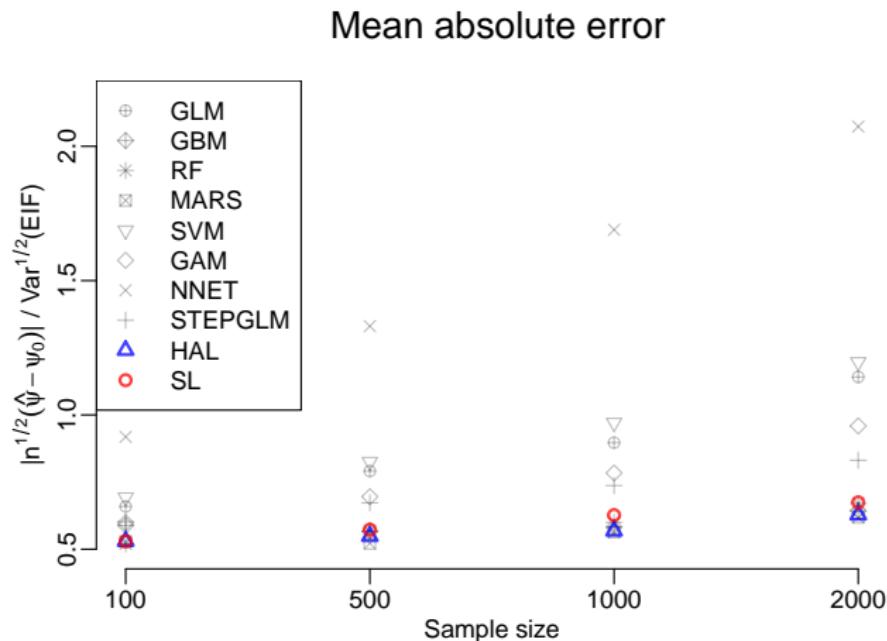
We computed TMLEs of the ATE based on different estimators of  $E_0(Y | A, W)$  and  $P_0(A = 1 | W)$ .

- 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.

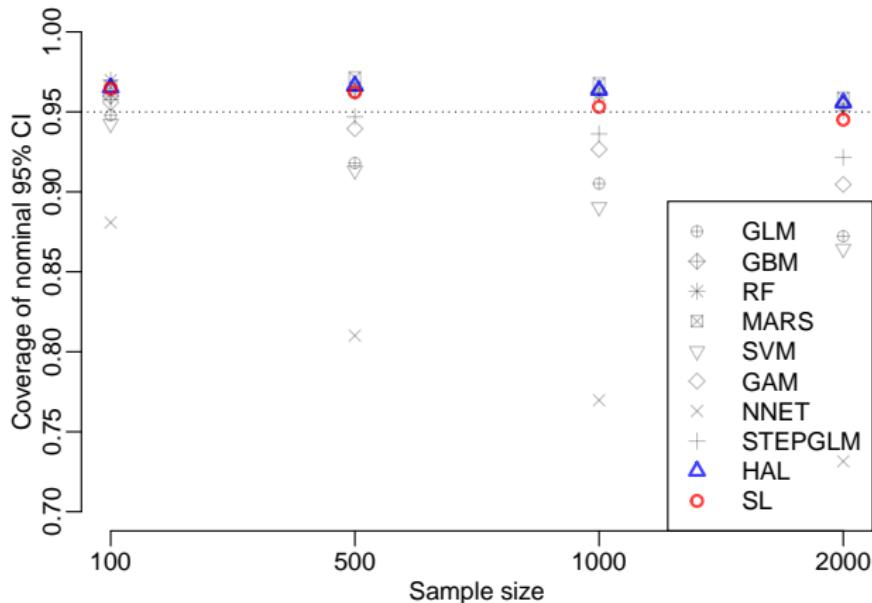
# Results – absolute error by sample size

HAL-TMLE exhibited excellent accuracy relative to competitors.



## Results – coverage by sample size

HAL-TMLE achieves approximate Normality in reasonable sample sizes.

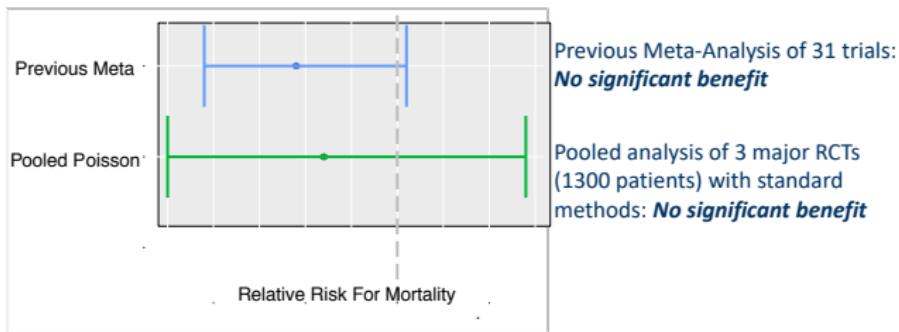


# Outline

## 1

## Better, cheaper trials

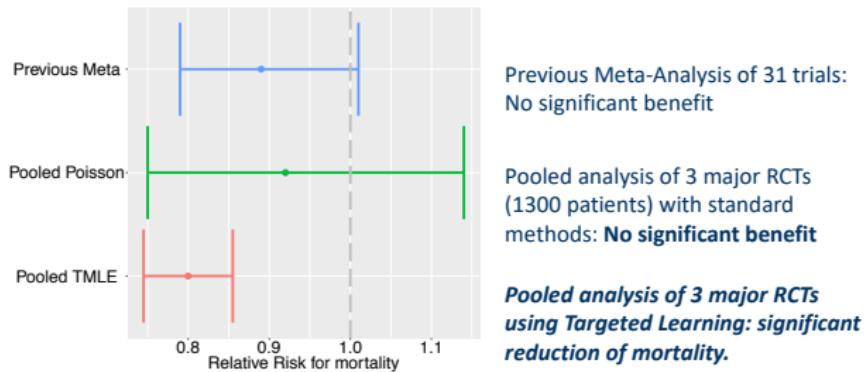
Do corticosteroids reduce mortality for adults with septic shock?



Pirracchio 2016

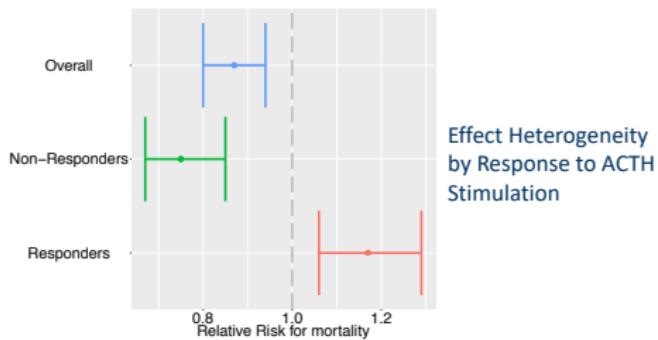
## Better, cheaper trials

Do corticosteroids reduce mortality for adults with septic shock?



## Not just is there an effect, but for whom?

- In Sepsis re-analysis: Targeted Learning showed **all benefit** occurred in a key subgroup
  - Heterogeneity in patient populations one cause of inconsistent results



# Estimating the causal effect of a community-level intervention in a clustered RCT

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

## HIV Testing and Treatment with the Use of a Community Health Approach in Rural Africa

D.V. Havlir, L.B. Balzer, E.D. Charlebois, T.D. Clark, D. Kwarisiima, J. Ayieko,  
J. Kabami, N. Sang, T. Liegler, G. Charmie, C.S. Camlin, V. Jain, K. Kadede,  
M. Atukunda, T. Ruel, S.B. Shade, E. Ssemmondo, D.M. Byonanebye,  
F. Mwangwa, A. Owaranaganise, W. Oolio, D. Black, K. Snyman, R. Burger,  
M. Getahun, J. Achando, B. Awuonda, H. Nakato, J. Kironde, S. Okiror,  
H. Thirumurthy, C. Koss, L. Brown, C. Marquez, J. Schwab, G. Lavoy, A. Plenty,  
E. Mugoma Wafula, P. Omanya, Y.-H. Chen, J.F. Rooney, M. Bacon,  
M. van der Laan, C.R. Cohen, E. Bukusi, M.R. Kamya, and M. Petersen

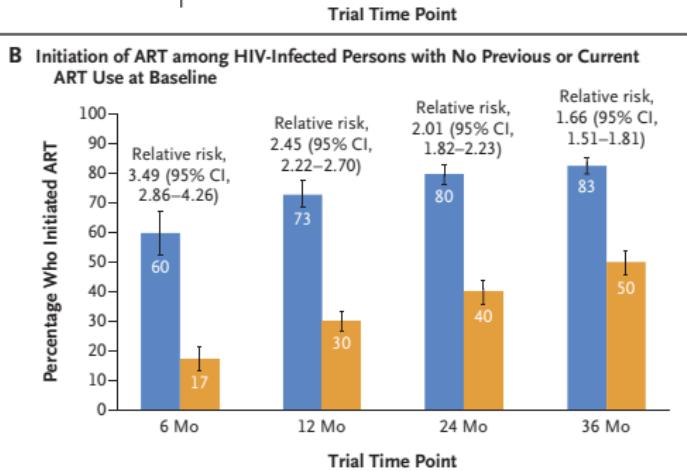
# Estimating the causal effect of a community-level intervention in a clustered RCT

The NEW ENGLAND JOURNAL of MEDICINE

## ORIGINAL ARTICLE

### HIV Testing and Treatment with the Use of a Community Health Approach in Rural Africa

D.V. Havlir, L.B. Balzer, E.D. Charlebois, T.D. Clark, D. Kwarisiima, J. A. J. Kabami, N. Sang, T. Liegler, G. Charmie, C.S. Camlin, V. Jain, K. Kad M. Atukunda, T. Ruel, S.B. Shade, E. Ssemmondo, D.M. Byonanyiby F. Mwangwa, A. Owaramanise, W. Oolio, D. Black, K. Snyman, R. Burg M. Getahun, J. Achando, B. Awuonda, H. Nakato, J. Kironde, S. Okir H. Thirumurthy, C. Koss, L. Brown, C. Marquez, J. Schwab, G. Lavoy, A. E. Mugoma Wafula, P. Omanya, Y.-H. Chen, J.F. Rooney, M. Bacon M. van der Laan, C.R. Cohen, E. Bukusi, M.R. Kamya, and M. Peters



# Increasing precision and accuracy by accounting for missing data in estimating impacts of HIV treatment program in clustered RCT

Research

JAMA | Original Investigation

## Association of Implementation of a Universal Testing and Treatment Intervention With HIV Diagnosis, Receipt of Antiretroviral Therapy, and Viral Suppression in East Africa

Maya Petersen, MD, PhD; Laura Balzer, PhD; Dalsone Kwartsima, MBChB, MPH; Norton Sang, MA; Gabriel Chamie, MD, MPH; James Ayieko, MBChB, MPH; Jane Kabami, MPH; Asiphas Owaraganise, MBChB; Teri Liegler, PhD; Florence Mwangwa, MBChB; Kevin Kadede, MA; Vivek Jain, MD, MAS; Albert Plenty, MS; Lillian Brown, MD, PhD; Geoff Lavoy; Joshua Schwab, MS; Douglas Black, BA; Mark van der Laan, PhD; Elizabeth A. Bukusi, MBChB, PhD; Craig R. Cohen, MD, MPH; Tamara D. Clark, MHS; Edwin Charlebois, MPH, PhD; Moses Kamya, MMed; Diane Havlir, MD

# Increasing precision and accuracy by accounting for missing data in estimating impacts of HIV treatment program in clustered RCT

Research

JAMA | Original Investigation

## Association of Implementation and Treatment Intervention of Antiretroviral Therapy

Table 2. Postbaseline HIV Viral Suppression in a Closed Cohort of HIV-Positive Stable Residents of 16 SEARCH Intervention Communities in Rural Uganda and Kenya Who Were Diagnosed At or Before Baseline (n = 7108)\*

Baseline Diagnosis, Treatment, and Suppression Status	No. of HIV-Positive Residents (%) <sup>a</sup>	Follow-up Year 1		Follow-up Year 2	
		No. of Residents With Viral Suppression/Total No. of Residents With Measured HIV RNA (%) <sup>a</sup>	Adjusted Proportion, % (95% CI) <sup>a</sup>	No. of Residents With Viral Suppression/Total No. of Residents With Measured HIV RNA (%) <sup>a</sup>	Adjusted Proportion, % (95% CI) <sup>a</sup>
Overall	7108 (100)	4682/5578 (83.9)	79.7 (78.7-80.8)	4602/5215 (88.2)	83.8 (82.8-84.9)
Newly diagnosed (HIV RNA≥500 copies/mL)	2080 (29.3)	963/1321 (72.9)	62.8 (60.4-65.2)	965/1205 (80.1)	68.8 (66.4-71.2)
Previously diagnosed with no ART (HIV RNA≥500 copies/mL)	990 (13.9)	649/812 (79.9)	78.1 (75.3-80.8)	685/778 (88.0)	86.5 (84.2-88.8)
Previous or current ART	4038 (56.8)	3070/3445 (89.1)	88.8 (87.7-89.9)	2952/3232 (91.3)	90.5 (89.4-91.6)
HIV RNA not measured	1063 (15.0)	732/846 (86.5)	86.6 (84.3-88.9)	685/779 (87.9)	87.2 (84.9-89.5)
HIV RNA≥500 copies/mL	426 (6.0)	175/355 (49.3)	49.5 (44.2-54.7)	204/325 (62.8)	62.2 (57.2-67.2)
HIV RNA<500 copies/mL	2549 (35.9)	2163/2244 (96.4)	96.3 (95.6-97.1)	2063/2128 (96.9)	96.8 (96.0-97.6)

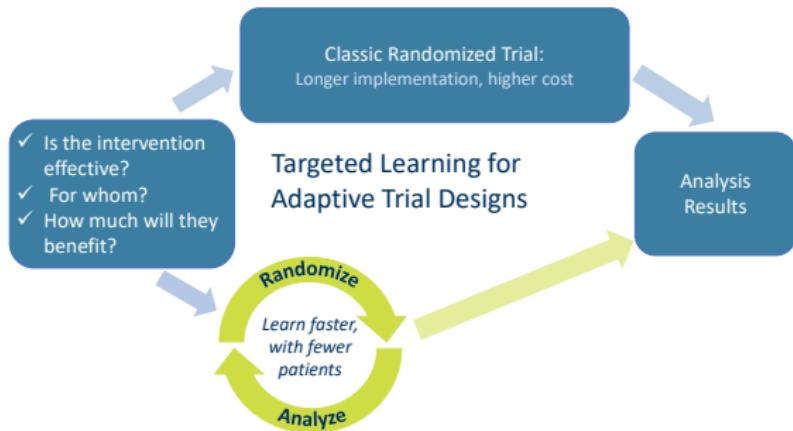
Maya Petersen, MD, PhD; Laura Balzer, PhD; Dalsone Kwasimba, MBChB, MPH; Norton Sang, MA; Gabriel Chamie, MD, MPH; James Ayleko, MBChB, MPH; Jane Kabami, MPH; Asiphas Owaraganise, MBChB; Teri Liegler, PhD; Florence Mwangwa, MBChB; Kevin Kadede, MA; Vivek Jain, MD, MAS; Albert Plenty, MS; Lillian Brown, MD, PhD; Geoff Lavoy; Joshua Schwab, MS; Douglas Black, BA; Mark van der Laan, PhD; Elizabeth A. Bukusi, MBChB, PhD; Craig R. Cohen, MD, MPH; Tamara D. Clark, MHS; Edwin Charlebois, MPH, PhD; Moses Kamya, MMed; Diane Havlir, MD



# Outline

# Robust estimation and inference for sequential designs adapting intervention allocation probabilities based on learning from past

## Optimal intervention allocation: “Learn as you go”



# Outline

# General Longitudinal Data Structure for Complex Observational Studies

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

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

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

**Survival outcome example:** For example,

$$A(t) = (A_1(t), A_2(t))$$

$A_1(t)$  = Indicator of being treated at time  $t$

$A_2(t)$  = Indicator of being right-censored at time  $t$

$Y(t)$  = Indicator of observing a failure by time  $t$

$L(t)$  Vector of time-dependent measurements

$Y(t) \subset L(t)$  and  $Y = Y(K + 1)$ .

# Comparing strategies for diabetes treatment intensification in Comparative Effectiveness Research (CER) study

- Data extracted from diabetes registries of 7 HMO research network sites: Kaiser Permanente, Group Health Cooperative, HealthPartners.
- Enrollment period: Jan 1<sup>st</sup> 2001 to Jun 30<sup>th</sup> 2009
- Enrollment criteria: past A1c < 7% (glucose level) while on 2+ oral agents or basal insulin and  $7\% \leq \text{latest A1c} \leq 8.5\%$  (study entry when glycemia was no longer reined in)

Longitudinal data:

- Follow-up til the earliest of Jun 30<sup>th</sup> 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.

# Comparing strategies for diabetes treatment intensification in Comparative Effectiveness Research (CER) study

Statistics  
in Medicine

Research Article

Received 24 May 2013,

Accepted 5 January 2014

Published online 17 February 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6099

## Targeted learning in real-world comparative effectiveness research with time-varying interventions

Romain Neugebauer,<sup>a\*†</sup> Julie A. Schmittiel<sup>a</sup> and  
Mark J. van der Laan<sup>b</sup>

# Comparing strategies for diabetes treatment intensification in Comparative Effectiveness Research (CER) study

Statistics  
in Medicine

Research Article

Received 24 May 2013,

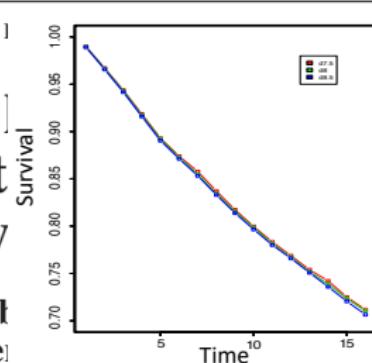
Accepted 5 January 2014

Published online 17 February 2014 in Wiley Online Library

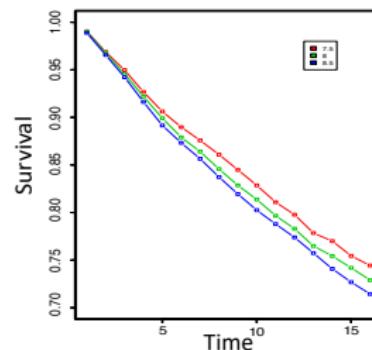
(wileyonlinelibrary.com)

Targeted  
comparat  
time-vary

Romain Neugeb  
Mark J. van de



**Standard methods:** No benefit to more aggressive intensification strategy



**Targeted Learning:** More aggressive intensification protocols result in better outcomes

# Estimating the cumulative, long-term impacts of environmental exposures

ORIGINAL ARTICLE

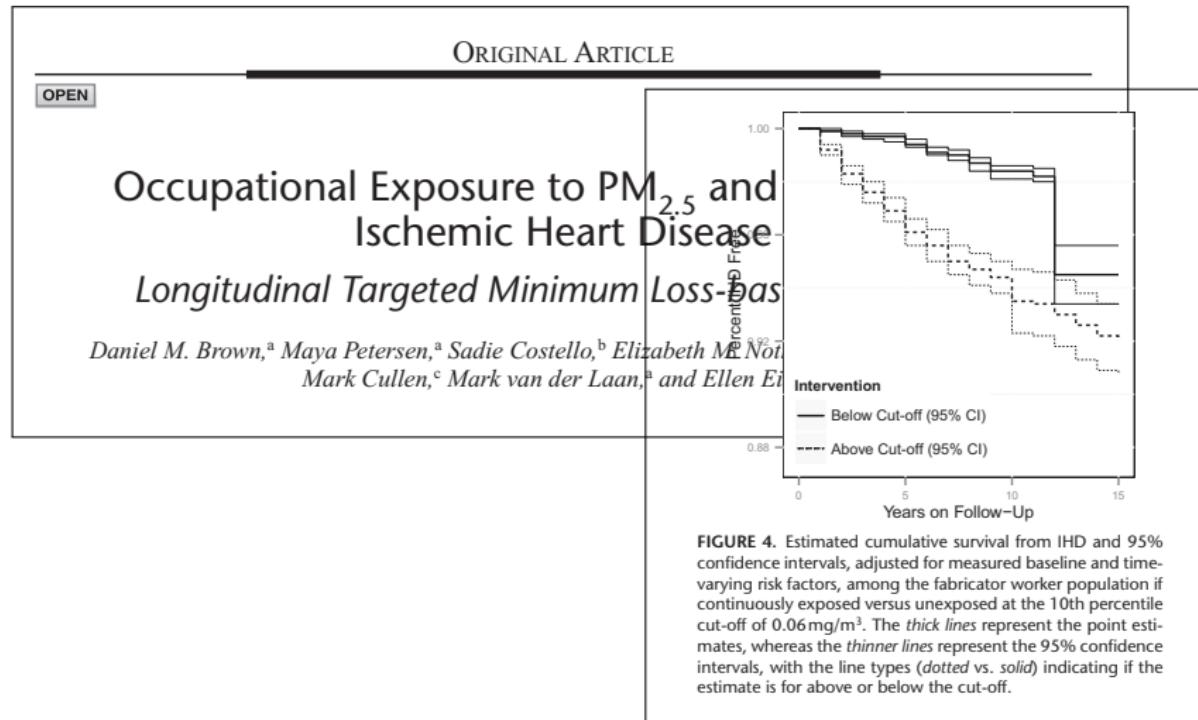
OPEN

## Occupational Exposure to PM<sub>2.5</sub> and Incidence of Ischemic Heart Disease

*Longitudinal Targeted Minimum Loss-based Estimation*

Daniel M. Brown,<sup>a</sup> Maya Petersen,<sup>a</sup> Sadie Costello,<sup>b</sup> Elizabeth M. Noth,<sup>b</sup> Katherine Hammond,<sup>b</sup>  
Mark Cullen,<sup>c</sup> Mark van der Laan,<sup>a</sup> and Ellen Eisen<sup>b</sup>

# Estimating the cumulative, long-term impacts of environmental exposures



# Estimating the impact of genetic polymorphisms on the efficacy of malaria vaccine on the time to infection

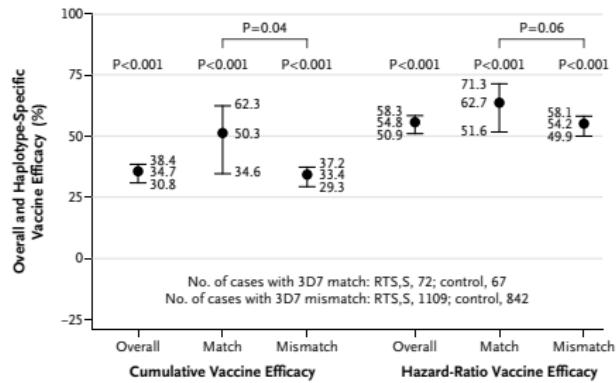
The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

## Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine

D.E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S.T. Agnandji, P. Aide, S. Anderson, D. Ansong, J.J. Aponte, K.P. Asante, P. Bejon, A.J. Birkett, M. Bruls, M.K. Connolly, U. D'Alessandro, C. Dobaño, S. Gesase, B. Greenwood, J. Grimesby, H. Tinto, M.J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P.G. Kremsner, A. Leach, B. Lell, N.J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J.T. Molet, E.L. Moss, P. Njuguna, C.F. Ockenhouse, B. Ragama Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D.J. Park, K. Pellé, D. Robbins, C. Russ, E.M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S.K. Volkman, Q. Yu, D. Lapierre, B.W. Birren, P.B. Gilbert, and D.F. Wirth

### D Cumulative and Hazard-Ratio Vaccine Efficacy



# Outline

## Inference with TMLE

- TMLE is **asymptotically linear with influence curve the canonical gradient**, so that Wald-type confidence intervals are based on estimating variance of its influence curve.
- The simple sample variance of influence curve can underestimate the variance if initial estimator is very adaptive or lack of positivity.
- Robust estimation of this variance by using sample splitting, or TMLE plug-in estimator corrects for this finite sample bias, and can be important (Tran et al, 19).
- One can also use the nonparametric bootstrap if one uses HAL as initial estimator (Cai, vdL, 19), resulting in better finite sample coverage by also picking up higher order behavior.

# Outline

## Advancing the vanilla TMLE: C-TMLE and extra targeting

- The least favorable parametric fluctuation model often depends on nuisance parameter (e.g., propensity score).
- C-TMLE targets estimation of this nuisance parameter based on criterion how well TMLE fits target estimand.
- Important for observational studies (vdL, Gruber, 2010 etc).
- By adding additional parameters to fluctuation model TMLE solves additional score equations that can be chosen to target second order remainder, and thereby improve finite sample performance.
- This has resulted in higher-order TMLE, double robust inference TMLE, etc (vdL, 14, Benkeser et al., Carone et al).

