Covariate Balancing Propensity Score

Kosuke Imai

Princeton University

May 8, 2013

Joint work with Marc Ratkovic and Christian Fong

References

1. **Main Paper**: Imai, K. and Ratkovic, M. (2013). "Covariate Balancing Propensity Score" *Journal of the Royal Statistical Society, Series B (Methodological)*, Forthcoming.
2. **Software**: Ratkovic, M., K. Imai, C. Fong. (2013). *CBPS: R Package for Covariate Balancing Propensity Score* available for download at the CRAN

These and other related materials available at
`http://imai.princeton.edu`

Motivation

- Causal inference is a central goal of scientific research
- Randomized experiments are not always possible
  $\Longrightarrow$ Causal inference in observational studies
- Experiments often lack external validity
  $\Longrightarrow$ Need to generalize experimental results
- Importance of statistical methods to adjust for confounding factors

Overview of the Talk

1. **Review:** Propensity score
    - propensity score is a covariate balancing score
    - matching and weighting methods
2. **Problem:** Propensity score tautology
    - sensitivity to model misspecification
    - adhoc specification searches
3. **Solution: Covariate balancing propensity score (CBPS)**
    - Estimate propensity score so that covariate balance is optimized
4. **Evidence:** Reanalysis of two prominent critiques
    - Improved performance of propensity score weighting and matching
5. **Software:** R package `CBPS`
6. **Extension:** General Treatment Regimes

Propensity Score

- Setup:
  - $T_i \in \{0, 1\}$: binary treatment
  - X_i: pre-treatment covariates
  - $(Y_i(1), Y_i(0))$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcomes
- Definition: conditional probability of treatment assignment

$$\pi(X_i) \ = \ \Pr(T_i = 1 \mid X_i)$$

- Balancing property (without assumption):

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

Rosenbaum and Rubin (1983)

- Assumptions:
  1. Overlap:
  $$0 < \pi(X_i) < 1$$
  2. Unconfoundedness:
  $$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i$$

- Propensity score as a dimension reduction tool:
$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \pi(X_i)$$

Matching and Weighting via Propensity Score

- Propensity score reduces the dimension of covariates
- But, propensity score must be estimated (more on this later)
- Once estimated, simple nonparametric adjustments are possible

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

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

  often, weights are 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

Propensity Score Tautology

- Propensity score is unknown
- Dimension reduction is purely theoretical: must model T_i given X_i
- Diagnostics: covariate balance checking
- In practice, adhoc specification searches are conducted
- Model misspecification is always possible

- Theory (Rubin *et al.*): ellipsoidal covariate distributions
$\Longrightarrow$ equal percent bias reduction
- Skewed covariates are common in applied settings

- Propensity score methods can be sensitive to misspecification

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified
- Setup:
  - 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 to be evaluated:
  1. Horvitz-Thompson
  2. Inverse-probability weighting with normalized weights
  3. Weighted least squares regression
  4. Doubly-robust least squares regression

Weighting Estimators Do Fine If the Model is Correct

Sample size	Estimator	Bias GLM	Bias True	RMSE GLM	RMSE 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

- Idea: Estimate the propensity score such that covariate balance is optimized

- Covariate balancing condition:

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

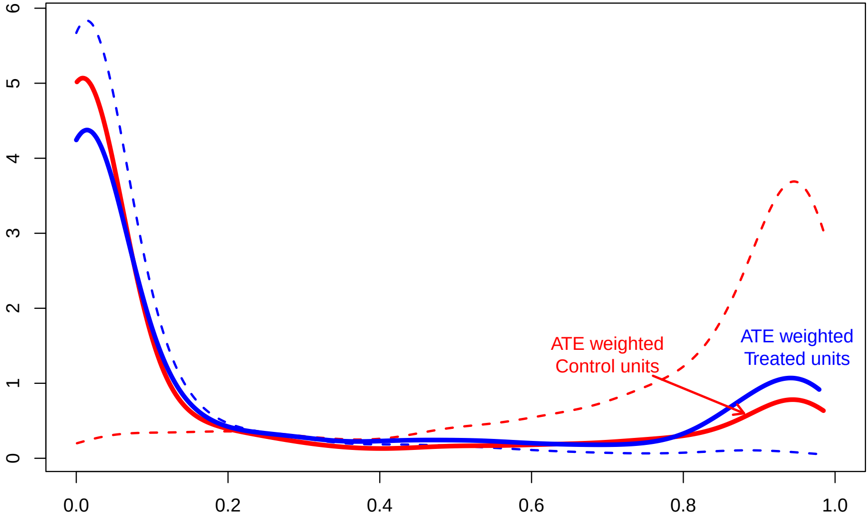
  where $\widetilde{X}_i = f(X_i)$ is any vector-valued function

