Distribution-Free Assessment of Population Overlap in Observational Studies

Lihua Lei (Stanford), Alexander D'Amour (Google Brain), Peng Ding (UC Berkeley), Avi Feller (UC Berkeley), and Jasjeet Sekhon (Yale)

Population overlap in observational studies

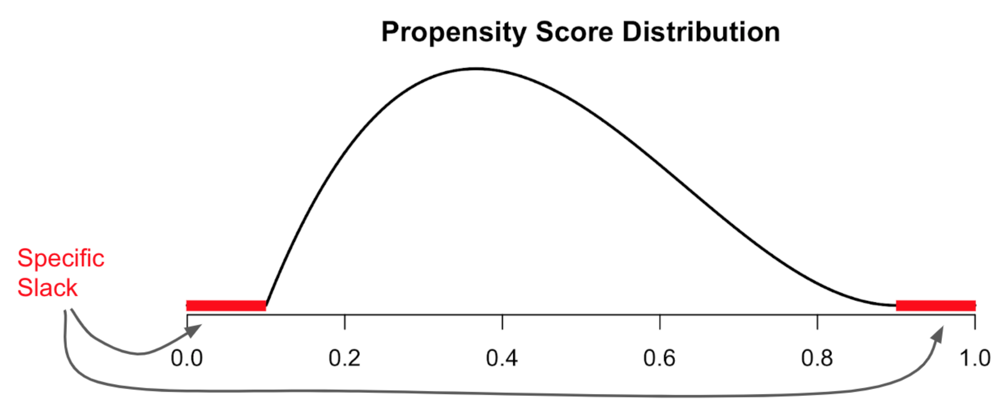
Setting: binary treatment T, baseline covariates X (arbitrary),

$$(T_i, X_i) \overset{i.i.d.}{\sim} (T, X) \quad \text{(the only assumption!)}$$

Strict overlap condition:

$$\exists \mathcal{O}_0 > 0, \quad \mathcal{O}_0 \le \underbrace{e(X) \triangleq \mathbb{P}(T = 1 \mid X)}_{\text{propensity score}} \le 1 - \mathcal{O}_0, \quad \text{a.s.}$$

Population overlap slack: $\mathcal{O}^* = \inf_x \min\{e(x), 1 - e(x)\}$



- Strict overlap condition $\Longleftrightarrow \mathcal{O}^* \ge \mathcal{O}_0$
- $n\mathcal{O}^*$ is the effective sample size without outcome restriction (Hong et al. '18)
- $1 - n\mathcal{O}^* / \min\{n_1, n_0\}$ measures the relative efficiency loss compared to an RCT
- In practice, high $\mathcal{O}^* \Longrightarrow$ stability of doubly robust estimators

Current approaches for assessing overlap:

- Informal comparisons or plug-in estimates based on estimated propensity scores
  - useful but lack of statistical guarantees
  - "sample overlap" $\neq$ population overlap
  - sensitive to model mis-specification or finite sample errors
- Standard two-sample test: testing the wrong null

$$H_0 : \mathbb{P}(X \mid T = 1) = \mathbb{P}(X \mid T = 0) \Longrightarrow H_0 : e(X) \equiv e_0$$

Major challenge: $\mathcal{O}^*$ is irregular (extreme of an unknown function)

O-value

Definition. *$\hat{\mathcal{O}}$ is an O-value if it is an upper confidence bound of $\mathcal{O}^*$, i.e.*

$$\mathbb{P}(\mathcal{O}^* \le \hat{\mathcal{O}}) \ge 1 - \alpha$$

Analogous to p-value:

- A small $\hat{\mathcal{O}}$ provides strong evidence against overlap
- A large $\hat{\mathcal{O}}$ does not necessarily imply sufficient overlap

Some practical implications:

- Strict overlap condition as a composite null hypothesis: reject if $\hat{\mathcal{O}} < \mathcal{O}_0$
- $(1 - n\hat{\mathcal{O}} / \min\{n_1, n_0\})_+$ estimates **efficiency loss** caused by the imbalance
- Assessing if trimming (say, at 0.1 and 0.9) is successful by comparing $\hat{\mathcal{O}}$ with 0.1
- Comparing different matches based on $\hat{\mathcal{O}}$

Main (and perhaps surprising) contribution:

We develop **distribution-free** O-values that are valid **in finite samples**!