# Outline

# Preparing Statistical Analysis Plan based on TMLE

- Prior data or **outcome blind** data can be used to decide on **target estimand** supported by data.
- Prior data can also be used to set up **realistic simulation** to benchmark *specifications* of TMLE implementation, where benchmarks includes confidence interval coverage and type I error control.
- These **specifications of TMLE** include deciding on library of SL; sample splitting version; C-TMLE for nuisance parameter; adaptive truncation; TMLE-update step (e.g, possible extra targeting).
- Once one commits, it freezes the **a priori-specified estimator** that can be submitted as part of SAP for FDA approval.

# Outline

## tlverse - Targeted Learning software ecosystem in R

- A curated collection of R packages for Targeted Learning
- Shares a consistent underlying philosophy, grammar, and set of data structures
- Open source
- Designed for generality, usability, and extensibility
- Microwave dinners for machine learning

# tlverse outreach to train and support practitioners

- May 2019 - Atlantic Causal Inference Conference (ACIC) Workshop
- June 2019 - tlverse book →
- October 2019 - University of Pittsburgh School of Public Health Workshop
- November 2019 - Bill & Melinda Gates Foundation Workshop
- December 2019 - Deming Conference on Applied Statistics Workshop



- February 2020 - Conference on Statistical Practice (CSP) Workshop
- March 2020 - Alan Turing Institute Workshop

# Outline

## Concluding Remarks

- **Targeted Learning** *optimally estimates* the causal impact of an intervention on an outcome for complex real-world data.
- It integrates **causal inference, machine learning, statistical theory**.
- Targeted Learning learns better answers to causal, actionable questions which result in improved policy, treatments, etc.
- 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.

# Data Generating Experiments

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 The role of statistics in science
- 2 Examples of observed data experiments
- 3 Randomized trials
- 4 Observational studies
- 5 Conclusion

## The role of statistics in science

Science is about discovering laws that govern the universe and understanding the impact/interplay of those laws. Progress in science is often achieved through analyzing experiments tailored to answer specific scientific questions about a stochastic system.

Statistics is about the analysis of real world experiments. Given a particular type of experiment on a stochastic system and a specific question, a statistical method aims to answer the question through a point estimate as well as through quantification of uncertainty (confidence interval).

Statistics quantifies our knowledge (or lack thereof) about scientific laws.

# The role of statistics in science

The sorts of scientific laws we are interested in are those involving biomedical or public health interventions.

“If I perform *some intervention* on *some population*, what *effect* will it have?”

The actual observed data experiment is a crucial factor in determining

- what scientific questions can be answered;
- what scientific assumptions are needed to justify answers.

# Outline

- 1 The role of statistics in science
- 2 Examples of observed data experiments
- 3 Randomized trials
- 4 Observational studies
- 5 Conclusion

## Examples of observed data experiments

We generally refer to the observed output/data of the experiment on  $n$  units as  $n$  independent and identically (i.i.d.) distributed copies of a random variable  $O$ . But "i.i.d." is already a real assumption!

Since  $O$  is a random variable it has a probability distribution which we will denote with  $P_0$  (i.e.  $_0$  standing for the true probability distribution).

In general, we will use  $W$  to denote baseline characteristics,  $A$  to denote some kind of treatment/intervention,  $\Delta$  to denote some form of missingness, and  $Y$  to denote an outcome of interest.

We will generally be interested in asking the question: what happens to  $Y$  if we perform an intervention on  $A$ .

## Examples of observed data experiments

One observes  $n$  i.i.d. copies of  $O = (W, A, Y)$ ,  $W$  baseline covariates,  $A$  binary treatment,  $Y$  final outcome.

Example: Vaccine development – immunogeneity experiments

- Volunteers/lab animals receive a vaccine/placebo and have their immune response measured.
- Does the vaccine create an immune response?

Big Data Example: Electronic Medical Records – drug efficacy

- Patients are recorded as receiving e.g., a heart medication
- Does the medication protect against cardiovascular events?

## Examples of observed data experiments

We observe  $O = (W, A, \Delta, \Delta Y)$ : the outcome is subject to missingness.

Example: Vaccine development – immunogeneity experiments

- In the lab, a fraction of immune response measures cannot be measured

Big Data Example: Electronic Medical Records – drug efficacy

- Some patients never appear again in EMR database.

## Examples of observed data experiments

We observe  $O = (W, \Delta, \Delta A, Y)$ : the treatment assignment is subject to missingness.

Example: Biomarkers studies in epidemiology

- A fraction of biomarkers cannot be measured

Big Data Example: Electronic Medical Records – drug efficacy

- Patients prescriptions may not be recorded

## Examples of observed data experiments

We observe  $O = (W, A, (\Delta(t), \Delta(t)Y(t) : t = 1, \dots, \tau))$ , where  $\Delta(t)$  is indicator of subject being monitored at time  $t$  and  $Y(t)$  is indicator of event/failure at time  $t$ . Special case:  $\Delta(t) = I(C > \min(t, T))$  and  $Y(t) = I(T \leq t)$  for a right-censoring/follow-up time  $C$ .

Example: Preventive HIV vaccine efficacy trial

- Individuals are tested for HIV at regular clinic visits (e.g., every six months). Some individuals miss clinic visits.

Big Data Example: Electronic health records – depression

- Patients depressive symptoms are assessed at regular clinic visits. Some individuals miss clinic visits.

## Examples of observed data experiments

We observe  $n_1$  observations of  $(W, A)$ , from population with  $Y = 1$ , and  $n_0$  observations of  $(W, A)$ , from a population with  $Y = 0$ . This is called standard case-control sampling.

Example: Rare forms of cancer

- Cohort studies are expensive and for rare diseases it is more cost-efficient to sample (e.g., from cancer registry) based on outcome.

Example: Immune responses in vaccine trials

- Measuring immune responses is expensive, so in practice we measure immune response on infected participants and a subset of uninfected participants.

# Outline

- 1 The role of statistics in science
- 2 Examples of observed data experiments
- 3 Randomized trials
- 4 Observational studies
- 5 Conclusion

## Simple Randomized Trials

Not all observed data are created equal. We must also consider what we know about how the data were generated.

In a randomized trial the assignment of  $A$  is controlled by the experiment. For example, we flip a coin and assign  $A = 1$  if it is head.

- Examples: Clinical trials, lab experiments

However, one can also assign  $A = 1$  with a probability depending on patient characteristics  $W$ .

- Examples: Clinical trials where randomization is stratified by site

However, even though a randomized trial controls assignment of  $A$ , it does usually not completely control missingness  $\Delta$  or censoring/monitoring  $\Delta(t)$ .

## Sequential RCT

One observes  $n$  i.i.d. observations of  $O = (L_0, A_0, L_1, A_1, L_2 = Y) \sim P_0$ .

In a sequential RCT, one controls the assignment of both initial treatment  $A(0)$  and subsequent treatment  $A(1)$ .

Example: Cancer treatment trial with multiple possible courses of treatment

- We assign a drug among 4 drugs, see how well the patient responds, and possibly assign a different treatment if the patient is not responding.

## Adaptive (Group) Sequential RCT

Patients enroll over time. Treatment assignment is controlled for each patient, but can be based on the observed past among all previously enrolled patients.

Example: Drug development

- If, e.g., women appear to be responding well to the drug, we may want to increase the probability of assigning treatment to newly enrolled women.

# Outline

- 1 The role of statistics in science
- 2 Examples of observed data experiments
- 3 Randomized trials
- 4 Observational studies
- 5 Conclusion

## Simple Observational Studies

In an observational study the data generating experiment does not control the assignment of  $A$ .

Example: Epidemiological cohort studies on drug use

- The treatment decision is in the hands of the individual's medical doctor.
- We try to collect all patient characteristics the doctor might take into account when making a treatment decision.

## Complex Observational Studies

One observes  $n$  i.i.d. observations of

$$(L_0, A_0, \Delta_0, \Delta_0 L_1, A_1, \Delta_1, \Delta_1 L_2, \dots, A_K, \Delta_K, \Delta_K Y),$$

where  $A_t$  represents treatment assignment at time  $t$ ,  $\Delta_t$  is an indicator of being monitored at time  $t + 1$ ,  $L_t$  are time dependent covariates measured at time  $t$ , and  $Y = L_{K+1}$  is the final outcome.

Example: Drug safety study, also called Phase IV-study, are of this type.

# Outline

- 1 The role of statistics in science
- 2 Examples of observed data experiments
- 3 Randomized trials
- 4 Observational studies
- 5 Conclusion

## Key Points

- The scientific question and data generating experiment determines what statistical methods are needed.
- Better: The scientific question determines the data generating experiment and statistical method needed.
- The design of a study has a large impact on what we know about the underlying distribution of the observed data.

# Traditional Data Analysis

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 Traditional Approach in Epidemiology and Clinical Medicine
- 2 Complications of Human Art in Statistics
- 3 Machine learning

# Traditional Approach in Epidemiology and Clinical Medicine

In general, conventional statistical practice lets the type of data at hand dictate the scientific question of interest:

Goal	Type of Data			
	Measurement (from Gaussian Population)	Rank, Score, or Measurement (from Non-Gaussian Population)	Binomial (Two Possible Outcomes)	Survival Time
Describe one group	Mean, SD	Median, Interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample t-test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired t-test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired t-test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*
Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression*

# Traditional Approach in Epidemiology and Clinical Medicine

The internet abounds with prescriptive statistical recommendations.

- continuous outcome + covariates → linear regression
- continuous and positive outcome + covariates → log-linear regression
- survival outcome and covariates → proportional hazards regression

The “parameter of interest” is defined as the regression coefficient for  $A$  in the chosen parametric model.

These parametric models are often too simplistic to reflect the underlying complexity, but remain popular because they are easy to implement.

## Traditional Approach

In this approach, statistical convenience is allowed to determine the scientific question of interest.

Several critical questions naturally arise:

- ① Is the “parameter of interest” truly of scientific interest?
- ② If the model is misspecified, what are you actually estimating?

Why not begin by defining the estimand of interest?

- Dialogue with collaborators to determine the true question of interest.
- The definition should not rely on a parametric model, i.e., the parameter should be defined nonparametrically.
- Nevertheless, we could use parametric or semiparametric models to inspire interesting parameters.

# Traditional Approach in Epidemiology and Clinical Medicine

Suppose we observe  $O = (A, Y)$  for some continuous exposure  $A$  and some continuous outcome  $Y$ .

Our "Statistics for Dummies" book recommends we use linear regression, i.e., assume something like

$$Y = \alpha + \beta A + \epsilon ,$$

where  $\epsilon$  is a mean zero, Normal random variable and  $(\alpha, \beta)$  are unknown real numbers.

# Traditional Approach in Epidemiology and Clinical Medicine

What if the relationship between  $A$  and  $Y$  is not linear?

- Example: threshold effect

It turns out that the least-squares estimator  $\beta_n$  of  $\beta$  is estimating

$$\tilde{\beta}_0 = \frac{\text{Cov}(A, Y)}{\text{Var}(A)} .$$

Is this a relevant parameter for assessing treatment efficacy?

# Traditional Approach in Epidemiology and Clinical Medicine

An example from survival analysis: Suppose the outcome of interest is  $T$ , a survival time we are interested in the effect of  $A$  on  $Y$ .

Over the course of the study, some subjects may be lost to followup, say at time  $C$ .

The observed data consist of  $n$  i.i.d. copies of  $(A, \Delta, Y)$  where  $Y = \min(T, C)$  and  $\Delta = I(T \leq C)$ .

Our "Statistics for Dummies" book recommends that we should use a Cox model:

$$\lambda(t|1) = \exp(\gamma^*)\lambda(t|0) ,$$

where  $\lambda(t|a)$  is the conditional hazard of  $T$  at time  $t$  given  $A = a$ .

# Traditional Approach in Epidemiology and Clinical Medicine

What if the hazards between the treatment groups are not actually proportional over time?

- Example: waning vaccine efficacy

It turns out that the estimator  $\gamma_n$  from a Cox model is converging to the solution in  $\gamma$  of the equation

$$0 = \int w(t) \left\{ \frac{\lambda(t|1)}{\lambda(t|0)} - \exp(\gamma) \right\} dt ,$$

where the weights are given by

$$\frac{\lambda(t|0)P(T \geq t|A=1)P(C \geq t|A=1)}{1 + \exp(\gamma) \frac{P(A=1)}{P(A=0)} \frac{P(T \geq t|A=1)P(C \geq t|A=1)}{P(T \geq t|A=0)P(C \geq t|A=0)}} .$$

In other words, the “parameter of interest” depends on the **censoring distribution**.

# Traditional Approach in Epidemiology and Clinical Medicine

What is the alternative to this approach? We could define the parameter of interest as

$$\gamma_0 = \operatorname{argmin}_{\gamma} \int \left\{ \frac{\lambda(t|1)}{\lambda(t|0)} - \exp(\gamma) \right\}^2 d\mu(t),$$

where  $\mu$  is some weight function that we choose such that  $\int d\mu(t) = 1$ .

We can easily check that

$$\gamma_0 = \log \int \frac{\lambda(t|1)}{\lambda(t|0)} d\mu(t).$$

If the Cox model holds, then the two parameters correspond. If not (as will almost always be the case),  $\gamma_0$  defines a user-specified weighted average of the log-hazard ratio over time.

## Traditional Approach in Epidemiology and Clinical Medicine

In both examples, the “recommended” statistical procedures are estimating some contrast across different levels of  $A$ , but probably not the contrast that your medical collaborators actually care about.

The difference between what you think you’re estimating and what you’re actually estimating could be drastic and amplified by a large sample size.

Inference, even for the true estimand, can also be misleading without modification. This has spawned an entire branch of statistics – so-called “robust” statistics.

# Outline

1 Traditional Approach in Epidemiology and Clinical Medicine

2 Complications of Human Art in Statistics

3 Machine learning

# Complications of Human Art in Statistics

Returning to the linear regression example, classical statistical practice encourages users to "check" models after they have been fit.

If one of these checks fails, choose a new model: add quadratic term, remove a covariate, choose a different link function, choose a different error distribution.

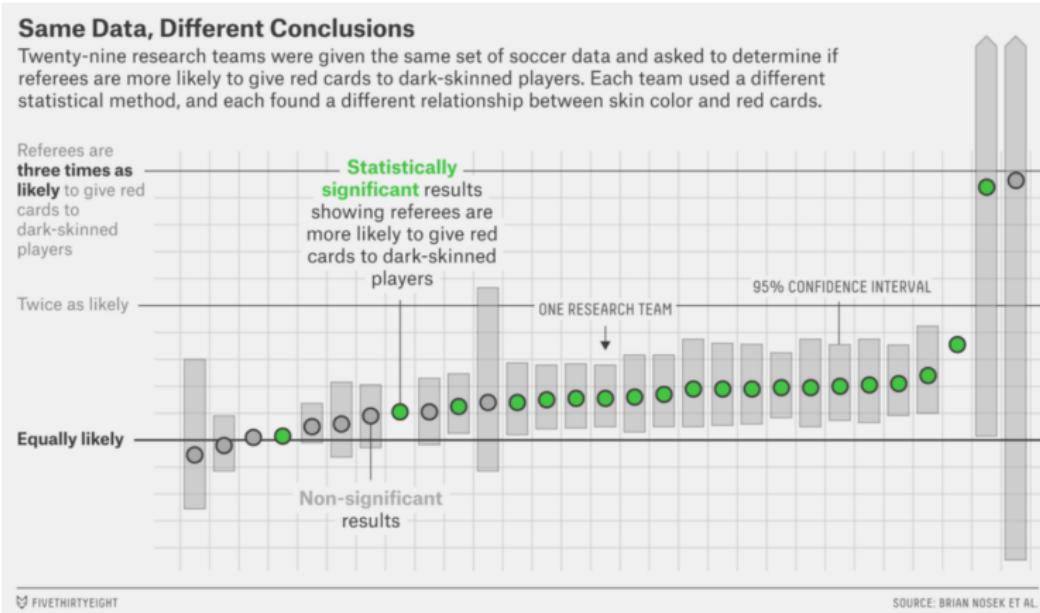
Problems with post-hoc procedures:

- ① Murky definition of target parameter; no correspondence with a scientific question.
- ② Even if model checks are performed honestly (a big if), inference does not account for ad-hocery!

# Complications of Human Art in Statistics

## Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



<http://fivethirtyeight.com/features/science-isnt-broken/# part1>

# Complications of Human Art in Statistics

*"Even the most skilled researchers must make subjective choices that have a huge impact on the result they find."*

- This should cause extreme discomfort!

# Summary

Problems with classical statistics:

- ① Parametric models are misspecified.
- ② Researchers often interpret the target parameter as if the parametric model is correct.
- ③ The parametric model is often data-adaptively (or worse!) selected, and this part of the estimation procedure is not accounted for in the variance.
- ④ Due to guaranteed bias in coefficient, if the null hypothesis of no effect is true, with probability tending to 1 as sample size converges to infinity, **one will falsely reject the null**.

# Outline

- 1 Traditional Approach in Epidemiology and Clinical Medicine
- 2 Complications of Human Art in Statistics
- 3 Machine learning

# Machine learning

Machine learning is garnering huge attention.

Machine learning primarily deals with training algorithms that can predict an outcome based on input features.

- Examples: self-driving cars, text recognition, image recognition, Alpha-Go

These algorithms are able to predict much better than the tools developed in classical statistics (e.g., prediction based on parametric regression models) owing to their flexibility and ability to adapt to underlying data.

However, this does not mean that we no longer have a need for statistics!

# Machine learning

Consider the (possibly apocryphal) story of the US Army using neural networks to train an algorithm to recognize camouflaged tanks.

- Training data consisted of pictures of tanks camouflaged in trees and trees without tanks.

The classifier was found to have incredible predictive accuracy on the training data, but essentially no accuracy when used in practice.

It turns out, all the tank pictures were taken on a sunny day and all the control pictures were taken on a cloudy day and this is what the neural net had learned.

- Understanding sampling and experimental design is important!

# Machine learning

Prediction is not always of scientific interest. Even in settings where prediction is of interest, we still need statistics to assess the performance of our predictions (and corresponding uncertainty!).

Nevertheless, these methods for learning from data are exciting and have been shown time and again to perform far better than the tools in our classical statistics tool box.

The challenge facing modern statisticians (and you, as students in this class) is how to develop statistical methods aimed at answering specific scientific questions of interest, while utilizing state-of-the-art algorithms.

# Structural Causal Model, Causal Quantity, Identification

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Statistical model

We are considering the general case that one observed  $n$  i.i.d. copies of a random variable  $O$  with probability distribution  $P_0$ .

The data-generating distribution  $P_0$  is also known to be an element of a statistical model  $\mathcal{M}$ :  $P_0 \in \mathcal{M}$ .

A **statistical model**  $\mathcal{M}$  is the set of possible probability distributions for  $P_0$ ; it is a collection of probability distributions.

If all we know is that we have  $n$  i.i.d. copies of  $O$ , this can be our statistical model, which we call a nonparametric statistical model

## Statistical model augmented with causal assumptions

A statistical model can be augmented with additional (nontestable causal) assumptions, allowing one to enrich the interpretation of  $\Psi(P_0)$ .

This does not change the statistical model.

We refer to the statistical model augmented with a possibly additional assumptions as the **model**.

- Causal assumptions made by the structural causal model (SCM)

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Defining the SCM

- We first specify a set of endogenous variables  $X = (X_j : j)$ .
- Endogenous variables are those variables for which the SCM will state that it is a (typically unknown) deterministic function of some of the other endogenous variables and an exogenous error.
- Typically, the endogenous variables  $X$  include the observables  $O$ , but might also include some non-observables that are meaningful and important to the scientific question of interest. Perhaps there was a variable you did not measure, but would have liked to, and it plays a crucial role in defining the scientific question of interest. This variable would then be an unobserved endogenous variable.

## Defining the SCM

- In a very simple example, we might have  $j = 1, \dots, J$ , where  $J = 3$ . Thus,  $X = (X_1, X_2, X_3)$ .
- We can rewrite  $X$  as  $X = (W, A, Y)$  if we say  $X_1 = W$ ,  $X_2 = A$ , and  $X_3 = Y$ .
- Let  $W$  represent the set of baseline covariates for a subject,  $A$  the treatment or exposure, and  $Y$  the outcome.
- All the variables in  $X$  are observed.

## Defining the SCM

- For each endogenous variable  $X_j$  one specifies the parents of  $X_j$  among  $X$ , denoted  $Pa(X_j)$ .
- The specification of the parents might be known by the time ordering in which the  $X_j$  were collected over time: the parents of a variable collected at time  $t$  could be defined as the observed past at time  $t$ .
- We can see the time ordering involved in this process: the baseline covariates occurred before the exposure LTPA, which occurred before the outcome of death:  $W \rightarrow A \rightarrow Y$ .

## Defining the SCM

- We denote a collection of exogenous variables by  $U = (U_{X_j} : j)$ .
- These variables in  $U$  are never observed and are not affected by the endogenous variables in the model, but instead they affect the endogenous variables.
- One assumes that  $X_j$  is some function of  $Pa(X_j)$  and an exogenous  $U_{X_j}$ :

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), \quad j = 1 \dots, J.$$

- The collection of functions  $f_{X_j}$  indexed by all the endogenous variables is represented by  $f = (f_{X_j} : j)$ .
- Together with the joint distribution of  $U$ , these functions  $f_{X_j}$ , specify the data-generating distribution of  $(U, X)$  as they describe a deterministic system of structural equations (one for each endogenous variable  $X_j$ ) that deterministically maps a realization of  $U$  into a realization of  $X$ .

## Defining the SCM

- In an SCM one also refers to some of the endogenous variables as intervention variables.
- The SCM assumes that intervening on one of the intervention variables by setting their value, thereby making the function for that variable obsolete, does not change the form of the other functions.
- The functions  $f_{X_j}$  are often unspecified, but in some cases it might be reasonable to assume that these functions have to fall in a certain more restrictive class of functions.
- Similarly, there might be some knowledge about the joint distribution of  $U$ .

## Defining the SCM

- The set of possible data-generating distributions of  $(U, X)$  can be obtained by varying the structural equations  $f$  over all allowed forms, and the distribution of the errors  $U$  over all possible error distributions defines the SCM for the full-data  $(U, X)$ , i.e., the SCM is a statistical model for the random variable  $(U, X)$ .
- An example of a fully parametric SCM would be obtained by assuming that all the functions  $f_{X_j}$  are known up to a finite number of parameters and that the error distribution is a multivariate normal distribution with mean zero and unknown covariance matrix. Such parametric structural equation models are not recommended.

## Defining the SCM

The corresponding SCM for the observed data  $O$  also includes specifying the relation between the random variable  $(U, X)$  and the observed data  $O$ , so that the SCM for the full data implies a parameterization of the probability distribution of  $O$  in terms of  $f$  and the distribution  $P_U$  of  $U$ . This SCM for the observed data also implies a statistical model for the probability distribution of  $O$ .

## Defining the SCM: Translation

We have the functions  $f = (f_W, f_A, f_Y)$  and the exogenous variables  $U = (U_W, U_A, U_Y)$ . The values of  $W$ ,  $A$ , and  $Y$  are deterministically assigned by  $U$  corresponding to the functions  $f$ . We specify our structural equation models, based on investigator knowledge, as

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{1}$$

where no assumptions are made about the true shape of  $f_W$ ,  $f_A$ , and  $f_Y$ . These functions  $f$  are nonparametric as we have not put a priori restrictions on their functional form.

## Defining the SCM: Translation

- We may assume that  $U_A$  is independent of  $U_Y$ , given  $W$ , which corresponds with believing that there are no unmeasured factors that predict both  $A$  and the outcome  $Y$ : this is often called the no unmeasured confounders assumption.
- This SCM represents a semiparametric statistical model for the probability distribution of the errors  $U$  and endogenous variables  $X = (W, A, Y)$ .
- We assume that the observed data structure  $O = (W, A, Y)$  is actually a realization of the endogenous variables  $(W, A, Y)$  generated by this system of structural equations.

This now defines the SCM for the observed data  $O$ .

## Defining the SCM: Translation

We have assumed that the underlying data were generated by the following actions:

- ① Drawing unobservable  $U$  from some probability distribution  $P_U$  ensuring that  $U_A$  is independent of  $U_Y$ , given  $W$ ,
- ② Generating  $W$  as a deterministic function of  $U_W$ ,
- ③ Generating  $A$  as a deterministic function of  $W$  and  $U_A$ ,
- ④ Generating  $Y$  as a deterministic function of  $W$ ,  $A$ , and  $U_Y$ .

## Defining the SCM

- Any probability distribution of  $O$  can be obtained by selecting a particular data-generating distribution of  $(U, X)$  in this SCM.
- Thus, the statistical model for  $P_0$  implied by this SCM is a nonparametric model.
- As a consequence, one cannot determine from observing  $O$  if the assumptions in the SCM contradict the data.
- One states that the SCM represents a set of nontestable causal assumptions we have made about how the data were generated in nature.

