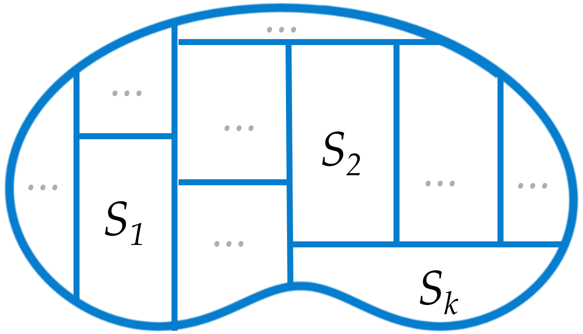
Smaller Confidence Intervals From IPW Estimators via Data-Dependent Coarsening



Alkis Kalavasis

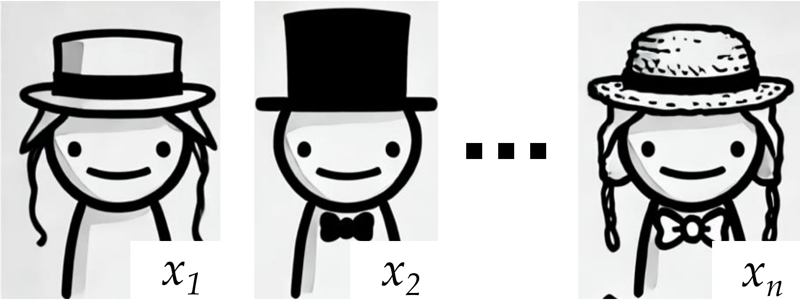
Anay Mehrotra

Manolis Zampetakis

Yale

Observational Studies

Randomized control trials (RCTs) are more powerful but also more costly than observational studies



x_1 x_2 ... x_n

Covariates $x \in \mathbb{R}^d$

$$t_i = \begin{cases} 1 & \text{with probability } e(xi), \\ 0 & \text{with probability } 1 - e(xi) \end{cases}$$

Propensity score $e(x) \in [0,1]$

Observations:

1. *treatments* $t_i \in \{0,1\}$ *and*
2. *outcomes* $y_i(t_i) \in [0,1]$

Does interacting with a sales representative increase in the average amount spent on a hat?

Treatments $t \in \{0,1\}$ *Average treatment effect* $\tau := \mathbb{E}[\ Y(1) - Y(0)\ ]$

Goal: Given a desired confidence level α, find an estimate $\hat{\tau}$ of τ with the *smallest confidence interval*

IPW Estimators

Widely used family of estimators of the average treatment effect (ATE)

- Economics (e.g., Dehejia and Wahba'98, Galiani, Gertler, Schargrodsky'05, Abadie and Imbens'06),
- Medicine (e.g., Rubin'97, Christakis and Iwashyna'03, Austin'08),
- Political Science (e.g., Brunell and Dinardo'04, Sekhon'04, Ho et al.'07)…

Given dataset $\mathcal{D} \coloneqq \{(x_i, y_i, t_i)\}_{i=1}^n$ and propensity scores $e(x): \mathbb{R}^d \to [0,1]$, the vanilla IPW estimator is

$$\mathrm{IPW}(\mathcal{D}; e) = \frac{1}{n} \sum_i \frac{t_i y_i}{e(x_i)} - \frac{(1 - t_i) y_i}{1 - e(x_i)}$$

IPW estimators have several desirable properties:

- Easy to describe,
- computationally efficient,
- unbiased (under standard assumptions),
- asymptotically normal → *For a fixed confidence level, confidence interval size ∝ RMSE (root-mean-squared error)*

Issues with IPW Estimators

$$\text{IPW}(\mathcal{D}; e) = \frac{1}{n}\sum_i \frac{t_i y_i}{e(x_i)} - \frac{(1-t_i)y_i}{1-e(x_i)}$$

Issue 1: Inaccurate propensity scores

Given scores $\hat{e}(\cdot)$ satisfying $\|\hat{e} - e\|_\infty \leq \varepsilon$, bias $|\tau_{\text{IPW}}(\mathcal{D}; \hat{e}) - \tau|$ can be *arbitrarily large!*

Issue 2: Outliers with extreme propensity scores

$\text{Var(IPW)} \propto \frac{1}{n}\mathbb{E}\left[\frac{1}{e(x)(1-e(x))}\right]$ which goes to ∞ with extreme propensity scores *(i.e., $e(x) \to 0$ or 1)*

Question: Are there estimators with "small" confidence intervals in the presence of above issues?