- Score condition from maximum likelihood:

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

Weighting to Balance Covariates

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



Generalized Method of Moments (GMM) Framework

- Just-identified CBPS: covariate balancing conditions alone
- Over-identified CBPS: combine them with score conditions

- GMM (Hansen 1982):

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

where

$$\bar{g}_\beta(T, X) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\begin{pmatrix} \text{score condition} \\ \text{balancing condition} \end{pmatrix}}_{g_\beta(T_i, X_i)}$$

- "Continuous updating" GMM estimator for Σ

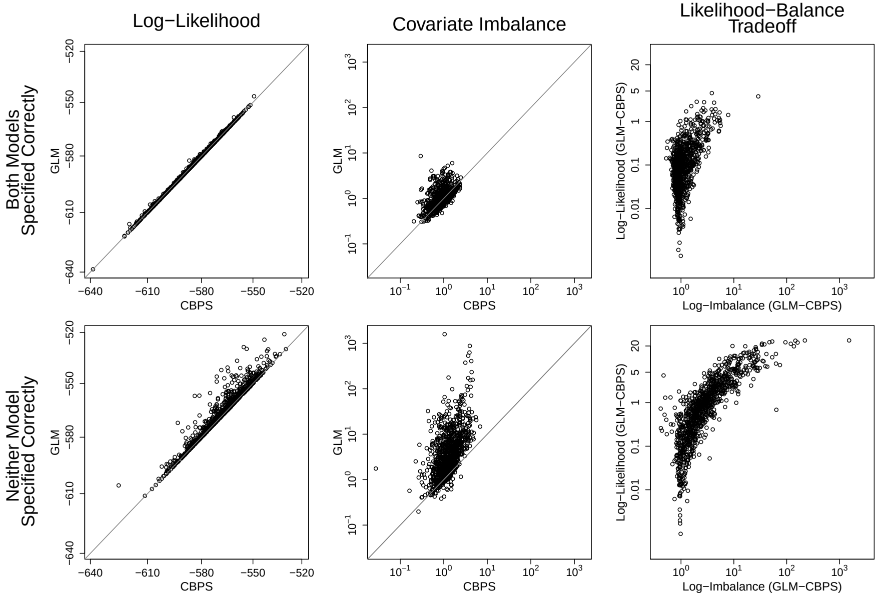
Revisiting Kang and Schafer (2007)

		Bias				RMSE			
	Estimator	GLM	CBPS1	CBPS2	True	GLM	CBPS1	CBPS2	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

CBPS Sacrifices Likelihood for Better Balance



Revisiting Smith and Todd (2005)

- Evaluation bias: "true" bias $= 0$
- CBPS improves propensity score matching across specifications and matching methods
- However, specification test rejects the null

	1-to-1 Nearest Neighbor			Optimal 1-to-N Nearest Neighbor		
Specification	GLM	CBPS1	CBPS2	GLM	CBPS1	CBPS2
Linear	−1209.15	−654.79	−505.15	−1209.15	−654.79	−130.84
	(1426.44)	(1247.55)	(1335.47)	(1426.44)	(1247.55)	(1335.47)
Quadratic	−1439.14	−955.30	−216.73	−1234.33	−175.92	−658.61
	(1299.05)	(1496.27)	(1285.28)	(1074.88)	(943.34)	(1041.47)
Smith & Todd	−1437.69	−820.89	−640.99	−1229.81	−826.53	−464.06
	(1256.84)	(1229.63)	(1757.09)	(1044.15)	(1179.73)	(1130.73)

