Covariate Balancing Propensity Score

Kosuke Imai

Princeton University

Winter Conference in Statistics

Borgafjäll, Sweden

March 9, 2015

Joint work with Christian Fong, and Marc Ratkovic

Motivation and Overview

- Central role of propensity score in causal inference
  - Adjusting for observed confounding in observational studies
  - Generalizing experimental and instrumental variables estimates

- Propensity score tautology
  - sensitivity to model misspecification
  - adhoc specification searches

- Covariate Balancing Propensity Score (CBPS)
  - Estimate the propensity score such that covariates are balanced
  - Inverse probability weights for marginal structural models

- Three cases:
  1. Binary treatment
  2. Time-varying binary treatments in longitudinal settings
  3. Multi-valued and continuous treatments

Propensity Score

- Notation:
  - $T_i \in \{0, 1\}$: binary treatment
  - X_i: pre-treatment covariates
- Dual characteristics of propensity score:
  1. Predicts treatment assignment:
  $$\pi(X_i) \;=\; \Pr(T_i = 1 \mid X_i)$$
  2. Balances covariates (Rosenbaum and Rubin, 1983):
  $$T_i \perp\!\!\!\perp X_i \;\mid\; \pi(X_i)$$
- But, propensity score must be estimated (more on this later)

Use of Propensity Score for Causal Inference

- Matching
- Subclassification
- Weighting (Horvitz-Thompson):

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right\}$$

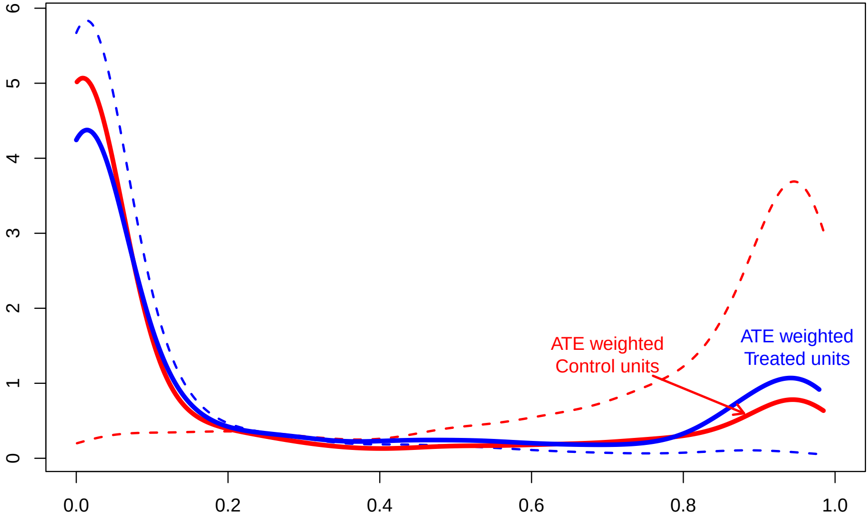
  where weights are often normalized
- Doubly-robust estimators (Robins *et al.*):

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\hat{\mu}(1,X_i) + \frac{T_i(Y_i-\hat{\mu}(1,X_i))}{\hat{\pi}(X_i)}\right\} - \left\{\hat{\mu}(0,X_i) + \frac{(1-T_i)(Y_i-\hat{\mu}(0,X_i))}{1-\hat{\pi}(X_i)}\right\}\right]$$

- They have become standard tools for applied researchers

Weighting to Balance Covariates

- Balancing condition: $\mathbb{E}\left\{\frac{T_i X_i}{\pi_\beta(X_i)} - \frac{(1-T_i)X_i}{1-\pi_\beta(X_i)}\right\} = 0$



Propensity Score Tautology

- Propensity score is unknown and must be estimated
  - Dimension reduction is purely theoretical: must model T_i given X_i
  - Diagnostics: covariate balance checking
- In theory: ellipsoidal covariate distributions
  $\Longrightarrow$ equal percent bias reduction
- In practice: skewed covariates and adhoc specification searches
- Propensity score methods are sensitive to model misspecification
- Tautology: propensity score methods only work when they work

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified

- 4 covariates X_i^*: all are *i.i.d.* standard normal
- Outcome model: linear model
- Propensity score model: logistic model with linear predictors
- Misspecification induced by measurement error:
  - $X_{i1} = \exp(X_{i1}^*/2)$
  - $X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)$
  - $X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3$
  - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$

Weighting Estimators Evaluated

1. Horvitz-Thompson (HT):

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right\}$$

2. Inverse-probability weighting with normalized weights (IPW): HT with normalized weights (Hirano, Imbens, and Ridder)
3. Weighted least squares regression (WLS): linear regression with HT weights
4. Doubly-robust least squares regression (DR): consistently estimates the ATE if *either* the outcome or propensity score model is correct (Robins, Rotnitzky, and Zhao)

Weighting Estimators Do Fine If the Model is Correct

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(1) Both models correct					
$n = 200$	HT	0.33	1.19	12.61	23.93
	IPW	−0.13	−0.13	3.98	5.03
	WLS	−0.04	−0.04	2.58	2.58
	DR	−0.04	−0.04	2.58	2.58
$n = 1000$	HT	0.01	−0.18	4.92	10.47
	IPW	0.01	−0.05	1.75	2.22
	WLS	0.01	0.01	1.14	1.14
	DR	0.01	0.01	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	−0.05	−0.14	14.39	24.28
	IPW	−0.13	−0.18	4.08	4.97
	WLS	0.04	0.04	2.51	2.51
	DR	0.04	0.04	2.51	2.51
$n = 1000$	HT	−0.02	0.29	4.85	10.62
	IPW	0.02	−0.03	1.75	2.27
	WLS	0.04	0.04	1.14	1.14
	DR	0.04	0.04	1.14	1.14

Weighting Estimators are Sensitive to Misspecification

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(3) Outcome model correct					
$n = 200$	HT	24.25	-0.18	194.58	23.24
	IPW	1.70	-0.26	9.75	4.93
	WLS	-2.29	0.41	4.03	3.31
	DR	-0.08	-0.10	2.67	2.58
$n = 1000$	HT	41.14	-0.23	238.14	10.42
	IPW	4.93	-0.02	11.44	2.21
	WLS	-2.94	0.20	3.29	1.47
	DR	0.02	0.01	1.89	1.13
(4) Both models incorrect					
$n = 200$	HT	30.32	-0.38	266.30	23.86
	IPW	1.93	-0.09	10.50	5.08
	WLS	-2.13	0.55	3.87	3.29
	DR	-7.46	0.37	50.30	3.74
$n = 1000$	HT	101.47	0.01	2371.18	10.53
	IPW	5.16	0.02	12.71	2.25
	WLS	-2.95	0.37	3.30	1.47
	DR	-48.66	0.08	1370.91	1.81

