Propensity Score Adapted Variable Selection for Causal Inference

Kangjie Zhou
Joint work with
Jinzhu Jia

August 16, 2021

Notation

- Let Y^T and Y^C denote potential outcomes, D denote treatment assignment.
- Let X denote covariates to be included in propensity score model, which may contain:
  - I: Instrumental variables that have no effect on Y, unless through D;
  - U: Confounders that are associated with both Y and D;
  - C: Outcome predictors that are are uncorrelated with D, but correlated with Y.

  Instrumental variable assumption: $I \perp (U, C)$.
- The propensity scores are denoted as:

$$p(X) = \mathbb{P}(D = 1|X).$$

Types of Covariates in Causal Inference

Y: outcome, D: treatment, $X = (I, U, C)$: covariates.

$$\begin{tikzcd} & & U \arrow[dl] \arrow[dr] & & \\ I \arrow[r] & D \arrow[rr] & & Y & C \arrow[l] \end{tikzcd}$$

Figure 1. The Directed Acyclic Graph for (Y, D, I, U, C).

- I: instrumental variables (treatment predictors)
- U: confounders
- C: outcome predictors

Goal of Covariate Selection

We aim to select covariates to be included in the propensity score model.

- Include all confounders:
  - Ensure validity of unconfoundedness assumption
  - Ensure consistency of estimators of ATE
- Include outcome predictors:
  - Improve statistical efficiency
- Exclude treatment predictors that are unassociated with outcome:
  - Can result in near-violation of overlap assumption
  - Can also decrease precision
- Exclude spurious variables (unrelated to both outcome and exposure)

Previous Results

- Outcome-Adaptive lasso (Shortreed et al., 2017)[2]: select covariates and estimate propensity scores simultaneously. Perform an adaptive lasso (Zou, 2006)[6] in logistic regression of exposure on covariates, using weights from coefficient estimates of relationship between outcome and covariates.
- Simultaneous penalization on regression models of propensity score and outcome (Ertefaie, 2018)[1].
- Bayesian methods, for examples, Wang et al. (2012)[4], Wilson and Reich (2014)[5], and Talbot et al. (2015)[3].

Outcome-Adaptive LASSO

Step 1: estimate outcome-covariates relationship.

$$(\tilde{\beta}, \tilde{\eta}) = \arg\min_{\beta,\eta} -L_n(\beta, \eta, Y, X, D). \tag{1}$$

Let $\hat{\omega}_j = |\tilde{\beta}_j|^{-\gamma}$, $\gamma > 1$, and assume a logit model for the PS.
Step 2: select covariates from propensity score model.

$$\hat{\alpha}(OAL) = \arg\min_{\alpha} \Big(\sum_{i=1}^{n} (-d_i(x_i^\top \alpha) + \log(1 + \exp(x_i^\top \alpha))) + \lambda_n \sum_{j=1}^{p} \hat{\omega}_j |\alpha_j|\Big).$$

Step 2 can estimate propensity scores simultaneously.

Outcome-Adaptive LASSO, continued

Theorem 1 of Shortreed et al. (2017)[2] demonstrated oracle properties of outcome-adaptive lasso under mild regularity conditions and

$$\frac{\lambda_n}{\sqrt{n}} \to 0, \ \lambda_n n^{\gamma/2-1} \to \infty. \tag{2}$$

To select λ_n, let $\hat{p}^{\lambda_n}(X)$ denote the estimated propensity scores, then choose λ_n to minimize the following weighted absolute mean difference (wAMD) between the exposure groups:

$$\text{wAMD}(\lambda_n) = \sum_{j=1}^{p} |\tilde{\beta}_j| \left| \frac{\sum_{i=1}^{n} \frac{D_i X_{ij}}{\hat{p}^{\lambda_n}(X_i)}}{\sum_{i=1}^{n} \frac{D_i}{\hat{p}^{\lambda_n}(X_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-D_i) X_{ij}}{1-\hat{p}^{\lambda_n}(X_i)}}{\sum_{i=1}^{n} \frac{1-D_i}{1-\hat{p}^{\lambda_n}(X_i)}} \right|, \tag{3}$$

then set $\gamma > 1$ in order to satisfy (2).