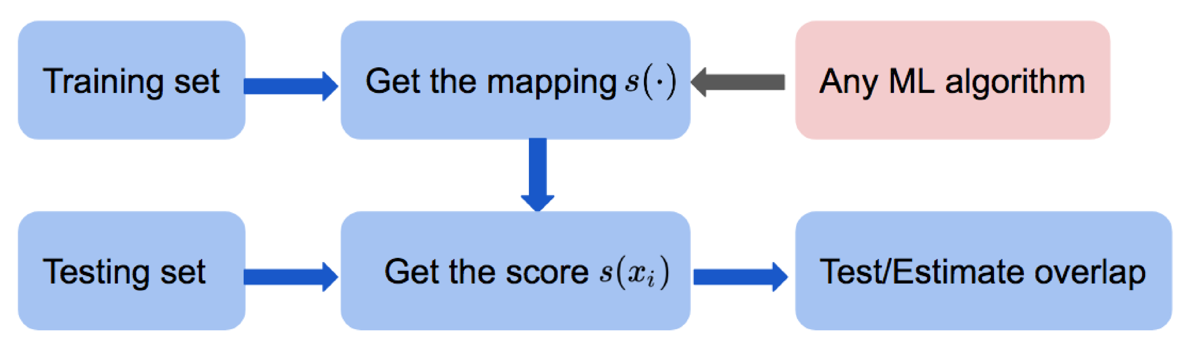
Step I: covariate standardization

Key observation: **overlap is preserved under transformation of X**

$$\mathcal{O}_0 \le e(x) \le 1 - \mathcal{O}_0 \Longrightarrow \mathcal{O}_0 \le \mathbb{P}(T = 1 \mid s(X) = s) \le 1 - \mathcal{O}_0$$

for any *fixed* function $s(\cdot)$

- Let $\mathcal{O}_s^*$ be the population overlap slack for $(T, s(X))$
- $\mathcal{O}^* = \mathcal{O}_e^*$ (Rosenbaum and Rubin, '83)
- **Data splitting** guarantees that $\hat{e}(\cdot) \perp\!\!\!\perp$ (second half of data)
- $\mathcal{O}^* \le \mathcal{O}_{\hat{e}}^*$ always holds even if $\hat{e}$ is bad; tight if $\hat{e}$ is good



From now on, we assume that

$$(S_i, T_i) \text{ are i.i.d. with } S_i = \hat{e}(X_i) \in [0, 1]$$

Goal: construct upper confidence bounds on $\mathcal{O}_{\hat{e}}^*$ (w/ **standardized covariates**)

Step II: careful balance check

Key observation: **overlap $\Longleftrightarrow$ bounded likelihood ratio**

$$b_{\min}(\mathcal{O}^*; \pi) \le \frac{dP_{S|T=1}}{dP_{S|T=0}}(s) \triangleq \frac{dP_1}{dP_0}(s) \le b_{\max}(\mathcal{O}^*; \pi), \quad \forall s \in [0, 1],$$

$$\text{where } \pi = \mathbb{P}(T = 1), \; b_{\min}(\mathcal{O}^*; \pi) = \frac{\mathcal{O}^*}{1 - \mathcal{O}^*} \frac{\pi}{1 - \pi}, \; b_{\max}(\mathcal{O}^*; \pi) = \frac{1 - \mathcal{O}^*}{\mathcal{O}^*} \frac{\pi}{1 - \pi}$$

Intuition: **larger $\mathcal{O}^* \Longrightarrow$ smaller discrepancy between P_1 and P_0**

A generic strategy:

- Find an estimable "discrepancy" $\Delta(P_0, P_1)$ and $B_\Delta(\mathcal{O}) \downarrow \mathcal{O}$

$$\Delta(P_0, P_1) \le B_\Delta(\mathcal{O}^*) \quad \textbf{(population property)}$$

- Compute a lower confidence bound on $\Delta(P_0, P_1)$

$$\mathbb{P}\left(\hat{\Delta}^- \le \Delta(P_0, P_1)\right) \ge 1 - \alpha \quad \textbf{(sample property)}$$

- $\hat{\mathcal{O}} = B_\Delta^{-1}(\hat{\Delta}^-)$ is a valid O-value:

$$\mathbb{P}\left(\hat{\mathcal{O}} \ge \mathcal{O}^*\right) = \mathbb{P}\left(\hat{\Delta}^- \le B_\Delta(\mathcal{O}^*)\right) \ge \mathbb{P}\left(\hat{\Delta}^- \le \Delta(P_0, P_1)\right) \ge 1 - \alpha$$

Summary of DiM/DiT/DiR/CE O-values

	Δ	$B_\Delta(\mathcal{O}^*)$	$\hat{\Delta}_-$	Two-sample test analogy
DiM	T-stat.	χ^2-divergence	Hedged capital bound Waudby-Smith-Ramdas ('20)	t-test
DiT	LR	Simple algebra	Line-crossing Dempster ('59) Simes' inequality Sarkar ('98)	Kolmogorov-Smirnov test
DiR	AUC	Generalized Neyman-Pearson	Hybrid bound for U-stat Bates-Candès-Lei-Romano-Sesia ('21)	Wilcoxon rank-sum test
CE	class. error	Formula of Bayes risk	Same as DiT O-values	Classification-based test

Example: (simplified) DiM O-value

Population property D'Amour-Ding-Feller-Lei-Sekhon ('17)

Theorem. *Let μ_t, σ_t^2 be the mean and variance of P_t. Then*

$$(\Delta(P_0, P_1) =) \; T_0 \triangleq \frac{|\mu_1 - \mu_0|}{\sigma_0} \le \sqrt{(1 - b_{\min}(\mathcal{O}^*; \pi))(b_{\max}(\mathcal{O}^*; \pi) - 1)} \; (= B_\Delta(\mathcal{O}^*)),$$

Sample property Maurer-Pontil ('09) (constants are not good enough!)

