Mixing Samples to Address Weak Overlap in Causal Inference

2025 한국통계학회 하계학술논문발표회

Suehyun Kim (김수현)

Department of Statistics, Seoul National University

20 June 2025

Collaborators

Kwonsang Lee
Seoul National University
kwonsanglee@snu.ac.kr

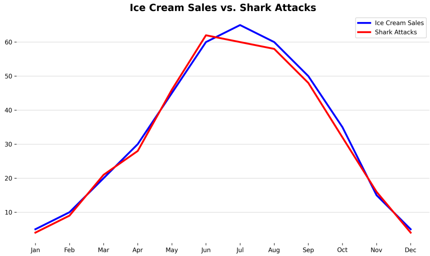
Jaehyuk (Jay) Jang
Seoul National University
bbq12340@snu.ac.kr

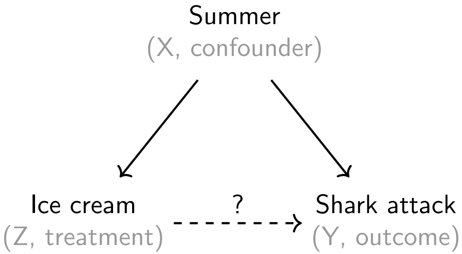
Overview

Overview

What is a causal effect?

- Association vs causation: Does ice cream cause shark attacks?

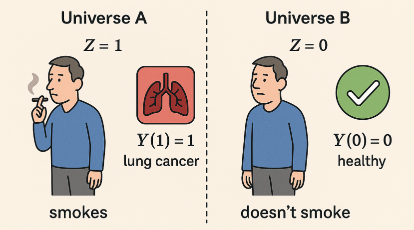




- We must control for confounders to infer causality from observational data.

What is a causal effect?

- Does smoking (Z, treatment) cause lung cancer (Y, outcome)?
  - Imagine a "parallel universe", where you observe what happens under *both* $Z = 1$ and $Z = 0$.
  - $Y(1), Y(0)$: Potential (counterfactual) outcomes



  - However, we can only observe $Y = ZY(1) + (1 - Z)Y(0)$.
- **Average Treatment Effect (ATE)**: $\mathbb{E}[Y(1) - Y(0)]$

Identification and estimation of causal effects

- Identification of ATE using weighting

$$\begin{aligned}\text{ATE} &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}\left[\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right]\end{aligned}$$

- **Propensity score (PS)**: Probability of treatment given covariates

$$e(X) = P(Z = 1 \mid X)$$

- Other causal effects, such as the ATT (Average Treatment Effect on the Treated), can be identified similarly using the PS.
  - Weights for ATE $= (\frac{1}{e}, \frac{1}{1-e})$
  - Weights for ATT $= (1, \frac{e}{1-e})$

Identification and estimation of causal effects

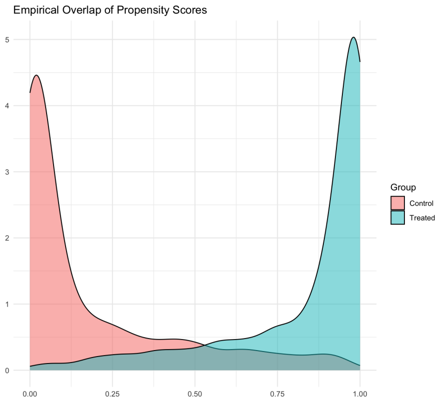
- **Inverse Probability Weighting (IPW)**: Based on the identification results, the ATE can be estimated as follows:

$$\widehat{\text{ATE}}^{IPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i Y_i}{\hat{e}_i(X)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}.$$

- The estimated propensity score $\hat{e}(X)$ can be obtained using any classification model, e.g. logistic regression.
- When $\hat{e}_i$ is close to 0 or 1, the weights $1/\hat{e}_i$ and $1/(1-\hat{e}_i)$ can be extremely large.
  - Can be highly unstable (large variance).
  - Happens when treated and control groups differ substantially $\rightarrow$ *"overlap issue"!*

Overview

Handling weak overlap in weighting methods



- **Overlap assumption**:
  $0 < e(x) = \mathbb{P}(Z = 1 \mid X = x) < 1$
- Weak overlap is problematic in weighting methods due to units with extreme weights such as $1/e \simeq \infty$ or $1/(1-e) \simeq \infty$.

