Data-adaptive variable selection for causal inference

Susan M Shortreed

Group Health Research Institute
Department of Biostatistics, University of Washington

shortreed.s@ghc.org

joint work with Ashkan Ertefaie
Department of Biostatistics and Computational Biology
University of Rochester

Oct 25, 2016

Outline

- Causal inference in observational settings
  - Estimating unbiased treatment effects
- Goals of variable selection
  - Prediction versus causal inference
- Outcome-adaptive lasso
  - Simulation results
  - Opioid use and depressive symptoms
- Discussion

Causal inference in observational setting

Goal: Unbiased treatment effect estimation from observational data

- Subject of several methodological advancements
- Propensity scores methods commonly implemented
  - Especially helpful when many confounders

- Several different propensity score approaches
  - Stratification
  - Matching
  - Adjustment in outcome model
  - Inverse probability weighting

Propensity scores as balancing scores

- Propensity score methods for causal inference in observational settings rely on the propensity score as a balancing score
- A balancing score is a summary measure of covariates
- At each level of balancing score, exposed and unexposed individuals can be compared directly
  - Rosenbaum, Rubin. *The central role of the propensity score in observational studies for causal effects*. Biometrika. 1983;70(1):41-55.

Propensity score & causal inference

- Propensity score: probability of exposure given covariates (assume binary exposure)
  - $p(A = 1|X_1, X_2, ..., X_{d_0})$
- Some assumptions required for propensity score to be a balancing score
  - No unmeasured confounders
  - Positivity
  - Stable unit value assumption
  - Consistency

Propensity score variable selection, key assumptions

1. No unmeasured confounding
   - $A \perp Y_A \mid X_1, ..., X_{d_0}$
   - All confounders of treatment effect measured and included in propensity score
2. Positivity
   - $0 < p(A = 1 | X_1, ..., X_{d_0}) < 1$

   Near-positivity violations: when propensity score very close to 0 or 1
   - Can result in really big weights

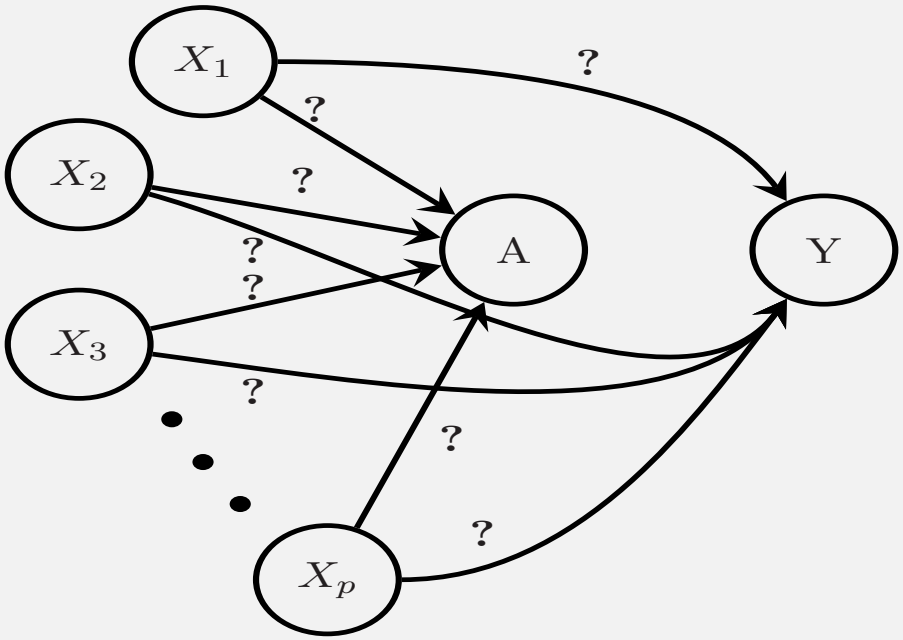
Propensity score variable selection

- Previously, 'throw-in-the-kitchen-sink' mentality
  - Concern excluding confounders leading to bias
- Literature shows statistical efficiency can be affected
  - Including variables related to exposure but not to the outcome can decrease precision
  - Both bias and precision important
- Ideal estimator is unbiased, while maintaining statistical efficiency