Smith and Todd (2005, *J. of Econometrics*)

- LaLonde (1986; *Amer. Econ. Rev.*):
  - Randomized evaluation of a job training program
  - Replace experimental control group with another non-treated group
  - Current Population Survey and Panel Study for Income Dynamics
  - Many evaluation estimators didn't recover experimental benchmark
- Dehejia and Wahba (1999; *J. of Amer. Stat. Assoc.*):
  - Apply propensity score matching
  - Estimates are close to the experimental benchmark
- Smith and Todd (2005):
  - Dehejia & Wahba (DW)'s results are sensitive to model specification
  - They are also sensitive to the selection of comparison sample

Propensity Score Matching Fails Miserably

- One of the most difficult scenarios identified by Smith and Todd:
  - LaLonde experimental sample rather than DW sample
  - Experimental estimate: \$886 (s.e. $=$ 488)
  - PSID sample rather than CPS sample
- Evaluation bias:
  - Conditional probability of being in the experimental sample
  - Comparison between experimental control group and PSID sample
  - "True" estimate $=0$
  - Logistic regression for propensity score
  - One-to-one nearest neighbor matching with replacement

Propensity score model	Estimates
Linear	-835
	(886)
Quadratic	-1620
	(1003)
Smith and Todd (2005)	-1910
	(1004)

Covariate Balancing Propensity Score (CBPS)

- Idea: Estimate propensity score such that covariates are balanced
- Goal: Robust estimation of parametric propensity score model

- Covariate balancing conditions:

$$\mathbb{E}\left\{\frac{T_i X_i}{\pi_\beta(X_i)} - \frac{(1-T_i)X_i}{1-\pi_\beta(X_i)}\right\} = 0$$

- Over-identification via score conditions:

$$\mathbb{E}\left\{\frac{T_i \pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)}\right\} = 0$$

- Can be interpreted as another covariate balancing condition

- Combine them with the Generalized Method of Moments

Revisiting Kang and Schafer (2007)

	Estimator	Bias GLM	Bias CBPS1	Bias CBPS2	Bias True	RMSE GLM	RMSE CBPS1	RMSE CBPS2	RMSE True
(1) Both models correct									
$n = 200$	HT	0.33	2.06	−4.74	1.19	12.61	4.68	9.33	23.93
	IPW	−0.13	0.05	−1.12	−0.13	3.98	3.22	3.50	5.03
	WLS	−0.04	−0.04	−0.04	−0.04	2.58	2.58	2.58	2.58
	DR	−0.04	−0.04	−0.04	−0.04	2.58	2.58	2.58	2.58
$n = 1000$	HT	0.01	0.44	−1.59	−0.18	4.92	1.76	4.18	10.47
	IPW	0.01	0.03	−0.32	−0.05	1.75	1.44	1.60	2.22
	WLS	0.01	0.01	0.01	0.01	1.14	1.14	1.14	1.14
	DR	0.01	0.01	0.01	0.01	1.14	1.14	1.14	1.14
(2) Propensity score model correct									
$n = 200$	HT	−0.05	1.99	−4.94	−0.14	14.39	4.57	9.39	24.28
	IPW	−0.13	0.02	−1.13	−0.18	4.08	3.22	3.55	4.97
	WLS	0.04	0.04	0.04	0.04	2.51	2.51	2.51	2.51
	DR	0.04	0.04	0.04	0.04	2.51	2.51	2.52	2.51
$n = 1000$	HT	−0.02	0.44	−1.67	0.29	4.85	1.77	4.22	10.62
	IPW	0.02	0.05	−0.31	−0.03	1.75	1.45	1.61	2.27
	WLS	0.04	0.04	0.04	0.04	1.14	1.14	1.14	1.14
	DR	0.04	0.04	0.04	0.04	1.14	1.14	1.14	1.14

CBPS Makes Weighting Methods Work Better

	Estimator	Bias GLM	Bias CBPS1	Bias CBPS2	Bias True	RMSE GLM	RMSE CBPS1	RMSE CBPS2	RMSE True
(3) Outcome model correct									
$n = 200$	HT	24.25	1.09	−5.42	−0.18	194.58	5.04	10.71	23.24
	IPW	1.70	−1.37	−2.84	−0.26	9.75	3.42	4.74	4.93
	WLS	−2.29	−2.37	−2.19	0.41	4.03	4.06	3.96	3.31
	DR	−0.08	−0.10	−0.10	−0.10	2.67	2.58	2.58	2.58
$n = 1000$	HT	41.14	−2.02	2.08	−0.23	238.14	2.97	6.65	10.42
	IPW	4.93	−1.39	−0.82	−0.02	11.44	2.01	2.26	2.21
	WLS	−2.94	−2.99	−2.95	0.20	3.29	3.37	3.33	1.47
	DR	0.02	0.01	0.01	0.01	1.89	1.13	1.13	1.13
(4) Both models incorrect									
$n = 200$	HT	30.32	1.27	−5.31	−0.38	266.30	5.20	10.62	23.86
	IPW	1.93	−1.26	−2.77	−0.09	10.50	3.37	4.67	5.08
	WLS	−2.13	−2.20	−2.04	0.55	3.87	3.91	3.81	3.29
	DR	−7.46	−2.59	−2.13	0.37	50.30	4.27	3.99	3.74
$n = 1000$	HT	101.47	−2.05	1.90	0.01	2371.18	3.02	6.75	10.53
	IPW	5.16	−1.44	−0.92	0.02	12.71	2.06	2.39	2.25
	WLS	−2.95	−3.01	−2.98	0.19	3.30	3.40	3.36	1.47
	DR	−48.66	−3.59	−3.79	0.08	1370.91	4.02	4.25	1.81