Remedies so far

There have been three main approaches to handle weak overlap in the literature:

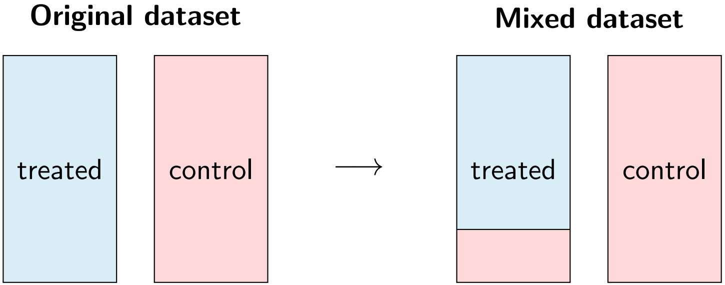
1. **Trimming/truncating units with extreme weights**
   - Loss of sample size, sensitivity to the choice of cutoff
2. **Targeting an alternative causal estimand**
   - Overlap weights and ATO (Average Treatment Effect of the Overlap Population)
   - Lack of interpretability
3. **Balancing weights**
   - Entropy Balancing, Covariate Balancing Propensity Score, etc.
   - Optimization may be infeasible under weak overlap

Our idea

We propose the **mixing framework**, which helps overcome the limitations of the above approaches by creating a synthetic sample of mixed treated and control units.

Main idea: Simple mixing strategy

Our strategy aims to intentionally increase overlap by mixing treated and control units.



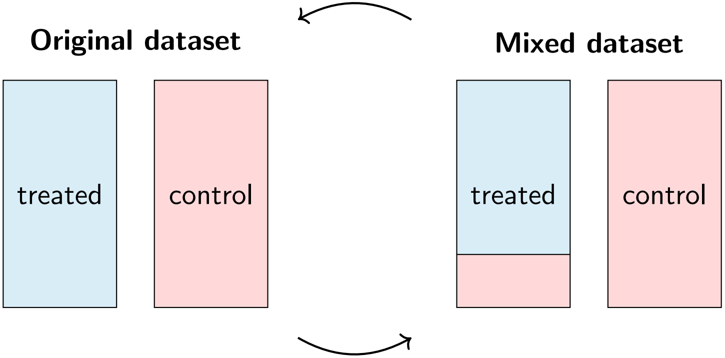
- Target population

- Stronger overlap
- More stable estimation within the mixed sample

Main idea: Simple mixing strategy

① Mixed distribution

Mixed PS, Mixed IPW and their properties



② Mixing implementation

(i) M-estimation (ii) Resampling algorithm

Overview

Notation & Setup

Under Rubin's Potential Outcomes Framework, our aim is to apply mixing to weighting estimators of the Average Treatment Effect on the Treated (ATT).

Assumptions

1. **Unconfoundedness**: $(Y(1), Y(0)) \perp\!\!\!\perp Z \mid X$
2. **Overlap**: $0 < e(x) = \mathbb{P}(Z = 1 \mid X = x) < 1$ for all x

- $(Y(0), Y(1))$: Potential (counterfactual) outcomes
- Y: Observed outcome
- X: Observed covariates
- Z: Binary treatment indicator
- $\tau = E\left[Y(1) - Y(0) \mid Z = 1\right]$: Target estimand (ATT)
- $f_{Y,X}$: Joint density of (Y, X)
- $f_{Y,X|Z=z}$: Joint density of (Y, X) given $Z = z$