Outcome-Adaptive LASSO, continued

Simulation studies in Shortreed et al. (2017)[2] reveals its stability in both covariate selection and ATE estimation.

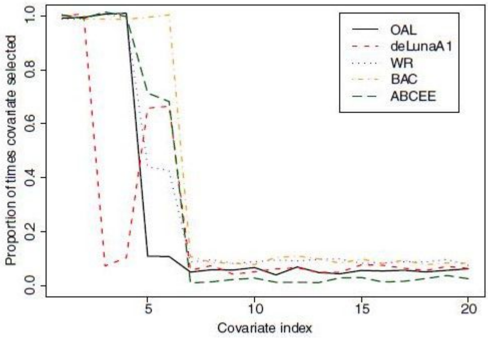

Figure 2. Frequency of covariates to be selected in propensity score model. $n = 500$, $p = 20$, covariate $1-4$: confounders and outcome predictors (Targ).

Outcome-Adaptive LASSO, continued

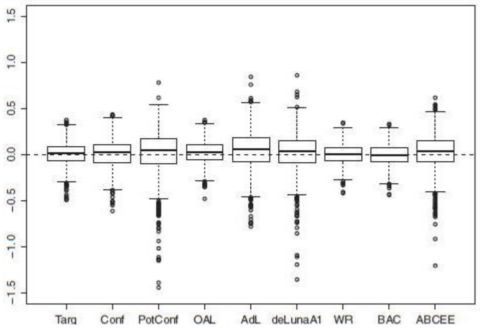

Figure 3. Box-plot of ATE estimates from different covariate selection methods. OAL, WR and BAC performs like Targ.

The key is covariate selection, not propensity score estimation!

Weaknesses of Outcome-Adaptive LASSO

- The outcome-adaptive lasso utilizes propensity score model to exclude instrumental variables, which are in fact included in the model, hence it requires more restrictive conditions on the tuning parameter than the adaptive lasso ($\lambda_n/\sqrt{n} \to 0,\ \lambda_n n^{(\gamma-1)/2} \to \infty$).
- The outcome-adaptive lasso must assume the correctness of both outcome model and propensity score model, we should propose a variable selection procedure that is robust to model misspecification.

We propose a Propensity Score Adapted Covariate Selection procedure (PACS) that enjoys robustness to outcome model misspecification, and performs better than OAL, especially when the sample size is large.

Propensity Score Adapted Covariate Selection

Step 1: $\sqrt{n}$-consistently estimate propensity scores from a generalized linear model, for examples, logit or probit. This step requires the propensity score model to be correctly specified.
Step 2: Adaptive lasso on weighted regression of Y on X in exposure groups, i.e., for the treatment group,

$$\hat{\beta}_T(PACS) = \arg\min_{\beta} \sum_{i \in T} \frac{1}{\hat{p}(X_i)} (Y_i - \eta - \beta^\top X_i)^2 + \lambda_n \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|,$$

where

$$\hat{\omega}_j = |\tilde{\beta}_j|^{-\gamma}, \ \gamma > 1, \tag{4}$$

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i \in T} \frac{1}{\hat{p}(X_i)} (Y_i - \eta - \beta^\top X_i)^2. \tag{5}$$

Similarly, we can define $\hat{\beta}_C(PACS)$ for the control group.