Causal Inference with Longitudinal Data

- Setup:
  - units: $i = 1, 2, \ldots, n$
  - time periods: $j = 1, 2, \ldots, J$
  - fixed J with $n \longrightarrow \infty$
  - time-varying binary treatments: $T_{ij} \in \{0, 1\}$
  - treatment history up to time j: $\overline{T}_{ij} = \{T_{i1}, T_{i2}, \ldots, T_{ij}\}$
  - time-varying confounders: X_{ij}
  - confounder history up to time j: $\overline{X}_{ij} = \{X_{i1}, X_{i2}, \ldots, X_{ij}\}$
  - outcome measured at time J: Y_i
  - potential outcomes: $Y_i(\bar{t}_J)$
- Assumptions:
  1. Sequential ignorability

     $$Y_i(\bar{t}_J) \perp\!\!\!\perp T_{ij} \mid \overline{T}_{i,j-1} = \bar{t}_{j-1}, \overline{X}_{ij} = \bar{x}_j$$

     where $\bar{t}_J = (\bar{t}_{j-1}, t_j, \ldots, t_J)$
  2. Common support

     $$0 < \Pr(T_{ij} = 1 \mid \overline{T}_{i,j-1}, \overline{X}_{ij}) < 1$$

Inverse-Probability-of-Treatment Weighting

- Weighting each observation via the inverse probability of its observed treatment sequence (Robins 1999)
- Inverse-Probability-of-Treatment Weights:

$$w_i = \frac{1}{P(\overline{T}_{iJ} \mid \overline{X}_{iJ})} = \prod_{j=1}^{J} \frac{1}{P(T_{ij} \mid \overline{T}_{i,j-1}, \overline{X}_{ij})}$$

- Stabilized weights:

$$w_i^* = \frac{P(\overline{T}_{iJ})}{P(\overline{T}_{iJ} \mid \overline{X}_{iJ})}$$

Marginal Structural Models (MSMs)

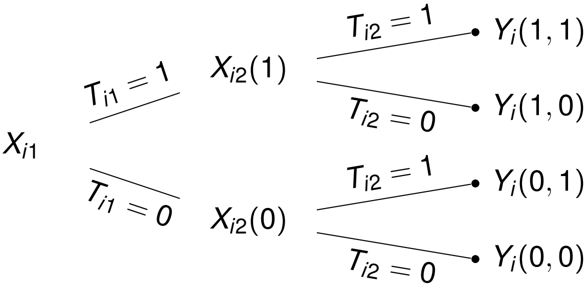
- Consistent estimation of the marginal mean of potential outcome:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\overline{T}_{iJ} = \overline{t}_J\} w_i Y_i \xrightarrow{p} \mathbb{E}(Y_i(\overline{t}_J))$$

- In practice, researchers fit a weighted regression of Y_i on a function of $\overline{T}_{iJ}$ with regression weight w_i
- Adjusting for $\overline{X}_{iJ}$ leads to post-treatment bias
- MSMs estimate the average effect of any treatment sequence

- **Problem:** MSMs are sensitive to the misspecification of treatment assignment model (typically a series of logistic regressions)
- The effect of misspecification can propagate across time periods
- **Solution:** estimate MSM weights so that covariates are balanced

Two Time Period Case



- time 1 covariates X_{i1}: 3 equality constraints

$$\mathbb{E}(X_{i1}) = \mathbb{E}[\mathbf{1}\{T_{i1} = t_1, T_{i2} = t_2\} w_i X_{i1}]$$

- time 2 covariates X_{i2}: 2 equality constraints

$$\mathbb{E}(X_{i2}(t_1)) = \mathbb{E}[\mathbf{1}\{T_{i1} = t_1, T_{i2} = t_2\} w_i X_{i2}(t_1)]$$

for $t_2 = 0, 1$

Orthogonalization of Covariate Balancing Conditions

	Treatment history: (t_1, t_2)				
Time period	(0,0)	(0,1)	(1,0)	(1,1)	Moment condition
time 1	$+$	$+$	$-$	$-$	$\mathbb{E}\left\{(-1)^{T_{i1}} w_i X_{i1}\right\} = 0$
	$+$	$-$	$+$	$-$	$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i1}\right\} = 0$
	$+$	$-$	$-$	$+$	$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i1}\right\} = 0$
time 2	$+$	$-$	$+$	$-$	$\mathbb{E}\left\{(-1)^{T_{i2}} w_i X_{i2}\right\} = 0$
	$+$	$-$	$-$	$+$	$\mathbb{E}\left\{(-1)^{T_{i1}+T_{i2}} w_i X_{i2}\right\} = 0$

GMM Estimator (Two Period Case)

- Independence across balancing conditions:

$$\hat{\beta} = \underset{\beta \in \Theta}{\text{argmin}} \ \mathrm{vec}(\mathbf{G})^\top \widehat{\mathbf{W}}^{-1} \mathrm{vec}(\mathbf{G})$$

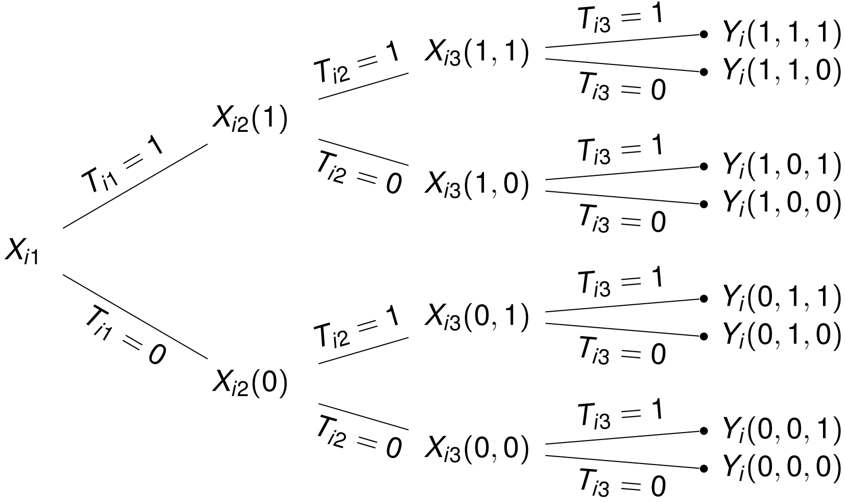
- Sample moment conditions **G**:

$$\frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} (-1)^{T_{i1}} w_i X_{i1} & (-1)^{T_{i2}} w_i X_{i1} & (-1)^{T_{i1}+T_{i2}} w_i X_{i1} \\ 0 & (-1)^{T_{i2}} w_i X_{i2} & (-1)^{T_{i1}+T_{i2}} w_i X_{i2} \end{bmatrix}$$