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

# Causal Graphs

$U_W[d] U_A[d]$

$W[r] @/_/[rrd] A[dr] U_Y[dl]$   
Y

**Figure:** A possible causal graph for (1).

## Causal Graphs

$U_W[d] @<--> @/ \cdot 5pc / [r] @<--> @/ \cdot 6pc / [drrr] U_A[d] @<--> @/ \cdot 1pc / [$

W [r] @/\_/[rrd] A[dr] U\_Y[dl]  
Y

Figure: A causal graph for (1) with no assumptions on the distribution of  $P_U$

# Causal Graphs

$U_W[d] @<--> @/2.5pc/[drrr] U_A[d]$   $U_W[d] @<--> @/5pc/[r] U_A[d]$

$W[r] @/_/[rrd] A[dr]$   $U_Y[dl] W[r] @/_/[rrd] A[dr]$   $U_Y[dl]$

Y      Y

Figure: Causal graphs for (1) with various assumptions about the distribution of  $P_U$

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Defining the Causal Target Parameter

We can explicitly define the target parameter of the probability distribution  $P_0$  as some function of  $P_0$ :  $\Psi(P_0)$ .

We are interested in estimating a parameter  $\Psi(P_0)$  of the probability distribution  $P_0 \in \mathcal{M}$ , which is known to be an element of a non-parameteric (or semiparametric) statistical model  $\mathcal{M}$ .

## Defining the Causal Target Parameter

- Formally, we denote the SCM for the full-data  $(U, X)$  by  $\mathcal{M}^F$ , a collection of possible  $P_{U,X}$  as described by the SCM.
- In other words,  $\mathcal{M}^F$ , a model for the full data, is a collection of possible distributions for the underlying data  $(U, X)$ .
- $\Psi^F$  is a mapping applied to a  $P_{U,X}$  giving  $\Psi^F(P_{U,X})$  as the target parameter of  $P_{U,X}$ .

## Defining the Causal Target Parameter

- This mapping needs to be defined for each  $P_{U,X}$  that is a possible distribution of  $(U, X)$ , given our assumptions coded by the posed SCM.
- We state  $\Psi^F : \mathcal{M}^F \rightarrow \mathbb{R}^d$ , where  $\mathbb{R}^d$  indicates that our parameter is a vector of  $d$  real numbers.
- The SCM  $\mathcal{M}^F$  consists of the distributions indexed by the deterministic function  $f = (f_{X_j} : j)$  and distribution  $P_U$  of  $U$ , where  $f$  and this joint distribution  $P_U$  are identifiable from the distribution of the full-data  $(U, X)$ .
- Thus the target parameter can also be represented as a function of  $f$  and the joint distribution of  $U$ .

## Defining the Causal Target Parameter

- Recall our example with data structure  $O = (W, A, Y)$  and SCM given in (1) with no assumptions about the distribution  $P_U$ .
- We can define  $Y_a = f_Y(W, a, U_Y)$  as a random variable corresponding with intervention  $A = a$  in the SCM.
- The marginal probability distribution of  $Y_a$  is thus given by

$$P_{U,X}(Y_a = y) = P_{U,X}(f_Y(W, a, U_Y) = y).$$

- The causal effect of interest for a binary  $A$  (suppose it is the causal risk difference) could then be defined as a parameter of the distribution of  $(U, X)$  given by

$$\Psi^F(P_{U,X}) = E_{U,X} Y_1 - E_{U,X} Y_0.$$

- In other words,  $\Psi^F(P_{U,X})$  is the difference of marginal means of counterfactuals  $Y_1$  and  $Y_0$ .

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Interventions on the causal model

- We will define our causal target parameter as a parameter of the distribution of the data  $(U, X)$  under an intervention on one or more of the structural equations in  $f$ .
- The intervention defines a random variable that is a function of  $(U, X)$ , so that the target parameter is  $\Psi^F(P_{U,X})$ .
- Intervening on the system defined by our SCM describes the data that would be generated from the system at the different levels of our intervention variable (or variables).

## Interventions

By assumption, intervening and changing the functions  $f_{X_j}$  of the intervention variables does not change the other functions in  $f$ . With the SCM given in (1) we can intervene on  $f_A$  and set  $a = 1$ :

$$\begin{aligned}W &= f_W(U_W), \\a &= 1, \\Y_1 &= f_Y(W, 1, U_Y).\end{aligned}$$

We can also intervene and set  $a = 0$ :

$$\begin{aligned}W &= f_W(U_W), \\a &= 0, \\Y_0 &= f_Y(W, 0, U_Y).\end{aligned}$$

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Counterfactuals

- We would ideally like to see each individual's outcome at all possible levels of exposure  $A$ . The study is only capable of collecting  $Y$  under one exposure, the exposure the subject experiences.
- $Y_a$  represents the outcome that would have been observed under this system for a particular subject under exposure  $a$ .
- In our example, for each realization  $u$ , which might correspond with an individual randomly drawn from some target population, by intervening on (1), we can generate so-called counterfactual outcomes  $Y_1(u)$  and  $Y_0(u)$ .

## Counterfactuals

- These counterfactual outcomes are implied by our SCM; they are consequences of it.
- That is,  $Y_0(u) = f_Y(W, 0, u_Y)$ , and  $Y_1(u) = f_Y(W, 1, u_Y)$ , where  $W = f_W(u_W)$  is also implied by  $u$ .
- The random counterfactuals  $Y_0 = Y_0(U)$  and  $Y_1 = Y_1(U)$  are random through the probability distribution of  $U$ .
- For example, the expected outcome of  $Y_1$  is the mean of  $Y_1(u)$  with respect to the probability distribution of  $U$ . Our target parameter is a function of the probability distributions of these counterfactuals:  
 $E_0 Y_1 - E_0 Y_0$ .

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Establishing Identifiability

Are the assumptions we have already made enough to express the causal parameter of interest as a parameter of the probability distribution  $P_0$  of the observed data?

We want to be able to write  $\Psi^F(P_{U,X,0})$  as  $\Psi(P_0)$  for some parameter mapping  $\Psi$ .

Since the true probability distribution of  $(U, X)$  can be any element in the SCM  $\mathcal{M}^F$ , and each such choice  $P_{U,X}$  implies a probability distribution  $P(P_{U,X})$  of  $O$ , this requires that we show that  $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$  for all  $P_{U,X} \in \mathcal{M}^F$ .

## Establishing Identifiability

This step involves establishing possible additional assumptions on the distribution of  $U$ , or sometimes also on the deterministic functions  $f$ , so that we can identify the target parameter from the observed data distribution.

Thus, for each probability distribution of the underlying data  $(U, X)$  satisfying the SCM with these possible additional assumptions on  $P_U$ , we have  $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$  for some  $\Psi$ .

$O$  is implied by the distribution of  $(U, X)$ , such as  $O = X$  or  $O \subset X$ , and  $P = P(P_{X,U})$ , where  $P(P_{U,X})$  is a distribution of  $O$  implied by  $P_{U,X}$ .

## Establishing Identifiability

Let us denote the resulting full-data SCM by  $\mathcal{M}^{F*} \subset \mathcal{M}^F$  to make clear that possible additional assumptions were made that were driven purely by the identifiability problem, not necessarily reflecting reality.

## Establishing Identifiability

Theorems exist that are helpful to establish such a desired identifiability result. For example, if  $O = (W, A, Y)$ , and the distribution of  $U$  is such that,  $A$  is independent of  $Y_1$ , given  $W$ , then the well-known g-formula expresses the distribution of  $Y_1$  in terms of the distribution of  $O$ :

$$P(Y_1 = y) = \int_w P(Y = y \mid A = 1, W = w) dP_W(w).$$

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter**
- 9 Positivity assumption
- 10 Target parameter

## Commit to a Statistical Model and Target Parameter

The identifiability result provides us with a purely statistical target parameter  $\Psi(P_0)$  on the distribution  $P_0$  of  $O$ .

The full-data model  $\mathcal{M}^{F^*}$  implies a statistical observed data model  $\mathcal{M} = \{P(P_{X,U}) : P_{X,U} \in \mathcal{M}^{F^*}\}$  for the distribution  $P_0 = P(P_{U,X,0})$  of  $O$ .

This now defines a target parameter  $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ .

## Commit to a Statistical Model and Target Parameter

The statistical observed data model for the distribution of  $O$  might be the same for  $\mathcal{M}^F$  and  $\mathcal{M}^{F*}$ .

If not, then one might consider extending the  $\Psi$  to the larger statistical observed data model implied by  $\mathcal{M}^F$ , such as possibly a fully nonparametric model allowing for all probability distributions.

If the more restricted SCM holds, our target parameter would still estimate the target parameter, but one now also allows the data to contradict the more restricted SCM based on additional doubtful assumptions.

## Commit to a Statistical Model and Target Parameter

The causal risk difference in our simple example, in terms of the corresponding statistical parameter  $\Psi(P_0)$ :

$$\begin{aligned}\Psi^F(P_{U,x,0}) &= E_0 Y_1 - E_0 Y_0 \\ &= E_0[E_0(Y | A = 1, W) - E_0(Y | A = 0, W)] \\ &\equiv \Psi(P_0)\end{aligned}$$

where the outer expectation in the definition of  $\Psi(P_0)$  is the mean across the strata for  $W$ .

## Commit to a Statistical Model and Target Parameter

This identifiability result for the additive causal effect as a parameter of the distribution  $P_0$  of  $O$  required making the randomization assumption stating that  $A$  is independent of the counterfactuals  $(Y_0, Y_1)$  within strata of  $W$ .

This assumption might have been included in the original SCM  $\mathcal{M}^F$ , but, if one knows there are unmeasured confounders, then the model  $\mathcal{M}^{F*}$  would be more restrictive by enforcing this “known to be wrong” randomization assumption.

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Positivity

Another required assumption is that  $P_0(A = 1, W = w) > 0$  and  $P_0(A = 0, W = w) > 0$  are positive for each possible realization  $w$  of  $W$ . Without this assumption, the conditional expectations of  $Y$  in  $\Psi(P_0)$  are not well defined. This positivity assumption is also called the experimental treatment assignment (ETA) assumption.

# Outline

- 1 The model
- 2 Causal model
- 3 Causal graphs
- 4 Causal target parameter
- 5 Interventions
- 6 Counterfactuals
- 7 Identifiability
- 8 Commit to a statistical model and target parameter
- 9 Positivity assumption
- 10 Target parameter

## Target Parameter

To be very explicit about how this parameter corresponds with mapping  $P_0$  into a number:

$$\begin{aligned}\Psi(P_0) &= \sum_w \left[ \sum_y y P_0(Y = y \mid A = 1, W = w) \right. \\ &\quad \left. - \sum_y y P_0(Y = y \mid A = 0, W = w) \right] P_0(W = w),\end{aligned}$$

where

$$P_0(Y = y \mid A = a, W = w) = \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}$$

is the conditional probability distribution of  $Y = y$ , given  $A = a, W = w$ , and

$$P_0(W = w) = \sum_{y,a} P_0(Y = y, A = a, W = w)$$

is the marginal probability distribution of  $W = w$ .

## Interpretation of the Target Parameter

The observed data parameter  $\Psi(P_0)$  can be interpreted in two possibly distinct ways:

- ①  $\Psi(P_0)$  with  $P_0 \in \mathcal{M}$  augmented with the truly reliable additional nonstatistical assumptions that are known to hold (e.g.,  $\mathcal{M}^F$ ). This may involve bounding the deviation of  $\Psi(P_0)$  from the desired target causal effect  $\Psi^F(P_{U,x,0})$  under a realistic causal model  $\mathcal{M}^F$  that is not sufficient for the identifiability of this causal effect.
- ② The truly causal parameter  $\Psi^F(P_{U,x}) = \Psi(P_0)$  under the more restricted SCM  $\mathcal{M}^{F*}$ , thereby now including all causal assumptions that are needed to make the desired causal effect identifiable from the probability distribution  $P_0$  of  $O$ .

## Example target parameter: Average causal effect

Causal risk difference:

$$\begin{aligned}\Psi(P_0) &= E_0[E_0(Y | A = 1, W) - E_0(Y | A = 0, W)] \\ &= E_0 Y_1 - E_0 Y_0\end{aligned}$$

## Example target parameter: Average Causal Effect Among the Treated

Consider the following modified system of structural equations:

$W = f_W(U_W)$ ,  $A = f_A(W, U_A)$ ,  $A^* = 1$ ,  $Y_1 = f_Y(W, A^*, U_Y)$ . Similarly, we can define this for  $A^* = 0$ . We can now define the causal quantity (i.e.,  $\Psi^F(P_{U,X})$ )

$$E_0(Y_1 - Y_0 \mid A = 1).$$

This is called the effect among the treated. Under RA it is identified by:

$$E_0(Y_1 - Y_0 \mid A = 1) = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 1).$$

Let  $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$  and  $g_0(a \mid W) = P_0(A = a \mid W)$ . Then,

$$E(Y_1 - Y_0 \mid A = 1) = \int_w \{\bar{Q}_0(1, w) - \bar{Q}_0(0, w)\} \frac{g_0(1 \mid w)}{P_0(A = 1)} dP_0(w).$$

Only positivity assumption needed is  $P(A = 1 \mid W) > 0$ .

## Example: Average treatment effect among compliers, using an instrument

- Suppose that we observe  $(W, R, A, Y)$ , where  $W$  is covariate vector,  $R$  is randomized treatment (instrument),  $A$  is actual treatment the subject took, and  $Y$  is outcome.
- This corresponds with a structural equation model  $W = f_W(U_W)$ ;  $R = f_R(W, U_R)$ ;  $A = f_A(W, R, U_A)$ ;  $Y = f_Y(W, A, U_Y)$ , assuming that the outcome is only affected by actual treatment.
- $R$  is an instrument since  $U_R$  is independent of  $(U_A, U_Y)$ , given  $W$ , i.e assignment of  $R$  is not affected by unmeasured confounders, while it strongly predicts  $A$  (strength of instrument).

- If we also assume  $P(A = 1 \mid R = 1, W) \geq P(A = 1 \mid R = 0, W)$ , then we have identification of the causal effect among the compliers

$$E(Y_1 - Y_0 \mid A_{R=1} = 1, A_{R=0} = 0)$$

by

$$\frac{E(E(Y \mid R = 1, W) - E(Y \mid R = 0, W))}{E(E(A \mid R = 1, W) - E(A \mid R = 0, W))}.$$

## Examples of Interventions: Optimal, Multiple Time-Point and Stochastic

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

1 Dynamic interventions

2 Stochastic interventions

3 Multiple time-point interventions

## Dynamic intervention

- $O = (W, A, Y)$ ,  $A$  binary treatment,  $Y$  indicator of bad outcome.
- $W = f_W(U_W)$ ;  $A = f_A(W, U_A)$ ;  $Y = f_Y(W, A, U_Y)$ .
- A dynamic intervention  $W \rightarrow d(W) \in \{0, 1\}$  is a rule that maps characteristics  $W$  of subject into a treatment decision  $A = d(W)$ .
- The counterfactual outcome under intervention  $d$  is given by  $Y_d \equiv f_Y(W, d(W), U_Y)$ .
- $EY_d$  is the mean of  $Y_d$ .

# Optimal Dynamic Intervention

The optimal rule  $W \rightarrow d_0(W)$  is defined by

$$d_0 = \arg \min_d E_0 Y_d.$$

It is given by:

$$d_0(W) = I(B_0(W) > 0),$$

where

$$B_0(W) = E_0(Y | A = 1, W) - E_0(Y | A = 0, W)$$

is the conditional additive treatment effect.

Both the rule  $d_0$  as well as its performance  $E_0 Y_{d_0}$  are quantities of interest in precision medicine.

# Outline

1 Dynamic interventions

2 Stochastic interventions

3 Multiple time-point interventions

## Stochastic Intervention

- Let  $G^*$  be a conditional probability distribution of  $A$ , given  $W$ .
- We could modify the structural equation model by replacing the equation  $A = f_A(W, U_A)$  by drawing  $A^* \sim G^*(\cdot | W)$ . One can also define  $A^* = d(W, U^*)$  for a rule  $d$  and random error  $U^*$ .
- This defines a counterfactual  $Y_{G^*} = f_Y(W, A^*, U_Y)$ .
- The mean outcome  $E_0 Y_{G^*}$  is the quantity of interest.

## Identification of mean outcome under stochastic intervention

- Recall  $\bar{Q}_0(A, W) = E_0(Y | A, W)$ ,  $Q_{W,0}$  is probability distribution of  $W$ .
- Under RA, it is identified by

$$E_0 Y_{G^*} = \int_{a,w} \bar{Q}_0(a, w) dG^*(a | w) dQ_{W,0}(w).$$

## Examples of Stochastic Interventions

- $A^* \sim Bernoulli(p)$  for some known  $p$ .
- $A^* \sim Bernoulli(p(W))$  for some known  $p(W)$ .
- $A^* = A + \delta$  for a deterministic rule. This corresponds with first drawing  $A$  from the treatment mechanism, and subsequently evaluating  $A + \delta$ :

$$g^*(a^* | W) = g_0(a^* - \delta | W).$$

- More generally, if  $A^* = d(A, W)$ , then

$$g^*(a^* | W) = g_0(d_W^{-1}(a^*) | W),$$

where  $d_W^{-1}$  is the inverse function of  $a \rightarrow d(a, W)$ .

## Missing and Censoring Indicators can be included in SCM as endogenous variables

- For example, suppose that our observed data structure on one unit is  $(W, A, \Delta, Y^* = \Delta Y) \sim P_0$ .
- We define structural equation model:  $W = f_W(U_W)$ ,  $A = f_A(W, U_A)$ ,  $\Delta = f_\Delta(W, A, U_\Delta)$ ,  $Y = f_Y(W, A, U_Y)$ ,  $Y^* = \Delta Y$ .
- The counterfactual  $Y_1^*$  of interest is now the one corresponding with intervention  $A = 1$  and  $\Delta = 1$ , and similarly  $Y_0^*$ .
- Under RA, the average causal effect  $E_0 Y_1^* - E_0 Y_0^*$  is identified by

$$E_0\{E_0(Y^* | A = 1, \Delta = 1, W) - E_0(Y^* | A = 0, \Delta = 1, W)\}.$$

# Outline

- 1 Dynamic interventions
- 2 Stochastic interventions
- 3 Multiple time-point interventions

## Identification of post-Intervention distribution for multiple time-point interventions: G-Computation Formula

Suppose  $O = (L(0), A(0), L(1), A(1), L(2) = Y) \sim P_0$ , where  
 $A(t) = (A_1(t), \Delta(t))$  with  $A_1(t)$  treatment and  $\Delta(t)$  monitoring indicator.  
We can define an SCM:

$$\begin{aligned}L(0) &= f_{L(0)}(U_{L(0)}) \\A(0) &= f_{A(0)}(L(0), U_{A(0)}) \\L(1) &= f_{L(1)}(L(0), A(0), U_{L(1)}) \\A(1) &= f_{A(1)}(L(0), A(0), L(1), U_{A(1)}) \\Y &= f_Y(L(0), A(0), L(1), A(1), U_Y).\end{aligned}$$

Consider a stochastic intervention  $g_{A(0)}^*, g_{A(1)}^*$  on  $(A(0), A(1))$ . This defines counterfactual  $O_{g^*} = (L(0), A^*(0), L_{g^*}(1), A^*(1), L_{g^*}(2))$ .

## Identification by G-computation formula under Sequential Randomization Assumption

- Assume SRA:  $A(j)$  is independent of  $Y_{g^*}$ , given  $\bar{L}(j)$ ,  $\bar{A}(j-1)$ ,  $j = 0, 1$ .
- Let  $q_{L(j)}$  be the conditional density of  $L(j)$ , given  $\bar{L}(j-1)$ ,  $\bar{A}(j-1)$ .  
The distribution  $P_{g^*}$  of  $L_{g^*}$  is identified by the density  
 $p^{g^*} = q_{L(0)} g_{A(0)}^* q_{L(1)} g_{A(1)}^* q_{L(2)}$ :

$$\begin{aligned} p_{g^*}(o) &= q_{L(0)}(I(0)) g_{A(0)}^*(a(0) | I(0)) \\ &\quad q_{L(1)}(I(1) | I(0), a(0)) g_{A(1)}^*(a(1) | I(0), a(0), I(1)) \\ &\quad q_{L(2)}(I(2) | I(0), a(0), I(1), a(1)). \end{aligned}$$

- The existence of this density relies on conditioning events having positive probability (the positivity assumption):

$$\frac{g_{A(j)}^*(a(j) \mid \bar{L}(j), \bar{A}(j-1))}{g_{0,A(j)}(a(j) \mid \bar{L}(j), \bar{A}(j-1))} < \infty,$$

across all possible histories  $\bar{L}(j), \bar{A}(j-1)$ .

- That is, if the probability that one assigns the value  $a(j)$  to a unit with history  $\bar{L}(j), \bar{A}(j-1)$  equals zero, then we also need that the stochastic intervention assigns this value  $a(j)$  with probability zero.

# Understanding Nonparametric Density Estimation: Super Learning of a Density

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Kernel density estimation

Suppose we observe  $n$  iid observations  $O_i \sim P_0$ ,  $i = 1, \dots, n$ , and that the statistical model  $\mathcal{M}$  consists of probability distributions dominated by a dominating measure  $\mu$ , and satisfying some smoothness assumptions.

For simplicity, let's consider the case that  $O$  a univariate and that  $\mu$  is the Lebesgue measure.

Suppose that our target parameter is  $p_0 = dP_0/d\mu$  is the density w.r.t. Lebesgue. Consider a kernel density estimator

$$p_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{O_i - x}{h}\right),$$

where  $K$  is a kernel,  $\int K(y)dy = 1$ , and  $h$  is a bandwidth.

## Multivariate kernel density estimator

If  $O$  would be  $d$ -dimensional, then

$$p_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{O_i - x}{h}\right),$$

where now  $K$  is a  $d$ -variate real valued function and  
 $(O - x)/h \equiv ((O_j - x_j)/h : j = 1, \dots, d)$ .

One could also naturally extend this latter definition to  $h = (h_1, \dots, h_d)$ , a bandwidth for each dimension.

## Variance of kernel density estimator

The variance of the kernel density estimator is given by:

$$\frac{1}{nh^2} \text{VAR} \left\{ K \left( \frac{O - x}{h} \right) \right\}.$$

It follows that the variance of the kernel is  $O(h)$ , so that we can conclude that

$$\text{VAR}(p_{n,h}(x)) = O \left( \frac{1}{nh} \right).$$

The simple intuition is that the kernel density estimator at a point is essentially an empirical mean over  $O(nh)$  iid observations, so that its variance is  $O(1/(nh))$ .

## Bias of kernel density estimator

The expectation of the kernel density estimator is given by:

$$E p_{n,h}(x) = \frac{1}{h} E \left\{ K \left( \frac{O - x}{h} \right) \right\}.$$

It follows that

$$E K \left( \frac{O - x}{h} \right) = \int K(y) p_0(x + hy) dy.$$

So that the bias of the kernel density estimator is given by:

$$\text{Bias}(p_{n,h}(x)) = \int K(y) \{ p_0(x + hy) - p_0(x) \} dy.$$

Tailor expansion of density at  $x$  to obtain alternative bias expression

Suppose that  $p_0$  is  $m$ -times continuously differentiable at  $x$ . Then,

$$p_0(x + hy) - p_0(x) = \sum_{j=1}^{m-1} \frac{(hy)^j}{j!} p_0^{(j)}(x) + \frac{(hy)^m}{m!} p_0^{(m)}(\xi(x, hy)),$$

for a  $\xi(x, hy)$  between  $x$  and  $x + hy$ . Thus, the bias of the kernel density estimator can be represented as:

$$\begin{aligned} \text{Bias}(p_{n,h}(x)) &= \int K(y) \sum_{j=1}^{m-1} \frac{(hy)^j}{j!} dy p_0^{(j)}(x) \\ &\quad + \int K(y) \frac{(hy)^m}{m!} p_0^{(m)}(\xi(x, hy)) dy. \end{aligned}$$

## Bias under smoothness assumptions and orthogonal kernel

Suppose that  $p_0$  is  $m$ -times continuously differentiable at  $x$  and that  $K$  satisfies  $\int K(y)dy = 1$ ,  $\int K(y)y^j dy = 0$  for  $j = 1, \dots, m-1$ . Then it follows that

$$\text{Bias}(p_{n,h}(x)) = \int K(y) \frac{(hy)^m}{m!} p_0^{(m)}(\xi(x, hy)) dy,$$

and thus that

$$\text{Bias}(p_{n,h}(x)) = O(h^m).$$

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Selection of bandwidth and kernel:

If one would know the underlying smoothness  $m$  of the true density  $p_0$  at  $x$ , then it would be good to select an  $m$ -orthogonal kernel. An optimal bandwidth could then be defined by

$$h_0 = \arg \min_h MSE(p_{n,h}(x)) = \arg \min_h E(p_{n,h}(x) - p_0(x))^2.$$

This MSE can be written as  $\text{VAR}(p_{n,h}(x)) + \text{BIAS}^2(p_{n,h}(x))$ .

## Selection of bandwidth and kernel:

Since the variance is order  $1/(nh)$  and the bias is order  $h^m$ , it follows that  $h_0 = O(n^{-1/(2m+1)})$ .

The MSE of the kernel density estimator using this bandwidth  $h_0$  would be  $O(n^{-2m/(2m+1)})$ .

## Selection of bandwidth and kernel

This is known to be an optimal minimax rate of convergence for the class of densities that are  $m$ -times continuously differentiable.

Note that if  $m$  gets larger and larger, this rate of convergence starts approximating the parametric model rate  $1/n^{0.5}$ .