Schisterman, Cole, Platt (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 20(4):488-95.
Rotnitzky, Li, and Li (2010). A note on overadjustment in inverse probability weighted estimation. Biometrika 97(4):997-1001.
Patrick, Schneeweiss, Brookhart, Glynn, Rothman, Avorn, Stürmer (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. Pharmacoepi and Drug Safety 20(6):551-9.

Estimating propensity score

- Which covariates needed to account for confounding?
  - Often do not know all confounders
    - Use scientific knowledge
    - Limited to covariates available
  - Electronic health records contain vast amounts of data



Goal Use data to select variables to include in propensity score

Variable selection for prediction

- Some notation
  - Continuous-valued outcome: Y
  - Covariates: X_j, $j = 1:d$
  - $E[Y|x] = \beta_1^* x_1 + \ldots + \beta_d^* x_d$
    - Where $d_0 < d$ of $\beta_j^* \neq 0$
- Prediction variable selection goal:
  - Estimate a parsimonious model to **predict** Y
  - Find and estimate $\beta_j^* \neq 0$

Adaptive lasso (for prediction)

Goal: Find and estimate $\beta_j^* \neq 0$

Optimize a weighted lasso equation:

$$\hat{\beta}(AL) = \underset{\beta}{\operatorname{argmin}} \left(\left\| \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} x_{i,j}\beta_j) \right\|^2 + \lambda_n \sum_{j=1}^{d} \hat{\omega}_j |\beta_j| \right)$$

$$\hat{\omega}_j = \frac{1}{|\hat{\beta}_j(ols)|^\gamma} \text{ such that } \gamma > 0$$

- Where $\hat{\beta}_j(ols)$ is unpenalized least squares estimates
- Smaller $\hat{\beta}_j(ols)$ means $\hat{\beta}_j(AL)$ penalized more
  - i.e. shrunk to 0
- Sparcity and consistency guarantees
  - Select λ_n appropriately as a function of n

Zou (2006) The adaptive lasso and its oracle properties.
J. Am Stat Assoc, 101(476):1418-29

Adaptive penalized likelihood - logistic

Goal: Estimate parsimonious relationship for A given $\mathbf{X}$

- A binary exposure
- X_i vector of d covariates for individual i

$$\hat{\eta}(AL) = \underset{\eta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(-a_i(\mathbf{x_i}^T \eta) + \log(1 + \exp^{\mathbf{x}_i^T \eta}) \right) + \lambda_n \sum_{j=1}^{p} \hat{\omega}_j |\eta_j| \right)$$

$$\hat{\omega}_j = \frac{1}{|\hat{\eta}_j(mle)|^\gamma} \text{ such that } \gamma > 0$$

- Where $\hat{\eta}_j(mle)$ is unpenalized MLE
- Same properties as linear adaptive lasso
  - Smaller $\hat{\eta}_j(mle)$ means $\hat{\eta}_j(AL)$ shrunk closer to 0
- Use to select variables for propensity score?

Variable selection for causal inference, some notation

- Continuous-valued outcome: Y
- Binary exposure: A
- Covariates: X_j, $j = 1:d$
  - Select $d_0 < d$ covariates to include in propensity score
  - Estimate propensity score using reduced model
- Estimate average treatment effect
  - Inverse probability weighted estimator

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{w}_i Y_i A_i}{\sum_{i=1}^{n} \hat{w}_i} - \frac{\sum_{i=1}^{n} \hat{w}_i Y_i (1 - A_i)}{\sum_{i=1}^{n} \hat{w}_i (1 - A_i)}$$

