# High Dimensional Propensity Score Estimation via Covariate Balancing

Kosuke Imai

Princeton University

Talk at Columbia University

May 13, 2017

Joint work with Yang Ning and Sida Peng

# Motivation

- Covariate adjustment in observational studies:
  1. Identify and measure a set of potential confounders
  2. Weight or match the treated and control units to make them similar
- Practical questions:
  1. Which covariates should we adjust?
  2. How should we adjust them?
  3. What should we do if we have many potential confounders?
- Major advances over the last several years in the literature:
  Tan, Hainmueller, Graham et al., Zubizarreta, Belloni et al., Chan et al., Farrell, Chernozhukov et al., Athey et al., Zhao, etc.
- Covariate Balancing Propensity Score (CBPS)
  - Imai & Ratkovic JRSSB; Fan et al.
  - Estimate propensity score such that covariates are balanced
  - Extensions: continuous treatment (Fong et al.), dynamic treatment (Imai & Ratkovic JASA), high-dimensional setting (Ning et al.)

# Overview

1. Review of propensity score and CBPS
   - Inverse probability weighting (IPW)
   - Propensity score tautology $\rightsquigarrow$ covariate balancing
   - Optimal covariate balancing for CBPS

2. High-dimensional CBPS (HD-CBPS)
   - Impossible to balance all covariates $\rightsquigarrow$ trade-off between covariates
   - Sparsity assumption $+$ Regularization
   - Balance covariates that are predictive of outcome
   - Remove bias (due to regularization) by covariate balancing

# Propensity Score and CBPS

- $T_i \in \{0, 1\}$: binary treatment
- $\mathbf{X}_i$: $d$-dimensional pre-treatment covariates
- Identification assumptions: SUTVA, overlap, unconfoundedness

- Dual characteristics of propensity score:
  1. Predicts treatment assignment:

$$\pi(\mathbf{X}_i) \;=\; \Pr(T_i = 1 \mid \mathbf{X}_i)$$

  2. Balances covariates:

$$\mathbb{E}\left\{\frac{T_i}{\pi(\mathbf{X}_i)} f(\mathbf{X}_i)\right\} \;=\; \mathbb{E}\left\{\frac{1 - T_i}{1 - \pi(\mathbf{X}_i)} f(\mathbf{X}_i)\right\} \;=\; \mathbb{E}\{f(\mathbf{X}_i)\}$$

- Propensity score tautology
- CBPS: estimate propensity score such that balance is optimized
- Covariate balancing as estimating equations

# Optimal Covariate Balancing Equations

- Which covariates and what functions of them should we balance?
- The outcome model matters
- Outcome model: $\mathbb{E}(Y_i(t) \mid \mathbf{X}_i) = K_t(\mathbf{X}_i)$ for $t = 0, 1$
- Possibly misspecified propensity score $\pi_\beta(\mathbf{X}_i) \rightsquigarrow \pi_{\beta^o}(\mathbf{X}_i)$
- Inverse probability weighting (IPW) estimator:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\pi_{\hat{\beta}}(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi_{\hat{\beta}}(\mathbf{X}_i)}\right\}$$

- Asymptotic bias for the ATE:

$$\mathbb{E}\left[\left(\frac{T_i}{\pi_{\beta^o}(\mathbf{X}_i)} - \frac{1-T_i}{1-\pi_{\beta^o}(\mathbf{X}_i)}\right)\underbrace{\{\pi_{\beta^o}(\mathbf{X}_i)K_0(\mathbf{X}_i) + (1-\pi_{\beta^o})K_1(\mathbf{X}_i)\}}_{\text{optimal } f(\mathbf{X}_i)}\right]$$

$$= \mathbb{E}\left[\underbrace{\left(1 - \frac{1-T_i}{1-\pi_{\beta^o}(\mathbf{X}_i)}\right)}_{\text{balance control group}} K_0(\mathbf{X}_i) + \underbrace{\left(\frac{T_i}{\pi_{\beta^o}(\mathbf{X}_i)} - 1\right)}_{\text{balance treatment group}} K_1(\mathbf{X}_i)\right]$$

# Covariate Balancing in High-Dimension

- We focus on the estimation of $\mathbb{E}(Y_i(1))$
- A similar estimation strategy can be applied to $\mathbb{E}(Y_i(0))$
- Low dimension $\rightsquigarrow$ strong covariate balancing

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}_i}-1\right)\mathbf{X}_i \;=\; 0$$

  asymptotic normality, semiparametric efficiency, double robustness (Fan et al.)