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Estimation of bias based on parametric working model and assuming smoothness

However, even when  $m$  is known, this is not very useful to know for the purpose of selecting a bandwidth for a particular sample. The variance can be estimated well, but as we noticed the bias depends on  $p_0^{(m)}$  in the neighborhood of  $x$ .

So in order to estimate the bias we would need to estimate the  $m$ -th derivative of the density, a much harder problem than estimation of  $p_0$  itself.

A possible approach is to derive the estimate of the bias (i.e.,  $m$ -th derivative of density) under a parametric working model and use this to estimate the MSE.

In that manner one still uses a bandwidth that converges to zero at the optimal rate, and if the working model is a great approximation for the  $m$ -derivative, then it might also do a reasonable job on the constant.

## Problem with bandwidth selection in this manner

- Firstly, even when we know the smoothness  $m$  of  $p_0$  at  $x$ , then we still have the problem of how to estimate the  $m$ -th derivative and thereby the bias term.
- Secondly, why would one know the underlying smoothness?
- So this whole approach is theoretically fun to talk about but is not practical.

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Adaptive estimation of the choice of kernel and bandwidth

This challenging problem can be beautifully solved with cross-validation in the following manner.

Let  $p_{n,h,m}$  be a kernel density estimator using an  $m$ -orthogonal kernel  $K_m$  and bandwidth  $h$ , where  $h$  varies over an interval and  $m = 0, 1, 2, \dots, M$  for some large  $M$ .

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Cross-validation: Split sample in training and validation

Suppose that we divide the sample of  $n$  observations into  $V$  equal size subgroups, which defines  $V$  splits of the sample into a training sample of size  $n(V - 1)/V$  and complementary validation sample  $\text{VAL}_v$  of size  $n/V$  defined as one of the  $V$  subgroups.

For a given  $v = 1, \dots, V$ , let  $p_{n,h,m}^v$  be the kernel density estimator applied to the  $v$ -th training sample.

## Loss function for density to define performance measure

Let  $(p, O) \rightarrow L(p)(O)$  be the so called log-likelihood loss defined as:

$$L(p)(O) = -\log p(O).$$

We have that

$$p_0 = \arg \min_{p \in \mathcal{M}} P_0 L(p).$$

$P_0 L(p) = \int L(p)(o) dP_0(o)$  is called the risk of candidate  $p$  and  $p_0$  is the unique density that minimizes this risk.

## Loss-based dissimilarity

The loss function defines a loss-based dissimilarity:

$$d_0(p, p_0) = P_0 L(p) - P_0 L(p_0).$$

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Performance assessment: Cross-validated log-likelihood

One can now define the cross-validated empirical risk w.r.t this loss function:

$$CV_n(h, m) \equiv \sum_{v=1}^V \sum_{i \in \text{VAL}_v} L(p_{n,h,m}^v)(O_i).$$

Since we use the log-likelihood loss, one calls this the cross-validated empirical log-likelihood.

# Outline

- 1 Kernel density estimation
- 2 Selection of bandwidth and kernel
- 3 Estimation of bias based on parametric working model and assuming smoothness
- 4 Adaptive estimation of the choice of kernel and bandwidth
  - Loss-based cross-validation
  - Performance assessment: Cross-validated log-likelihood
  - Discrete super learner

## Cross-validation selector

The cross-validation selector of the bandwidth and kernel is now given by:

$$(h_n, m_n) \equiv \arg \min_{h,m} CV_n(h, m).$$

The resulting density estimator is now defined as:

$$p_n = p_{n,h_n,m_n}.$$

## This super-learner adapts to underlying smoothness

This estimator  $p_n$  is an example of a (discrete) super-learner of the true density  $p_0$ , where the candidate estimators are indexed by the choice of kernel  $m$  and the bandwidth  $h$ .

In the next lecture we discuss the theoretical oracle inequality this cross-validation selector satisfies.

It shows that the density estimator  $p_n$  is asymptotically equivalent with the kernel density estimator  $p_{n,h_{0,n},m_{0,n}}$  using an oracle selector  $(h_{0,n}, m_{0,n})$  that for the given sample minimizes the average over  $v$  of the Kullback-Leibner dissimilarity between the candidate kernel density estimator applied to the  $v$ -th training sample and the true density  $p_0$ .

## Asymptotic equivalence with oracle selected bandwidth and kernel

- The oracle selector is doing the perfect trade-off between bias and variance, and somehow this data adaptive density estimator does not only converge at the same optimal rate but cannot even be distinguished up till the constant.
- The only assumptions under which this holds is that the loss function at the candidate density estimators needs to be bounded by some universal  $M < \infty$ , that the number of candidate estimators is bounded by some polynomial power in sample size  $n$ .

## Example of adaptation to underlying smoothness

Suppose that  $p_0$  is  $m_0$ -times continuously differentiable at all  $x$ , but not  $m_0 + 1$ , but that this maximal smoothness is unknown.

Since this super learner is asymptotically equivalent with the oracle selector, it follows that it will achieve the rate of convergence of the kernel density estimator using the  $m_0$ -orthogonal kernel and the corresponding oracle bandwidth.

So we achieve the same performance as if we would have known the underlying smoothness and we would have been told the best possible bandwidth choice of the given sample.

# Super Learning Theory and Applications

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 Super learning and the oracle inequality
- 2 Prediction of survival based on right-censored data
- 3 Super learning of the optimal individualized treatment rule
- 4 Super learning of a conditional density

## Loss-based dissimilarity

Let  $L(\psi)(o)$  be a loss function for  $\psi_0 = \arg \min_{\psi} \int L(\psi)(o) dP_0(o)$ . We can define a loss-based dissimilarity between a candidate  $\psi$  and true parameter value  $\psi_0$ :

$$\begin{aligned}d_0(\psi, \psi_0) &= P_0 L(\psi) - P_0 L(\psi_0) \\&= \int_o L(\psi)(o) dP_0(o) - \int_o L(\psi_0)(o) dP_0(o).\end{aligned}$$

## Cross-validation selector

- Given a library of candidate estimator mappings  $P_n \rightarrow \hat{\Psi}_k(P_n)$ ,  $k = 1, \dots, K_n$ , we will define a cross-validation selector of  $k$ .
- Consider a  $V$ -fold cross-validation scheme that defines  $V$  sample splits in training and validation sample. For each sample split  $v$ , let  $P_{n,v}$  be the empirical probability distribution of the training sample, and let  $Val(v)$  be the set of indices that are in the validation sample.
- Let  $p = 1/V$  the proportion of observations in validation sample.
- The cross-validation selector is defined by

$$k_n = \arg \min_k \frac{1}{V} \sum_{v=1}^V \frac{1}{np} \sum_{i \in Val(v)} L(\hat{\Psi}_k(P_{n,v}))(O_i).$$

## Discrete super learner

The discrete super learner is defined as the estimator

$$\hat{\Psi}(P_n) = \hat{\Psi}_{k_n}(P_n).$$

## Oracle inequality for quadratic loss-based dissimilarities

Suppose that  $\sup_{\psi,o} |L(\psi) - L(\psi_0)|(o) < M_1$  and

$$\sup_{\psi} \frac{P_0 \{L(\psi) - L(\psi_0)\}^2}{P_0 L(\psi) - P_0 L(\psi_0)} < M_2.$$

Let  $p = 1/V$ , and  $C(M_1, M_2, \delta) = O(M_1 + M_2/\delta)$ . Then, for each  $\delta > 0$ , we have

$$\begin{aligned} E_0 \frac{1}{V} \sum_v d_0(\hat{\Psi}_{k_n}(P_{n,v}), \psi_0) &\leq (1 + \delta) E_0 \min_k \frac{1}{V} \sum_v d_0(\hat{\Psi}_k(P_{n,v}), \psi_0) \\ &\quad + C(M_1, M_2, \delta) \frac{\log K_n}{np} \end{aligned}$$

# Asymptotic equivalence of cross-validation selector and oracle selector

Suppose that

$$\frac{\log K_n/n}{E_0 \min_k \frac{1}{V} \sum_v d_0(\hat{\Psi}_k(P_{n,v}), \psi_0)} \rightarrow 0.$$

Then,

$$\frac{E_0 \frac{1}{V} \sum_v d_0(\hat{\Psi}_{K_n}(P_{n,v}), \psi_0)}{E_0 \min_k \frac{1}{V} \sum_v d_0(\hat{\Psi}_k(P_{n,v}), \psi_0)} \rightarrow 1.$$

## Asymptotic equivalence of cross-validation selector and oracle selector

In words, if  $K_n = n^m$  for some finite  $m$ , and the oracle selected estimator converges at a slower rate than  $\log n/n$  (i.e., rate for a correctly specified parametric model), then the ratio of the dissimilarity of the cross-validated selected estimator and the truth and the dissimilarity of the oracle selected estimator and the truth converges to 1.

If, one of the candidate estimators happens to be based on a correctly specified parametric model, then the dissimilarity of the cross-validated selected estimator and the truth converges at rate  $\log n/n$ .

## Oracle inequality for non-quadratic loss-based dissimilarities

Suppose that  $\sup_{\psi,o} | L(\psi) - L(\psi_0) | (o) < M_1$ . Let  $p = 1/V$  and  $C(M_1) = O(M_1)$ . Then,

$$\begin{aligned} E_0 \frac{1}{V} \sum_v d_0(\hat{\Psi}_{k_n}(P_{n,v}), \psi_0) &\leq E_0 \min_k \frac{1}{V} \sum_v d_0(\hat{\Psi}_k(P_{n,v}), \psi_0) \\ &\quad + C(M_1) \frac{(\log K_n)^{0.5}}{(np)^{0.5}}. \end{aligned}$$

# Outline

- 1 Super learning and the oracle inequality
- 2 Prediction of survival based on right-censored data
- 3 Super learning of the optimal individualized treatment rule
- 4 Super learning of a conditional density

## Causal effect of treatment on right-censored survival time

Let  $O = (W, A, \Delta = I(T \leq C), \tilde{T} = \min(T, C))$ ,  $T$  is a survival time,  $C$  is a right-censoring time,  $A$  is a binary treatment,  $W$  are baseline covariates. Suppose that  $C$  is independent of  $T$ , given  $A, W$ . Consider the conditional survivor function:

$$S(t_0 | A, W) = P(T > t_0 | A, W).$$

Let

$$\lambda(t | A, W) = P(T = t | A, W, T \geq t)$$

be the conditional hazard of survival at time  $t$ . We have

$$S(t_0 | A, W) = \prod_{s \leq t_0} (1 - \lambda(s | A, W)).$$

If  $T$  is continuous, then this writes as

$$S(t_0 | A, W) = \prod_{s \leq t_0} (1 - d\Lambda(s | A, W)).$$

## Identification of conditional hazard

If  $C$  is independent of  $T$ , given  $A, W$ , then we can identify the conditional hazard as follows:

$$\lambda(t \mid A, W) = P(\tilde{T} = t, \Delta = 1 \mid \tilde{T} \geq t, A, W).$$

Let's denote the right-hand side with  $\tilde{\lambda}(t \mid A, W)$ .

## Treatment specific survival

Define

$$\Psi_a(P) = E_P S(t_0 \mid A = a, W).$$

Under a causal model in which  $T = T_A$ ,  $T_0, T_1$  treatment specific counterfactual survival times, and the assumption that  $A$  is independent of  $T_0, T_1$ , given  $W$ , we have

$$\Psi_a(P) = P(T_a > t_0),$$

the counterfactual survival function at  $t_0$  under intervention  $A = a$ .

## Prediction of survival based on right-censored data

Suppose that we want to estimate  $\psi_0(t_0, A, W) = S_0(t_0 \mid A, W)$  at a given point  $t_0$ , based on the right-censored data structure  $O = (W, A, \Delta, \tilde{T})$ . If there would not be right-censoring, then a valid loss function would be

$$L^F(\psi)(W, A, T) = (I(T > t_0) - \psi(t_0, A, W))^2.$$

## Inverse probability of censoring weighted loss function

Let  $\bar{G}(t | A, W) = P(C \geq t | A, W)$ . We can define the Inverse probability of censoring weighted loss function:

$$L_{G_0}(\psi)(O) = \frac{L^F(\psi)(W, A, \tilde{T})\Delta}{\bar{G}(\tilde{T} | A, W)}.$$

## Improved IPCW-loss

An improved IPCW-loss is given by:

$$L_{G_0}(\psi)(O) = \frac{L^F(\psi)(W, A, \tilde{T})\{I(\tilde{T} \leq t_0, \Delta = 1) + I(\tilde{T} > t_0)\}}{\bar{G}_0(\min(\tilde{T}, t_0) \mid A, W)}.$$

## Cross-validation selector

The cross-validation selector is now defined by:

$$k_n = \arg \min_k \frac{1}{V} \sum_v \sum_{i \in VAL(v)} L_{G_{n,v}}(\hat{\Psi}_k(P_{n,v})(O_i)),$$

where  $G_{n,v}$  is an estimator of  $G_0$  based on the training sample  $P_{n,v}$ .

## Candidate estimators of the conditional survival function: Direct estimators

IPC-weighted logistic regression estimators regressing  $I(T > t_0)$  on  $A, W$  using weights

$$\frac{I(\tilde{T} \leq t_0, \Delta = 1) + I(\tilde{T} > t_0)}{\bar{G}_0(\min(\tilde{T}, t_0) \mid A, W)}.$$

## Candidate estimators of conditional survival function: Hazard based estimators

Consider the case that  $T$  is discrete. For each  $t$ , we can estimate the hazard  $\tilde{\lambda}(t | A, W)$  with logistic regression of the binary outcome  $I(\tilde{T} = t, \Delta = 1)$ , given  $A, W$ , among observations with  $\tilde{T} \geq t$ .

We can also run a pooled logistic regression by creating a row of data for each  $s \leq \tilde{T}$ , defined by  $(W, A, I(\tilde{T} = s, \Delta = 1))$ ,  $s = 1, \dots, \tilde{T}$ , thereby simultaneously fitting the conditional hazard as a function of  $(t, A, W)$ .

## How to handle continuous survival?

- If  $T$  is continuous, we can discretize  $T$  using an interval partition of its range. We can now use the above pooled or  $t$ -specific logistic regression approach to estimate the corresponding discrete conditional hazard.
- By dividing this discrete hazard at time  $t$  by the width  $h(t)$  of the interval that contains  $t$ , we obtain an estimate of the actual continuous conditional hazard.
- The choice of interval partition, width of the interval, can now be a tuning parameter, defining a candidate estimator.
- One could also use Cox-proportional hazard based estimators.

# Outline

- ① Super learning and the oracle inequality
- ② Prediction of survival based on right-censored data
- ③ Super learning of the optimal individualized treatment rule
- ④ Super learning of a conditional density

## Super learning of optimal individualized treatment rule

- $O = (W, A, Y)$ , nonparametric model only assumptions on  $g_0(A | W)$ .
- Target: Optimal treatment rule  $\psi_0(W) = I(B_0(W) > 0)$ , where  $B_0(W) = E_0(Y | A = 1, W) - E_0(Y | A = 0, W)$ .
- Possible loss function for  $\psi_0$  is an IPCW-loss:

$$L_{g_0}(\psi) = \frac{I(A = \psi(W))}{g(A | W)} Y.$$

- Indeed,  $\psi_0$  is the minimizer of  $EL_{g_0}(\psi)$  over all rules  $\psi$ .
- Loss-based dissimilarity:  $d_0(\psi, \psi_0) = E_0 Y_\psi - E_0 Y_{\psi_0}$ .

## Super learning of the optimal individualized treatment rule

- Construct library of candidate estimators of  $\psi_0 = I(B_0 > 0)$ . This can include estimators based on plugging in an estimator of  $B_0$ .
- One could also include a candidate estimator  $I(B_n > 0)$  where  $B_n$  is a super learner of  $B_0$ , e.g. based on loss function

$$L_{g_0}(B) = ((2A - 1)/g(A | W)Y - B(W))^2$$

that directly targets  $B_0 = \arg \min_B P_0 L_{g_0}(B)$ .

- Estimate  $g_0$  if not known.
- Compute cross-validation selector:

$$k_n = \arg \min_k E_{B_n} P_{n, B_n}^1 L_{\hat{g}(P_{n, B_n}^0)}(\hat{\Psi}_k(P_{n, B_n}^0)).$$

- super learner of optimal rule  $\psi_0$ :  $\hat{\Psi}_{k_n}(P_n)$ .

# Outline

- ① Super learning and the oracle inequality
- ② Prediction of survival based on right-censored data
- ③ Super learning of the optimal individualized treatment rule
- ④ Super learning of a conditional density

## Super learner of a multinomial conditional distribution

Suppose we want to construct a super learner of the conditional probability distribution  $g_0(a | W) = P_0(A = a | W)$ , where  $a \in \mathcal{A}$ . Let's denote the values of  $a$  with  $\{0, 1, \dots, K\}$ . A valid loss function for the conditional density is

$$L(g)(O) = -\log g(A | W).$$

That is,  $g_0 = \arg \min_g P_0 L(g)$ , i.e.,  $g_0$  is the minimizer of the expectation of the log-likelihood loss.

## Candidate estimators

Let  $\hat{g}_k(P_n)$ ,  $k = 1, \dots, K$ , be a collection of candidate estimators of  $g_0$ . The discrete super learner is defined by

$$g_n = \hat{g}_{k_n}(P_n),$$

where

$$k_n = \arg \min_k E_{B_n} P_{n, B_n}^1 L(\hat{g}(P_{n, B_n}^0)) = E_{B_n} \frac{1}{np} \sum_{i: B_n(i)=1} L(\hat{g}(P_{n, B_n}^0))(O_i),$$

and  $B_n \in \{0, 1\}^n$  is a random sample split in training sample  $\{i : B_n(i) = 0\}$  and validation sample  $\{i : B_n(i) = 1\}$ . Here  $p$  is the proportion of observations that are in the validation sample,  $P_{n, B_n}^1$ ,  $P_{n, B_n}^0$  are the empirical probability distributions of the validation and training sample, respectively.

## Weighted combinations

We can also define a parametric family of candidate estimators  $\hat{g}_\alpha(P_n)$ , indexed by a vector  $\alpha$ , such as

$$\hat{g}_\alpha = \sum_{k=1}^K \alpha(k) \hat{g}_k$$

under the constraint that  $\alpha(k) \geq 0$ ,  $k = 1, \dots, K$ , and  $\sum_k \alpha(k) = 1$ . This choice of family is contained in the class of probability distributions.

## Super learner

The super learner for this family of candidate estimators is given by

$$g_n = \hat{g}_{\alpha_n}(P_n),$$

where

$$\alpha_n = \arg \min_{\alpha} E_{B_n} P_{n, B_n}^1 L(\hat{g}_{\alpha}(P_{n, B_n}^0)).$$

One might have to program this optimization over  $\alpha$ , but existing routines are available for doing such constrained maximization problems. This step is often referred to as the meta-learner step.

## Candidate estimators

**Candidate estimators based on multinomial logistic regression:** To start with one can use existing parametric model based MLE and machine learning algorithms in *R* that fit a multinomial regression.

**Candidate estimators based on machine learning for multinomial logistic regression:** Secondly, one can use a machine learning algorithm such as `polyclass()` in R that adaptively fits a multinomial logistic regression, which itself has tuning parameters, again generating a class of candidate estimators.

## Incorporating screening

- Note that one can also marry any of these choices with a screening algorithm, thereby creating more candidate estimators of interest.
- The screening can be particularly important when there are many variables.

## Candidate estimators by fitting separate logistic regressions and using post-normalization:

- Code  $A$  in terms of Bernoullis  $B_k = I(A = k)$ ,  $k = 0, \dots, K$ .
- Construct an estimator  $\bar{g}_{nk}$  of  $\bar{g}_{0k}(W) \equiv P_0(B_k = 1 | W)$  using any of the logistic regression algorithms, for all  $k = 0, \dots, K$ .
- This implies an estimator

$$g_n(a | W) = \frac{\bar{g}_{na}(W)}{\sum_{k=0}^K \bar{g}_{nk}(W)}.$$

- In other words, we simply normalize these separate logistic regression estimators so that we obtain a valid conditional distribution.
- This generates an enormous amount of interesting algorithms, since we have available the whole machine learning literature for binary outcome regression.

## Candidate estimators by estimating the conditional "hazard with" pooled logistic regression

Finally, we have used the following strategy in our research for construction of candidate estimators. Note that

$$g_0(a | W) = \lambda_0(a | W)S_0(a | W),$$

where

$$\lambda_0(a | W) = P_0(A = a | A \geq a, W),$$

and  $S_0(a | W) = \prod_{s \leq a} (1 - \lambda_0(s | W))$  is the conditional survivor function  $P_0(A > a | W)$ . So we have now parameterized the conditional distribution of  $A$ , given  $W$ , by a conditional hazard  $\lambda_0(a | W)$ :  $g_0 = g_{\lambda_0}$ .

## Candidate estimators of conditional "hazard"

- We could now focus on constructing candidate estimators of  $\lambda_0(a | W)$ , which implies candidate estimators of  $g_0$ .
- For every observation  $A_i$ , we can create  $A_i + 1$  rows of data  $(W, s, I(A_i = s))$ ,  $s = 0, \dots, A_i$ ,  $i = 1, \dots, n$ . We now run a logistic regression estimator based on the pooled data set, ignoring ID, where we regress the binary outcome  $I(A_i = s)$  on the covariates  $(W, s)$ .

## Pooling across levels or not

- If one assumes a parametric model, then this is nothing else than using the maximum likelihood estimator, demonstrating that ignoring the ID is not inefficient.
- This defines now an estimator of  $\lambda_0(s | W) = P_0(A = s | W, A \geq s)$  as a function of  $(s, W)$ .
- Different choices of logistic regression based estimators will define different estimators.
- The pooling across  $s$  is not very sensible if  $A$  is not an ordered variable. If  $A$  is categorical, we recommend to compute a separate logistic regression estimator of  $\lambda_0(a | W)$  for each  $a$  (i.e., stratify by  $s$  in the above pooled data set).
- For non-categorical  $A$ , one could include both stratified (by level) as well as pooled (across levels) based logistic regression estimators.

# Targeted Maximum Likelihood Estimation

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Statistical Estimation Problem

**Data:** We observe  $O_1, \dots, O_n \sim_{iid} P_0$ .

**Statistical Model:** We know  $P_0 \in \mathcal{M}$ .

**Target Parameter:** Let  $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$  denote the target parameter.

**Plug-in estimator:** We want to construct an efficient plug-in estimator  $\Psi(P_n^*)$  of the estimand  $\Psi(P_0)$  that is asymptotically linear, so that we also obtain asymptotic (maximally narrow) confidence intervals.

## Canonical gradient

- Let  $P$  be given.
- Consider class of paths  $\{P_\epsilon : \epsilon\} \subset \mathcal{M}$  through  $P$  at  $\epsilon = 0$  with score  $S$ , and show that for each path

$$\frac{d}{d\epsilon} \Psi(P_\epsilon) \Big|_{\epsilon=0} = E_P D(P) S$$

for some  $D(P) \in L^2(P)$ .

- $O \rightarrow D(P)(O)$  is called a gradient at  $P$ , and the unique gradient that is an element of the tangent space  $T(P)$  (closure of linear span of all scores  $S$ ) is the *canonical gradient* at  $P$ .
- We say  $\Psi$  is pathwise differentiable at  $P$  with canonical gradient  $D^*(P)$ .

## Exact second order expansion for target parameter

- Let

$$R_2(P, P_0) \equiv \Psi(P) - \Psi(P_0) - (P - P_0)D^*(P).$$

- This will behave as a second order difference of  $P$  and  $P_0$ .
- By definition (note  $PD^*(P) = 0$ ), this yields following exact second order expansion of  $\Psi$ :

$$\Psi(P) - \Psi(P_0) = -P_0 D^*(P) + R_2(P, P_0).$$

## Efficient estimator

- An estimator  $\psi_n$  of  $\psi_0 = \Psi(P_0)$  is called asymptotically efficient at  $P_0$  if it is asymptotically linear with influence curve  $D^*(P_0)$ :

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

- Canonical gradient is also called efficient influence curve.
- If this holds, then  $\psi_n$  is also a regular estimator, and it is the asymptotic best estimator among all regular estimators.

## Initial estimator for TMLE

- Let  $\mathbf{P}_n$  be an initial estimator of  $P_0$ .
- If  $\Psi(P) = \Psi(Q(P))$  only depends on  $P$  through a functional parameter  $Q(P)$ , then one only needs  $Q(\mathbf{P}_n)$ , so direct estimators of  $Q(\mathbf{P}_n)$  of  $Q(P_0)$  can (should) be used.

## Loss function

- Let  $L(P)(O)$  be a loss for  $P_0$ , such as  $L(P) = -\log p(O)$ .
- More generally, one can use a loss  $L(Q)(O)$  for relevant part  $Q(P)$  of  $P$ .
- So  $P_0 L(Q_0) = \min_{P \in \mathcal{M}} P_0 L(Q(P))$ .