Variable selection for propensity score

- For unbiased treatment effect estimation

~~Goal~~: Parsimonious prediction model for exposure

Goal: Parsimonious balancing score to account for bias, while maintaining statistical efficiency

- Estimate propensity score to get a balancing score
  - Propensity score not simply predict exposure
- Which covariates include in propensity score model?
  - Need valid assumptions for causal inference
    - No unmeasured confounding and positivity

Variable selection for causal inference

Goal Select variables to include in propensity score

- Include all confounders
  - Ensure validity of no umeasured confounders
- Include predictors of outcome
  - Even if not related to exposure
  - Can improve precision
- Exclude variables that predict exposure, but not outcome
  - Can result in unnecessary near-violations to positivity assumption
  - Results in large weights and decreased precision
- Exclude spurious variables

Schisterman, Cole, Platt (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 20(4):488-95.
Rotnitzky, Li, and Li (2010). A note on overadjustment in inverse probability weighted estimation. Biometrika 97(4):997-1001.
Patrick, Schneeweiss, Brookhart, Glynn, Rothman, Avorn, Stürmer (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. Pharmacoepi and Drug Safety 20(6):551-9.

Outcome-adaptive lasso for causal inference

- Estimate propensity score for binary exposure, A
  - Include confounders and predictors of the outcome
  - Exclude predictors of exposure and spurious variables

$$\hat{\alpha}(OAL) = \underset{\alpha}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(-a_i(\mathbf{x_i}^T \alpha) + \log(1 + e^{\mathbf{x}_i^T \alpha}) \right) + \lambda_n \sum_{j=1}^{d} \hat{\omega}_j |\alpha_j| \right)$$

Define $\hat{\omega}_j = \frac{1}{|\hat{\beta}_j(ols)|^\gamma}$, where $\hat{\beta}_j(ols)$ is the estimate from:

$$\hat{\beta}(ols) = \underset{\beta}{\operatorname{argmin}} \left\| (\mathbf{y} - \beta_A \mathbf{a} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j) \right\|^2$$

Outcome-adaptive lasso for causal inference

- Estimate propensity score for binary exposure, A
  - Include confounders and predictors of the outcome
  - Exclude predictors of exposure and spurious variables

$$\hat{\alpha}(OAL) = \underset{\alpha}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(-a_i(\mathbf{x_i}^T \alpha) + \log(1 + e^{\mathbf{x}_i^T \alpha}) \right) + \lambda_n \sum_{j=1}^{d} \hat{\omega}_j |\alpha_j| \right)$$

Define $\hat{\omega}_j = \frac{1}{|\hat{\beta}_j(ols)|^\gamma}$, where $\hat{\beta}_j(ols)$ is the estimate from:

$$\hat{\beta}(ols) = \underset{\beta}{\operatorname{argmin}} \left\| (\mathbf{y} - \beta_A \mathbf{a} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j) \right\|^2$$

Outcome-adaptive lasso for causal inference

- Smaller $\hat{\beta}(ols)$ means $\hat{\alpha}(OAL)$ shrunk closer to 0
  - Spurious variables and variables that predict exposure, but not the outcome have small $\hat{\beta}(ols)$

$$\hat{\alpha}(OAL) = \underset{\alpha}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(-a_i(\mathbf{x_i}^T \alpha) + \log(1 + e^{\mathbf{x}_i^T \alpha}) \right) + \lambda_n \sum_{j=1}^{d} \hat{\omega}_j |\alpha_j| \right)$$

Define $\hat{\omega}_j = \frac{1}{|\hat{\beta}_j(ols)|^\gamma}$, where $\hat{\beta}_j(ols)$ is the estimate from:

$$\hat{\beta}(ols) = \underset{\beta}{\operatorname{argmin}} \left\| (\mathbf{y} - \beta_A \mathbf{a} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j) \right\|^2$$

Properties of outcome-adaptive lasso

- If certain criteria regarding mild regularity conditions, λ_n, and γ are met, outcome-adaptive lasso approach:
  - Includes confounders
  - Includes predictors of the outcome in finite samples
  - Excludes variables that predict exposure, but not outcome
  - Excludes spurious variables