Mixed distribution

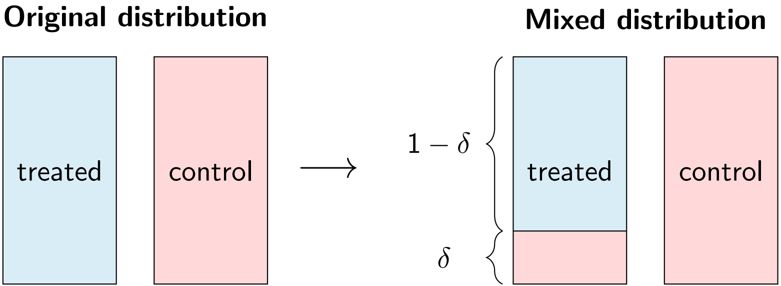
Definition (Mixed distribution)

We define the distribution of (Y^*, Z^*, X^*) as the distribution with the conditional joint densities of (Y^*, X^*) given $Z^* = 1, 0$, respectively,

$$
\begin{aligned}
f_{Y^*,X^*|Z^*=1} &= (1-\delta) f_{Y,X|Z=1} + \delta f_{Y,X|Z=0} \\
f_{Y^*,X^*|Z^*=0} &= f_{Y,X|Z=0}
\end{aligned}
$$

for a fixed constant $0 < \delta < 1$ and Z^* to satisfy $\mathbb{P}(Z^* = 1) = \mathbb{P}(Z = 1) =: \pi$. We refer to the mixed distribution with a constant δ as the **simple mixed distribution**.

Mixed distribution



Mixed propensity score

Lemma 1 (Mixed propensity score and its robustness)

Let $e^*(x) = \mathbb{P}(Z^* = 1 \,|\, X^* = x)$ *be the propensity score of the mixed distribution. Then,*

$$\frac{e^*}{1-e^*}(x) = (1-\delta)\frac{e}{1-e}(x) + \delta\frac{\pi}{1-\pi} \quad \textit{for all } x.$$

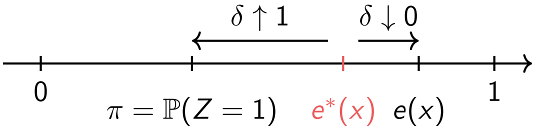


Figure 1: Behavior of $e^*(x)$ with respect to δ

Mixed IPW (MIPW) estimator

Theorem 1 (MIPW estimator and its consistency)

Using the mixed propensity score

$$\frac{e}{1-e}(x) = \frac{\frac{e^*}{1-e^*}(x) - \delta\frac{\pi}{1-\pi}}{1-\delta},$$

we define the **Mixed IPW (MIPW)** *estimator as follows:*

$$\hat{\tau} := \frac{\sum_i Z_i Y_i}{\sum_i Z_i} - \frac{\sum_i \left(\frac{e^*}{1-e^*}(X_i^*) - \delta\frac{\pi}{1-\pi}\right)(1-Z_i^*)Y_i^*}{\sum_i \left(\frac{e^*}{1-e^*}(X_i^*) - \delta\frac{\pi}{1-\pi}\right)(1-Z_i^*)}$$

Under the strong ignorability assumptions, $\hat{\tau}$ is a consistent estimator of τ.

Mixing implementation 1: M-estimation

Proposition 1 (Asymptotic normality based on observed samples)

Under the strong ignorability assumptions, $\hat{\theta} = Solve_{\theta}\left[\sum_i \psi^{**}(\theta; Y_i, X_i, Z_i) = 0\right]$ *is an M-estimator of* $\theta = (\beta, \pi, E[Y(1) \mid Z = 1], E[Y(0) \mid Z = 1])$*, where, for* $0 < \delta < 1$*,*

