2013

A Targeted Confounder Selection Strategy for Propensity Score Estimation

Susan Gruber, *Harvard School of Public Health*

A Targeted Confounder Selection Strategy for Propensity Score Estimation

Susan Gruber

Department of Epidemiology, Harvard School of Public Health

A Targeted Confounder Selection Strategy for Propensity Score Estimation

I. Motivation
   i. Propensity scores to address confounding bias
   ii. Impact of alternative propensity score models

II. Data-adaptive propensity score modeling using Collaborative Targeted Maximum Likelihood Estimation
   Gruber and van der Laan, 2010

III. Biomarker discovery using C-TMLE

IV. Concluding remarks

Causal Effect Estimation in Observational Data

- Propensity score methods
  - Model conditional distribution of treatment given covariates
    π_i = subject i's probability of receiving treatment
  - Matching, stratification
  - Inverse probability weighting (IPW)
- Outcome regression methods
  - Model expected value of outcome given treatment and covariates
- Double-Robust methods

  consistent if *either* outcome regression *or* pscore model correct
  - Targeted maximum likelihood estimation (TMLE)

Running Example

- Estimate marginal additive treatment effect (ATE)
  outcome Y, binary point treatment A, covariate vector W

$$\begin{aligned} \psi &= E[Y_1 - Y_0] \\ &= E[E(Y \mid A = 1, W) - E(Y \mid A = 0, W)] \end{aligned}$$

  when appropriate causal assumptions hold

- Y_1, Y_0 *potential* or *counterfactual* outcomes
  obtained by intervening to set A to 1 or 0, respectively
- For clarity, focus on selection bias, binary treatment, no dropout, no unmeasured confounding

Propensity Score Modeling

- Data $= n$ i.i.d. copies of $O = (W, A, Y)$
  W = covariate vector, A = binary treatment, Y = outcome
- IPW Estimator of ATE

$$\begin{aligned} \hat{\psi}_{IPW} &= \frac{1}{n}\sum_{i=1}^{n} wt_i Y_i \\ wt_i &= \frac{A}{P(A_i = 1 \mid \widetilde{W}_i)} - \frac{1 - A_i}{1 - P(A_i = 1 \mid \widetilde{W}_i)} \end{aligned}$$

- $\pi_i = P(A_i = 1 \mid \widetilde{W}_i)$ is pscore for subject i

What covariates should be included in $\widetilde{W}$?

Propensity Score Modeling: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

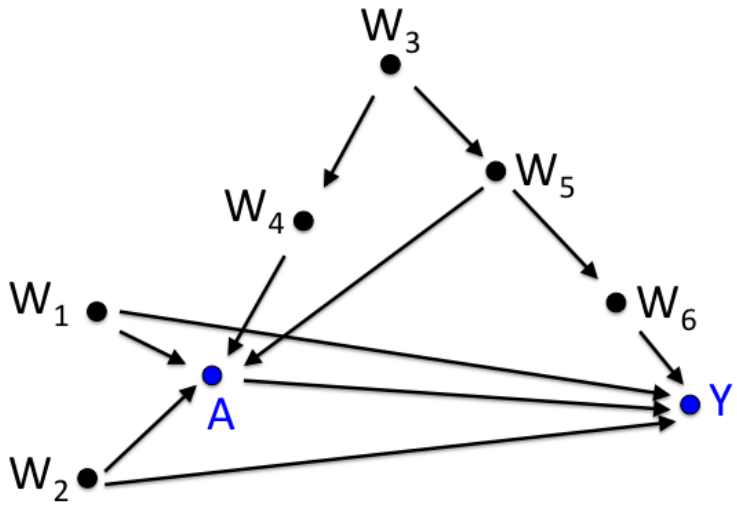
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

How to choose?