- High dimension $\rightsquigarrow$ weak covariate balancing

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\hat{\pi}_i}-1\right)\boldsymbol{\alpha}^{*\top}\mathbf{X}_i \;=\; 0$$

# Assumptions for Estimating $\mu_1^* = \mathbb{E}(Y_i(1))$

1. Linear outcome model:

$$K_1(\mathbf{X}_i) \;=\; \boldsymbol{\alpha}^{*\top}\mathbf{X}_i$$

2. Logistic propensity score model:

$$\Pr(T_i = 1 \mid \mathbf{X}_i) \;=\; \pi(\boldsymbol{\beta}^{*\top}\mathbf{X}_i) \;=\; \frac{\exp(\boldsymbol{\beta}^{*\top}\mathbf{X}_i)}{1+\exp(\boldsymbol{\beta}^{*\top}\mathbf{X}_i)}$$

3. Sparsity of both models:

$$\max(s_1, s_2)\log d\sqrt{\log n/n} = o(1)$$

where $s_1 = \#\{\boldsymbol{\beta}_j^* > 0\}$ and $s_2 = \#\{\boldsymbol{\alpha}_j^* > 0\}$

4. Tuning parameters for Lasso:

$$\lambda = a\sqrt{\log(d)/n} \quad \text{and} \quad \lambda' = a'\sqrt{\log(d)/n}$$

for some unknown constants $a$ and $a'$

# The Proposed Methodology

1. Fit the penalized logistic regression model for propensity score:

$$\hat{\beta} \;=\; \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\{ T_i(\beta^\top \mathbf{X}_i) - \log(1 + \exp(\beta^\top \mathbf{X}_i)) \right\} + \lambda \|\beta\|_1,$$

2. Fit the penalized linear regression model for the outcome:

$$\widetilde{\alpha} \;=\; \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} T_i \{ Y_i - \alpha^\top \mathbf{X}_i \}^2 + \lambda' \|\alpha\|_1,$$

3. Calibrate the estimated propensity score by balancing covariates:

$$\widetilde{\gamma} \;=\; \underset{\gamma \in \mathbb{R}^{|\widetilde{S}|}}{\operatorname{argmin}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \frac{T_i}{\pi(\gamma^\top \mathbf{X}_{i\widetilde{S}} + \hat{\beta}_{\widetilde{S}^c}^\top \mathbf{X}_{i\widetilde{S}^c})} - 1 \right) \mathbf{X}_{i\widetilde{S}} \right\|_2^2$$

   where $\widetilde{S} = \{j : |\tilde{\alpha}_j| > 0\}$

4. Use the estimated propensity score $\tilde{\pi}_i = \pi(\widetilde{\gamma}^\top \mathbf{X}_{i\widetilde{S}} + \hat{\beta}_{\widetilde{S}^c}^\top \mathbf{X}_{i\widetilde{S}^c})$ for the IPW estimator $\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} T_i Y_i / \tilde{\pi}_i$

# Theoretical Properties

1. $\sqrt{n}$-consistency

$$\hat{\mu}_1 - \mu_1^* = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i}{\pi_i^*}(Y_i(1) - \boldsymbol{\alpha}^{*\top}\mathbf{X}_i) + \boldsymbol{\alpha}^{*\top}\mathbf{X}_i - \mu_1^*\right] + o_p(n^{-1/2})$$

2. asymptotic normality and semiparametric efficiency

$$\sqrt{n}(\hat{\mu}_1 - \mu_1^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\ \mathbb{E}\left[\frac{1}{\pi^*}\mathbb{E}(\epsilon_{i1}^2 \mid \mathbf{X}) + (\boldsymbol{\alpha}^{*\top}\mathbf{X}_i - \mu_1^*)^2\right]\right)$$

   $\rightsquigarrow$ valid under $K$-fold cross-validation

3. sample boundedness

$$\hat{\mu}_1 \geq \frac{\min_{i:T_i=1} Y_i}{n}\sum_{i=1}^{n}\frac{T_i}{\tilde{\pi}_i} = \min_{i:T_i=1} Y_i$$

   and similarly, $\hat{\mu}_1 \leq \max_{i:T_i=1} Y_i$

# Robustness and Generalizations

1. Multi-valued treatment regime
   - use the penalized multinomial logistic regression
   - balance selected covariates for each treatment group
2. Generalized linear models for the outcome
   - exponential family

$$p(Y \mid \mathbf{X}) = h(Y, \psi) \exp[\{Y \boldsymbol{\alpha}^{*\top} \mathbf{X} - b(\boldsymbol{\alpha}^{*\top} \mathbf{X})\}/a(\psi)]$$