- Covariance matrix **W**:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left\{ \begin{bmatrix} 1 & (-1)^{T_{i1}+T_{i2}} & (-1)^{T_{i2}} \\ (-1)^{T_{i1}+T_{i2}} & 1 & (-1)^{T_{i1}} \\ (-1)^{T_{i2}} & (-1)^{T_{i1}} & 1 \end{bmatrix} \otimes w_i^2 \begin{bmatrix} X_{i1} X_{i1}^\top & X_{i1} X_{i2}^\top \\ X_{i2} X_{i1}^\top & X_{i2} X_{i2}^\top \end{bmatrix} \,\middle|\, \mathbf{X}_i \right\}$$

Extending Beyond Two Period Case



Generalization of the proposed method to J periods is in the paper

Orthogonalized Covariate Balancing Conditions

Design matrix			Treatment History Hadamard Matrix: (t_1, t_2, t_3)								Time		
			(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)			
T_{i1}	T_{i2}	T_{i3}	h_0	h_1	h_2	h_{12}	h_{13}	h_3	h_{23}	h_{123}	1	2	3
$-$	$-$	$-$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	✗	✗	✗
$+$	$-$	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$	✓	✗	✗
$-$	$+$	$-$	$+$	$+$	$-$	$-$	$+$	$+$	$-$	$-$	✓	✓	✗
$+$	$+$	$-$	$+$	$-$	$-$	$+$	$+$	$-$	$-$	$+$	✓	✓	✗
$-$	$-$	$+$	$+$	$+$	$+$	$+$	$-$	$-$	$-$	$-$	✓	✓	✓
$+$	$-$	$+$	$+$	$-$	$+$	$-$	$-$	$+$	$-$	$+$	✓	✓	✓
$-$	$+$	$+$	$+$	$+$	$-$	$-$	$-$	$-$	$+$	$+$	✓	✓	✓
$+$	$+$	$+$	$+$	$-$	$-$	$+$	$-$	$+$	$+$	$-$	✓	✓	✓

- The mod 2 discrete Fourier transform:

$$\mathbb{E}\{(-1)^{T_{i1}+T_{i3}} w_i X_{ij}\} = 0 \quad (\text{6th row})$$

- Connection to the fractional factorial design
  - "Fractional" = past treatment history
  - "Factorial" = future potential treatments

GMM in the General Case

- The same setup as before:

$$\hat{\beta} = \underset{\beta \in \Theta}{\operatorname{argmin}} \ \mathrm{vec}(\mathbf{G})^\top \widehat{\mathbf{W}}^{-1} \mathrm{vec}(\mathbf{G})$$

  where

$$\mathbf{G} = \frac{1}{n} \sum_{i=1}^{n} \left(M_i^\top \otimes w_i X_i \right) \mathbf{R}$$

$$\mathbf{W} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left(M_i M_i^\top \otimes w_i^2 X_i X_i^\top \mid X_i \right)$$

- M_i is the $(2^J - 1)$th row of *model matrix* based on the design matrix in Yates order
- For each time period j, define the *selection matrix* $\mathbf{R}$

$$\mathbf{R} = [\mathbf{R}_1 \ \ldots \mathbf{R}_J] \quad \text{where} \quad \mathbf{R}_j = \begin{bmatrix} \mathbf{0}_{2^{j-1} \times 2^{j-1}} & \mathbf{0}_{2^{j-1} \times (2^J - 2^{j-1})} \\ \mathbf{0}_{(2^J - 2^{j-1}) \times 2^{j-1}} & \mathbf{I}_{2^J - 2^{j-1}} \end{bmatrix}$$

Low-rank Approximation

- When the number of time periods J increases, the dimensionality of optimal $\mathbf{W}$, which is equal to $(2^J - 1) \times JK$, exponentially increases

- Low-rank approximation:

$$\widetilde{\mathbf{W}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I} \otimes \widetilde{X}_i \widetilde{X}_i^\top = \mathbf{I} \otimes \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$$

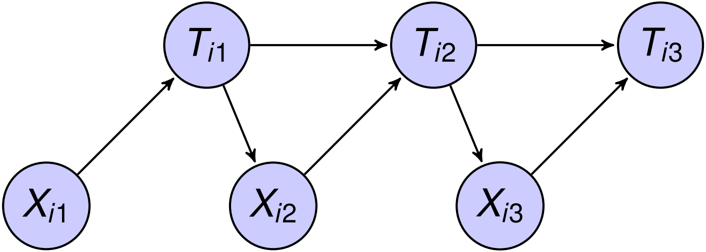
  where $\widetilde{X}_i = w_i X_i$

- Then,

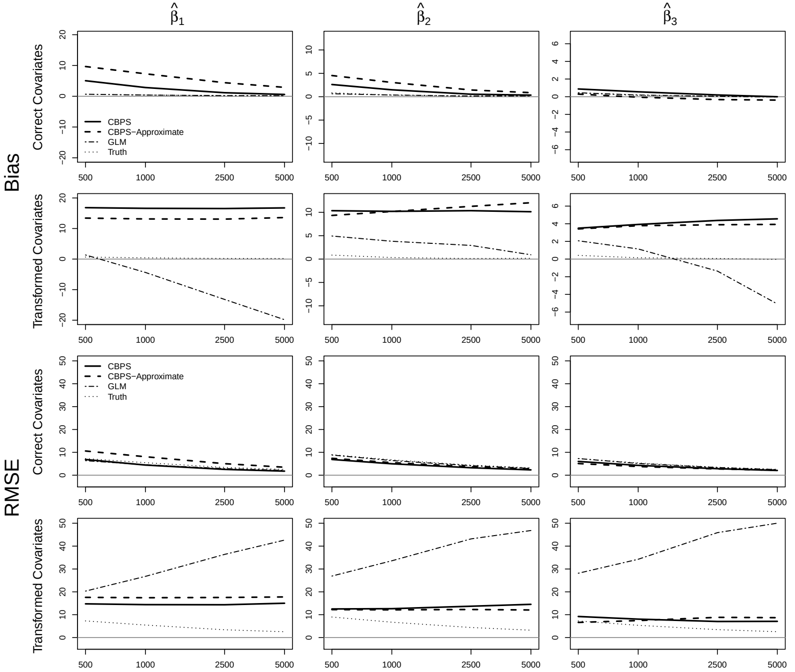
$$\begin{aligned}
\hat{\beta} &= \underset{\beta \in \Theta}{\operatorname{argmin}} \quad \operatorname{vec}(\mathbf{G})^\top \{\mathbf{I} \otimes \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\}^{-1} \operatorname{vec}(\mathbf{G}) \\
&= \underset{\beta \in \Theta}{\operatorname{argmin}} \quad \operatorname{trace}\{\mathbf{R}^\top \mathbf{M}^\top \widetilde{\mathbf{X}} (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \mathbf{M} \mathbf{R}\}
\end{aligned}$$