Selecting λ_n

- Minimize weighted absolute mean distance
  - $\hat{w}_j^{\lambda_n}$ are weights estimated using λ_n
  - $\hat{\beta}_j(ols)$ are OLS estimates from outcome model

$$\text{wAMD}(\lambda_n) = \sum_{j=1}^{d} \left|\hat{\beta}_j(ols)\right| \left| \frac{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n} X_{ij} A_i}{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n} A_i} - \frac{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n} X_{ij}(1-A_i)}{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n}(1-A_i)} \right|,$$

- Large λ_n forces all propensity score coefficients to zero
- Small coefficients in propensity score may cause differences in covariate means b/w treatment groups
  - If X_j impacts outcome, increase values of wAMD
  - If X_j does not impact outcome, will not impact wAMD

Simulation set-up

- Continuous-valued outcome, Y, generated from $Y_i = \beta_a A + \sum_{j=1}^{d} \beta_j X_j + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$ and $\beta_a = 0$
- $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})$ generated from multivariate standard normal
- Binary exposure, A, generated from Bernouli with $\text{logit}[P(A=1)] = \left[\sum_{j=1}^{d} \nu_j X_j\right]$
- Investigated several scenarios varying magnitude of β_j and ν_j, sample size, n, and number of covariates, d.
  - Modeled simulations after those performed in Zigler, Dominici (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. J Am Stat Assoc, 109:95-107.

Simulation set-up

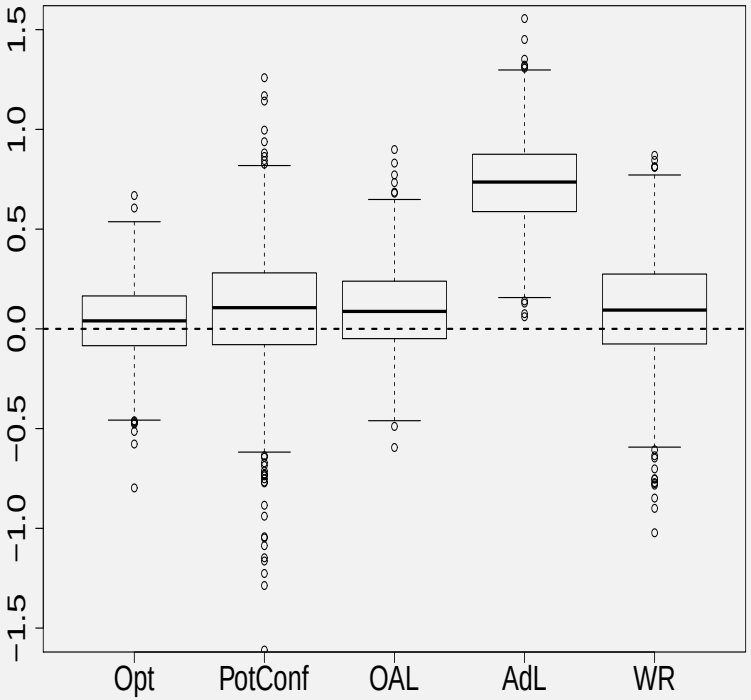
- $\lambda_n \in \{n^{-5}, n^{-1}, n^{-0.75}, n^{-0.5}, n^{-0.25}, n^{0.25}, n^{0.49}\}$
  - Select λ_n^{opt} using wAMD
- Select γ s.t. properties of outcome-adaptive lasso hold
- Perform 1,000 simulations and calculate

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n^{\text{opt}}} Y_i A_i}{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n^{\text{opt}}} A_i} - \frac{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n^{\text{opt}}} Y_i (1 - A_i)}{\sum_{i=1}^{n} \hat{w}_i^{\lambda_n^{\text{opt}}} (1 - A_i)}.$$