   - balance $f(\mathbf{X}_i) = (b'(\widetilde{\boldsymbol{\alpha}}^\top \mathbf{X}_i), b''(\widetilde{\boldsymbol{\alpha}}^\top \mathbf{X}_i)\mathbf{X}_{i\widetilde{S}})$

$$b'(\boldsymbol{\alpha}^{*\top} \mathbf{X}_i) \;\approx\; b'(\widetilde{\boldsymbol{\alpha}}^\top \mathbf{X}_i) + b''(\widetilde{\boldsymbol{\alpha}}^\top \mathbf{X}_i)(\boldsymbol{\alpha}^* - \widetilde{\boldsymbol{\alpha}})\mathbf{X}_i^\top$$

3. Misspecified propensity score model
   - theoretical properties hold so long as the outcome model is correct
   - sure screening property must hold for $\widetilde{S}$ in the outcome model

# Comparison with Other Methods

1. Augmented inverse probability weighting (AIPW)
   - Belloni et al. (2014), Farrell (2015), Chernozhukov et al. (2016)
   - Extends Robins' doubly-robust estimator to high-dimension

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_1(\mathbf{X}_i) + \frac{T_i(Y_i - \hat{m}_1(\mathbf{X}_i))}{\hat{\pi}_i(\mathbf{X}_i)}\right\}$$

   - "Double selection" method

2. Residual balancing (RB)
   - Athey, Imbens, and Wagner (2016)
   - Assumes sparsity only for the outcome model
   - Critically relies on linearity assumption

# Simulation Studies

- Inspired by Kang and Schafer (2007)
- Covariates: $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ where $\Sigma_{jk} = \rho^{|j-k|}$
- Propensity score model:

$$\Pr(T_i \mid \mathbf{X}_i) = \text{logit}^{-1}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.1X_{i4})$$

- Outcome model:

$$Y_i(1) = 2 + 0.137X_{i7} + 0.137X_{i8} + 0.137X_{i9} + \epsilon_{1i}$$
$$Y_i(0) = 1 + 0.291X_{i5} + 0.291X_{i6} + 0.291X_{i7} + 0.291X_{i8} + 0.291X_{i9} + \epsilon_{0i}$$

  where $\epsilon_{ti} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for $t = 0, 1$
- Add irrelevant covariates so that a total number of covariates, $d$, varies from 10 to 2000
- Comparison with Residual Balancing (RB) and regularized augmented inverse probability weighting (AIPW)

# Both Models are Correct

- Standardized root-mean-squared error $\sqrt{\mathbb{E}(\hat{\mu} - \mu^*)^2}/\mu^*$

| | $n = 100$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|
| $d$ | HD-CBPS | RB | AIPW | HD-CBPS | RB | AIPW |
| 10 | **0.2163** | 0.2301 | 0.2209 | **0.0707** | 0.0720 | 0.0950 |
| 100 | **0.2272** | 0.2421 | 0.2273 | 0.0765 | 0.0787 | **0.0692** |
| 500 | 0.3262 | **0.3132** | 0.3859 | **0.0729** | 0.0810 | 0.1009 |
| 1000 | 0.2083 | 0.2413 | **0.2072** | **0.0817** | 0.0894 | 0.0992 |
| 2000 | 0.2327 | 0.2212 | **0.2058** | **0.0716** | 0.0760 | 0.0903 |

# Propensity Score Model is Misspecified

- Misspecification through transformation:
  $\mathbf{X}_{mis} = (\exp(X_1/2), X_2(1+\exp(X_1))^{-1} + 10, (X_1 X_3/25 + 0.6)^3, (X_2 + X_4 + 20)^2, X_5, \cdots, X_d)$

| | $n = 100$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|
| $d$ | HD-CBPS | RB | AIPW | HD-CBPS | RB | AIPW |
| 10 | 0.2230 | 0.2348 | **0.2174** | **0.0714** | 0.0734 | 0.0934 |
| 100 | 0.2163 | 0.2185 | **0.2104** | 0.0720 | 0.0784 | **0.0688** |
| 500 | **0.3255** | 0.3256 | 0.3742 | **0.0731** | 0.0852 | 0.1063 |
| 1000 | 0.2273 | 0.2462 | **0.2239** | **0.0892** | 0.0981 | 0.1074 |
| 2000 | 0.2232 | 0.2348 | **0.2061** | **0.0764** | 0.0844 | 0.0924 |