A Simulation Study with Correct Lag Structure

- 3 time periods
- Treatment assignment process:

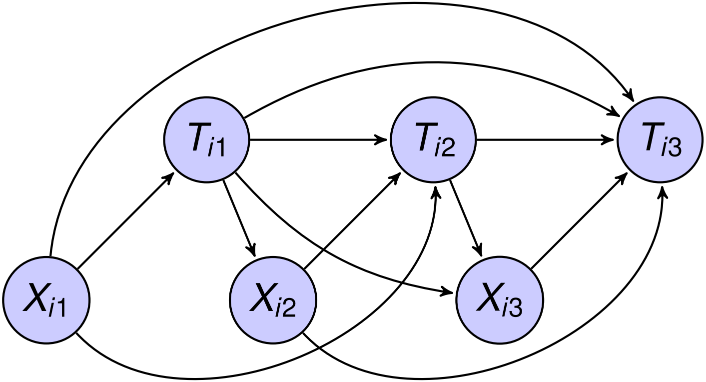


- Outcome: $Y_i = 250 - 10 \cdot \sum_{j=1}^{3} T_{ij} + \sum_{j=1}^{3} \delta^\top X_{ij} + \epsilon_i$
- Functional form misspecification by nonlinear transformation of X_{ij}

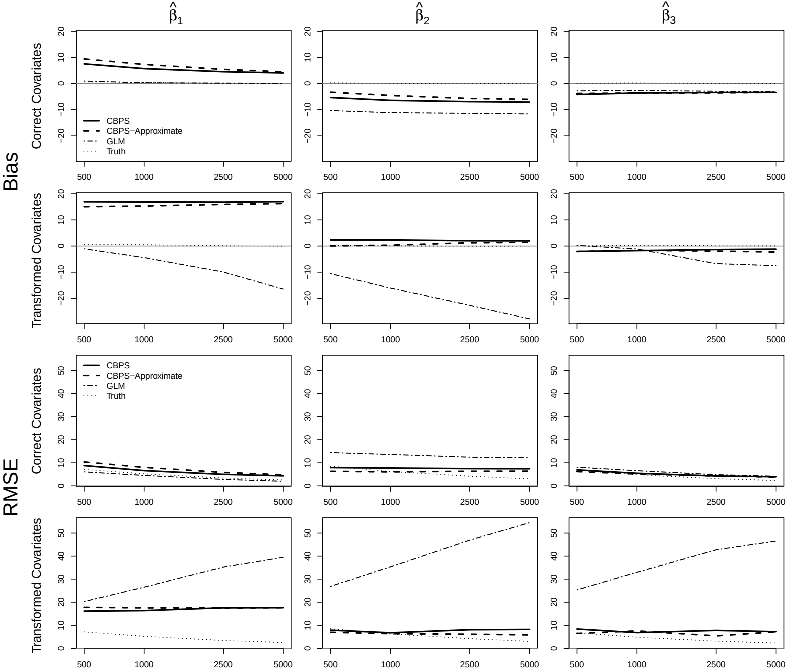

$\hat{\beta}_1$
$\hat{\beta}_2$
$\hat{\beta}_3$
Bias
RMSE
Correct Covariates
Transformed Covariates
CBPS
CBPS−Approximate
GLM
Truth
500
1000
2500
5000

A Simulation Study with Incorrect Lag Structure

- 3 time periods
- Treatment assignment process:



- The same outcome model
- Incorrect lag: only adjusts for previous lag but not all lags
- In addition, the same functional form misspecification of X_{ij}


$\hat{\beta}_1$
$\hat{\beta}_2$
$\hat{\beta}_3$
Bias
RMSE
Correct Covariates
Transformed Covariates
CBPS
CBPS−Approximate
GLM
Truth
500
1000
2500
5000

Empirical Illustration: Negative Advertisements

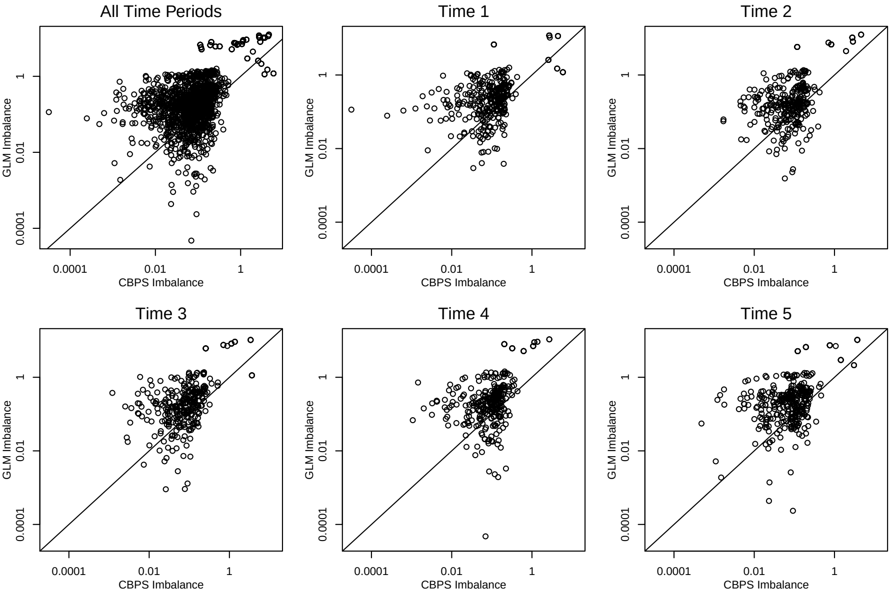
- Electoral impact of negative advertisements (Blackwell, 2013)
- For each of 114 races, 5 weeks leading up to the election

- Outcome: candidates' voteshare
- Treatment: negative ($T_{it} = 1$) or positive ($T_{it} = 0$) campaign

- Time-varying covariates: Democratic share of the polls, proportion of voters undecided, campaign length, and the lagged and twice lagged treatment variables for each week
- Time-invariant covariates: baseline Democratic voteshare, baseline proportion undecided, and indicators for election year, incumbency status, and type of office

- Original study: pooled logistic regression with a linear time trend
- We compare period-by-period GLM with CBPS