True causal DAG

Propensity Score Modeling: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

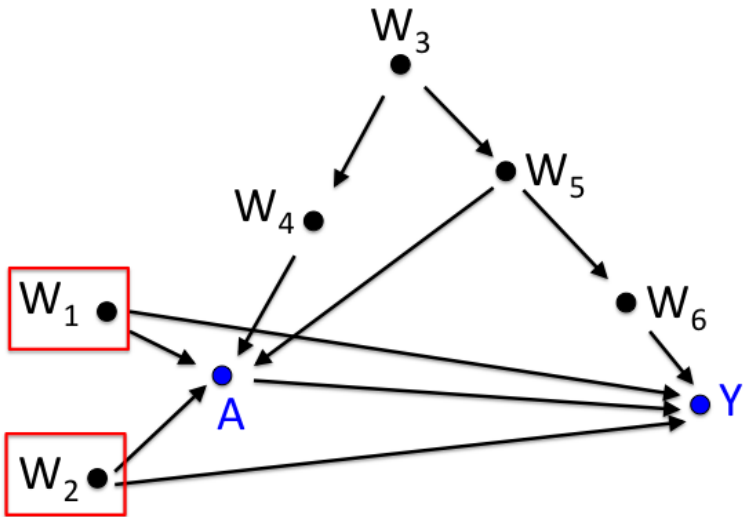
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

How to choose?



True causal DAG

Need to condition on common causes of A and Y

Propensity Score Modeling: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

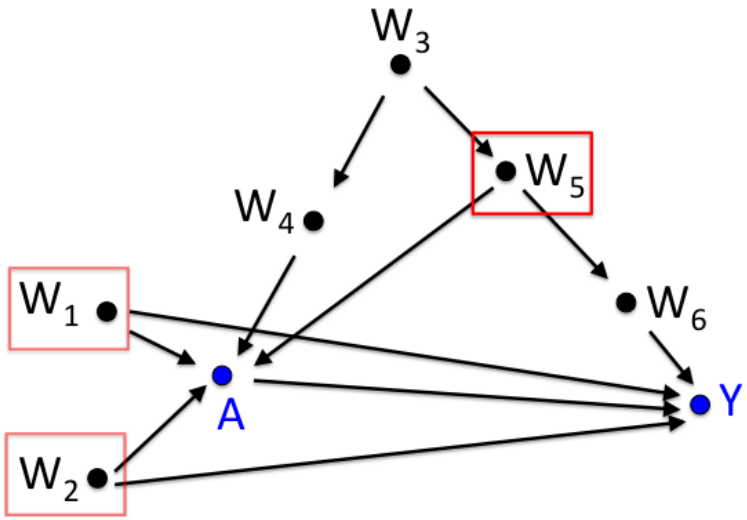
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

How to choose?



True causal DAG

Need to condition on common causes of A and Y

Propensity Score Modeling: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

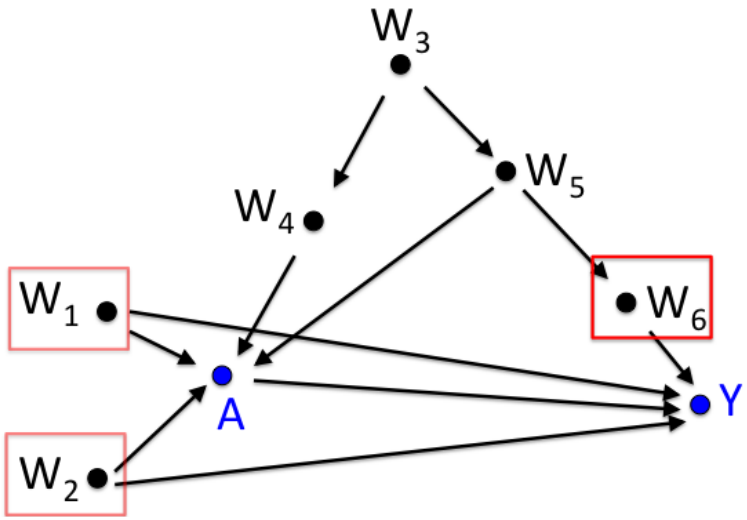
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

How to choose?