Variants of Estimators to IPW

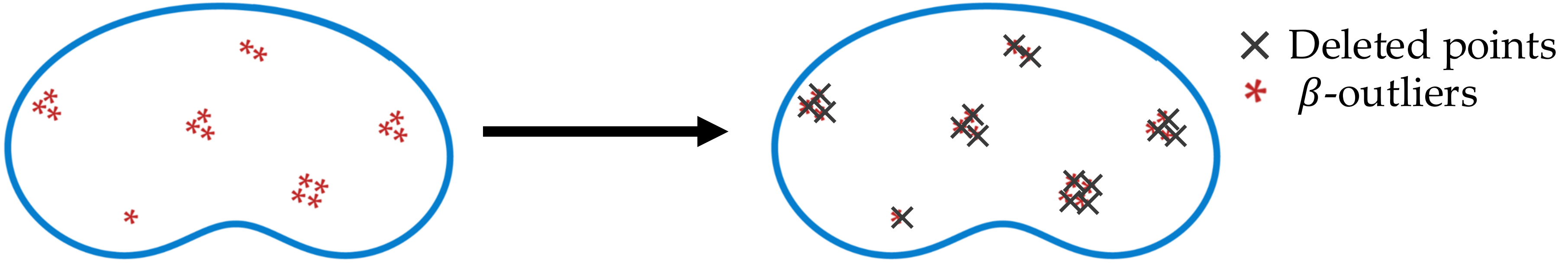
Doubly robust (DR) estimators (*Foster and Syrgkanis'23, Chernozhukov et al.'18, Bang and Robins'05…*)

- They reduce IPW estimator's RMSE by combining it with $\mu_t(x) \coloneqq \mathbb{E}[Y(t) \mid X = x]$ for each $t \in \{0,1\}$
- **Issue:** RMSE of DR estimators can be arbitrarily large with outliers and inaccurate propensity scores!

Trimmed IPW estimator (*Crump'09, Li, Thomas, and Li'19, …*)

Definition (Outliers). A covariate x is a *β-outlier* if $e(x) \cdot (1 - e(x)) < \beta$

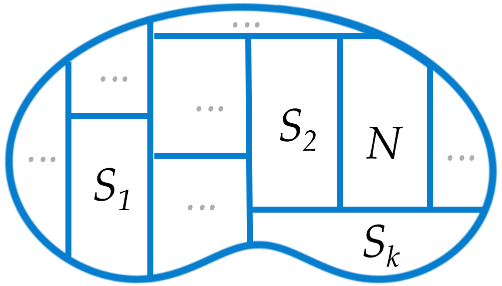
- IPW estimators which "remove" all β-outliers for some $\beta = \Omega(1)$
- **Issue:** Trimmed IPW estimator can have a bias of $\Omega\left(\rho + \frac{1}{\sqrt{\beta n}}\right)$, where ρ is the mass of the β-outliers



Most existing variants of IPW estimators are data independent

A New Family of Coarse-IPW Estimators

- Defined as the IPW estimator on a "coarse" covariate space – where many covariates are merged



Definition (CIPW Estimators) Given a partition ($\mathcal{S} = \{S_1, S_2, \dots\}, N$) of $\mathbb{R}^d$, a dataset $\mathcal{D}$, and propensity scores $e(\cdot)$, the Coarse-IPW CIPW estimator is

$$\mathrm{CIPW}_{\mathcal{S},N}(\mathcal{D}; e) := \frac{1}{|\{i \in [n]: x_i \notin N\}|} \cdot \sum_{S \in \mathcal{S}} \sum_{i: x_i \in S} \frac{t_i y_i}{e(S)} - \frac{(1 - t_i) y_i}{1 - e(S)}$$

where $e(S)$ is the average propensity score over set S

- Captures almost all existing variants of IPW estimators as special cases

Assumptions on Data

1 (Lipschitzness) The expected outcome under treatment t, conditioned on covariates, is *L-Lipschitz, i.e.,*

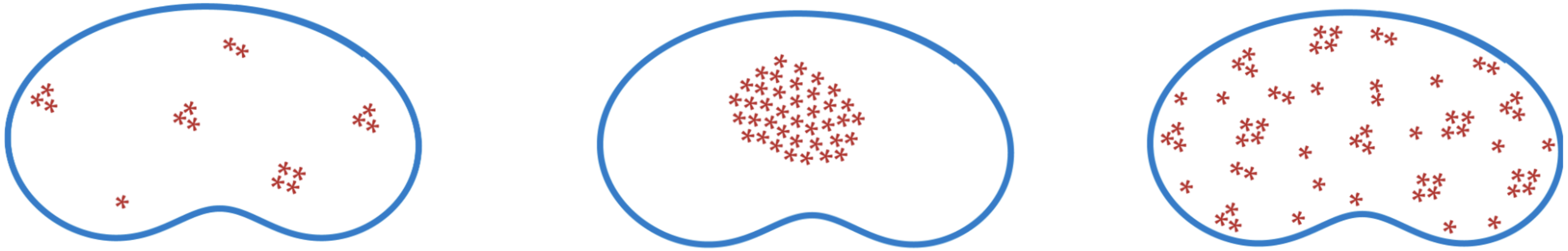
$$\mu_t(x) \coloneqq \mathbb{E}[Y(t) \mid X = x] \textit{ is L-Lipschitz for each } t \in \{0,1\}$$