Covariate Balance

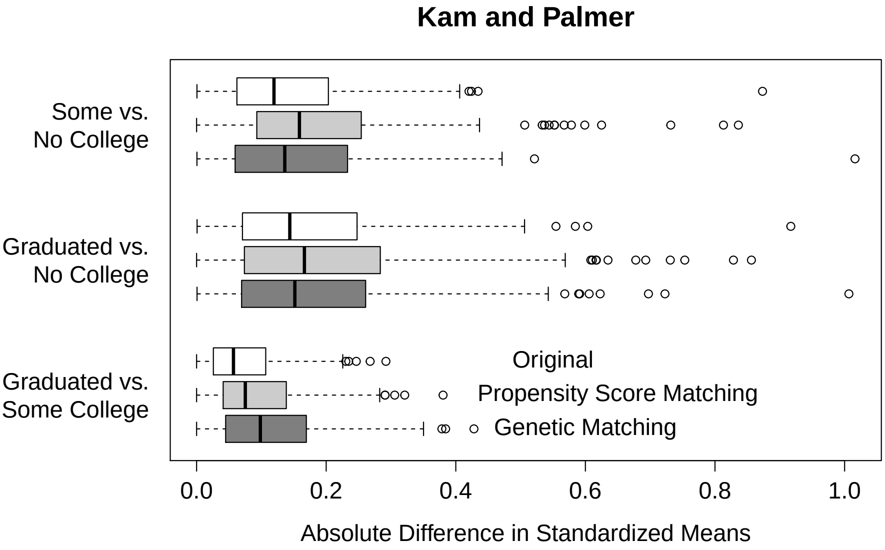


	GLM	CBPS	CBPS (approx.)	GLM	CBPS	CBPS (approx.)
(Intercept)	55.69*	57.15*	57.94*	55.41*	57.06*	57.73*
	(4.62)	(1.84)	(2.12)	(3.09)	(1.68)	(1.88)
Negative	2.97	5.82	3.15			
(time 1)	(4.55)	(5.30)	(3.76)			
Negative	3.53	2.71	5.02			
(time 2)	(9.71)	(9.26)	(8.55)			
Negative	−2.77	−3.89	−3.63			
(time 3)	(12.57)	(10.94)	(11.46)			
Negative	−8.28	−9.75	−10.39			
(time 4)	(10.29)	(7.79)	(8.79)			
Negative	−1.53	−1.95*	−2.13*			
(time 5)	(0.97)	(0.96)	(0.98)			
Negative				−1.14	−1.35*	−1.51*
(cumulative)				(0.68)	(0.39)	(0.43)
R^2	0.04	0.14	0.13	0.02	0.10	0.10
F statistics	0.95	3.39	3.32	2.84	12.29	12.23

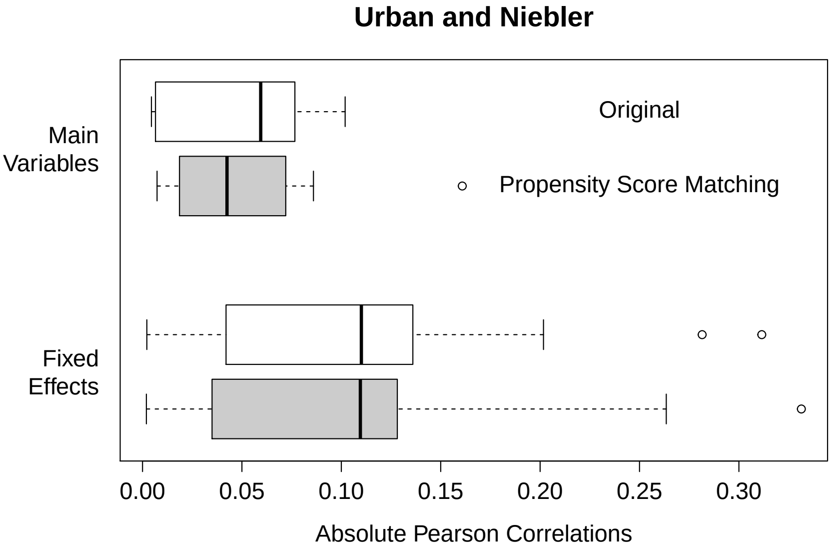
Two Motivating Examples for Multi-valued Treatments

1. Effect of education on political participation
    - Education is assumed to play a key role in political participation
    - T_i: 3 education levels (graduated from college, attended college but not graduated, no college)
    - Original analysis $\rightsquigarrow$ dichotomization (some college vs. no college)
    - Propensity score matching
    - Critics employ different matching methods
2. Effect of advertisements on campaign contributions
    - Do TV advertisements increase campaign contributions?
    - T_i: Number of advertisements aired in each zip code
    - ranges from 0 to 22,379 advertisements
    - Original analysis $\rightsquigarrow$ dichotomization (over 1000 vs. less than 1000)
    - Propensity score matching followed by linear regression with an original treatment variable

Balancing Covariates for a Dichotomized Treatment



May Not Balance Covariates for the Original Treatment



Propensity Score for a Multi-valued Treatment

- Consider a multi-valued treatment: $\mathcal{T} = \{0, 1, \ldots, J-1\}$
- Standard approach: MLE with multinomial logistic regression

$$\pi^j(X_i) \;=\; \Pr(T_i = j \mid X_i) \;=\; \frac{\exp\left(X_i^\top \beta_j\right)}{1 + \exp\left(\sum_{j'=1}^{J} X_i^\top \beta_{j'}\right)}$$

where $\beta_0 = 0$ and $\sum_{j=0}^{J-1} \pi^j(X_i) = 1$

- Covariate balancing conditions with inverse-probability weighting:

$$\mathbb{E}\left(\frac{\mathbf{1}\{T_i = 0\}X_i}{\pi_\beta^0(X_i)}\right) = \mathbb{E}\left(\frac{\mathbf{1}\{T_i = 1\}X_i}{\pi_\beta^1(X_i)}\right) = \cdots = \mathbb{E}\left(\frac{\mathbf{1}\{T_i = J-1\}X_i}{\pi_\beta^{J-1}(X_i)}\right)$$

which equals $\mathbb{E}(X_i)$

- Idea: estimate $\pi^j(X_i)$ to optimize the balancing conditions

CBPS for a Multi-valued Treatment

- Consider a 3 treatment value case as in our motivating example
- Sample balance conditions with orthogonalized contrasts:

$$\bar{g}_\beta(T, X) = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} 2\frac{\mathbf{1}\{T_i=0\}}{\pi_\beta^0(X_i)} - \frac{\mathbf{1}\{T_i=1\}}{\pi_\beta^1(X_i)} - \frac{\mathbf{1}\{T_i=2\}}{\pi_\beta^2(X_i)} \\ \frac{\mathbf{1}\{T_i=1\}}{\pi_\beta^1(X_i)} - \frac{\mathbf{1}\{T_i=2\}}{\pi_\beta^2(X_i)} \end{pmatrix} X_i$$