## Local least favorable submodel (LFM) through initial estimator

- Let  $\{\mathbf{P}_{n,\epsilon} : \epsilon\} \subset \mathcal{M}$  be a finite dimensional submodel so that the linear span of the components of its score at  $\epsilon = 0$  include  $D^*(\mathbf{P}_n)$ .
- Such a submodel is called a local least favorable submodel.
- More generally, let  $\{\mathbf{P}_{n,\epsilon} : \epsilon\} \subset \mathcal{M}$  be a finite dimensional submodel so that the linear span of the components of its generalized score

$$\frac{d}{d\epsilon} L(Q(\mathbf{P}_{n,\epsilon})) \Big|_{\epsilon=0}$$

includes  $D^*(\mathbf{P}_n)$ .

## Uniform least favorable submodel (ULFM) through initial estimator

- Let  $\{\mathbf{P}_{n,\epsilon} : \epsilon\} \subset \mathcal{M}$  be a finite dimensional submodel so that the linear span of the components of its score at *any*  $\epsilon$  include  $D^*(\mathbf{P}_{n,\epsilon})$ .
- Such a submodel is called a *uniform* least favorable submodel.
- Similarly, as above we can generalize this definition to loss function  $L(Q)$  for a parameter  $Q(P)$  chosen so that  $\Psi(P)$  only depends on  $P$  through  $Q(P)$ .

## Update initial estimator with MLE along LFM

- Let

$$\epsilon_n^0 = \arg \min_{\epsilon} P_n L(\mathbf{P}_{n,\epsilon})$$

be the MLE.

- Define the first TMLE-update as  $\mathbf{P}_n^1 = \mathbf{P}_{n,\epsilon_n^0}$ .

## Iterating TMLE-updating till efficient score equation is solved approximately

- Iterate this updating process

$$\epsilon_n^k = \arg \min_{\epsilon} P_n L(\mathbf{P}_{n,\epsilon}^k)$$

and  $\mathbf{P}_n^{k+1} = \mathbf{P}_{n,\epsilon_n^k}$ ,  $k = 1, \dots, \dots, K_n$ .

- One iterates till  $K_n$  satisfies

$$P_n D^*(\mathbf{P}_n^{K_n}) = r(n)$$

for a user-supplied sequence  $r(n) = o(n^{-1/2})$ .

- If one iterates till  $\epsilon_n^K = 0$ , then we have

$$P_n D^*(\mathbf{P}_n^K) = 0.$$

## Universal least favorable submodel

We define a 1-d universal least favorable submodel at  $P$  as a submodel  $\{P_\epsilon : \epsilon\}$  so that for all  $\epsilon$

$$\frac{d}{d\epsilon} \log \frac{dP_\epsilon}{dP} = D^*(P_\epsilon).$$

This acts as a local least favorable submodel at any point  $\epsilon$  on its path.

## One-step TMLE

- If we use a universal LFM, then the above TMLE algorithm converges in one-step, so that  $P_n^* = \mathbf{P}_n^1$ .
- Then,

$$P_n D^*(\mathbf{P}_n^1) = 0.$$

- If the initial estimator  $\mathbf{P}_n$  converges at a rate faster than  $n^{-1/4}$  so that the second order remainder  $R_2(\mathbf{P}_n, P_0) = o_P(n^{-1/2})$ , then under regularity conditions, a first step TMLE using local LFM will satisfy  $P_n D^*(\mathbf{P}_n^1) = o_P(n^{-1/2})$ .

# TMLE

- Let  $P_n^*$  be final update.
- $\psi_n^* = \Psi(P_n^*)$  is the TMLE of  $\psi_0$ .

## TMLE using $n^{-1/4}$ -initial estimator is efficient

- Combining  $P_n D^*(P_n^*) = 0$  with exact second order expansion for  $\Psi(P_n^*) - \Psi(P_0)$  yields

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^*(P_n^*) + R_2(P_n^*, P_0).$$

- Use  $o(n^{-1/4})$ -estimator  $\mathbf{P}_n$  (Highly Adaptive Lasso!) so that  $R_2(P_n^*, P_0) = o_P(n^{-1/2})$ .
- Control overfitting so that  $D^*(P_n^*)$  falls in a Donsker class with probability tending to 1 (e.g, sectional variation norm of  $D^*(P_n^*)$  is bounded with probability tending to 1).
- Then, also  $(P_n - P_0)D^*(P_n^*) = (P_n - P_0)D^*(P_0) + o_P(n^{-1/2})$  so that  $\Psi(P_n^*)$  is asymptotically efficient.

## Inference

- Thus, inference can be based on working model  $\Psi(P_n^*) \sim N(\psi_0, \Sigma_0)$ , where  $\Sigma_0 = P_0 D^*(P_0) D^*(P_0)^\top$ .
- $\Sigma_0$  can be consistently estimated with  $\Sigma_n = P_n D^*(P_n^*) D^*(P_n^*)^\top$ .
- For example,  $\Psi_j(P_n^*) \pm 1.96 \Sigma_n(j,j)^{0.5} / n^{1/2}$  is an asymptotic 0.95-confidence interval for  $\Psi_j(P_0)$ ,  $j = 1, \dots, d$ .

# Targeted Minimum Loss-Based Estimation of the Treatment Specific Survival Function for Right-Censored Survival Data

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 Causal model for the counterfactual treatment specific survival curve
- 2 Efficient influence curve
- 3 Super-learner of conditional hazard
- 4 TMLE of treatment specific survival curve
- 5 One-step TMLE of treatment specific survival curve
- 6 Example: HAL-MLE of conditional hazard

## Observed data

- We observe

$$O = (W, A, \tilde{T} = \min(T, C), \Delta = I(T \leq C)) \sim P_0$$

- $W$  baseline covariates.
- $A$  binary treatment.
- $C, T$ , censoring and survival time.

## Causal formulation of observed data

- $T_0, T_1$  potential survival times under control and treatment.
- $C_0, C_1$  potential censoring times under control and treatment.
- $T = T_A, C = C_A, \tilde{T} = \tilde{T}_A, \Delta = \Delta_A$ .
- $O = (W, A, \tilde{T}_A = \min(T_A, C_A), \Delta_A = I(T_A \leq C_A))$ .

## Causal quantity: Treatment specific survival curve

- Let  $W \rightarrow d(W) \in \{0, 1\}$  be a dynamic treatment rule.
- Let  $S_d(t) = P(T_d > t)$ , where  $T_d = T_{d(W)}$  be quantity of interest.

## Coarsening at random assumption on treatment and censoring

- Randomization of treatment: Assume that  $A$  is independent of  $T_d$ , given  $W$ .
- Let  $A_2(t) = I(\tilde{T} \leq t, \Delta = 0)$  is censoring process, jumps at observed censoring time.
- Let  $N(t) = I(\tilde{T} \leq t, \Delta = 1)$  is failure process that jumps at observed failure time.
- $(W, A, \bar{N}(t), \bar{A}_2(t-))$  is available history right before  $A_2(t)$ .
- Non-informative censoring (CAR): Assume that at each time  $t$ ,  $A_2(t)$  is independent of  $T_d$ , given  $(W, A, \bar{N}(t), \bar{A}_2(t-))$ .
- These assumptions are non-testable: put no restrictions on data distribution  $P_0$ .

## Identifiability of treatment specific survival curve

- Under the above CAR assumption we have

$$S_d(t_0) = E_P S(t_0 \mid A = d(W), W).$$

- $S(t_0 \mid A, W) = P(T > t_0 \mid A, W)$  conditional survival curve of  $T$ .
- For any data distribution  $P$ , let  $\Psi(P) = E_P S(t_0 \mid A = d(W), W)$ .

## Longitudinal formulation of observed data

- We can represent the observed data as

$$O = (W, A, \bar{A}_2(\tau), \bar{N}(\tau)),$$

where  $\tau$  is a maximal follow up time.

- This data is time-ordered as:

$$O = (W, A, A_2(0), N(1), A_2(1), N(2), \dots, A_2(\tau - 1), N(\tau)).$$

## Conditional probability distributions of data distribution

- Let  $Q_W$  be probability measure of  $W$ .
- Let  $Q_{N(t)}$  be conditional probability measure of  $N(t)$ , given  $W, A, \bar{N}(t-1), \bar{A}_2(t-1)$ .
- Let  $Q_N = (Q_{N(t)} : t)$ .
- Let  $G_1$  be conditional probability measure of  $A$ , given  $W$ .
- Let  $G_{A_2(t)}$  be conditional probability measure of  $A_2(t)$ , given  $W, A, \bar{A}_2(t-1), \bar{N}(t)$ .
- Let  $G_2 = (G_{A_2(t)} : t)$ .

## Density of data distribution

- The density of a data distribution  $P$  of  $O$  can be factorized as: for  $o = (w, a, \bar{a}_2(\tau), \bar{n}(\tau))$ ,

$$p(o) = q_W(w) \prod_{t=0}^{\tau} q_{N(t)}(n(t) | w, a, \bar{a}_2(t-1), \bar{n}(t-1)) \\ g_1(a | W) \prod_{t=0}^{\tau-1} g_{A_2(t)}(a_2(t) | w, a, \bar{a}_2(t-1), \bar{n}(t)).$$

- $g_1(a | W) = P(A = a | W)$  and  $g_{A_2(t)}$  is conditional probability distribution of  $A_2(t)$ , given  $Pa(A_2(t))$ .
- $q_W(w)$  density of  $W$ , and  $q_{N(t)}$  is conditional probability distribution of  $N(t)$  given  $Pa(N(t))$ .

## Conditional densities become degenerate after failure or censoring event

- Note that if  $N(t) = 1$  or  $A_2(t - 1) = 1$ , then the conditional density  $g_{A_2(t)}$  of  $A_2(t)$  is degenerate at 0 or 1, respectively.
- Note that if  $N(t - 1) = 1$  or  $A_2(t - 1) = 1$ , then the conditional density  $q_{N(t)}$  of  $N(t)$  is degenerate at 1 or 0, respectively.
- Thus, the product over  $t$  for  $q_{N(t)}$  only includes  $t$  with  $n(t - 1) = a_2(t - 1) = 0$ .
- The product over  $t$  for  $g_{A_2(t)}$  only includes  $t$  with  $n(t) = a_2(t - 1) = 0$ .

## Hazard of censoring

- The conditional density  $g_{A_2(t)}(1 \mid W, A, A_2(t-1) = N(t) = 0)$  is a conditional hazard  $\lambda_C(t \mid W, A, N(t) = 0)$ .
- Under CAR, this equals conditional hazard  $\lambda_C(t \mid W, A)$ .
- If censoring  $C$  is continuous, we can replace  $g_{A_2(t)}(1 \mid W, A, \bar{A}_2(t-1), \bar{N}(t))$  by an intensity

$$E(dA_2(t) \mid W, A, \bar{A}_2(t-), \bar{N}(t-)) = I(\tilde{T} \geq t) \lambda_C(t \mid W, A) dt.$$

## Hazard of failure

- The conditional density  $q_{N(t)}(1 \mid W, A, A_2(t-1) = N(t-1) = 0)$  is a conditional hazard  $\lambda(t \mid W, A, A_2(t-1) = 0)$ .
- Under CAR, this equals conditional hazard  $\lambda(t \mid W, A)$ .
- If  $T$  is continuous, we can replace  $q_N(1 \mid W, A, \bar{A}_2(t-1), \bar{N}(t-1))$  by an intensity

$$E(dN(t) \mid W, A, \bar{A}_2(t-), \bar{N}(t-)) = I(\tilde{T} \geq t) \lambda(t \mid W, A) dt.$$

## Observed data density in terms of hazards

- Thus,

$$\begin{aligned} p(o) &= q_W(w) \prod_{t=0}^{\tilde{t}} \lambda(t \mid W, A)^{dn(t)} (1 - \lambda(t \mid W, A))^{1-dn(t)} \\ &\quad \times g_1(a \mid W) \prod_{t=0}^{\tilde{t}} \lambda_C(t \mid w, a)^{da_2(t)} (1 - \lambda_C(t \mid w, a))^{1-da_2(t)}. \end{aligned}$$

## Statistical model

- Recall  $P = P_{Q_W, Q_N, G_1, G_2}$ .
- Parameter space for distribution of  $W$  is nonparametric.
- Parameter space for conditional hazard of  $T$  is nonparametric.
- Parameter space  $\mathcal{G}_1$  for  $G_1$  and parameter space  $\mathcal{G}_2$  for  $G_2$  are possibly restricted.
- The statistical model

$$\mathcal{M} = \{P_{Q_W, Q_N, G_1, G_2} : G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\}.$$

## Statistical estimation problem

- We observe  $n$  i.i.d. copies of  $O \sim P_0$ .
- We know  $P_0 \in \mathcal{M}$ .
- $\Psi : \mathcal{M} \rightarrow \mathbb{R}$  is statistical target parameter,  
$$\Psi(P) = E_P P(T > t \mid A = d(W), W).$$
- We want to estimate  $\Psi(P_0)$ .

# Outline

- ① Causal model for the counterfactual treatment specific survival curve
- ② Efficient influence curve
- ③ Super-learner of conditional hazard
- ④ TMLE of treatment specific survival curve
- ⑤ One-step TMLE of treatment specific survival curve
- ⑥ Example: HAL-MLE of conditional hazard

## Tangent space

- $T_{Q_W} = L_0^2(P_W)$  is all functions of  $W$ .
- $T_{Q_{N(t)}}$  is all functions of  $N(t)$  and  
 $Pa(N(t)) = (W, A, \bar{A}_2(t-1), \bar{N}(t-1))$  with conditional mean zero,  
given  $Pa(N(t))$ .
- $T_{G_1}$  is all function of  $A, W$  with conditional mean zero, given  $W$ .
- $T_{G_{A_2(t)}}$  is all functions of  $A_2(t)$  and  
 $Pa(A_2(t)) = (W, A, \bar{A}_2(t-1), \bar{N}(t))$  with conditional mean zero,  
given  $Pa(A_2(t))$ .

## Tangent space is orthogonal sum of all tangent spaces

- By factorization of  $p(O)$ , the tangent space  $T(P)$  at  $P$  is an orthogonal sum of all these tangent spaces:

$$T(P) = T_{Q_W} \oplus T_{G_1} \oplus \sum_t T_{Q_{N(t)}} \oplus \sum_t T_{G_{A_2(t)}}.$$

Canonical gradient is projection on tangent space for model  $G$  known

- Let  $\mathcal{M}(G) = \{P_{Q_W, Q_N, G_1, G_2} : Q_W, Q_N\}$  be submodel of  $\mathcal{M}$  defined by  $G = (G_1, G_2)$  being known.
- Let  $D$  be a gradient of  $\Psi : \mathcal{M}(G) \rightarrow \mathbb{R}$  at  $P \in \mathcal{M}(G)$ .
- Let  $T_Q(P)$  be the tangent space of  $Q = (Q_W, Q_N)$ .
- Then the canonical gradient equals  $D^* P = \Pi(D \mid T_Q(P))$ .

## Canonical gradient is sum of projections

- So

$$D^*(P) = \Pi(D \mid T_{Q_W}) + \sum_t \Pi(D \mid T_{Q_{N(t)}}).$$

- This equals:

$$D^*(P) = E(D \mid W) + \sum_t \{E(D \mid N(t), Pa(N(t))) - E(D \mid Pa(N(t)))\}.$$

## Initial gradient

- Let

$$D(P) = \frac{I(\tilde{T} > t_0, A = d(W), A_2(t_0) = 0)}{g_1(A | W) \prod_{t \leq t_0} (1 - \lambda_C(t | A, W))} - \Psi(P).$$

- This is influence curve of RAL estimator in model  $\mathcal{M}(G)$  and therefore a gradient.

# Canonical gradient

- Let  $D_0^*(W) = S(t_0 \mid A = d(W), W) - \Psi(P)$ .
- For  $t = 1, \dots, \min(\tau, t_0)$ , define

$$C_t(Q, G) = \frac{I(A = d(W), \bar{A}_2(t-1) = 0)}{g_1(A \mid W) \prod_{s \leq t-1} (1 - \lambda_C(s \mid A, W))} \frac{S(t_0 \mid A, W)}{S(t \mid A, W)}.$$

- and

$$D_t^*(P) = C_t(Q, G)(dN(t) - q_{N(t)}(1 \mid Pa(N(t))).$$

- This can also be written as

$$D_t^*(P) = I(\tilde{T} \geq t) C_t(Q, G)(dN(t) - q_{N(t)}(1 \mid Pa(N(t))).$$

- We have

$$D^*P = \sum_{t=0}^{\min(\tau, t_0)} D_t^*(P).$$

## Clever covariate times residual representation of each score component

- Note that for  $t = 1, \dots, t_0$ , we have

$$D_t^*(P) = C_t(Q_N, G)(W, A, N(t), A_2(t-1))(dN(t) - \lambda(t | A, W)).$$

- We often refer to  $C_t(Q_N, G)$  as the clever covariate (that will target fit of  $\lambda(t | A, W)$  towards  $\Psi(P)$ ).

## Recursive structure in canonical gradient

- Note that for  $t < t_0$ ,

$$\frac{S(t_0 \mid A, W)}{S(t \mid A, W)} = \prod_{t+1 \leq s \leq t_0} (1 - \lambda(s \mid A, W)).$$

- The clever covariate  $C_{t_0}(G)$  does not depend on  $\lambda$ .
- The clever covariate  $C_{t_0-1}(G, Q_{N(t_0)})$  depends  $\lambda(t_0 \mid \cdot)$ .
- In general, the clever covariate  
 $C_{t_0-k}(G, (Q_{N(s)} : t_0 - k + 1 \leq s \leq t_0))$ .
- This structure allows us to compute a closed form TMLE starting to target the last  $\lambda(t_0 \mid \cdot)$ , going backwards, each time being able to compute clever covariate from previous targeted fits of  $\lambda$ .

# Outline

- ① Causal model for the counterfactual treatment specific survival curve
- ② Efficient influence curve
- ③ Super-learner of conditional hazard
- ④ TMLE of treatment specific survival curve
- ⑤ One-step TMLE of treatment specific survival curve
- ⑥ Example: HAL-MLE of conditional hazard

## Loss function for conditional hazard

- The log-likelihood loss for  $\lambda$  is given by:

$$L(\lambda)(O) = -\log \left\{ \prod_{t \leq \tilde{T}} \lambda(t | A, W)^{dN(t)} (1 - \lambda(t | A, W))^{1-dN(t)} \right\}.$$

- The true failure time hazard

$$\lambda_0 = \arg \min_{\lambda} P_0 L(\lambda).$$

## Estimation of hazard with pooled logistic regression

- $\lambda(t | A, W) = E(dN(t) | A, W, \tilde{T} \geq t)$  is a regression of binary outcome  $dN(t)$  on  $(A, W)$ , given  $\tilde{T} \geq t$ .
- Thus, one can estimate this function with any logistic regression estimator based on pooled data set in which each unit  $i$  has  $\tilde{T}_i$  rows of data with covariates  $(W_i, A_i, t)$  and outcome  $dN(t)$ ,  $t = 1, \dots, \tilde{T}_i$ .
- If  $\tilde{T}$  is continuous, then one could discretize time to create such estimators, or use Cox-proportional hazard regression type estimators.

## Library of candidate estimators of hazard

- Let  $\hat{\lambda}_j : \mathcal{M}_{NP} \rightarrow \sim$  be a candidate estimator,  $j = 1, \dots, J$ .
- This library of  $J$  estimators can include parametric logistic regression, and large variety of machine learning algorithms.

## Cross-validation to evaluate performance of candidate estimators

- Let  $B_n \in \{0, 1\}^n$  be a random split of sample.
- Let  $P_{n, B_n}^1, P_{n, B_n}^0$  be empirical measures for validation sample  $\{O_i : B_n(i) = 1\}$  and training sample  $\{O_i : B_n(i) = 0\}$ , respectively.
- We evaluate performance of  $\hat{\lambda}_j$  with its cross-validated log-likelihood:

$$R(\hat{\lambda}_j, P_n) = E_{B_n} P_{n, B_n}^1 L(\hat{\lambda}_j(P_{n, B_n}^0)).$$

## Super-learner of hazard

- Let

$$J_n = \arg \min_j R(\hat{\lambda}_j, P_n).$$

- The super-learner is defined as

$$\hat{\lambda}_{SL}(P_n) = \hat{\lambda}_{J_n}(P_n).$$

## Oracle inequality

- Let  $\tilde{J}_n(P_0) = \arg \min_j E_{B_n} P_0 L(\hat{\Psi}_j(P_{n,B_n}^0))$  be the oracle selector.
- Let  $d_0(\lambda, \lambda_0) = P_0 L(\lambda) - P_0 L(\lambda_0)$  be the loss-based dissimilarity.
- By the oracle inequality for the cross-validation selector, the super learner is asymptotically equivalent with the oracle selected estimator:

$$\frac{E_0 d_0(\hat{\lambda}_{\tilde{J}_n}(P_{n,B_n}^0), \lambda_0)}{E d_0(\hat{\lambda}_{\tilde{J}_n(P_0)}(P_{n,B_n}^0), P_0)} \rightarrow 1$$

as  $n \rightarrow \infty$ .

- This remains true if the number  $J$  of candidate estimators grows as fast with sample size as  $n^p$  for some finite integer  $p$ .

# Outline

- ① Causal model for the counterfactual treatment specific survival curve
- ② Efficient influence curve
- ③ Super-learner of conditional hazard
- ④ TMLE of treatment specific survival curve
- ⑤ One-step TMLE of treatment specific survival curve
- ⑥ Example: HAL-MLE of conditional hazard

## Recap estimation problem

- We want to construct a TMLE of  $\Psi(P_0) = E_{P_0} S_0(t_0 \mid A = d(W), W)$ .
- $S_0(t_0 \mid A, W) = \prod_{t \in [0, t_0]} (1 - \lambda(t \mid A, W))$ .
- Thus,  $\Psi(P) = \Psi(Q_W, \lambda)$ , where  $Q_W$  is probability measure of  $W$ .
- TMLE will be a plug-in estimator  $\Psi(Q_{W,n}, \lambda_n)$ ,  $Q_{W,n}$  empirical measure of  $W_1, \dots, W_n$ .

## Recap loss function for conditional hazard

- The log-likelihood loss for  $\lambda$  is given by:

$$L(\lambda)(O) = -\log \left\{ \prod_{t \leq \tilde{T}} \lambda(t | A, W)^{dN(t)} (1 - \lambda(t | A, W))^{1-dN(t)} \right\}.$$

- The log-likelihood loss for  $\lambda_C$  is given by:

$$L(\lambda_C)(O) = -\log \left\{ \prod_{t \leq \tilde{T}} \lambda(t | A, W)^{dA_2(t)} (1 - \lambda(t | A, W))^{1-dA_2(t)} \right\}.$$

## Loss function for conditional hazard at single time-point $t$

- Loss function for  $\lambda_t = \lambda(t | A, W)$  at fixed  $t$ :

$$L_t(\lambda_t)(O) = -I(t \leq \tilde{T}) \log \left( \lambda(t | A, W)^{dN(t)} (1 - \lambda(t | A, W))^{1-dN(t)} \right).$$

## Recap canonical gradient

- We derived canonical gradient  $D^*(P) = D^*(Q, G)$ ,  $Q = (Q_W, Q_N)$ ,  $Q_N$  is determined by  $\lambda$ .
- $G = (G_1, G_C)$ ,  $G_1$  is determined by  $g_1(a | W)$ ,  $G_C$  is determined by conditional hazard  $\lambda_C(t | A, W)$  of  $C$ .
- $D^*(P) = D_0^*(Q) + \sum_{t=1}^{\tau} D_t^*(Q, G)$ .
- $D_0^*(Q)$  is score of  $Q_W$ .
- $D_t^*(Q, G)$  score of  $\lambda(t | A, W)$ .

## $\lambda$ -component of canonical gradient

- $D_t^*(Q, G) = C_t(Q, G)I(\tilde{T} \geq t)(dN(t) - \lambda(t | A, W)).$
- Clever covariate:

$$C_t(Q, G) = \frac{I(A = d(W), \bar{A}_2(t-1) = 0)}{g_1(A | W) \prod_{s \leq t-1} (1 - \lambda_C(s | A, W))} \frac{S(t_0 | A, W)}{S(t | A, W)}.$$

## Initial estimator

- We estimate  $Q_{0,W}$  with the empirical measure of  $W_1, \dots, W_n$ .
- Let  $\lambda_n$  be an initial estimator of  $\lambda_0$  such as the super-learner presented in previous lecture.
- Let  $\lambda_{Cn}$  be an initial estimator (e.g., super-learner) of the conditional hazard  $\lambda_{0C}(t | A, W)$  of censoring
- Let  $g_{1n}(a | W)$  be estimator of treatment mechanism  $g_{10} = P_0(A = a | W)$ .

## Least favorable submodel for conditional hazard

- Let

$$\text{Logit}\lambda_{n,\epsilon}(t \mid A, W) = \text{Logit}\lambda_n(t \mid A, W) + \epsilon C_t(Q_n, G_n).$$

## Iterative TMLE

- Let  $\epsilon_n^0 = \arg \min_{\epsilon} P_n L(\lambda_{n,\epsilon})$ .
- This is equivalent with running a univariate logistic regression of  $dN(t)$  on  $C_t(Q_n, G_n)(A, W)$  based on pooled data set in which each subject contributes  $\tilde{T}$  rows of data, using as off-set  $\text{Logit}\lambda_n(t | A, W)$ .
- Let  $\lambda_n^1 = \lambda_{n,\epsilon_n^0}$  and  $Q_n^1 = (Q_{W,n}, Q_N(\lambda_n^1))$ .
- Set  $k = 1$ ;

$$\text{Logit}\lambda_{n,\epsilon}^k(t | A, W) = \text{Logit}\lambda_n^k(t | A, W) + \epsilon C_t(Q_n^k, G_n).$$

- $\epsilon_n^k == \arg \min_{\epsilon} P_n L(\lambda_{n,\epsilon}^k); \lambda_n^{k+1} = \lambda_{n,\epsilon_n^k}^k$ .
- Iterate till  $\epsilon_n^k \approx 0$ .