- Lipschitzness holds under standard parametric assumptions, e.g., when $\mu_t(x) \approx w^\top x$

2 (Sparsity) For any ℓ_2-ball of diameter $> \alpha$ at least $\Omega(1)$ fraction of covariates in the ball are *not* β outliers

3 (Isolation) There are k ℓ_2-balls of diameter $> \alpha$ which are pairwise $\Omega(\alpha)$ far and that partition the outliers

- Sparsity and isolation are ***testable*** from data given estimates $\hat{e}(\cdot)$ of the propensity scores $e(\cdot)$



(a) Sparse and Isolated Instance (b) Non-Sparse Instance (c) Non-Isolated Instance

Our Main Results

RMSE guarantee of CIPW estimators

Theorem 1 (Informal). Suppose Assumptions 1-3 hold with constants α, β, L and β-outliers have mass ρ. There is an algorithm that given

1. propensity scores $\hat{e}$ with $\|\hat{e} - e\|_\infty \leq \varepsilon$, and
2. $n = \Omega(d/\varepsilon^2)$ independent samples,

outputs a value τ_A in $O(n^3)$ time such that $\mathbb{E}[|\tau - \tau_A|^2]^{1/2} = O\left(\varepsilon + \alpha\rho L + \frac{1}{\sqrt{n}}\right)$

Comparison to baseline estimators

Theorem 2 (Informal). For each $\eta > 0$ there is an example satisfying Assumptions 1-3 where

1. The RMSE of IPW and doubly robust estimators is $\Omega(1/\eta)$
2. The RMSE of ε-Trimmed IPW estimator is $\Omega(1)$
3. The RMSE of the CIPW estimator in Theorem 1 is $O(\varepsilon + 1/\sqrt{n})$

Finding CIPW Estimator with Small RMSE

Computational hardness results

Proposition 1 (Informal). Given dataset $\mathcal{D}$ and propensity scores $e(\cdot)$, it is **NP-hard** to find the minimum RMSE of a CIPW estimator on the dataset

In fact, it is **NP-hard to approximate** it within any *exponential-in-bit-complexity factor*

Since finding the best CIPW estimator is hard, we focus on finding a "good" CIPW estimator

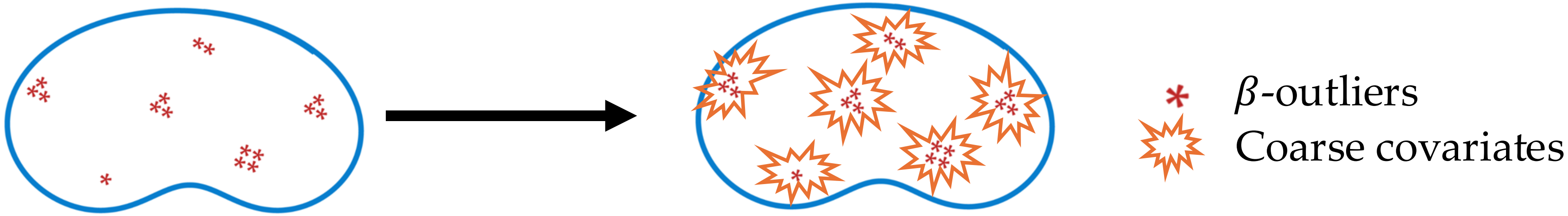
$$\begin{aligned} \text{Bias}\left(\text{CIPW}_{\mathcal{S},N}\right) &\leq \Pr[x \in N] + \frac{O(L)}{\Pr[x \notin N]} \cdot \sum_{S \in \mathcal{S}} \text{diam}(S) \cdot \Pr[x \in S], \\ \text{Var}\left(\text{CIPW}_{\mathcal{S},N}\right) &\leq \frac{1}{n} \cdot \frac{1}{\Pr[x \notin N]} \cdot \sum_{S \in \mathcal{S}} \frac{\Pr[x \in S]}{e(s)\left(1 - e(S)\right)}. \end{aligned}$$