- Generalized method of moments (GMM) estimation:

$$\hat{\beta}_{\text{CBPS}} = \underset{\beta}{\operatorname{argmin}}\ \bar{g}_\beta(T, X)\, \Sigma_\beta(T, X)^{-1}\, \bar{g}_\beta(T, X)$$

where $\Sigma_\beta(T, X)$ is the covariance of sample moments

Score Conditions as Covariate Balancing Conditions

- Balancing the first derivative across treatment values:

$$
\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{N} s_\beta(T_i, X_i) \\
= \quad & \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \left(\frac{\mathbf{1}\{T_i=1\}}{\pi_\beta^1(X_i)} - \frac{\mathbf{1}\{T_i=0\}}{\pi_\beta^0(X_i)}\right) \frac{\partial}{\partial\beta_1}\pi_\beta^1(X_i) + \left(\frac{\mathbf{1}\{T_i=2\}}{\pi_\beta^2(X_i)} - \frac{\mathbf{1}\{T_i=0\}}{\pi_\beta^0(X_i)}\right) \frac{\partial}{\partial\beta_1}\pi_\beta^2(X_i) \\ \left(\frac{\mathbf{1}\{T_i=1\}}{\pi_\beta^1(X_i)} - \frac{\mathbf{1}\{T_i=0\}}{\pi_\beta^0(X_i)}\right) \frac{\partial}{\partial\beta_2}\pi_\beta^1(X_i) + \left(\frac{\mathbf{1}\{T_i=2\}}{\pi_\beta^2(X_i)} - \frac{\mathbf{1}\{T_i=0\}}{\pi_\beta^0(X_i)}\right) \frac{\partial}{\partial\beta_2}\pi_\beta^2(X_i) \end{pmatrix} \\
= \quad & \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \mathbf{1}\{T_i = 1\} - \pi_\beta^1(X_i) \\ \mathbf{1}\{T_i = 2\} - \pi_\beta^2(X_i) \end{pmatrix} X_i
\end{aligned}
$$

- Can be added to CBPS as over-identifying restrictions
- Generalizable to more treatment values

Propensity Score for a Continuous Treatment

- Standardize X_i and T_i such that
  - $\mathbb{E}(X_i^*) = \mathbb{E}(T_i^*) = \mathbb{E}(X_i^* T_i^*) = 0$
  - $\mathbb{V}(X_i) = \mathbb{V}(T_i) = 1$
- The stabilized weights:

$$w_i = \frac{f(T_i^*)}{f(T_i^* \mid X_i^*)}$$

- Covariate balancing condition:

$$\begin{aligned}\mathbb{E}(w_i T_i^* X_i^*) &= \int \left\{ \int \frac{f(T_i^*)}{f(T_i^* \mid X_i^*)} T_i^* dF(T_i^* \mid X_i^*) \right\} X_i^* dF(X_i^*) \\ &= \mathbb{E}(T_i^*)\mathbb{E}(X_i^*) = 0.\end{aligned}$$

- Again, estimate the generalized propensity score such that covariate balance is optimized

CBPS for a Continuous Treatment

- Standard approach (e.g., Robins *et al.* 2000):

$$T_i^* \mid X_i^* \overset{\text{indep.}}{\sim} \mathcal{N}(X_i^{*\top}\beta,\ \sigma^2)$$
$$T_i^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\ \sigma^2)$$

where further transformation of T_i can make these distributional assumptions more credible

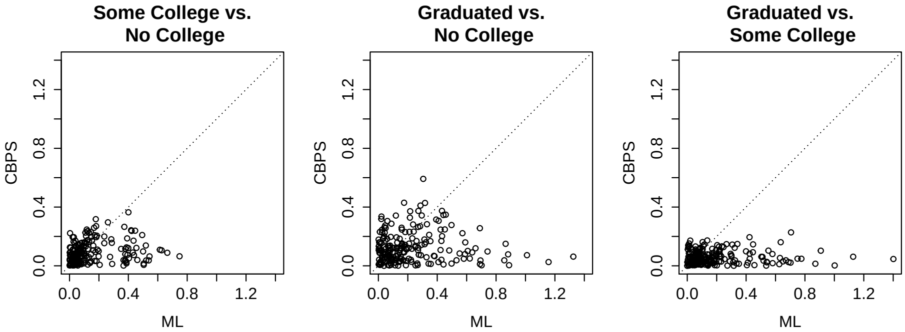
- Sample covariate balancing conditions:

$$\bar{g}_\theta(T, X) = \begin{pmatrix} \bar{s}_\theta(T, X) \\ \bar{w}_\theta(T, X) \end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N} \begin{pmatrix} \frac{1}{\sigma^2}(T_i^* - X_i^{*\top}\beta)X_i^* \\ -\frac{1}{2\sigma^2}\left\{1 - \frac{1}{\sigma^2}(T_i^* - X_i^{*\top}\beta)^2\right\} \\ \exp\left[\frac{1}{2\sigma^2}\left\{-2X_i^{*\top}\beta + (X_i^{*\top}\beta)^2\right\}\right] T_i^* X_i^* \end{pmatrix}$$

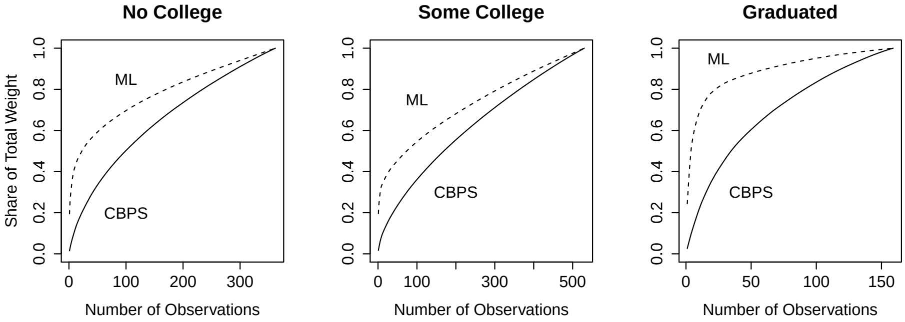
- GMM estimation: covariance matrix can be analytically calculated