True causal DAG

Need to condition on common causes of A and Y

Propensity Score Modeling: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

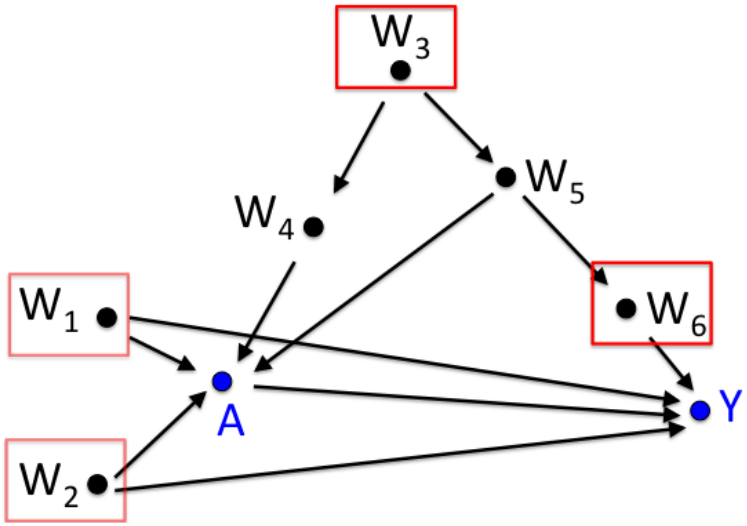
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

How to choose?



True causal DAG

Need to condition on common causes of A and Y

Propensity Score Modeling: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

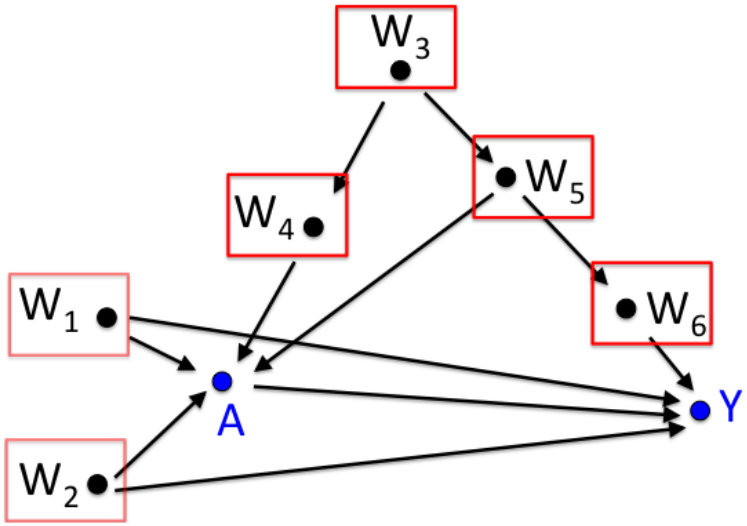
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