Definition (Good-local partition). A partition is (α, β)-good-local if (i) $\Pr[x \in N] \leq \alpha$, (ii) for each $S \in \mathcal{S}$, $e(S)\left(1 - e(S)\right) \geq \beta$ and $\text{diam}(S) \leq \alpha$

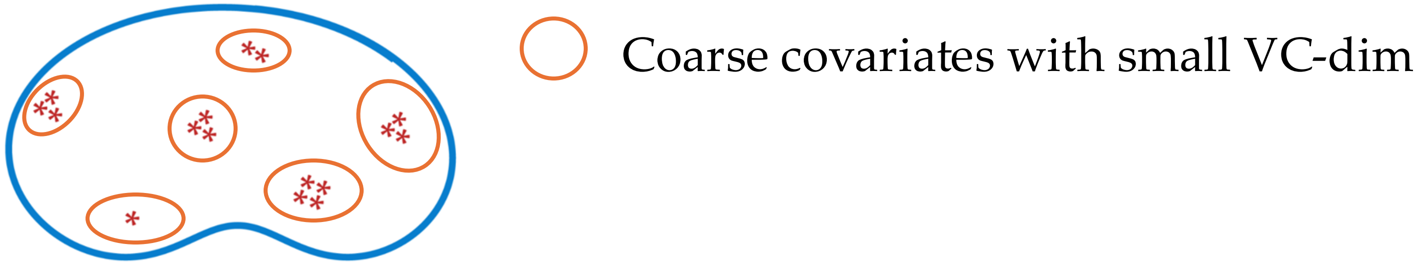
Proposition (Informal). Suppose $O(1)$-Lipschitzness holds. Any CIPW estimator defined by an (α, β)-good-local partition has RMSE $\leq O\left(\alpha + \frac{\varepsilon}{\beta}\right)$

Additional Challenges

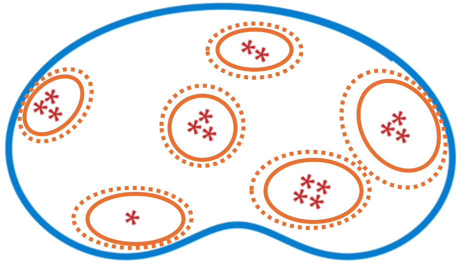
If propensity scores $e(\cdot)$ were exactly known:



Since $e(\cdot)$ are learned from data and affect "coarsening" $(\mathcal{S}, N)$, the learned $(\mathcal{S}, N)$ must generalize



Since $e(\cdot)$ are learned from data has errors, the learned coarse covariates must have "margin"



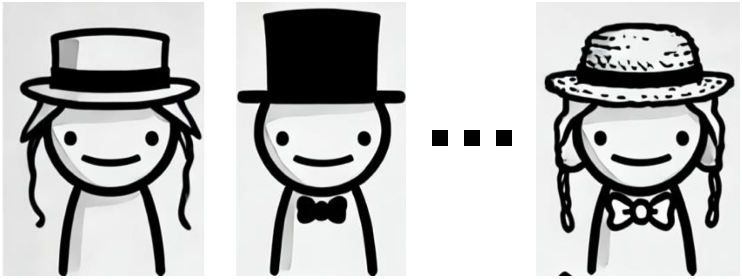
The complete algorithm overcomes a few other challenges, e.g., uses fractional $(\mathcal{S}, N)$ to reduce RMSE further

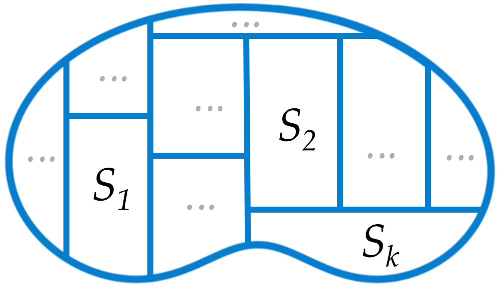
Conclusion

(Downsides of IPW) Confidence intervals arising from IPW Estimators can be arbitrarily poor with

1. Inaccuracies in propensity scores, and
2. Outliers (with propensity scores close to 0 or 1)

(CIPW Estimators) We introduce a family of Coarse IPW (CIPW) estimators which captures IPW estimators





(Main algorithmic result) We give an algorithm for data-dependent coarsening such that the resulting CIPW estimator has smaller confidence intervals than IPW Estimators with (1) inaccuracies and (2) outliers

(Further results) We explore the statistical and computational complexity of finding CIPW estimators with "small" confidence intervals