## LFM submodel for recursive TMLE

- For each  $t = 1, \dots, \tau$ , given a current estimator  $\lambda_{n,t}$  of  $\lambda_0(t | A, W)$ ,

$$\text{Logit}\lambda_{n,t,\epsilon}(A, W) = \text{Logit}\lambda_{n,t}(A, W) + \epsilon C_t(Q_n, G_n).$$

- This is a submodel to fluctuate  $\lambda_{n,t}(A, W)$  at fixed  $t$ .
- Its score spans  $D_t^*(Q_n, G_n)$ .

Carry out one-step updates recursively starting at end point  $\tau$

- Let  $\epsilon_n^\tau = \arg \min_\epsilon P_n L_\tau(\lambda_{\tau,n,\epsilon})$ .
- Let  $\lambda_{\tau,n}^* = \lambda_{\tau,n,\epsilon_n^\tau}$ .
- Let  $\lambda_{\tau-1,n,\epsilon}$  be above submodel through  $\lambda_{\tau-1,n}$  with covariate  $C_t(Q_n^*, G_n)$  using latest update  $\lambda_{\tau,n}^*$ .
- Let  $\epsilon_n^{\tau-1} = \arg \min_\epsilon P_n L_{\tau-1}(\lambda_{\tau-1,n,\epsilon})$ .
- Let  $\lambda_{\tau-1,n}^* = \lambda_{\tau-1,n,\epsilon_n^{\tau-1}}$ .

## Closed form recursive TMLE

- Let  $t = \tau - 2$ .
- Let  $\lambda_{t,n,\epsilon}$  be above submodel through  $\lambda_{t,n}$  with covariate  $C_t(Q_n^*, G_n)$  using latest update  $(\lambda_{t+1:\tau,n}^*$ .
- Let  $\epsilon_n^t = \arg \min_\epsilon P_n L_t(\lambda_{t,n,\epsilon})$ .
- Let  $\lambda_{t,n}^* = \lambda_{t,n,\epsilon_n^t}$ .
- Let  $t = t - 1$  and repeat above three steps, iterate, till we end up at  $t = 1$ .
- Then, we have the TMLE  $\lambda_n^* = (\lambda_{t,n}^* : t = 1, \dots, \tau)$ .

## Solves efficient influence curve equation

- The TMLE  $Q_n^* = (Q_{W,n}, Q_N(\lambda_n^*))$  solves

$$0 = P_n D_t^*(Q_n^*, G_n)$$

for each  $t = 0, \dots, \tau$ .

- In particular, it solves  $P_n D^*(Q_n^*, G_n) = 0$ .
- The iterative TMLE will solve approximately  $P_n D^*(Q_n^*, G_n) \approx 0$ .

# Outline

- ① Causal model for the counterfactual treatment specific survival curve
- ② Efficient influence curve
- ③ Super-learner of conditional hazard
- ④ TMLE of treatment specific survival curve
- ⑤ One-step TMLE of treatment specific survival curve
- ⑥ Example: HAL-MLE of conditional hazard

## From local least favorable submodel to universal least favorable submodel

- A local least favorable submodel (LLFM) for  $S_d(t)$  around initial estimator of conditional hazard:

$$\text{logit}(\lambda_{n,\varepsilon}(\cdot|A=1, W)) = \text{logit}(\lambda_n(\cdot|A=1, W)) + \varepsilon h_t.$$

- Similarly, we have this local least favorable submodel for a vector  $(S_d(t) : t)$  by adding vector  $(h_t : t)$  extension.
- These imply, as above, universal least favorable submodels for single and multidimensional survival function.

# Simulations for one-step TMLE of survival curve

We investigated the performance of one-step TMLE for treatment specific survival curve in two simulation settings.

## Data structure

- $O = (W, A, T) \sim P_0$
- $A \in \{0, 1\}$
- treatment intervention:  $W \rightarrow d(W) = 1$
- $S_d(t)$  is defined by

$$\Psi(P)(t) = E_P [P(T > t | A = d(W), W)]$$

# Candidate estimators

- ① Kaplan Meier
- ② Iterative TMLE for each single  $t$  separately
- ③ One-step TMLE targeting the whole survival curve  $S_d$

# Results

Figure: Based on one data set

## Monte-carlo results ( $n = 100$ )

**Figure:** Relative efficiency against iterative TMLE, as a function of t

# Outline

- ① Causal model for the counterfactual treatment specific survival curve
- ② Efficient influence curve
- ③ Super-learner of conditional hazard
- ④ TMLE of treatment specific survival curve
- ⑤ One-step TMLE of treatment specific survival curve
- ⑥ Example: HAL-MLE of conditional hazard

## Example: HAL-MLE of conditional hazard

- Suppose that  $O = (W, A, \tilde{T} = \min(T, C), \Delta = I(T \leq C))$ , and that we are interested in estimating the conditional hazard  $\lambda(t | A, W)$ .
- Let  $L(\lambda)$  be the log-likelihood loss.
- If  $T$  is continuous, we could parametrize  $\lambda(t | A, W) = \exp(\psi(t, A, W))$ , or, if  $T$  is discrete,  $\text{Logit}\lambda(t | A, W) = \psi(t, A, W)$ .
- We can represent  $\psi = \sum_{s \in \{1, \dots, d\}} \beta_{s,j} \phi_{u_{s,j}}$  as linear combination of indicator basis functions, where  $L^1$ -norm of  $\beta$  represents the sectional variation norm of  $\psi$ .
- Therefore, we can compute the HAL-MLE of  $\lambda$  with either Cox-Lasso or logistic Lasso regression (`glmnet()`).

# Targeted Minimum Loss-Based Estimation for Longitudinal Data

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 Causal model for longitudinal data with multiple time-point interventions
- 2 Sequential regression representation of target parameter
- 3 Efficient influence curve
- 4 TMLE of multiple time-point intervention specific survival curve based on sequential regression
- 5 TMLE package for multiple time-point interventions in longitudinal studies

## Longitudinal data

- Let

$$O = (L(0), A(0), \dots, L(K), A(K), Y = L(K + 1)) \sim P_0.$$

- $A(t) = (A_1(t), A_2(t))$ ,  $A_1(t)$  discrete valued treatment,  
 $A_2(t) = I(\tilde{T} \leq t, \Delta = 0)$  jumps at right censoring, where  
 $\tilde{T} = \min(T, C)$ .
- $L(t) = (N(t), L_1(t))$ ,  $N(t) = I(\tilde{T} \leq t, \Delta = 1)$ .
- $L_1(t)$  time-dependent covariates.
- $Y = I(\tilde{T} > K + 1)$  indicator of no event by time  $K + 1$ .
- Note that this process is degenerate after failure or censoring.

## Observed data distribution and notation

$$P_0(do) = \prod_{k=1}^{K+1} P_{L(k)}(dl(k) | \bar{l}(k-1), \bar{a}(k-1)) \prod_{k=1}^K P_{A(k)}(a(k) | \bar{a}(k-1), \bar{l}(k)).$$

- Let  $Q_{L(k)}$  denote  $P_{L(k)}$  and  $q_{L(k)}$  be its density,  $k = 0, \dots, K + 1$ .
- Let  $G_{A(k)}$  denote  $P_{A(k)}$ , and  $g_{A(k)}$  be its density (w.r.t. counting measure),  $k = 0, \dots, K$ ,
- Then, density of  $P$ :

$$p(o) = \prod_k q_{L(k)} \prod_k g_{A(k)}.$$

- Let  $Q = (Q_{L(k)} : k)$ ,  $G = (G_{A(k)} : k)$ .

## Statistical model

- Let  $\mathcal{Q}$  be a nonparametric parameter space for  $Q$ .
- Let  $\mathcal{G}$  be a possibly restricted parameter space for  $G$ .
- This allows incorporating assumptions on treatment and censoring mechanism.

$$\mathcal{M} = \{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}.$$

## Stochastic intervention

- Let  $G^* = (G_{A^*(1)}, \dots, G_{A^*(k)})$  be user supplied choice of conditional distributions.
- $g_{A^*(k)} = g_{A_1^*(k)} g_{A_2^*(k)}$  factorizes in treatment and censoring conditional distribution.
- Let censoring distribution  $G_{A_2^*(k)}$  be degenerate at 0, whatever conditioning event.
- We refer to  $G^*$  as a stochastic intervention if  $g_{A_1^*}$  is non-degenerate, and static or dynamic otherwise.

## G-computation formula for post-intervention distribution

Then,

$$p_{Q,G^*}(do) = \prod_k q_{L(k)}(I(k) | \bar{I}(k-1), \bar{a}(k-1)) \prod_k g_{A^*(k)}(a(k) | \bar{a}(k-1), \bar{I}(k)).$$

- The latter density is called the *G*-computation formula for  $G^*$ -post-intervention distribution.
- Note, it represents a modification of a data distribution  $P_{Q,G}$  obtained by replacing  $G$  by  $G^*$ .

## Statistical target parameter

- Let  $\Psi_{g^*} : \mathcal{M} \rightarrow \mathbb{R}$  be defined by

$$\Psi_{g^*}(P_{Q,G}) = E_{P_{Q,G^*}} Y.$$

- Could be evaluated by Monte-Carlo simulation.

## Causal model

- Define equations  $L(k) = f_{L(k)}(\bar{L}(k-1), \bar{A}(k-1), U_{L(k)})$ ,  
 $k = 1, \dots, K+1$ ,  $A(k) = f_{A(k)}(\bar{A}(k-1), \bar{L}(k), U_{A(k)})$ ,  $k = 1, \dots, K$ .
- Let  $U \sim P_U$  be collecting of exogenous error-variables  $U_{L(k)}$ ,  $U_{A(k)}$ .
- Let  $f$  be collection of functions  $f_{L(k)}$  and  $f_{A(k)}$ .
- Given  $P_U$  and  $f$ , this defines a probability distribution for  $O$ .
- Thus, this defines a parameterization  $P = P_{P_U, f}$ .

## Post-intervention distribution of $P$

- Given  $f$  and  $P_U$  (and thus given  $P$ ).
- Replace evaluation of  $f_{A(k)}$  in this data experiment by drawing from  $G_{A(k)}^*$ .
- This defines a new data distribution.
- We denote the random variable with  $O_{g^*}$  and its distribution with  $P_{g^*}$ .
- $P_{g^*}$  is a post-intervention distribution of  $P$ .

## Intervention specific survival curve under multiple time-point intervention

- Let  $Y_{g^*}$  be final outcome in  $O_{g^*}$ .
- Full-data quantity  $S_{0g^*}(K + 1) = E_0 Y_{g^*}$  treatment specific survival curve under intervention  $g^*$  at time  $K + 1$ .

## Strong sequential randomization assumption

- Assume, under  $P$ ,  $A(t)$  is independent of  $(Y_{a_1 0} : a_1)$ , given  $\bar{L}(t), \bar{A}(t - 1)$ .
- This strong SRA assumption on  $P$  allows identification of any full-data parameter of distribution of  $Y_{g^*}$  from observed data distribution  $P$ .

## Identification

- If strong SRA holds and  $\sup_o |g^*(o)/g(o)| < \infty$  on a support of  $P$ , then

$$p_{g^*} = p_{Q, G^*}.$$

- Thus, then  $S_{g^*}(K + 1) = \Psi_{g^*}(P)$ .
- Note that  $\Psi_{g^*}(P)$  only depends on  $P$  through  $Q(P)$ : we will also use notation  $\Psi_{g^*}(Q)$ .

## Statistical estimation problem

- Given a data set  $O_1, \dots, O_n \sim P_{Q_0, G_0} \in \mathcal{M}$ , we want to construct a TMLE (i.e., efficient plug-in estimator) of  $\Psi_{g^*}(P_0)$ , where  $\Psi_{g^*} : \mathcal{M} \rightarrow \mathbb{R}$  is defined by  $\Psi_{g^*}(P) = E_{P_{Q, G^*}} Y$ .

# Outline

- 1 Causal model for longitudinal data with multiple time-point interventions
- 2 Sequential regression representation of target parameter
- 3 Efficient influence curve
- 4 TMLE of multiple time-point intervention specific survival curve based on sequential regression
- 5 TMLE package for multiple time-point interventions in longitudinal studies

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  $bar{Q}_{A(K-1)} = E_{g_{A^*(K-1)}} \bar{Q}_{L(K)}$ .

## 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  $\bar{Q} = (\bar{Q}_{L(0)}, \bar{Q}_{L(K+1)})$ .
- 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  $\bar{Q}_{L(k)}^{g^*}$ ,  $\bar{Q}_{A(k)}^{g^*}$  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.

# Outline

- 1 Causal model for longitudinal data with multiple time-point interventions
- 2 Sequential regression representation of target parameter
- 3 Efficient influence curve
- 4 TMLE of multiple time-point intervention specific survival curve based on sequential regression
- 5 TMLE package for multiple time-point interventions in longitudinal studies

## Recap computation of canonical gradient

- A previous lecture on calculations of canonical gradients shows that

$$D^*(P) = \sum_{k=0}^{K+1} \left\{ E_P(D \mid \bar{L}(k), \bar{A}(k-1)) - E_P(D \mid \bar{L}(k-1), \bar{A}(k-1)) \right\},$$

where  $D = D(P)$  is a gradient of  $\Psi_{g^*}$  at  $P$  in model  $\mathcal{M}(G)$  with  $G$  known

## Initial gradient for model with $G$ known

- Let

$$D(P) = \frac{\prod_{k=1}^K g_{A^*(k)}(O)}{\prod_{k=1}^K g_{A(k)}(O)} Y - \Psi_{g^*}(P).$$

- Note that  $E_P D(P) = 0$ . Thus

$$\psi_n = \frac{1}{n} \sum_{i=1}^n \frac{\prod_{k=1}^K g_{A^*(k)}(O_i)}{\prod_{k=1}^K g_{A(k)}(O_i)} Y_i$$

is RAL (unbiased even) estimator of  $\psi_0$ , and its influence curve equals  $D(P)$ .

- Therefore,  $D(P)$  is a gradient of pathwise derivative of  $\Psi_{g^*}$  at  $P$ .

## Canonical gradient

- Plugging in and some algebra now yields following expression for  $D^*(P)$ .
- We have

$$D^*(P) = \sum_{t=0}^{K+1} D_t^*(P),$$

where  $D_0^*(P) = \bar{Q}_{A(0)} - \bar{Q}_{L(0)}$ , and for  $k = 1, \dots, K + 1$ ,

$$D_k^*(P) = \frac{g_{1:k-1}^*(O)}{g_{1:k-1}(O)} (\bar{Q}_{A(k)} - \bar{Q}_{L(k)}).$$

- Let

$$C_k(g) = \frac{g_{1:k-1}^*(O)}{g_{1:k-1}(O)}.$$

- Here  $g_{1:k} = \prod_{j=1}^k g_{A(j)}$  and similarly we define  $g_{1:k}^* = \prod_{j=1}^k g_{A^*(j)}$ .

# Outline

- 1 Causal model for longitudinal data with multiple time-point interventions
- 2 Sequential regression representation of target parameter
- 3 Efficient influence curve
- 4 TMLE of multiple time-point intervention specific survival curve based on sequential regression
- 5 TMLE package for multiple time-point interventions in longitudinal studies

## Sequential regression based TMLE

- We focus on estimation of the sequential regression  $\bar{Q}_{L(k)}$ .
- The TMLE will target each of these regressions sequentially in same order.
- The final targeted fit of  $\bar{Q}_{L(0)}$  will represent the TMLE of  $\Psi_{g^*}(P_0)$ .

## Submodel

- For  $k = K + 1, \dots, 1$ , let

$$\text{Logit } \bar{Q}_{L(k),\epsilon} = \text{Logit } \bar{Q}_{L(k)} + \epsilon C_k(g),$$

where

$$C_k(g) = \frac{\prod_{j=1}^{k-1} g_{A^*(j)}}{\prod_{j=1}^{k-1} g_{A(j)}} (\bar{L}(k-1), \bar{A}(k-1)).$$

- Let

$$\text{Logit } \bar{Q}_{L(0),\epsilon} = \text{Logit } \bar{Q}_{L(0)} + \epsilon$$

be the submodel through  $\bar{Q}_{L(0)}$ .

## Loss-function

- The loss for  $\bar{Q}_{L(K+1)}$  is

$$L(\bar{Q}_{L(K+1)}) = - \left\{ Y \log \bar{Q}_{L(k)} + (1 - Y) \log(1 - \bar{Q}_{L(k)}) \right\}.$$

- For  $k = K, \dots, 1$ , we define the following loss function for  $\bar{Q}_{L(k)}$ , given that we already computed  $\bar{Q}_{A(k)}$ :

$$L_{\bar{Q}_{A(k)}}(\bar{Q}_{L(k)}) = - \left\{ \bar{Q}_{A(k)} \log \bar{Q}_{L(k)} + (1 - \bar{Q}_{A(k)}) \log(1 - \bar{Q}_{L(k)}) \right\}.$$

- We also define the loss for  $\bar{Q}_{L(0)}$ :

$$L_{\bar{Q}_{A(0)}}(\bar{Q}_{L(0)}) = - \left\{ \bar{Q}_{A(0)} \log \bar{Q}_{L(0)} + (1 - \bar{Q}_{A(0)}) \log(1 - \bar{Q}_{L(0)}) \right\}.$$

- These loss functions can be used in sequence for both sequential (super) learning as well as the sequential targeting.

## Score of submodel generate canonical gradient

- For  $k = K, \dots, 1$ , we have

$$\frac{d}{d\epsilon} L_{\bar{Q}_{A(k)}}(\bar{Q}_{L(k),\epsilon}) \Big|_{\epsilon=0} = C_k(g)(\bar{Q}_{A(k)} - \bar{Q}_{L(k)}),$$

which thus equals  $D_k^*(P)$ . We make the convention that  
 $\bar{Q}_{A(K+1)} = Y$

- In addition, for  $k = 0$ -term

$$\frac{d}{d\epsilon} L_{\bar{Q}_{A(0)}}(\bar{Q}_{L(0),\epsilon}) \Big|_{\epsilon=0} = \bar{Q}_{A(0)} - \bar{Q}_{L(0)},$$

which equals  $D_0^*(P)$ .

- Thus, the linear span of the  $K + 2$  scores of the  $K + 2$  submodels corresponding with  $k = 0, \dots, K + 1$ , includes

$$D^*(P) = -\sum_{k=0}^{K+1} D_k^*(P).$$

## Estimator of treatment and censoring mechanism

- We can use loss-based super-learning to estimate the conditional probability distributions  $g_{A_1(k)}$  and  $g_{A_2(k)}$  of treatment and censoring, respectively.
- One uses the log-likelihood loss.
- One could use a separate estimator for each  $k$ , but one could also treat  $k$  as a covariate and carry out a super-learner of the  $(g_{A_1(k)} : k)$  with its log-likelihood loss applied to pooled data set in which each subject contributes  $\tilde{T}$  rows.

## First step of TMLE

- Obtain estimator  $\bar{Q}_{L(K+1),n}$ .
- Construct submodel  $\bar{Q}_{L(K+1),n,\epsilon}$  and fit  
 $\epsilon_{n,K+1} = \arg \min_{\epsilon} P_n L(\bar{Q}_{L(K+1),n,\epsilon})$ .
- Let  $\bar{Q}_{L(K+1),n}^* = \bar{Q}_{L(K+1),n,\epsilon_{n,K+1}}$ , which is the TMLE of  $\bar{Q}_{L(K+1)}$ .
- Compute  $\bar{Q}_{A(K),n}^*$  by evaluating conditional expectation of  $\bar{Q}_{L(K+1),n}^*$  w.r.t.  $g_{A^*(K)}$ .

## k-th step TMLE

- Let  $k = K$ .
- Obtain estimator  $\bar{Q}_{L(k),n}$ .
- Construct submodel  $\bar{Q}_{L(k),n,\epsilon}$  and fit
$$\epsilon_{n,k} = \arg \min_{\epsilon} P_n L_{\bar{Q}_{n,A(k)}^*}(\bar{Q}_{L(K+1),n,\epsilon}).$$
- Let  $\bar{Q}_{L(k),n}^* = \bar{Q}_{L(k),n,\epsilon_{n,k}}$ , which is the TMLE of  $\bar{Q}_{L(k)}$ .
- Compute  $\bar{Q}_{A(k-1),n}^*$  by evaluating conditional expectation of  $\bar{Q}_{L(k),n}^*$  w.r.t.  $g_{A^*(k-1)}$ .

# TMLE

- Set  $k = k - 1$ , and iterate the above steps for computing the next TMLE  $\bar{Q}_{L(k),n}^*$  and  $\bar{Q}_{A(k-1),n}^*$ , till  $k = 1$ .
- Estimate  $\bar{Q}_{L(0)}$  with the empirical mean  $\bar{Q}_{L(0),n}^*$  of  $\bar{Q}_{A(0),n}^*(L_i(0))$  over  $L_i(0)$ .
- No need to do a targeting step for the latter (would just set  $\epsilon_{n,k=0} = 0$ ).
- This finalizes the TMLE  $\bar{Q}_n^*$  of  $\bar{Q}_0 = (\bar{Q}_{0,L(k)} : k = 0, \dots, K + 1)$ .
- The TMLE of  $\psi_0$  is thus  $\psi_n^* = \Psi(\bar{Q}_n^*) = \bar{Q}_{L(0),n}^*$ .

## Inference

- We estimate  $D^*(P_0)$  with its plug-in estimator  $D^*(\bar{Q}_n^*, G_n)$ .
- We estimate the variance of  $\psi_n^*$  with  $\sigma_n^2 = P_n D^*(\bar{Q}_n^*, G_n)^2/n$ .
- $\psi_n^* \pm 1.96\sigma_n$  is an asymptotic 0.95-confidence interval.

# Outline

- 1 Causal model for longitudinal data with multiple time-point interventions
- 2 Sequential regression representation of target parameter
- 3 Efficient influence curve
- 4 TMLE of multiple time-point intervention specific survival curve based on sequential regression
- 5 TMLE package for multiple time-point interventions in longitudinal studies

`ltmle` package (Petersen et al. (2013), van der Laan, Gruber (2012), Schitzer et al. (2013))

R package: `ltmle`

- Causal effect estimation with multiple intervention nodes
  - Intervention specific mean under longitudinal dynamic interventions
  - Dynamic marginal structural working models
- General longitudinal data structures
  - Repeated measures outcomes, survival
  - Right censoring
  - Inference: different variance estimators.
- Estimators
  - IPTW
  - Non-targeted MLE
  - TMLE (two algorithms for MSM)
- Options include nuisance parameter estimation via `glm` regression formulas or calling `SuperLearner()`

## Appendix 1 - Highly Adaptive Lasso (HAL)

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

## Representation of cadlag function as linear combination of indicators

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

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

where  $\psi_s(u) = \psi(u_s, 0_{s^c})$  is the section of  $\psi$  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  $\{1, \dots, d\}$ .

## Variation norm

The variation norm of  $\psi$  can be defined as:

$$\| \psi \|_v = \sum_{s \in \{1, \dots, d\}} \int | d\psi_s(u_s) | .$$

## 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:

$$\psi(x) = \sum_{s \in \{1, \dots, d\}} \sum_j \beta_{s,j} \phi_{u_{s,j}}(x),$$

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

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

## 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  $\psi_{n,M} = \sum_{s \in \{1, \dots, d\}} \beta_{n,M,s,j} \phi_{u_{s,j}}$ , where

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

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  $\psi_n$  as the Highly Adaptive Lasso estimator (HAL-E).