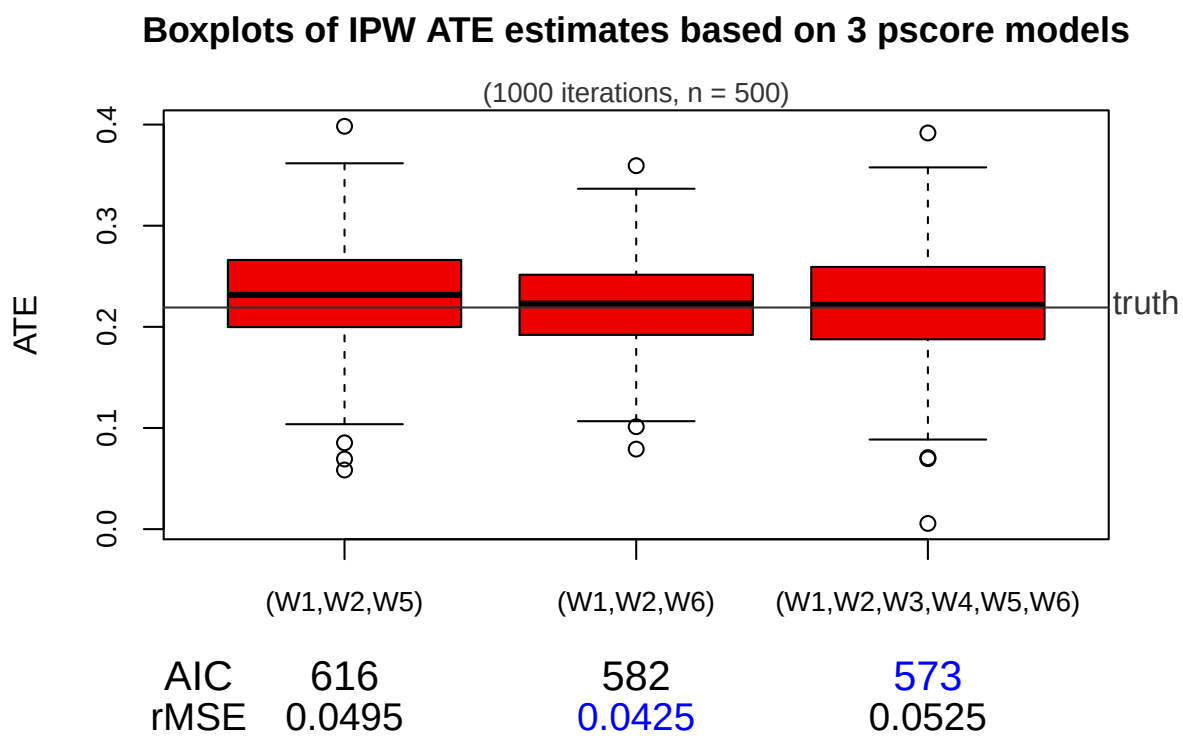
How to choose?



True causal DAG

Need to condition on common causes of A and Y

Demonstration: Choice of Adjustment Set Matters



- IPW estimates using pscore models based on three of the many *sufficient* adjustment sets (sufficient for consistent estimation of ATE)
- rMSE selects (W_1, W_2, W_6), AIC selects $(W_1, W_2, W_3, W_4, W_5, W_6)$

Confounder Selection: Point Treatment Example

Estimate additive effect of treatment A on outcome Y

Candidate Adjustment Sets

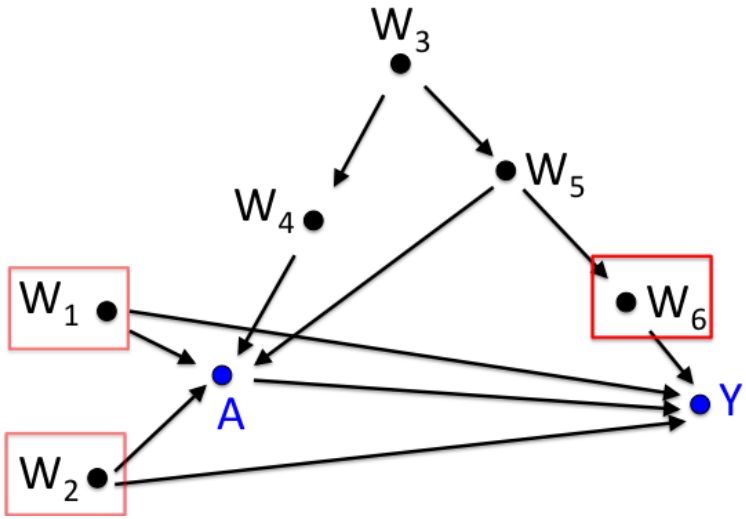
(W_1, W_2, W_5)

(W_1, W_2, W_6)

$\vdots$

$(W_1, W_2, W_3, W_4, W_5, W_6)$

How to choose?