Linear Association Conditions

Let $X = (I, S, U, C)$, we aim to include (U, C) and exclude (I, S). Denote Σ the covariance matrix of (U, C), Σ_T the covariance matrix between (U, C) and Y^T, and Σ_C the covariance matrix between (U, C) and Y^C. We do not need the linear model to be correctly specified, but only requires the following linear association conditions:

Condition (A)

Linear association condition for the treatment group:

$$(\Sigma^{-1}\Sigma_T)_j \neq 0, \ \forall j. \tag{6}$$

Linear association condition for the control group:

$$(\Sigma^{-1}\Sigma_C)_j \neq 0, \ \forall j. \tag{7}$$

Oracle Properties

Theorem (1)

Suppose $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$, for $\gamma > 0$, then under mild regularity conditions, if linear association condition for potential outcome Y^T (Condition 1) holds, then

1. $\lim_{n\to\infty} \mathrm{P}(\hat{\beta}^T_{PACS,j} \neq 0, \ \forall j \in \mathcal{A} = \mathcal{U} \cup \mathcal{C}) = 1.$
2. $\lim_{n\to\infty} \mathrm{P}(\hat{\beta}^T_{PACS,j} = 0, \ \forall j \in \mathcal{A}^c = \mathcal{I} \cup \mathcal{S}) = 1.$
3. *The limiting distribution of $\sqrt{n}(\hat{\beta}^T_{PACS} - \beta^{T*})$ is normal.*

If linear association condition for potential outcome Y^C (Condition 2) holds, then

1. $\lim_{n\to\infty} \mathrm{P}(\hat{\beta}^C_{PACS,j} \neq 0, \ \forall j \in \mathcal{A} = \mathcal{U} \cup \mathcal{C}) = 1.$
2. $\lim_{n\to\infty} \mathrm{P}(\hat{\beta}^C_{PACS,j} = 0, \ \forall j \in \mathcal{A}^c = \mathcal{I} \cup \mathcal{S}) = 1.$
3. *The limiting distribution of $\sqrt{n}(\hat{\beta}^C_{PACS} - \beta^{C*})$ is normal.*

Implement the PACS

If condition (A), i.e. linear association conditions for both treatment group and control group hold, then both $\hat{\beta}_T(PACS)$ and $\hat{\beta}_C(PACS)$ enjoy oracle properties. Therefore, covariate j should be selected into the propensity score model if

$$\hat{\beta}_{T,j}(PACS)\hat{\beta}_{C,j}(PACS) \neq 0. \tag{8}$$

As for the choice of tuning parameters (λ_n, γ), we appeal to cross-validation. The PACS is implemented using LARS algorithm after reweighing on Y and X with inversed propensity scores. Then we use selected covariates to estimate propensity scores and calculate the IPW estimator:

$$\hat{\beta}_{IPW} = \frac{\sum_{i=1}^{n} \frac{D_i Y_i}{\hat{p}(X_i)}}{\sum_{i=1}^{n} \frac{D_i}{\hat{p}(X_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}}{\sum_{i=1}^{n} \frac{1-D_i}{1-\hat{p}(X_i)}}. \tag{9}$$

Simulation Studies

- Scenario 1. We first illustrate PACS's robustness to model misspecification, compared to OAL. We set $n = 500$ and $p = 20$, the first two covariates are confounders, covariates $3 - 4$ are outcome predictors, covariates $5 - 8$ are instrumental variables, others are spurious variables. Assuming a logit model for the propensity score, with coefficients α, and linear models for potential outcomes, with coefficients β_T and β_C, respectively.
- Scenario 2. If the linear model is right, we show that PACS is more powerful than OAL in excluding instrumental variables and spurious covariates. We set a linear outcome model with $\beta = (0.6, 0.6, 0.6, 0.6, 0, \cdots, 0)^\top$, and $\eta = 0$.
- We perform $m = 200$ recurrent trials in both scenarios.

References I

[1] Ashkan Ertefaie, Masoud Asgharian, and David A Stephens. Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, 6(1), 2018.

[2] Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

[3] Denis Talbot, Geneviève Lefebvre, and Juli Atherton. The bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2):207–236, 2015.

[4] Chi Wang, Giovanni Parmigiani, and Francesca Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–671, 2012.

References II

[5] Ander Wilson and Brian J Reich. Confounder selection via penalized credible regions. *Biometrics*, 70(4):852–861, 2014.

[6] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.