# Both Models are Misspecified

- Outcome model misspecification through another transformation: $\mathbf{X}_{mis} = (\exp(X_1/2), X_2(1+\exp(X_1))^{-1} + 10, (X_1 X_3/25 + 0.6)^3, (X_2 + X_4 + 20)^2, X_6, \exp(X_6 + X_7), X_9^2, X_7^3 - 20, X_9, \cdots, X_d)$

| | $n = 100$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|
| $d$ | HD-CBPS | RB | AIPW | HD-CBPS | RB | AIPW |
| 10 | **0.2386** | 0.2462 | 0.2614 | **0.0655** | 0.0661 | 0.0724 |
| 100 | **0.2385** | 0.2556 | 0.2422 | **0.0647** | 0.0680 | 0.0693 |
| 500 | **0.3324** | 0.3400 | 0.3696 | 0.0790 | 0.0861 | **0.0765** |
| 1000 | 0.2226 | 0.2369 | **0.2209** | **0.0725** | 0.0822 | 0.0765 |
| 2000 | 0.2108 | 0.2157 | **0.1974** | **0.0843** | 0.0878 | 0.0972 |

# Coverage of 95% Confidence Intervals

- Both models are correctly specified
- Average length of confidence intervals in parentheses

| $d$ | $n = 100$ | | $n = 500$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|
| 10 | 0.9450 | (0.8550) | 0.9500 | (0.3674) | 0.9550 | (0.2766) |
| 100 | 0.9000 | (0.7649) | 0.9450 | (0.3767) | 0.9700 | (0.2747) |
| 500 | 0.9000 | (0.8027) | 0.9300 | (0.3800) | 0.9650 | (0.2617) |
| 1000 | 0.9350 | (0.7268) | 0.9700 | (0.3718) | 0.9350 | (0.2697) |
| 2000 | 0.9050 | (0.7882) | 0.9450 | (0.4000) | 0.9350 | (0.2609) |

# Logistic Outcome Model

$$\Pr(Y_i(1) = 1 \mid \mathbf{X}_i)$$
$$= \text{logit}^{-1}(1 + \exp(2 + 0.137X_{i1} + 0.137X_{i2} + 0.137X_{i3}))$$
$$\Pr(Y_i(0) = 1 \mid \mathbf{X}_i)$$
$$= \text{logit}^{-1}(1 + \exp(1 + 0.291X_{i1} + 0.291X_{i2} + 0.291X_{i3} + 0.291X_{i4} + 0.291X_{i5}))$$

| | $n = 100$ | | $n = 500$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|
| $d$ | HD-CBPS | AIPW | HD-CBPS | AIPW | HD-CBPS | AIPW |
| 10 | 0.0947 | **0.0908** | **0.0398** | 0.0441 | **0.0252** | 0.0297 |
| 100 | **0.0745** | 0.0759 | **0.0352** | 0.0367 | **0.0239** | 0.0252 |
| 500 | **0.1075** | 0.1082 | **0.0351** | 0.0354 | **0.0303** | 0.0367 |
| 1000 | **0.0729** | 0.0730 | **0.0350** | 0.0358 | **0.0294** | 0.0320 |
| 2000 | **0.2113** | 0.2144 | **0.0357** | 0.0378 | **0.0232** | 0.0255 |

# Empirical Illustration

- Effect of college attendance on political participation (Kam and Palmer 2008 *JOP*)
- Data: Youth-Parent Socialization Panel Study ($n = 1051$)
- Outcome: a total number of (up to 8) "participatory acts"
- Covariates: 81 pre-adult characteristics
- Propensity score: a total of 204 predictors
- The authors found little effect of college attendance

- Henderson and Chatfield (2011) and Mayer (2011) use genetic matching and find a positive and statistically significant effect

# Empirical Results

- Propensity score: logistic regression with 204 predictors
- Outcome model: binomial logistic regression with 204 predictors
- 27 covariates are selected by HD-CBPS

| | HD-CBPS | CBPS | AIPW |
|---|---|---|---|
| Overall (ATE) | 0.8293 | 1.0163 | 0.8796 |
| | (0.1247) | (0.2380) | (0.1043) |
| Overall (ATT) | 0.8439 | 1.1232 | |
| | (0.1420) | (0.3094) | |
| Whites (ATE) | 0.8445 | | 0.8977 |
| | (0.1279) | | (0.1089) |

# Concluding Remarks

- Covariate balancing as an efficient and robust way to estimate propensity score
- Challenges in high-dimension $=$ cannot balance all covariates
- High-dimensional covariate balancing propensity score (HD-CBPS)
  - sparsity assumption $+$ weak covariate balancing
  - efficient and robust to misspecification of propensity score model
  - more reliance on the outcome model specification
    $\rightsquigarrow$ loss of double-robustness
- HD-CBPS is implemented in the R package CBPS