True causal DAG

Best adjustment set depends on
strength of relationships + causal structure

Collaborative Targeted Maximum Likelihood Estimator (C-TMLE)

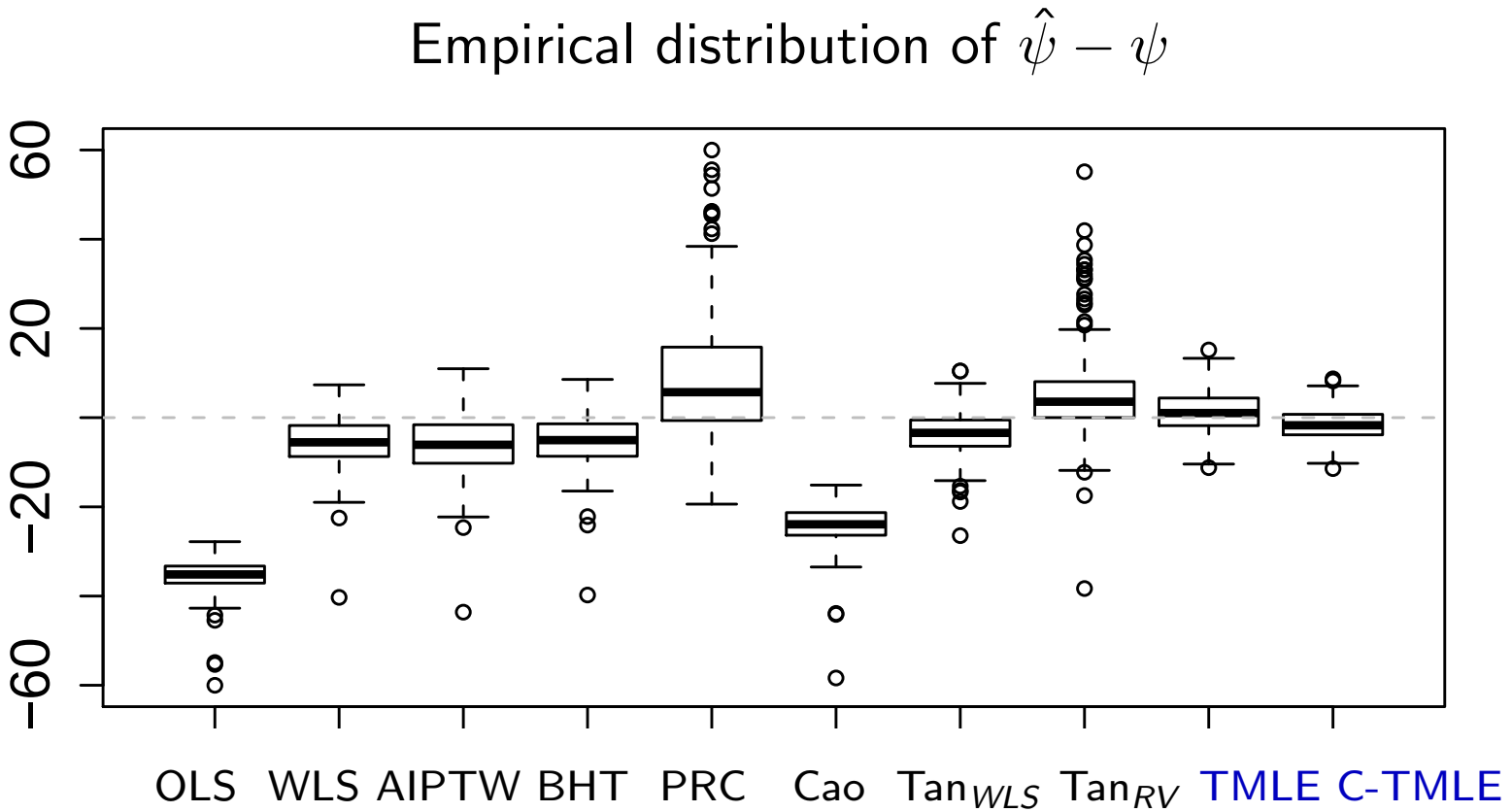
- If there exists a sufficient adjustment set, C-TMLE will asymptotically select this set
- If multiple sufficient adjustment sets exist with different asymptotic variance, C-TMLE will with high probability make the oracle selection
- Algorithm iteratively updates an ordering over covariates
  - sequence of nested pscore models, $\{\hat{\pi}^1, \ldots, \hat{\pi}^K\}$
  - corresponding sequence of double robust TMLEs

$$\left\{(\widehat{OR}, \hat{\pi}^1), \quad \ldots, \quad (\widehat{OR}, \hat{\pi}^K)\right\}$$

- Cross-validation to select the best TMLE in the sequence

Simulation Results

Misspecified outcome regression, correct pscore model (poor overlap)



Porter, Gruber, Sekhon and van der Laan. *The International Journal of Biostatistics*, 2011.

