Doubly Robust Estimators with Weak Overlap

A Robust Strategy for Causal Inference

Based on the paper by:
Yukun Ma, Pedro H.C. Sant'Anna, Yuya Sasaki, Takuya Ura

Department of Economics, UC3M

December 15, 2025

Outline

The Dilemma: Doubly Robustness vs. Weak Overlap

The Appeal of Doubly Robust (DR) Estimators

DR estimators (e.g., AIPW) are popular in causal inference because:

- They remain consistent if **either** the Propensity Score Model **or** the Outcome Regression Model is correctly specified.
- They offer "double protection" against misspecification.

The Problem: Weak Overlap

However, when propensity scores $P(X)$ are close to 0 or 1:

- The IPW weights ($\frac{1}{P(X)}$ or $\frac{1}{1-P(X)}$) become extremely large.
- **Consequence:** The estimator suffers from **Variance Explosion** and instability.

The Cost of Existing Solutions

How do researchers typically handle weak overlap?

Current Practice: Fixed Trimming

- *Method:* Drop observations with scores outside $[0.1, 0.9]$.
- *Literature:* Crump et al. (2009), Yang & Ding (2018).

The Hidden Cost

- It **changes the estimand**.
- You are no longer estimating the ATE for the whole population, but only for the sub-population with good overlap.
- Risk: You might exclude the individuals who need the intervention the most.

Goal: Solve the variance explosion without changing the target parameter.

Core Intuition: "Trim-then-Bias-Correct"

The authors propose a two-step strategy to handle extreme weights:

1. **Dynamic Trimming:**
   - Use a moving threshold h_n that shrinks to 0 as sample size $n \to \infty$.
   - Temporarily remove observations causing variance explosion.
   - *Function: Stabilizes Variance.*
2. **Bias Correction:**
   - Acknowledge that trimming introduces bias.
   - Use **Sieve Regression** to estimate the contribution of the trimmed tails based on the behavior of the remaining data.
   - *Function: Recovers the Original Estimand (ATE/ATT).*

Methodology: General Framework

The authors observe that most causal estimands (ATE, LATE, ATT) can be expressed as functions of **Moments of Ratios**.

General Structure

$$\theta_0 = \Lambda\left(E\left[\frac{B_1}{A_1}\right], \ldots, E\left[\frac{B_L}{A_L}\right]\right)$$

- B_l: Numerator term (e.g., Outcome $\times$ Treatment).
- A_l: The problematic **denominator** (e.g., Propensity Score $P(X)$).
- When $A_l \approx 0$, we face weak overlap and variance explosion.

Strategy: Construct a robust estimator $\hat{\alpha}(h, \hat{\gamma})$ for each moment $E[B/A]$ using the "Trim-then-Bias-Correct" approach.

Examples from the Paper

The framework $\theta_0 = \Lambda(E[B/A])$ covers many standard designs.

Example 1: ATE (Unconfoundedness)

Standard DR estimator involves terms like:

$$E\left[\frac{(Y - \nu(1, X))D}{P(X)}\right]$$

Here, $B = (Y - \nu)D$ and A = P(X).

Example 2: LATE (Instrumental Variables)

Ratio of two expectations (Wald-type estimator):

$$\theta_{LATE} = \frac{E[\text{Effect on Outcome}]}{E[\text{Effect on Treatment}]} \approx \frac{E[B_1/A_1]}{E[B_2/A_2]}$$

Here, A involves the instrument propensity score. Our method robustifies both the numerator and denominator separately.

The Estimator: Trim-then-Bias-Correct

For each ratio moment, the estimator $\hat{\alpha}$ consists of two parts:

$$\hat{\alpha}(h) = \underbrace{E_n\left[\frac{B}{A} \cdot \mathbb{I}\{|A| \geq h_n\}\right]}_{\textbf{Part 1: Trimmed Mean}} + \underbrace{\text{Bias Correction Term}}_{\textbf{Part 2: Recovery}}$$

Part 1: The Trimmed Mean

- h_n: A threshold shrinking to 0 as $n \to \infty$.
- **Role:** Forcibly cuts off the extremely small denominators.
- **Result:** Stabilizes variance but introduces bias (ignores $|A| < h_n$).

Methodology: How Bias Correction Works

Part 2: The Bias Correction Term

$$\sum_{\kappa=1}^{k} \frac{E_n[A^{\kappa-1}\mathbb{I}\{|A| < h_n\}]}{\kappa!} \cdot \hat{m}^{(\kappa)}(0)$$

- **Intuition:** We threw away data where $|A| < h_n$, but we can use data where $|A| \approx h_n$ (boundary region) to **extrapolate** the tail behavior.

Mechanism: Sieve Regression

1. Define conditional expectation $m(a) = E[B|A = a]$.
2. Perform a **Taylor expansion** of $m(A)/A$ near zero.
3. Use **Sieve Regression** (Legendre polynomials) to estimate the derivatives $\hat{m}^{(\kappa)}(0)$.
4. "Fill in" the missing information using this polynomial fit.

Application: Difference-in-Differences (DiD)

This framework extends the DR-DiD method by Sant'Anna & Zhao (2020).

The Challenge in DR-DiD

- The estimator relies on weights like $\frac{1}{1-P(X)}$.
- If some control units look very similar to treated units ($P(X) \approx 1$), the weights explode.

The Improvement

- Applies the trimming + correction strategy to the ATT estimator.
- Allows for robust estimation of ATT even when covariate overlap between treated and control groups is poor.

It acts as a "safety net" for DR-DiD against data irregularity.

Simulation Results

Comparing the **Conventional (CON)** method vs. the **New (NEW)** method under weak overlap (DGP2):

Statistic	Conventional (CON)	**New (NEW)**	Interpretation
Bias	-3.317	**-0.000**	Both have acceptable bias
Std. Dev (SD)	268.8	0.257	**Variance reduced by ∼1000x!**
RMSE	268.8	**0.257**	Massive gain in precision
95% Coverage	0.958	**0.947**	Accurate inference for both

Table: Performance under Severe Weak Overlap (df=10)

* Note: CON refers to the standard Sant'Anna & Zhao (2020) estimator.

Conclusion

1. **Awareness:** Be cautious of variance explosion in IPW/DR methods when propensity scores are extreme.
2. **Innovation:** The **"Trim-then-Bias-Correct"** strategy solves instability without sacrificing the target parameter (Estimand).
3. **Practical Value:**
   - Provides a robust alternative when reviewers question "Common Support".
   - Allows researchers to retain the full sample for causal inference.

Thank You!