## Guaranteed rate of convergence faster than $n^{-1/4}$

We have

$$d_0(\psi_{n,M}, \psi_{0,M}) = o_P(n^{-(1/2+\alpha(d)/4)}),$$

where  $\alpha(d) = 1/(d+1)$ . Thus, if we select  $M > \| \psi_0 \|_\nu$ , then

$$d_0(\psi_{n,M}, \psi_0) = o_P(n^{-(1/2+\alpha(d)/4)}).$$

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

$$d_0(\psi_n = \psi_{n,M_n}, \psi_0) = o_P(n^{-(1/2+\alpha(d)/4)}).$$

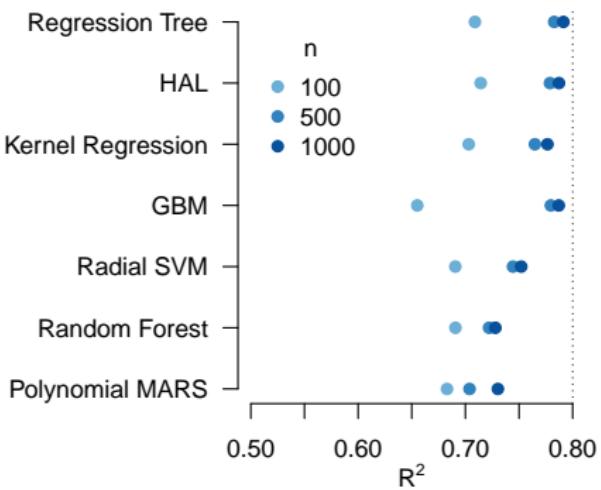
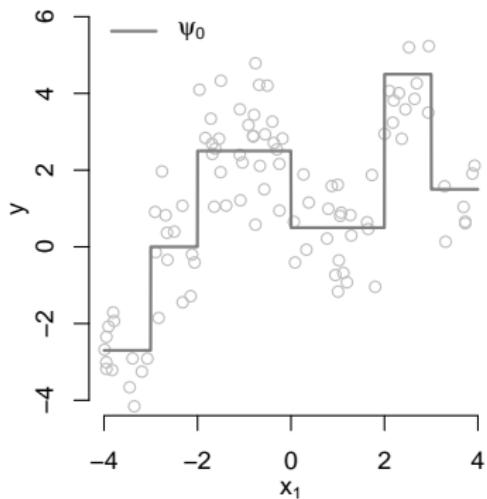
Super Learner should use HAL to guarantee this minimal rate of convergence: HAL-SL

## 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  
 $R^2 = 0.8$ .

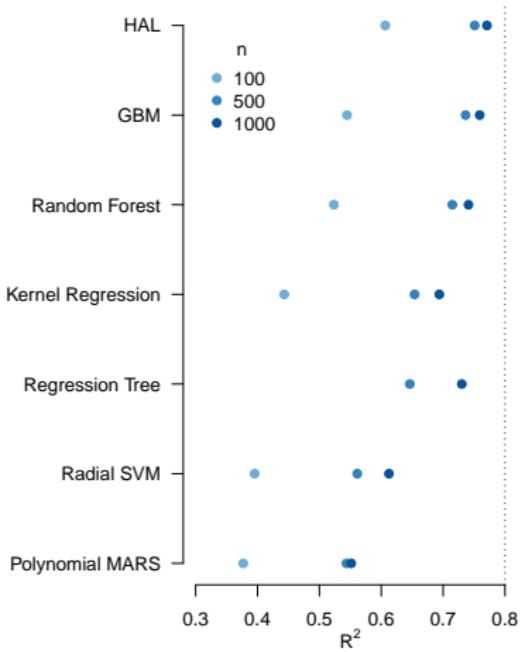
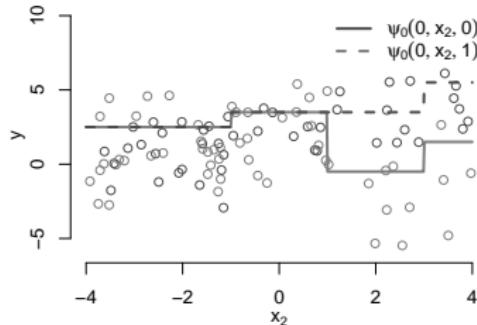
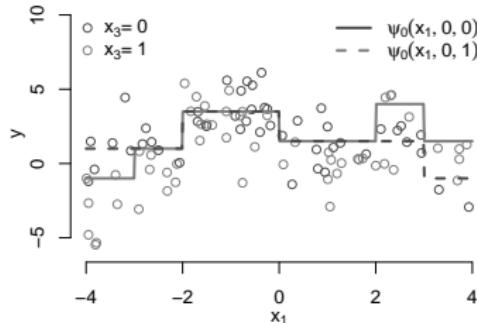
HAL was evaluated against competitor algorithms based on  $R^2$  calculated  
on an independent evaluation data set of size 10,000 averaged across the  
20 data sets.

# Jump functions, $d = 1$



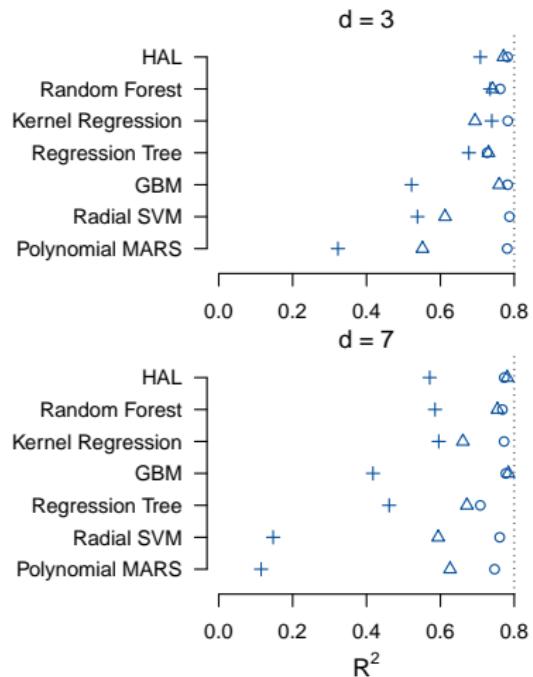
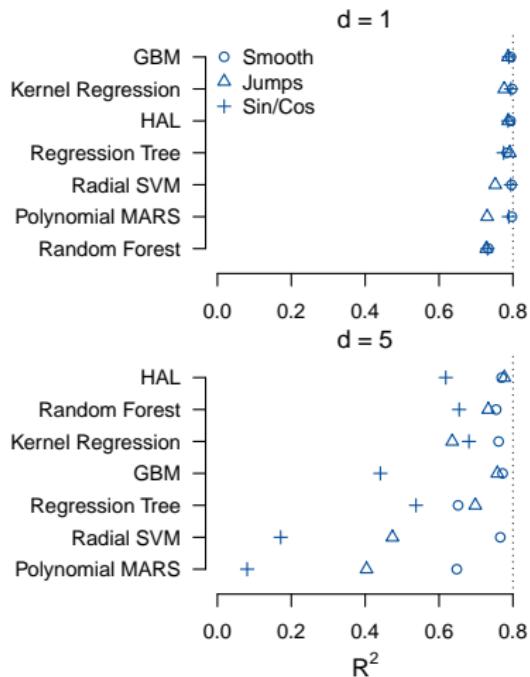
# Jump functions, $d = 3$

$$\begin{aligned}\psi_0(x) = & -2x_3 \mathbb{I}(x_1 < -3) + 2.5 \mathbb{I}(x_1 > -2) - 2 \mathbb{I}(x_1 > 0) + 2.5x_3 \mathbb{I}(x_1 > 2) \\ & -2.5 \mathbb{I}(x_1 > 3) + \mathbb{I}(x_2 > -1) - 4x_3 \mathbb{I}(x_2 > 1) + 2 \mathbb{I}(x_2 > 3)\end{aligned}$$



# Overall Performance

Different Data Generating Mechanisms and Dimensions, n=1000



## Appendix 2 – Online Super Learning

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

## Online estimation

- Let  $\mathbf{O}_k = (O_{n_{k-1}+1}, \dots, O_{n_k})$  represent the  $m = n_k - n_{k-1}$  observations making up batch  $k$ ,  $k = 1, 2, \dots, K$ , where  $n_0 = 0$ .
- An online estimator of  $\psi_0$  uses the next batch  $\mathbf{O}_k$  to update a current estimator  $\psi_{k-1}$  *without* revisiting the past data set  $\mathbf{O}_1, \dots, \mathbf{O}_{k-1}$ .
- For parametric (working) models  $\{p_\theta : \theta\}$ , and corresponding efficient score  $S_k^*(\theta)(\mathbf{O}_k) = c(\theta)S_k(\theta)(\mathbf{O}_k)$ , online estimators of  $\theta_0$  are of the form:

$$\theta_K = \frac{1}{K} \sum_{k=1}^K \theta_{k-1} + \frac{1}{K} \sum_{k=1}^K S_k^*(\theta_{k-1})(\mathbf{O}_k),$$

or equivalently,

$$\theta_K = \theta_{K-1} + \frac{1}{K} S_K^*(\theta_{K-1})(\mathbf{O}_K).$$

## Second and first order Stochastic Gradient Descent algorithm

- This is called the second order stochastic gradient descent algorithm (SGD), and it differs by the MLE by an asymptotically negligible term  $o_P(1/\sqrt{K})$ .
- By using diagonal standardizing matrix  $c(\theta)$ , different first order SGD are obtained that scale very well for high dimensional  $\theta$
- These first order SGD approximate the MLE up till  $O_P(1/\sqrt{K})$ -term.

## Online super learner of $P_0$

- Generate a library of online estimators: e.g.
  - Use an online SGD for a particular parametric model
  - Use a current estimator  $P_{k-1}$  based on  $\mathbf{O}_1, \dots, \mathbf{O}_{k-1}$  as off-set in a machine learning algorithm applied to the next batch  $\mathbf{O}_k$ .
  - Similarly, use an MLE for a parametric model  $\{P_{k-1,\theta} : \theta\}$  through the current estimator  $P_{k-1}$ , applied to next batch  $\mathbf{O}_k$ .
- The best linear combination of the candidate estimators  $\hat{P}^j$ ,  $j = 1, \dots, J$ , would be the  $\alpha_{K,online-MLE}$  that minimizers the following online-cross-validated risk:

$$\alpha \rightarrow \sum_{k=1}^K L \left( \sum_{j=1}^J \alpha(j) \hat{P}_{k-1}^j \right) (\mathbf{O}_k).$$

## Continued: Online super learner of $P_0$

- Instead, we could also use an SGD-online estimator of  $\alpha$  approximating the MLE  $\alpha_{K,online-MLE}$ , so that at step  $k$ , our next  $\alpha_k$  is based on the current  $\alpha_{k-1}$  and new batch  $\mathbf{O}_k$ .
- The online super-learner is defined as:

$$\hat{P}_k^{SL} = \sum_{j=1}^J \alpha_k(j) \hat{P}_{k-1}^j.$$

## Online learning for a time-series

- We have developed the theory (e.g. oracle inequality for cross-validation selector) for the online-super-learner for times series data  $O(t)$ ,  $t = 1, \dots, N$ .
- The statistical model assumes  $P(O(t) | \bar{O}(t - 1))$  depends on past through fixed summary measure  $Z_{t-1}$  of past with bounded memory.
- It also assumes this conditional probability distribution is indexed by common parameter across time  $t$ .
- For example, the density of  $O(t)$ , given  $\bar{O}(t - 1)$  is given by  $\bar{p}(o(t) | Z_{t-1})$  for common conditional density  $\bar{p}$ .

## Appendix 3 – Theoretical Foundations for TMLE

Mark van der Laan

Division of Biostatistics, University of California at Berkeley

**Deming Conference on Applied Statistics**

December 4-6 2019, Atlantic City NJ

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Finite dimensional results for the empirical measure

- Notation:  $P_n f \equiv \int f(o) dP_n(o) = \frac{1}{n} \sum_{i=1}^n f(O_i)$ .
- Empirical means converge to true means as  $n \rightarrow \infty$ :

$$P_n f \rightarrow_p P_0 f.$$

- Standardized empirical means converge in probability distribution to a normal:

$$\sqrt{n}(P_n - P_0)\vec{f} \Rightarrow_d N\left(0, \Sigma_0 = P_0(\vec{f} - P_0\vec{f})(\vec{f} - P_0\vec{f})^\top\right).$$

## Not sufficient

- Estimators or (e.g., test) statistics are not always a function of a vector of empirical means.
- Estimators and statistics are often a function of an infinite collection of empirical means of functions ranging over class  $\mathcal{F}$ .
- To analyze such an estimator or statistic, one will need to understand the empirical process uniformly in that class  $\mathcal{F}$ .

## Example of inference relying on empirical process for class of functions

- $S_0(t) = P_0(T > t)$ .
- $O = T$ .
- $\sup_t |S_n(t) - S_0(t)|$ , the supremum norm difference between empirical survival and true survival function, is a statistic depending on an infinite dimensional class of functions  $\mathcal{F} = \{T \rightarrow I(T > t) : t\}$ .
- In order to construct a simultaneous confidence band for  $S_0$ , one will need to understand the sampling distribution of this statistic.
- Same for right-censored data and Kaplan-Meier.

## Empirical process indexed by class of functions

- $G_n = \sqrt{n}(P_n - P_0)$
- $G_n(f) = \sqrt{n}(P_n - P_0)f = \frac{1}{n^{1/2}} \sum_{i=1}^n (f(O_i) - P_0 f).$
- Let  $\mathcal{F}$  a class of real valued functions of  $O$ .
- $(G_n(f) : f \in \mathcal{F})$  is an empirical process indexed by class  $\mathcal{F}$ .

## Uniform consistency of empirical measure

- $\| G_n \|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} | G_n(f) |$  supremum norm.
- A Glivenko-Cantelli class  $\mathcal{F}$  satisfies

$$\| P_n - P_0 \|_{\mathcal{F}} \rightarrow_p 0.$$

## Embedding empirical process in function space with supremum norm

- $\ell^\infty(\mathcal{F})$  is Banach space of functionals that map  $\mathcal{F}$  into real line, endowed with supremum norm.
- $G_n \in \ell^\infty(\mathcal{F})$  is a random element in this space, and it is a measurable function from underlying sample space to closed ball sigma algebra of  $\ell^\infty(\mathcal{F})$ .
- Thus, one can talk about the probability that  $G_n$  falls in a subset of  $\ell^\infty(\mathcal{F})$  in this sigma-algebra.
- $G_n$  is not measurable w.r.t. Borel sigma-algebra.
- Hofman-Jorgensen weak convergence theory handles Borel sigma-algebra through replacing probability by outer and inner probability.

## Weak convergence of random elements/function

- $G_n$  converges weakly to a  $G_0$  if, for each measurable set  $\mathcal{A} \subset \ell^\infty(\mathcal{F})$  for which  $G_0$  has zero probability to fall on edge of  $\mathcal{A}$

$$\Pr(G_n \in A) \rightarrow \Pr(G_0 \in A).$$

- Another equivalent definition for  $G_n \Rightarrow_d G_0$ : for each bounded continuous function  $h : \ell^\infty(\mathcal{F}) \rightarrow \mathbb{R}$ , we have  $Eh(G_n) \rightarrow Eh(G_0)$ .

## Continuous mapping theorem

- If  $g : \ell^\infty(\mathcal{F} \rightarrow (B, \|\cdot\|))$  is continuous for some banach space  $(B, \|\cdot\|)$ , and  $G_n \Rightarrow_d G_0$  in  $\ell^\infty(\mathcal{F})$ , then  $g(G_n) \Rightarrow_d g(G_0)$  in  $(B, \|\cdot\|)$ .
- More general: we can have that  $g_n : \ell^\infty(\mathcal{F}) \rightarrow (B, \|\cdot\|)$  is a sequence of functions approximating a fixed  $g$ . In that case, we need that  $(g_n : n)$  is continuous in the sense that  $g_n(x_n) \rightarrow g(x)$  for any sequence  $x_n \rightarrow x$ .
- In that case, we also have that  $G_n \Rightarrow_d G_0$  implies  $g_n(G_n) \Rightarrow_d g(G_0)$ .
- The latter result is called the **Extended Continuous Mapping Theorem**.

## Weak convergence of empirical process

- Let  $G_0$  be the Gaussian process implied by  $(G_0(f_1), \dots, G_0(f_m)) \sim N(0, \Sigma)$  for any subset  $\{f_1, \dots, f_m\} \subset \mathcal{F}$ , where  $\Sigma(i, j) = P_0 f_i f_j - P_0 f_i P_0 f_j$  is covariance of  $f_i(O)$  and  $f_j(O)$ .
- We want to know if  $G_n \Rightarrow_d G_0$ . Under what conditions on  $\mathcal{F}$  is this true?

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Equivalence of weak convergence with asymptotic equicontinuity

- In general, weak convergence is equivalent with convergence of all the finite dimensional distributions and an asymptotic equicontinuity/tightness condition.
- For empirical processes we already have the convergence of the finite dimensional distributions by the multivariate CLT.
- Asymptotic equicontinuity:

If  $f_n$  converges to 0 s.t.  $P_0 f_n^2 \rightarrow_p 0$ , then  $G_n(f_n) \rightarrow_p 0$ .

## Donsker class

- If  $G_n \Rightarrow_d G_0$  in  $\ell^\infty(\mathcal{F})$ , then  $\mathcal{F}$  is called a Donsker class.

## Example of Donsker class: Class of cadlag functions with bounded sectional variation norm

- Let  $O \in [0, \tau] \subset \mathbb{R}^d$ . Let  $D[0, \tau]$  be space of  $d$ -variate real valued cadlag functions.
- Let  $\| f \|_v^*$  be the sectional variation norm of  $f$  which computes the maximum variation norm of the measure  $df$  and the measures generated by the sections  $x(s) \rightarrow f(x(s), 0(-s))$  that sets various coordinates equal to 0.
- Let  $\mathcal{F} = \{f \in D[0, \tau] : \| f \|_v^* < M\}$  for some  $M < \infty$ .
- $\mathcal{F}$  is a Donsker class.

## Hilbert space $L^2(P)$

- For a given measure  $\mu$ ,  $L^2(P)$  is Hilbert space of real valued functions of  $O$  with inner product  $\langle f, g \rangle_P = Pf g$ .
- Thus,  $L^2(P)$ -norm is  $\| f \|_P = (Pf^2)^{1/2}$ .

## Entropy integral condition for Donsker class

- Let  $N(\epsilon, \mathcal{F}, L^2(P))$  be number of balls of size  $\epsilon$  needed to cover  $\mathcal{F}$  embedded in Hilbert space  $L^2(P)$ .
- If  $\int_0^1 \sup_P \sqrt{\log N(\epsilon, \mathcal{F}, L^2(P))} d\epsilon < \infty$ , then  $\mathcal{F}$  is (for any  $P_0$ ) a Donsker class.
- Thus  $N(\epsilon, \mathcal{F}, L^2(P)) \sim \exp(1/\epsilon^{2-\delta})$  for some  $\delta > 0$  is allowed.

## Important implication of Donsker class

If  $\mathcal{F}$  is a Donsker class, then

- $\sup_{f \in \mathcal{F}} |(P_n - P_0)f| = O_P(n^{-1/2}).$
- For any sequence  $f_n \in \mathcal{F}$  (possibly random by depending on  $O_1, \dots, O_n$ ) satisfying  $P_0 f_n^2 \rightarrow_p 0$ ,  
$$\sqrt{n}(P_n - P_0)f_n \rightarrow 0.$$
- The latter is an enormously important ingredient for any formal analysis of estimators (as we will see).

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Linear estimators for Linear Parameters

- If  $\Psi(P_0) = E_{P_0} f(O)$  for some function  $f$  (parameter is linear function of  $P_0$ ), then one can estimate it with a sample mean:

$$\hat{\Psi}(P_n) = \Psi(P_n) = P_n f = \frac{1}{n} \sum_{i=1}^n f(O_i).$$

- Such an estimator is unbiased and is a linear estimator (linear in  $P_n$ ).
- Unbiased estimators typically do not exist for non-linear parameters.

## Asymptotically linear estimators

- An estimator  $\hat{\Psi}(P_n)$  of  $\psi_0 \in \mathbb{R}^d$  is asymptotically linear with influence curve  $IC(P_0)$  if

$$\hat{\Psi}(P_n) - \Psi(P_0) = \frac{1}{n} \sum_{i=1}^n IC(P_0)(O_i) + o_P(n^{-1/2}).$$

- Then,

$$n^{1/2}(\hat{\Psi}(P_n) - \psi_0) \Rightarrow_d N_d(0, \Sigma_0),$$

where  $\Sigma_0$  is the  $d \times d$ -covariance matrix of the influence curve  $IC_0(O)$ :

$$\Sigma_0 = E_{P_0} IC(P_0) IC(P_0)^\top.$$

- An influence curve should have mean zero:  $E_0 IC_0(O) = 0$ .

## Variance estimation

- One can estimate the covariance matrix of  $\hat{\Psi}(P_n)$  with  $\Sigma_n/n$ , where  $\Sigma_n$  is the empirical covariance matrix of  $IC_0(O_i)$ ,  $i = 1, \dots, n$ :

$$\Sigma_n = P_n IC_0 IC_0^\top.$$

## Confidence interval

- An asymptotic 0.95-confidence interval is given by:

$$\psi_n(j) \pm 1.96 \Sigma_n(j,j)^{1/2} / n^{1/2}.$$

## Simultaneous confidence interval

- Let  $\rho$  be the correlation matrix of the covariance matrix  $\Sigma$ .
- Let  $q_{0.95}$  be the 0.95-quantile of  $\max_j |Z(j)|$  for  $Z \sim N(0, \rho)$ .
- A simultaneous 0.95-confidence interval for  $\psi_0(j)$ ,  $j = 1, \dots, d$ , is given by:

$$\psi_n(j) \pm q_{0.95} \Sigma_n(j, j)^{1/2} / n^{1/2}.$$

- Good homework exercise!

## Robustness analysis

- Let  $IC_n$  be an estimator of  $IC_0$ .
- $IC_n(O_i)$  measures the influence of observation  $O_i$  on the estimator  $\hat{\psi}_n = \hat{\Psi}(P_n)$ .
- Plotting  $IC_n(O_i)$  provides a tool to discover outliers.
- The function  $O \rightarrow IC_0(O)$  tells us how robust the estimator is.

## Theoretical insight from influence curve

- Assumptions on  $P_0$  that guarantee that  $IC(P_0)$  is a uniformly bounded function of  $O$  guarantee that the estimator will be well behaved in first order.

## Relative efficiency of two estimators

- Consider two asymptotically linear estimators with influence curves  $IC_{10}$  and  $IC_{20}$  at  $P_0$ .
- The relative efficiency of the two estimators is given by:

$$\frac{VAR_{P_0} IC_{10}(O)}{VAR_{P_0} IC_{20}(O)}.$$

## Efficient estimator

- In a future lecture we define an efficient estimator as an estimator that is asymptotically linear with influence curve equal to a so called **efficient influence curve**.
- The efficient influence curve is the best possible influence curve a regular asymptotically linear estimator can have.
- It represents a mathematical object that can be computed based on knowing the model  $\mathcal{M}$  and target parameter mapping  $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ .
- More later.

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Kaplan-Meier estimator as a functional of empirical measure

- $$S_n(t_0) = \prod_{s \in [0, t_0]} \left( 1 - \frac{dP_{1n}(s)}{\bar{P}_n(s-)} \right).$$
- $$S_0(t_0) = \prod_{s \in [0, t_0]} \left( 1 - \frac{dP_{10}(s)}{\bar{P}_0(s-)} \right).$$

## Define functional

- Define

$$\Phi(P_1, \bar{P}) = \prod_{s \in [0, t_0]} \left( 1 - \frac{dP_1(s)}{\bar{P}(s-)} \right).$$

- Then  $S_n(t_0) = \Phi(P_{1n}, \bar{P}_n)$  and  $S_0(t_0) = \Phi(P_{10}, \bar{P}_0)$ .

## Directional derivative of functional

- Define the functional derivative of  $\Phi$  at  $(P_{10}, \bar{P}_0)$  in direction  $(h_1, \bar{h})$  as

$$d\Phi(P_0)(h_1, \bar{h}) \equiv \left. \frac{d}{d\epsilon} \Phi(P_{10} + \epsilon h_1, \bar{P}_0 + \epsilon \bar{h}) \right|_{\epsilon=0}.$$

## Can define derivative as sum of partial derivatives

- Derivative of  $\Phi$  w.r.t.  $P_{10}$  in direction  $h_1$

$$\frac{d\Phi(P_0)}{dP_{10}}(h_1) \equiv \left. \frac{d}{d\epsilon} \Phi(P_{10} + \epsilon h_1, \bar{P}_0) \right|_{\epsilon=0}.$$

- Derivative of  $\Phi$  w.r.t  $\bar{P}_0$  in direction  $\bar{h}$

$$\frac{d\Phi(P_0)}{d\bar{P}_0}(\bar{h}) \equiv \left. \frac{d}{d\epsilon} \Phi(P_{10}, \bar{P}_0 + \epsilon \bar{h}) \right|_{\epsilon=0}.$$

- Derivative of  $\Phi$  in joint direction  $(h_1, \bar{h})$  equals sum of two partial derivatives:

$$d\Phi(P_0)(h_1, \bar{h}) = \frac{d\Phi(P_0)}{dP_{10}}(h_1) + \frac{d\Phi(P_0)}{d\bar{P}_0}(\bar{h}).$$

## Functional delta-method conditions

- $\Phi : D[0, \tau] \times D[0, \tau] \rightarrow \mathbb{R}$  functional on cartesian product of cadlag function spaces endowed with supremum norm.
- **Stochastic Convergence of input:** By empirical proces theory (indicators are a Donsker class):  
 $(n^{1/2}(P_{1n} - P_{10}), n^{1/2}(\bar{P}_n - \bar{P}_0)) \Rightarrow_d (Z_1, \bar{Z})$  for a joint Gaussian process  $(Z_1, \bar{Z})$ .
- **Differentiability of functional:** Suppose that for any sequence  $n^{1/2}(P'_{1n} - P_{10}) \rightarrow Z'_1$  and  $n^{1/2}(\bar{P}'_n - \bar{P}_0) \rightarrow Z'_2$  (analytically!), we have

$$\Phi(P'_{1n}, \bar{P}'_n) - \Phi(P_{10}, \bar{P}_0) - d\Phi(P_0)(P'_{1n} - P_{10}, \bar{P}'_n - \bar{P}_0) = o(n^{-1/2}).$$

## Functional delta-method statement

- Then,

$$\Phi(P_{1n}, \bar{P}_n) - \Phi(P_{10}, \bar{P}_0) = d\Phi(P_0)(P_{1n} - P_{10}, \bar{P}_n - \bar{P}_0) + o_P(n^{-1/2}).$$

## Implicaton I: Asymptotic linearity and normality

- Recall  $P_{1n}(t) = 1/n \sum_i I(\tilde{T}_i \leq t, \Delta_i = 1)$  and  $\bar{P}_n(t) = \frac{1}{n} \sum_i I(\tilde{T}_i > t)$ .
- Since  $d\Phi(P_0) : D[0, \tau] \times D[0, \tau] \rightarrow \mathbb{R}$  is a linear operator, we have

$$d\Phi(P_0)(P_{1n} - P_{10}, \bar{P}_n - \bar{P}_0) = \frac{1}{n} \sum_{i=1}^n d\Phi(P_0)(f_{1,O_i} - P_0 f_{1,O_i}, \bar{f}_{O_i} - P_0 \bar{f}_{O_i}),$$

where for a given  $O_i$

$$\begin{aligned} f_{1,O_i}(s) &= I(\tilde{T}_i \leq s, \Delta_i = 1) \\ \bar{f}_{O_i}(s) &= I(\tilde{T}_i \geq s). \end{aligned}$$

## Implication II: Influence curve of estimator as functional derivative

- Thus, the influence curve of  $S_n(t_0)$  as estimator of  $S_0(t_0)$  is given by:

$$IC_0(O_i) = d\Phi(P_0)(f_{1,O_i} - P_0 f_{1,O_i}, \bar{f}_{O_i} - P_0 \bar{f}_{O_i}).$$

- In general, indeed, the influence curve of an estimator is the functional derivative of the estimator  $\hat{\Psi}$  at  $P_0$  in the direction of  $(P_{n=1,O_i} - P_0)$ , the empirical process for a single observation  $O_i$ .