TMLE for locally efficient double-robust estimation

Two-stage procedure

Stage 1: Obtain initial outcome regression fit

parametric model, data-adaptive modeling

Stage 2: Targeting Stage

- Fluctuate initial fit to reduce bias in estimate of ψ
- Fluctuation parameter ϵ fit by maximum likelihood

fitting ϵ solves $P_n D^*(P) = 0$

efficient influence curve estimating equation for ψ

TMLE Algorithm for Estimating ATE Parameter

1. Obtain predictions from initial outcome regression, $\hat{Y}_0$, $\hat{Y}_1$
2. Fit fluctuation parameter ϵ
   - *regress residuals on parameter-specific covariate: $h = \frac{A}{\pi} - \frac{(1-A)}{(1-\pi)}$

   ```
   glm(Y ~ -1 + h, offset = Yhat, family = "binomial")
   ```
3. Update initial predictions

$$\hat{Y}_1^* = \hat{Y}_1 + \hat{\epsilon}\left(\frac{1}{\hat{\pi}}\right), \quad \hat{Y}_0^* = \hat{Y}_0 + \hat{\epsilon}\left(\frac{-1}{1-\hat{\pi}}\right)$$

4. Evaluate estimate $\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_1^* - \hat{Y}_0^*)$

*oversimplification

TMLE Algorithm for Estimating ATE Parameter

1. Obtain predictions from initial outcome regression, $\hat{Y}_0$, $\hat{Y}_1$
2. Fit fluctuation parameter ϵ
   - *regress residuals on parameter-specific covariate: $h = \frac{A}{\pi} - \frac{(1-A)}{(1-\pi)}$

   ```
   glm(Y ~ -1 + h, offset = Yhat, family = "binomial")
   ```
3. Update initial predictions

$$\hat{Y}_1^* = \hat{Y}_1 + \hat{\epsilon}\left(\frac{1}{\hat{\pi}}\right), \quad \hat{Y}_0^* = \hat{Y}_0 + \hat{\epsilon}\left(\frac{-1}{1-\hat{\pi}}\right)$$

4. Evaluate estimate $\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_1^* - \hat{Y}_0^*)$

How to estimate the pscores?

*oversimplification

C-TMLE Algorithm

Example: $O = (W_1, W_2, W_3, A, Y)$

1. Predicted values from initial OR fit, $\hat{Y} = \hat{E}(Y|A, W)$
2. Collaborative targeted forward selection

   $k = 1$: Intercept model, $\hat{\pi}^{k=1} = P(A = 1)$

   - Use $\hat{\pi}^{k=1}$ to create candidate TMLE$^{k=1}$

     $$\widehat{Y}^{k=1} = \widehat{Y} + \hat{\epsilon} h$$

   - parameter specific covariate, $h = \frac{A}{\pi} - \frac{(1-A)}{(1-\pi)}$
   - regress residuals on h to fit ϵ
   - h chosen so that fitting ϵ by maximum likelihood solves efficient influence curve estimating equation for ψ.
   - Evaluate targeted predicted values, $\hat{Y}_a^{k=1} = \hat{Y}_a + \hat{\epsilon} h_a$

Collaborative Targeted Forward Selection (continued)

$k = 2$: Pscore model with one covariate

For each covariate in turn, W_x

- Calculate $\hat{\pi}^{k=2,W_x} = P(A \mid W_x)$
- Evaluate $h^{k=2,W_x} = \frac{A}{\pi^{k=2}} - \frac{(1-A)}{(1-\pi^{k=2})}$

Collaborative Targeted Forward Selection (continued)