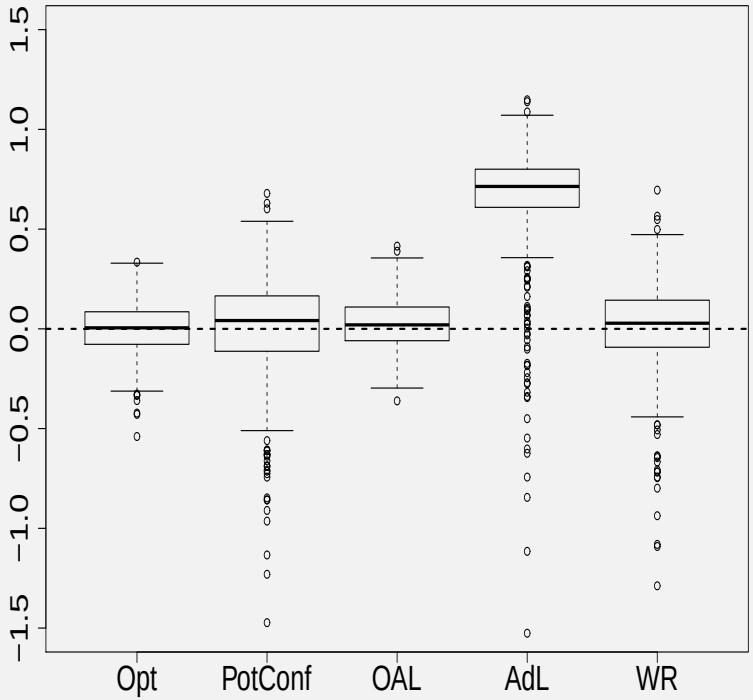
Simulations: large d, modest n

- outcome model: $\beta = (0.6, 0.6, 0.6, 0.6, 0, 0, 0, \ldots, 0)$
- exposure model: $\nu = (1, 1, 0, 0, 1, 1, 0, \ldots, 0)$