## Proof of functional delta-method

- Let  $\Phi(P_n)$  be estimator of  $\Phi(P_0)$ .
- Let  $Z_n = n^{1/2}(P_n - P_0)$ , which, is assumed to converge weakly to  $Z_0$  in function space.
- Let  $g_n(Z_n) = n^{1/2}(\Phi(P_0 + n^{-1/2}Z_n) - \Phi(P_0))$  and notice that it equals  $n^{1/2}(\Phi(P_n) - \Phi(P_0))$ .
- Let  $g(Z) = d\Phi(P_0)(Z)$  be functional derivative at  $P_0$  in direction  $Z$ .
- By analytic differentiability of  $\Phi$ , we have  $g_n(Z_n) \rightarrow g(Z_0)$  for any (analytic) sequence  $Z_n \rightarrow Z_0$ .
- Extended continuous mapping theorem says that  $g_n(Z_n) \Rightarrow_d g(Z_0)$ .

## General purpose of functional delta method

- $\theta_n$  is an asymptotically linear estimator of  $\theta_0$  (in some Banach space):

$$\theta_n - \theta_0 = (P_n - P_0)IC_\theta + o_P(n^{-1/2}).$$

- $\Phi$  is differentiable at  $\theta_0$  in above defined sense.
- Then,

$$\Phi(\theta_n) - \Phi(\theta_0) = \frac{1}{n} \sum_{i=1}^n d\Phi(\theta_0)(IC_\theta(O_i)) + o_P(n^{-1/2}).$$

- Thus, functional delta method is a method that tells us how to map influence curve of inputted estimator into influence curve of our estimator (and thereby providing inference).
- It also translates consistency and convergence in distribution of inputted estimator into these same properties for the estimator of interest.

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Class of parametric submodels through data distribution

- Let  $\mathcal{M}_h(P) = \{P_\epsilon^h : \epsilon \in (-\delta, \delta)\} \subset \mathcal{M}$  be a one-dimensional submodel/path through  $P$  at  $\epsilon = 0$ .
- Let  $S_h$  be its score:

$$S_h(O) = \frac{d}{d\epsilon} \log dP_\epsilon^h/dP(O) \Big|_{\epsilon=0}.$$

- We define whole class  $\mathcal{M}_h(P)$ ,  $h \in \mathcal{H}$ , of such submodels.
- Let  $\mathcal{S} = \{S_h : h \in \mathcal{H}\}$  be the set of all scores.

## Hilbert space for scores

- Let  $L_0^2(P)$  be Hilbert space of functions of  $O \sim P$  with mean zero with inner product

$$\langle f, g \rangle_P = E_P f(O)g(O).$$

- A score is an element of  $L_0^2(P)$ .

# Projection in Hilbert Space

- Let  $S$  be an element of  $L_0^2(P)$ .
- Let  $H$  be a sub-Hilbert space of  $L_0^2(P)$ .
- Then the projection  $\Pi(S | H)$  of  $S$  onto  $H$  is defined by 1) being an element in  $H$  and 2)  $S$  minus projection is orthogonal to any element in  $H$ :

$$\begin{aligned}\Pi(S | H) &\in H \\ S - \Pi(S | H) &\perp H.\end{aligned}$$

## Example

- $\mathcal{M}$  is nonparametric.
- $dP_\epsilon(o) = (1 + \epsilon h(o))dP(o)$ ,  $h$  uniformly bounded and  $E_P h(O) = 0$ .
- For  $\epsilon \in (-\delta, \delta)$  with  $\delta = 1/\|h\|_\infty$ , this is a submodel  $\mathcal{M}_h(P)$ .
- Score  $S = h$ .
- $\mathcal{S}$  is all  $h \in L_0^2(P)$  with  $\|h\|_\infty < \infty$ .

# Tangent Space

- Let  $T(P) \subset L_0^2(P)$  be the closure of the linear span of the set of scores  $\mathcal{S}$  of our class of paths.
- This is a sub-Hilbert space of  $L_0^2(P)$ .
- It is called the tangent space at  $P$ .

## Tangent space for nonparametric model

- The tangent space is the whole  $L_0^2(P)$ .
- We say that the model is locally saturated at  $P$ .

## Problem with standard directional derivative of target parameter

- We want to define a type of differentiability of  $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ .
- We could use the definition of a directional derivative in direction  $h$ :

$$d\Psi(P)(h) = \left. \frac{d}{d\epsilon} \Psi(P + \epsilon h) \right|_{\epsilon=0}.$$

- However,  $P + \epsilon h$  is not a path within  $\mathcal{M}$ , so this could be ill defined.
- Therefore, we define a derivative along paths that are submodels of  $\mathcal{M}$ .

## Pathwise derivative

- The pathwise derivative is defined as:

$$d\Psi(P)(S_h) = \left. \frac{d}{d\epsilon} \Psi(P_\epsilon^h) \right|_{\epsilon=0}.$$

- This is linear operator in its score  $S_h$ .
- Thus,  $d\Psi(P) : L_0^2(P) \rightarrow \mathbb{R}^d$  is a real valued linear operator on a Hilbert space  $L_0^2(P)$ .

## Pathwise differentiability and gradient

- $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$  is pathwise differentiable at  $P$  if its pathwise derivative is a **bounded** linear operator.
- By the Riesz-representation theorem, then  $d\Psi(P) : L_0^2(P) \rightarrow \mathbb{R}^d$  can be represented as an inner product of gradient with score:

$$d\Psi(P)(S_h) = E_P D(P)(O) S_h(O) = \langle D(P), S_h \rangle_P.$$

- $D(P)$  is called a gradient of the pathwise derivative.

## Class of gradients

- A gradient is not necessarily unique.
- Let  $T(P)^\perp = \{S \in L_0^2(P) : P \perp T(P)\}$  be orthogonal complement of  $T(P)$ .
- If  $D(P)$  is a gradient, then  $D(P) + S$  with  $S \in T(P)^\perp$  is also a gradient.

## Canonical gradient is projection of gradient on tangent space

- There is one unique gradient  $D^*(P) \in T(P)$  in the tangent space.
- This is called the canonical gradient.
- The set of all gradients is  $D^*(P) + S$  with  $S \in T(P)^\perp$ .
- If  $D$  is gradient, then canonical gradient  $D^*(P)$  is the projection of  $D(P)$  onto tangent space.

## Example

- $O = T$ ,  $\mathcal{M}$  nonparametric model,  $\Psi(P) = P(T > 5)$ .
- $dP_\epsilon(T) = (1 + \epsilon S(T))dP(T)$ ,  $S(T)$  is score.
- 

$$\frac{d}{d\epsilon} \Psi(P_\epsilon^h) \Big|_{\epsilon=0} = E_P D(P)(T) S_h(T),$$

where gradient

$$D(P)(T) = I(T > 5) - \Psi(P).$$

## Nonparametric model has only one gradient

- This gradient  $D(P)$  is also the canonical gradient.

## Finding canonical gradient in non-saturated models

- First find a gradient  $D(P)$  by computing the pathwise derivative.
- The canonical gradient equals the projection of  $D(P)$  onto the tangent space  $T(P)$ :

$$D^*(P) = \Pi(D(P) \mid T(P)).$$

- We will learn some powerful tools for computing such projections in a next lecture.

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Conditional Expectation is a Projection

- Let  $O = (X, Y) \sim P$ .
- Let  $H = \{S \in L^2(P) : S(X, Y) = S(X)\}$  be space of all functions of  $O$  that only depend on  $X$ .
- The projection of a function  $D(O)$  onto  $H$  equals the conditional expectation of  $D(O)$ , given  $X$ :

$$\Pi(D | H) = E_P(S(O) | X).$$

## Projection of functions with conditional mean equal to zero

- Let  $H = \{S(X, Y) : E_P(S(X, Y) | X) = 0\}$  be all functions with conditional mean, given  $X$ , equal to 0.
- Then, the projection of  $D$  onto  $H$  equals  $D$  minus its conditional mean:

$$\Pi(D | H) = D(X, Y) - E_P(D(X, Y) | X).$$

- More general, if  $O = (X, Y, Z)$  and  $H = \{S : S(X, Y, Z) = S(X, Y), E(S(X, Y) | X) = 0\}$ , then

$$\Pi(D | H) = E(D(O) | X, Y) - E(D(O) | X).$$

- Good homework assignments.

## Tangent space for model with variation independent parameterization

- $\mathcal{M} = \{P_{\theta_1, \theta_2} : \theta_1 \in \Theta_1, \theta_2 \in \Theta_2\}.$
- Tangent space  $T_1(\theta)$  of  $\theta_1$  is obtained by paths  $P_{\theta_1(\epsilon), \theta_2}$ , where  $\theta_1(\epsilon)$  is a path through  $\theta_1$ .
- Similarly, tangent space  $T_2(\theta)$  of  $\theta_2$ .
- The tangent space  $T(\theta)$  is the (closure of) sum  $T_1(\theta) + T_2(\theta)$  of the individual tangent spaces.
- This generalizes to  $T(\theta) = \sum_{j=1}^J T_j(\theta)$  for model  $\{P_{\theta_1, \dots, \theta_J} : \theta_j \in \Theta_j, j = 1, \dots, J\}.$

## Tangent spaces are orthogonal when density factorizes

- Consider model above but density factorizes:

$$p_\theta = p_{1,\theta_1} p_{2,\theta_2} \cdots$$

- Now tangent spaces of  $\theta_1, \theta_2$  are orthogonal:  $T_1(\theta) \perp T_2(\theta)$ .
- $T(\theta) = T_1(\theta) \oplus T_2(\theta)$  is an orthogonal sum of the two tangent spaces.
- In general, if  $p_\theta = \prod_{j=1}^J p_{j,\theta_j}$ , then  $T(\theta) = T_1(\theta) \oplus \dots \oplus T_J(\theta)$ .

## Projection on orthogonal sum of tangent spaces

- If  $T(\theta) = T_1(\theta) \oplus \dots \oplus T_J(\theta)$ , then the projection equals the sum of projections:

$$\Pi(D \mid T(\theta)) = \sum_{j=1}^J \Pi(D \mid T_j(\theta)).$$

## General model for longitudinal data

- Let  $O = (L(0), A(0), \dots, L(K), A(K), Y = L(K + 1))$ .
- Let the density of  $O$  factorize as

$$p(O) = \prod_{k=0}^{K+1} p_{L(k)}(O) \prod_{k=0}^K p_{A(k)}(O),$$

where  $p_{L(k)}$  is unspecified conditional density of  $L(k)$ , given  $Pa(L(k))$ , and  $p_{A(k)}$  is unspecified conditional density of  $A(k)$ , given  $Pa(A(k))$ .

- The parent set  $Pa(L(k))$  is specified subset of history  $\bar{L}(k - 1), \bar{A}(k - 1)$
- The parent set  $Pa(A(k))$  is specified subset of history  $\bar{L}(k), \bar{A}(k - 1)$ .
- This defines a statistical model  $\mathcal{M}$ .

## Tangent space

- Tangent space of  $p_{L(k)}$

$$T_{L(k)}(P) = \{S(L(k), Pa(L(k)) : E(S | Pa(L(k)) = 0\}.$$

- Tangent space of  $p_{A(k)}$ :

$$T_{A(k)}(P) = \{S(A(k), Pa(A(k)) : E(S | Pa(A(k)) = 0\}.$$

- Tangent space is orthogonal sum of  $T_{L(k)}$ ,  $k = 1, \dots, K+1$  and  $T_{A(k)}$ ,  $k = 1, \dots, K$ .

# Projection on Tangent Space

- Thus,

$$\Pi(D \mid T(P)) = \sum_{k=1}^{K+1} D_{L(k)}(P) + \sum_{k=1}^K D_{A(k)}(P).$$

- Projection on  $T_{L(k)}(P)$ :

$$D_{L(k)}(P) = E_P(D \mid L(k), Pa(L(k))) - E_P(D \mid Pa(L(k))).$$

- Projection on  $T_{A(k)}(P)$ :

$$D_{A(k)}(P) = E_P(D \mid A(k), Pa(A(k))) - E_P(D \mid Pa(A(k))).$$

## Application to find canonical gradients for causal quantities

- Define causal quantity, e.g.  $EY_a$  for static intervention  $a = (a(0), \dots, a(K))$ .
- Identify with statistical target parameter  $\Psi(P)$  under causal assumptions.
- Find initial gradient.
- Use the above formula to obtain the canonical gradient.
- Will do this concretely later.

# Outline

- 1 Empirical probability measure
- 2 Functional Central Limit Theorem for empirical processes
- 3 Asymptotic linearity of an estimator and its influence curve
- 4 Functional delta-method to obtain inference for estimators
- 5 Canonical gradient/efficient influence curve
- 6 Tools for computing projections / canonical gradient
- 7 Efficiency theory

## Asymptotically linear estimator

- For convenience, the true data distribution is now  $P$ .
- Suppose that  $\hat{\Psi}(P_n)$  is asymptotically linear at  $P$  with influence curve  $D(P)$ :

$$\hat{\Psi}(P_n) - \Psi(P) = P_n D(P) + R_2(P_n, P),$$

where  $R_2(P_n, P) = o_P(n^{-1/2})$  is second order remainder.

- This estimator is well behaved under i.i.d. sampling from  $P$ .
- Note the expansion simply defines  
 $R_2(P_n, P) = \hat{\Psi}(P_n) - \Psi(P) - R_2(P_n, P).$
- This expansion can be applied to any empirical measure/data set.

## Efficiency theory requires estimators to be regular

- There is no best sensible estimator under sampling from a fixed  $P$ , since I could select as estimator  $\hat{\Psi}(P_n) = \psi$ , which is perfect at true  $P$ , but horrible elsewhere.
- Therefore, for a sensible efficiency theory at  $P$  we need to restrict to estimators that behave well in neighborhood of  $P$ .
- Such estimators are called regular estimators, and that will be formally defined.

## Perturb data distribution

- Let  $\epsilon_n = n^{-1/2}$ .
- Take one of our paths  $\{P_\epsilon^h : \epsilon\} \subset \mathcal{M}$  through  $P$  at  $\epsilon = 0$  with score  $S_h$  and set  $\epsilon = \epsilon_n$ :

$$P_{\epsilon_n}^h.$$

- Consider a sample of  $n$  draws  $O_i^{\epsilon_n} \sim P_{\epsilon_n}^h$ .
- Let  $P_{\epsilon_n, n}^h$  be the empirical probability measure for this sample.
- We want our estimator to also behave well under sampling from  $P_{\epsilon_n, n}^h$ , for each path choice  $h$ .

Use the expansion of estimator at  $P$  but apply to perturbed sample

- Let's suppress  $h$  in notation.
- The expansion for  $\hat{\Psi}(P_n)$  at  $P$  applied to the perturbed data  $P_{\epsilon_n, n}$  yields

$$\hat{\Psi}(P_{\epsilon_n, n}) - \Psi(P) = P_{\epsilon_n, n} D(P) + R_2(P_{\epsilon_n, n}, P).$$

## Require robustness of estimator in terms of second order remainder

- An asymptotically linear estimator satisfies  $R_2(P_n, P_0) = o_P(n^{-1/2})$  and  $E\hat{\Psi}(P_n) - \Psi(P) = o(n^{-1/2})$ .
- We want this behavior to hold up under sampling from  $P_{\epsilon_n}$  as well
- Since  $P_{\epsilon_n}$  converges to  $P$  at rate  $n^{-1/2}$ , it is more than reasonable to require that the second order remainder is still  $o_P(n^{-1/2})$ :

**ConditionA :**  $R_2(P_{\epsilon_n, n}, P) = O_P(n^{-1/2})$ .

- Note that  $P_{\epsilon_n}$  approximates  $P$  in a very strong sense: convergence of densities at rate  $\epsilon_n$ , making this condition trivially satisfied for reasonable estimators.

## Implication of Condition A

- Then, our expansion reduces to:

$$\hat{\Psi}(P_{\epsilon_n, n}) - \Psi(P) = P_{\epsilon_n, n} D(P) + o_P(n^{-1/2}).$$

## Condition B: No asymptotic bias

- **Condition B:** For each path, we assume that  $\sqrt{n}(\hat{\Psi}(P_{\epsilon_n,n}) - \Psi(P_{\epsilon_n})) \Rightarrow_d N(0, \Sigma)$ .
- That is, we assume that our estimator preserves asymptotic normality with mean zero bias under sampling from  $P_{\epsilon_n}$ , for all paths.
- Alternatively ,we could also assume the stronger **condition B\***: for all paths,

$$\epsilon_n^{-1} \left\{ E\hat{\Psi}(P_{\epsilon_n,n}) - \Psi(P_{\epsilon_n}) \right\} \rightarrow 0$$

and

$$\epsilon_n^{-1} ER_2(P_{\epsilon_n,n}, P) \rightarrow 0.$$

- The latter states that the finite sample bias is  $o(n^{-1/2})$  under sampling from  $P_{\epsilon_n}$ , for all paths.

## What are we trying to show?

- Under conditions A and B, we will show that the influence curve  $D(P)$  of  $\hat{\Psi}$  at  $P$  is a gradient.
- This then shows that the best estimator among this class satisfying conditions A and B is an estimator whose influence curve  $D(P)$  is the canonical gradient.

## Center expansion

- Center estimator w.r.t its target:

$$\hat{\Psi}(P_{\epsilon_n, n}) - \Psi(P) = \hat{\Psi}(P_{\epsilon_n, n}) - \Psi(P_{\epsilon_n}) + \Psi(P_{\epsilon_n}) - \Psi(P).$$

- Center empirical sum w.r.t. its target

$$P_{\epsilon_n, n} D(P) = (P_{\epsilon_n, n} - P_{\epsilon_n}) D(P) + P_{\epsilon_n} D(P).$$

- The first term is empirical mean of mean zero random variables with variance converging to variance of  $D(P)$ . So by CLT this converges to  $N(0, \Sigma(P))$ , same limit distribution of  $\hat{\Psi}(P_n)$ .

## Isolate bias term

- So

$$\begin{aligned}\epsilon_n^{-1} \left\{ \hat{\Psi}(P_{\epsilon_n, n}) - \Psi(P_{\epsilon_n}) \right\} &= \epsilon_n^{-1} (P_{\epsilon_n, n} - P_{\epsilon_n}) D(P) + o_P(n^{-1/2}) \\ &\quad + \epsilon_n^{-1} P_{\epsilon_n} D(P) - \epsilon_n^{-1} (\Psi(P_{\epsilon_n}) - \Psi(P)).\end{aligned}$$

- The last term is a pure bias term, and represents the asymptotic bias of the standardized estimator applied to perturbed sample.

Asymptotic zero bias condition B will thus require bias term to be negligible

- Thus, in order to have that the standardized estimator under sampling from  $P_{\epsilon_n}$  still converges in distribution to  $N(0, \Sigma(P))$  the bias will need to disappear:

$$\epsilon_n^{-1} P_{\epsilon_n} D(P) - \epsilon_n^{-1} (\Psi(P_{\epsilon_n}) - \Psi(P)) = o(\epsilon_n) = o(1).$$

- Thus, **Condition B**: implies that this bias term is  $o(1)$ .

## Understanding negligible bias condition for single path

- Thus, we have

$$\epsilon_n^{-1}(P_{\epsilon_n} - P)D(P) - \epsilon_n^{-1}(\Psi(P_{\epsilon_n}) - \Psi(P)) \rightarrow 0.$$

- By pathwise differentiability with canonical gradient  $D^*(P)$ , this is equivalent with

$$\epsilon_n^{-1}(P_{\epsilon_n} - P)D(P) - PD^*(P)S \rightarrow 0,$$

where  $S$  is score of path.

- But,

$$\int D(P)\epsilon_n^{-1}(dP_{\epsilon_n} - dP) = \int D(P)\epsilon_n^{-1}\frac{(dP_{\epsilon_n} - dP)}{dP}dP,$$

and that converges to  $ED(P)S$ .

- Thus, the negligible bias condition is equivalent with requiring  $PD(P)S = PD^*(P)S$ .

## Understanding negligible bias condition uniformly in all paths

- Since we need this for all paths, we thus require that  $D(P) - D^*(P) \perp T(P)$ :

$$P(D(P) - D^*(P))S_h = 0 \text{ for all paths } h \in \mathcal{H}.$$

- This implies  $D(P) = D^*(P) + S$  for a  $S \in T(P)^\perp$ .
- Thus, the influence curve  $D(P)$  of  $\hat{\Psi}$  is a gradient of the pathwise derivative at  $P$ .

# Cramer-Rao lower bound for asymptotic variance of regular asymptotically linear estimators

This proves the following Theorem: **Theorem:**

- Consider an asymptotically linear estimator at  $P$  with influence curve  $D(P)$  and define its remainder

$$R_2(P_n, P) \equiv \hat{\Psi}(P_n) - \Psi(P) - P_n D(P) = o_P(n^{-1/2}).$$

- Regularity condition at  $P$ :** Assume that for each path  $\{P_\epsilon^h : \epsilon\} \subset \mathcal{M}$ ,  $h \in \mathcal{H}$ ,  $\epsilon_n = n^{-1/2}$ , empirical distribution  $P_{n,\epsilon_n}$  under i.i.d. sampling from  $P_{\epsilon_n}$ :

$$R_2(P_{\epsilon_n, n}, P) = o_P(n^{-1/2}),$$

and  $\sqrt{n}(\hat{\Psi}(P_{\epsilon_n, n}) - \Psi(P_{\epsilon_n})) \Rightarrow N(0, \Sigma(P))$ .

- Then,  $D(P)$  is a gradient of  $\Psi : \mathcal{M} \rightarrow \mathbb{R}$  at  $P$ .
- In particular,  $\text{VARD}(P)(O) \geq \text{VARD}^*(P)(O)$ .

## Generalized Cramer-Rao lower bound

- $\text{VARD}^*(P)(O)$  is called the generalized Cramer-Rao lower bound for the asymptotic variance of RAL-estimators.

## Defining efficient estimator

- Thus, an estimator that is asymptotically linear at  $P$  with influence curve equal to canonical gradient  $D^*(P)$  has the smallest asymptotic variance among all asymptotically linear estimator satisfying the above regularity condition at  $P$ .
- Therefore, an estimator is asymptotically efficient (among this class of regular asymptotically linear (RAL)-estimators) if it is asymptotically linear with influence curve equal to the canonical gradient  $D^*(P)$ .
- This explains why the canonical gradient is also called the efficient influence curve.

## Convolution Theorem

- There is a more powerful convolution theorem that applies to all estimators that converge in distribution at  $n^{1/2}$ -rate to limit distribution, and preserve this convergence to same limit distribution under sampling from  $P_{\epsilon_n}$  for all paths. Among this class of regular estimators it shows that the limit distribution the estimator is a convolution  $Z + E$  of two independent random variables  $Z$  and  $E$ , where  $Z \sim N(0, \text{VARD}^*(P))$ .
- Thus this shows that the most concentrated limit distribution is the normal with mean zero and variance the variance of canonical gradient.
- This convolution theorem is more general then our theorem above, by not requiring that the estimators are asymptotically linear.