$k = 2$: Pscore model with one covariate

For each covariate in turn, W_x

- Calculate $\hat{\pi}^{k=2,W_x} = P(A \mid W_x)$
- Evaluate $h^{k=2,W_x} = \frac{A}{\pi^{k=2}} - \frac{(1-A)}{(1-\pi^{k=2})}$
- Fit $\epsilon^{k=2,W_x}$ by regressing $(Y - \hat{Y}^{k-1})$ on $h^{k=2,W_x}$

Collaborative Targeted Forward Selection (continued)

$k = 2$: Pscore model with one covariate

For each covariate in turn, W_x

- Calculate $\hat{\pi}^{k=2,W_x} = P(A \mid W_x)$
- Evaluate $h^{k=2,W_x} = \frac{A}{\pi^{k=2}} - \frac{(1-A)}{(1-\pi^{k=2})}$
- Fit $\epsilon^{k=2,W_x}$ by regressing $(Y - \hat{Y}^{k-1})$ on $h^{k=2,W_x}$
- Obtain targeted predicted values: $\hat{Y}^{k=2,W_x} = \hat{Y}^{k=1} + \hat{\epsilon} h^{k=2,W_x}$

Collaborative Targeted Forward Selection (continued)

$k = 2$: Pscore model with one covariate

For each covariate in turn, W_x

- Calculate $\hat{\pi}^{k=2,W_x} = P(A \mid W_x)$
- Evaluate $h^{k=2,W_x} = \frac{A}{\pi^{k=2}} - \frac{(1-A)}{(1-\pi^{k=2})}$
- Fit $\epsilon^{k=2,W_x}$ by regressing $(Y - \hat{Y}^{k-1})$ on $h^{k=2,W_x}$
- Obtain targeted predicted values: $\hat{Y}^{k=2,W_x} = \hat{Y}^{k=1} + \hat{\epsilon} h^{k=2,W_x}$
- Evaluate $\text{RSS}^{k=2,W_x} = \sum (Y - \hat{Y}^{k=2,W_x})^2$

Collaborative Targeted Forward Selection (continued)

$k = 2$: Pscore model with one covariate

For each covariate in turn, W_x

- Calculate $\hat{\pi}^{k=2,W_x} = P(A \mid W_x)$
- Evaluate $h^{k=2,W_x} = \frac{A}{\pi^{k=2}} - \frac{(1-A)}{(1-\pi^{k=2})}$
- Fit $\epsilon^{k=2,W_x}$ by regressing $(Y - \hat{Y}^{k-1})$ on $h^{k=2,W_x}$
- Obtain targeted predicted values: $\hat{Y}^{k=2,W_x} = \hat{Y}^{k=1} + \hat{\epsilon} h^{k=2,W_x}$
- Evaluate $\text{RSS}^{k=2,W_x} = \sum (Y - \hat{Y}^{k=2,W_x})^2$

Select W_x that minimizes RSS

$k = 3$: Pscore model with two covariates

.
.
.

Continue until all covariates have been incorporated

C-TMLE Algorithm

$$\{\hat{\pi}^{k=1}, \quad \ldots, \quad \hat{\pi}^{k=K}\}$$

$$\{\hat{Y}^{k=1}, \quad \ldots, \quad \hat{Y}^{k=K}\}$$

- Sequence of nested pscore models and corresponding candidate TMLEs
- V-fold cross-validation to select candidate TMLE that minimizes cross-validated RSS for targeted outcome

$$RSS_{cv}^{k} = \sum_{v \in V} \sum_{i \in v} (Y - \hat{Y}_i^k)^2$$

Remarks: Collaborative targeted forward selection

- Select covariate that most improves *targeted outcome regression fit* at each step, *not* best predictor of treatment
- Strongest confounder chosen first
- Covariates re-ordered at each step, k
- Goes beyond univariate associations
- Covariates not associated with outcome avoided
- Delays incorporation of a highly correlated covariate