$n = 200, d = 100$



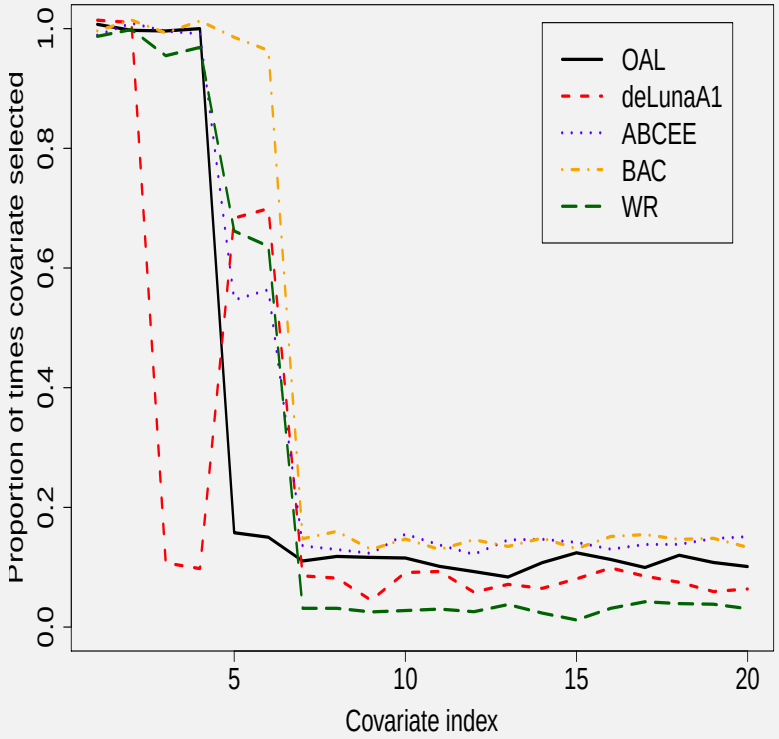
$n = 500, d = 200$



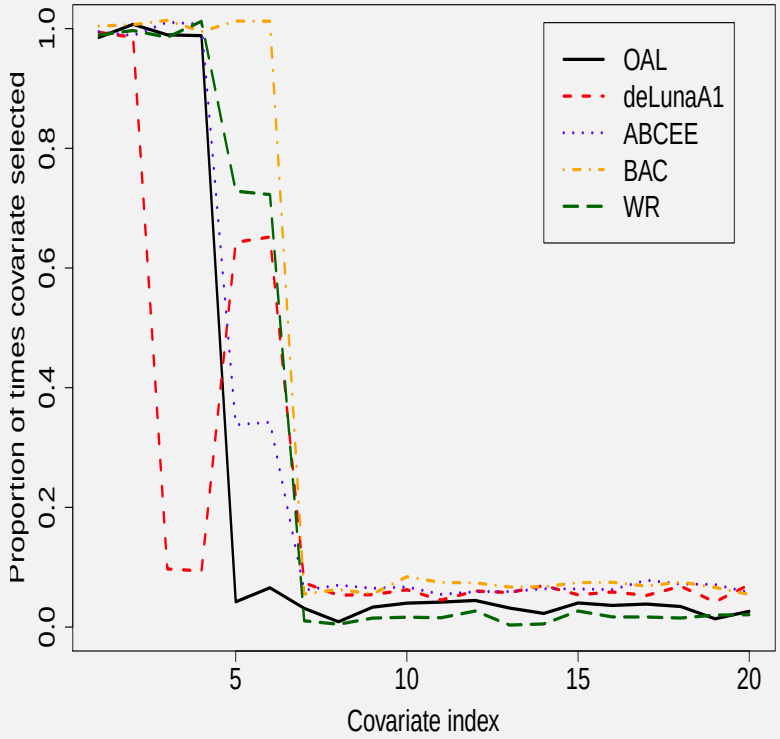
Simulations: modest d, vary n

- outcome model: $\beta = (0.6, 0.6, 0.6, 0.6, 0, 0, 0, \ldots, 0)$
- exposure model: $\nu = (1, 1, 0, 0, 1, 1, 0, \ldots, 0)$

$n = 200, d = 20$



$n = 1000, d = 20$



MASCOT study

- Some chronic pain patients take opioids long-term
- Some evidence opioids increase depressive symptoms
- MASCOT study of long-term opioid therapy patients
  - Middle-Aged/Seniors Chronic Opioid Therapy
  - Collected information from survey (self-report) and electronic medical records
- Depression symptoms measured by PHQ-8
  - Measured at baseline and 4 months
- Compare 4 month depressive symptoms in two exposure groups based on opioid use between baseline and 4 month follow-up
  - Lower dose and higher dose
- 37 covariates considered for propensity score

Opioids and depressive symptoms

Baseline covariates	Lower dose	Higher dose	% Selected
PHQ-8	7.1 (5.7)	8.2 (5.9)	100.0
Anxiety symptoms	1.6 (1.8)	1.7 (1.8)	84.3
# pain days (6 mo)	144.5 (53.7)	143.4 (53.2)	34.0
Pain scale	6.0 (2.3)	6.4 (2.0)	34.0

- 10,000 bootstrap replicates to calculate standard error and selection percentage
- PHQ-8 4 month scores in lower dose group 5.93 (sd=5.10); higher dose 6.79 (sd=5.79)
- IPTW estimate comparing lower and higher does group 0.13 (0.10,0.17)

Efron. (2014). Estimation and accuracy after model selection. J Am Stat Assoc 109:991-1007.

Discussion

- Variable selection for prediction and causal inference have different goals
  - Approaches from one setting may not directly apply to the other
- Outcome-adaptive lasso for causal inference
  - Good theoretical and empirical properties
  - Current approach designed for $d < n$
    - Future work to expand to settings with $d > n$ and with rare binary outcome
    - Efficient approaches for calculating accurate standard errors after model selection