Proposition (Empirical Berstein's inequality)**.** *Let $Z_1, \ldots, Z_n \in [0, 1]$ be i.i.d. with $\mathbb{E} Z_i = \mu, \mathrm{Var}(Z_i) = \sigma^2$. Then with probability $1 - \delta$,*

$$|\hat{\mu} - \mu| \le \hat{\sigma} \sqrt{\frac{2 \log\left(\frac{3}{\delta}\right)}{n}} + \frac{7 \log\left(\frac{3}{\delta}\right)}{3(n - 1)}, \quad \sigma - \hat{\sigma} \le \sqrt{\frac{2 \log\left(\frac{3}{\delta}\right)}{n - 1}}.$$

$\Longrightarrow$ a lower confidence bound on T_0 (with Bonferroni correction over (μ_1, μ_0, σ_0)):

$$\hat{T}_0^- = \frac{\hat{\mu}_1^- - \hat{\mu}_0^+}{\hat{\sigma}_0^+}$$

(Simplified) DiM O-value

$\mathcal{C}^\pi \leftarrow (1 - \alpha/2)$-CI for π, $\hat{T}_0^- \leftarrow (1 - \alpha/2)$-lower confidence bound on T_0

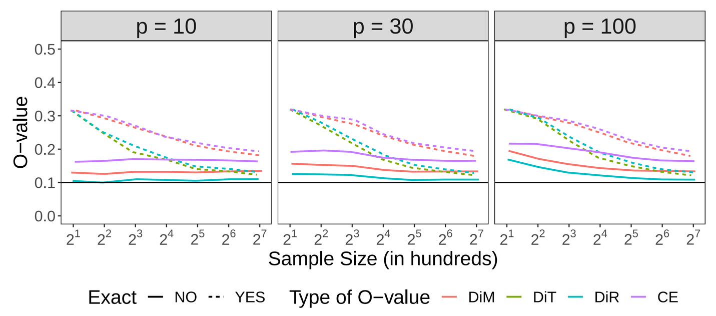
$$\hat{\mathcal{O}}_{\mathrm{DiM}} = \sup_{\pi \in \mathcal{C}^\pi} \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\pi(1 - \pi)}{\pi^2 (\hat{T}_0^-)^2 + 1}}$$

Theorem. *With solely the i.i.d. assumption,* $\mathbb{P}(\mathcal{O}^* \le \hat{\mathcal{O}}_{\mathrm{DiM}}) \ge 1 - \alpha$

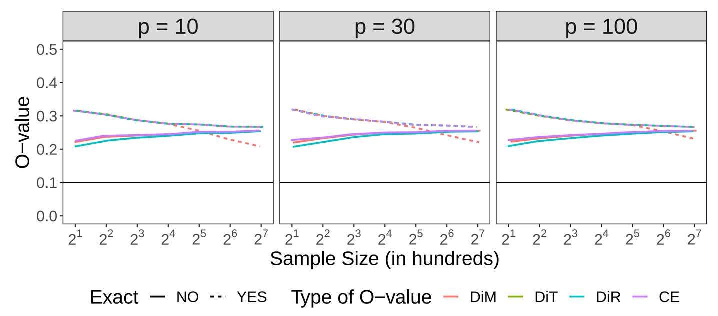
Comparisons of O-values

An illustrative simulation study

- $X \sim N(0, I_p)$ with $p \in \{10, 30, 100\}$
- $e(x) = f(x^T \beta), \quad f(y) = \begin{cases} 0.1 \; (y < c) \\ 0.9 \; (y > c) \end{cases}$
- β sparse; c is chosen such that $\mathbb{P}(e(X) = 0.1) = 0.8$



Gradient Boosting



Random Forest

Practical recommendation based on extensive numerical experiments

- Algorithm to estimate propensity scores: **gradient boosting**
- Type of O-value: **DiT**

O-values for Lalonde data

- National Supported Work Demonstration program Lalonde ('86)
- Treatment group has $n_1 = 185$ units
- 7 control groups: 6 from observational studies, 1 from an RCT
- Apply gradient boosting for DiT O-values

	CPS			PSID			RCT		
	n_0	$\hat{\mathcal{O}}$	$\hat{L}$	n_0	$\hat{\mathcal{O}}$	$\hat{L}$	n_0	$\hat{\mathcal{O}}$	$\hat{L}$
Raw	15992	0.003	77%	2490	0.018	75%	260	0.483	0%
V2	2369	0.021	71%	253	0.234	45%			
V3	429	0.143	53%	128	0.313	23%			