Comparison with the Experimental Benchmark

- LaLonde, Dehejia and Wahba, and others did this comparison
- Experimental estimate: \$866 (s.e. $=$ 488)
- LaLonde$+$PSID pose a challenge: e.g., GenMatch -571 (1108)

	1-to-1 Nearest Neighbor			Optimal 1-to-N Nearest Neighbor		
Specification	GLM	CBPS1	CBPS2	GLM	CBPS1	CBPS2
Linear	-304.92	423.30	183.67	-211.07	423.30	138.20
	(1437.02)	(1295.19)	(1240.79)	(1201.49)	(1110.26)	(1161.91)
Quadratic	-922.16	239.46	1093.13	-715.54	307.51	185.57
	(1382.38)	(1284.13)	(1567.33)	(1145.82)	(1158.06)	(1247.99)
Smith & Todd	-734.49	-269.07	423.76	-439.54	-617.68	690.09
	(1424.57)	(1711.66)	(1404.15)	(1259.28)	(1438.86)	(1288.68)

Software: R Package CBPS

```
## upload the package
library("CBPS")
## load the LaLonde data
data(LaLonde)
## Estimate ATT weights via CBPS
fit <- CBPS(treat ~ age + educ + re75 + re74 +
                    I(re75==0) + I(re74==0),
            data = LaLonde, ATT = TRUE)
summary(fit)
## matching via MatchIt
library(MatchIt)
## one to one nearest neighbor with replacement
m.out <- matchit(treat ~ 1, distance = fitted(fit),
                 method = "nearest", data = LaLonde,
                 replace = TRUE)
summary(m.out)
```

Extensions to Other Causal Inference Settings

- Propensity score methods are widely applicable
- This means that CBPS is also widely applicable

- Non-binary treatment regimes
- Imai, K. and van Dyk, D. (2004). "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score" *Journal of the American Statistical Association*

- Challenge: many treatment groups $\Longrightarrow$ covariate balance checking is difficult
- Estimate the generalized propensity score such that covariate is balanced across *all* treatment groups

Multi-valued Categorical Treatment

- Propensity score for each value:

$$\pi_\beta(t, X_i) \;=\; \Pr(T_i = t \mid X_i)$$

- Commonly used model: multinomial logistic regression
- CBPS: balance covariates across all groups

$$\mathbb{E}\left\{\frac{\mathbf{1}\{T_i = t\}X_i}{\pi_\beta(t, X_i)}\right\} \;=\; \mathbb{E}\left\{\frac{\mathbf{1}\{T_i = t'\}X_i}{\pi_\beta(t', X_i)}\right\}$$

- Orthogonalize the conditions when the number of groups is 2^J
- Estimation of ATE: weighting or multi-dimensional matching/subclassification

Continuous and Other Treatments

- Generalized propensity score:

$$\pi_\beta(t, X_i) \; = \; p(T_i = t \mid X_i)$$

- Propensity function: $\psi_\beta(X_i)$ where $p_\psi(T_i = t \mid X_i)$
- Commonly used models: linear regression, GLMs

$$\pi_\beta(t, X_i) \; = \; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t - X_i^\top \beta)^2\right\}, \quad \psi_\beta(X_i) = X_i^\top \beta$$

- CBPS: balance covariates across discretized treatment categories

- Estimation of causal effects:
  - subclassification on propensity function (Imai and van Dyk)
  - subclassification on treatment (Zhao, van Dyk, and Imai)
  - smooth coefficient model (Zhao, van Dyk, and Imai)

Concluding Remarks

- Covariate balancing propensity score:
  1. simultaneously optimizes prediction of treatment assignment and covariate balance under the GMM framework
  2. is robust to model misspecification
  3. improves propensity score weighting and matching methods

- Extensions:
  1. Non-binary treatment regimes
  2. Dynamic treatment regimes in longitudinal analysis
  3. Generalizing experimental estimates
  4. Generalizing instrumental variable estimates
  5. Weighting methods for causal mediation analysis
  6. Model and confounder selection in a high-dimensional setting