Back to the Education Example: CBPS vs. ML

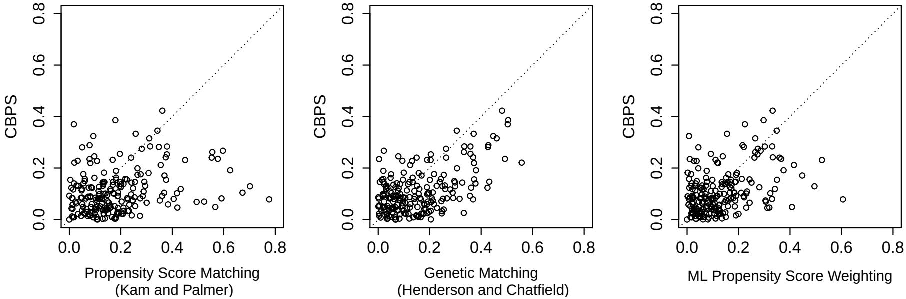
- CBPS achieves better covariate balance



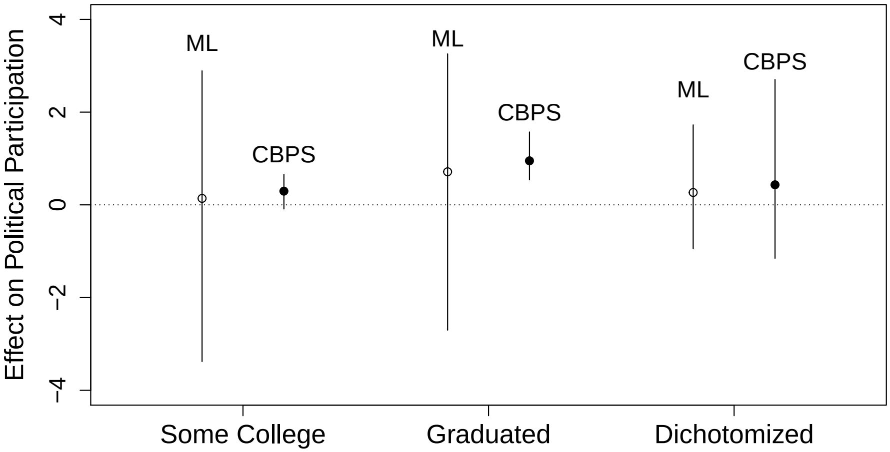
CBPS Avoids Extremely Large Weights



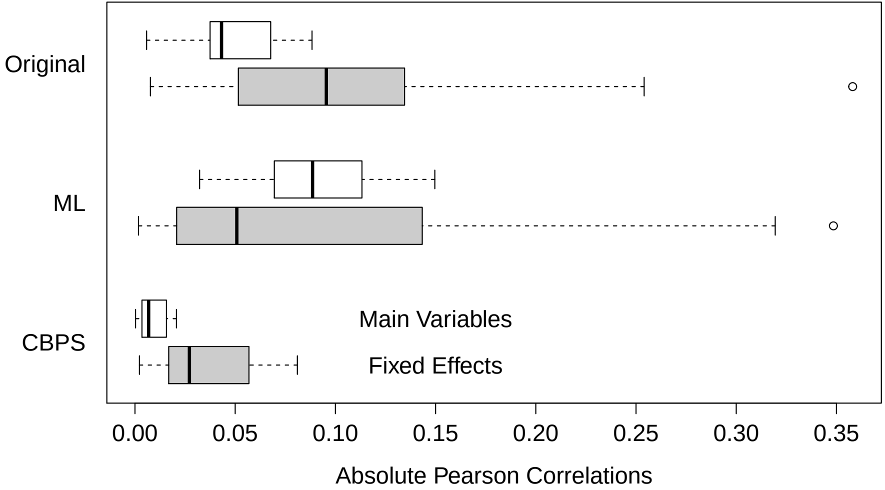
CBPS Balances Well for a Dichotomized Treatment



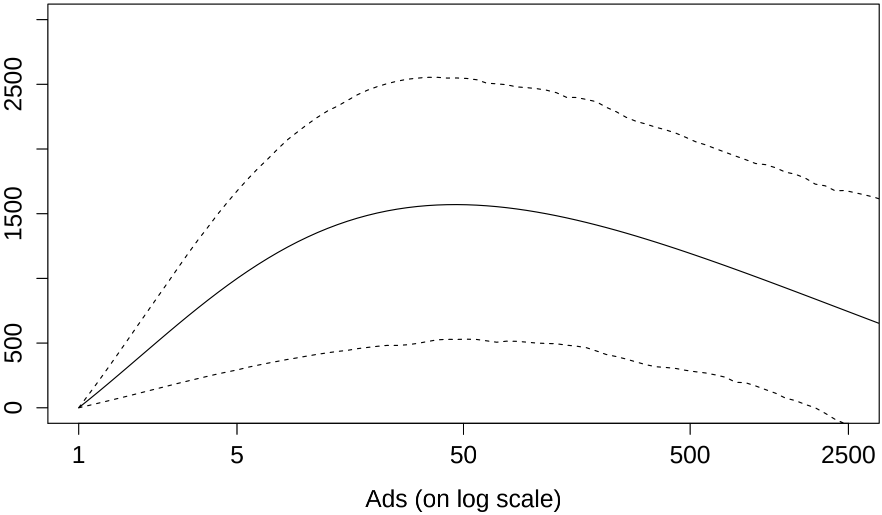
Empirical Results: Graduation Matters, Efficiency Gain



Onto the Advertisement Example



Empirical Finding: Some Effect of Advertisement



Concluding Remarks

- Covariate balancing propensity score:
  1. optimizes covariate balance under the GMM/EL framework
  2. is robust to model misspecification
  3. improves inverse probability weighting methods

- Ongoing work:
  1. Nonparametric estimation via empirical likelihood
  2. Generalizing experimental and instrumental variable estimates
  3. Confounder selection, moment selection

- Open-source software, `CBPS: R Package for Covariate Balancing Propensity Score`, is available at CRAN

References

1. "Covariate Balancing Propensity Score" *J. of the Royal Statistical Society, Series B (Methodological)*. (2014) Vol. 76, No. 1 (January), pp. 243–263.
2. "Robust Estimation of Inverse Probability Weights for Marginal Structural Models" *Journal of the American Statistical Association*, Forthcoming
3. "Covariate Balancing Propensity Score for General Treatment Regimes" Working paper available at **http://imai.princeton.edu**

Send comments and suggestions to kimai@Princeton.Edu