Collaborative Double Robustness Theorem

- Information in pscore only has to account for *residual* confounding not addressed by initial OR
- Pscore model conditioning on fewer covariates can often
  - produce less extreme propensity scores
  - reduce variance in estimate of parameter of interest
- Multiple sufficient $\widetilde{W}$ for pscore model may exist, depending on initial OR and true joint distribution of the data

Any DR estimator that solves $P_n D^*(P) = 0$ *can (at times) have more than two chances to get it right!*

Biomarker Discovery using C-TMLE

Goal: Identify HIV mutations affecting response to *lopinavir*

Data: 401 observations $O = (Y, A, W)$ on 372 subjects

Y = Change in $\log_{10}$(viral load)
between baseline and post-treatment follow-up*

A = Mutation indicator (binary)

W = Potential confounders
51 baseline characteristics, treatment history, 25 other mutations

Challenge: High correlations among some of the covariates and/or low probability of observing a given mutation make it difficult to obtain stable, low variance estimates of the association between A and Y

Data courtesy of Robert Shafer, Stanford University; Bembom, Petersen, & van der Laan, UC Berkeley

Data Analysis: Effect of HIV Mutation on resistance to lopinavir

Mutation	Score	Estimate	Mutation	Score	Estimate
p50V	20	1.70*	p53LY	3	0.21
p82AFST	20	0.39*	p73CSTA	2	0.64*
p54VA	11	0.51*	p24IF	2	0.23
p54LMST	11	0.37*	p10FIRVY	2	−0.27
p84AV	11	0.10	p71TVI	2	0.02
p46ILV	11	0.05	p23I	0	0.82
p82MLC	10	1.61*	p36ILVTA	0	0.27
p47V	10	0.81*	p16E	0	0.24
p84C	10	0.60*	p20IMRTVL	0	0.18
p32I	10	0.54*	p63P	0	−0.13
p48VM	10	0.31	p88DTG	0	−0.43*
p90M	10	0.21	p30N	0	−0.44*
p33F	5	0.30	p88S	0	−0.47*

*95% CI excludes the NULL

Multicollinearity: C-TMLE correctly classifies most mutations

Stanford score from hivdb.stanford.edu accessed Sept, 2007,

Computational Challenges

Strategies to reduce computational expense

- "Embarrassingly parallel"
- Reduce high dimensional covariate space

  pre-screen, create summary measures
- Run computationally expensive pieces on subset of data

Summary

- *Targeted* confounder selection strategy for pscore estimation is guided by effect on $\hat{\psi}$
- C-TMLE
  - marriage of collaborative double robustness + TMLE
  - performs well in challenging scenarios
- Continued refinements, extensions, collaborations
  - Pscore estimation in other settings, e.g. MSM, non-DR
  - triple-robust TMLE
- Broad applicability
  - censoring, multiple time point interventions
  - CER, environmental epi, pharmacoepi, social epi, etc.

Acknowledgements

Miguel Hernán and Jamie Robins, Harvard School of Public Health

Mark van der Laan, UC Berkeley

NIH grants R01 AI074345-04 and R01 AI073127

R Core Development Team

References

S Gruber and MJ van der Laan, An Application of Collaborative Targeted Likelihood Estimation in Causal Inference and Genomics. *The International Journal of Biostatistics*, 6(1), 2010.

S. Gruber and M.J. van der Laan. Consistent Causal Effect Estimation Under Dual Misspecification and Implications for Confounder Selection Procedures. *Statistical Methods in Medical Research* [epub ahead of print February, 2012].

S Gruber and MJ van der Laan. tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software* 2012; 51(13).

HIV-CAUSAL Collaboration. The effect of combined antiretroviral therapy on overall mortality of HIV-infected individuals. *AIDS* 2010; 24, 123-137.

J Robins, M Hernán, B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11(5), 550560

J Robins, A Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer." *Statistica Sinica*, 2001; 11(4):920-36.

MJ van der Laan and S Gruber, Collaborative Double Robust Targeted Maximum Likelihood Estimation. *The International Journal of Biostatistics*, 6(1), 2010.

C-TMLE demo code: www.stat.berkeley.edu/∼laan/Software