$$\psi^{**}(\theta; Y, X, Z) = \begin{pmatrix} \left\{ \frac{1-\delta}{e^*(X;\beta)} Z + \left(\frac{\delta\pi}{(1-\pi)e^*(X;\beta)} - \frac{1}{1-e^*(X;\beta)} \right)(1-Z) \right\} \nabla_{\beta} e^*(X;\beta) \\ Z - \pi \\ ZY - ZE[Y(1) \mid Z = 1] \\ \frac{e(X;\beta)}{1-e(X;\beta)}(1-Z)Y - \frac{e(X;\beta)}{1-e(X;\beta)}(1-Z)E[Y(0) \mid Z = 1] \end{pmatrix}.$$

$$\begin{tikzcd}
\{(Y_i, Z_i, X_i)\}_{i=1}^{n} \arrow[dr, dashed] & & \\
\{(Y_i, Z_i, X_i, Y_i^*, Z_i^*, X_i^*)\}_{i=1}^{n} \arrow[r, "\psi^*"'] & \hat{\tau} \arrow[r, "n \to \infty"] & \tau
\end{tikzcd}$$

Mixing implementation 1: M-estimation

Proposition 2 (Asymptotic normality based on observed samples)

Under the strong ignorability assumptions, $\hat{\theta} = \mathit{Solve}_\theta \left[\sum_i \psi^{**}(\theta; Y_i, X_i, Z_i) = 0\right]$ *is an M-estimator of* $\theta = (\beta, \pi, E[Y(1) \mid Z = 1], E[Y(0) \mid Z = 1])$, *where, for* $0 < \delta < 1$,

$$\psi^{**}(\theta; Y, X, Z) = \begin{pmatrix} \left\{ \frac{1-\delta}{e^*(X;\beta)} Z + \left(\frac{\delta\pi}{(1-\pi)e^*(X;\beta)} - \frac{1}{1-e^*(X;\beta)} \right)(1-Z) \right\} \nabla_\beta e^*(X;\beta) \\ Z - \pi \\ ZY - ZE[Y(1) \mid Z = 1] \\ \frac{e(X;\beta)}{1-e(X;\beta)}(1-Z)Y - \frac{e(X;\beta)}{1-e(X;\beta)}(1-Z)E[Y(0) \mid Z = 1] \end{pmatrix}.$$

$$\begin{tikzcd}
\{(Y_i, Z_i, X_i)\}_{i=1}^n \arrow[dr, "\psi^{**}"] & & \\
\{(Y_i, Z_i, X_i, Y_i^*, Z_i^*, X_i^*)\}_{i=1}^n \arrow[r, "\psi^*"'] & \hat{\tau} \arrow[r, "n \to \infty"] & \tau
\end{tikzcd}$$

Mixing implementation 2: Resampling algorithm

Another way to implement mixing is to use a **resampling algorithm** that directly estimates $\hat{f}^*_{Y,X|Z=z}$ from the observed dataset.

Observed dataset

$$\{(Y_i, Z_i, X_i)\}_{i=1}^n \sim \hat{f}_{Y,X|Z=z} \Longrightarrow \hat{f}^*_{Y,X|Z=z}$$

Resample M times

$$\left(\{(Y_i^*, Z_i^*, X_i^*)\}_{i=1}^n\right)_{m=1}$$

$$\vdots$$

$$\vdots$$

$$\left(\{(Y_i^*, Z_i^*, X_i^*)\}_{i=1}^n\right)_{m=M}$$

$$\xrightarrow{\text{Aggregate}} \hat{\tau}$$

Mixing implementation 2: Resampling algorithm

The resampling algorithm allows for extensions to various weighting schemes, such as Entropy Balancing or Covariate Balancing Propensity Score.

Proposition 3 (Extension to balancing weights)

Suppose W^* *is a* **balancing weight** *for* X^**, satisfying*

$$E[X^* \mid Z^* = 1] = E[W^* X^* \mid Z^* = 0].$$

Then,

$$W := \frac{W^* - \delta \frac{\pi}{1-\pi}}{1 - \delta}$$

is a balancing weight for X*.*

Overview

Simulation study

Simulation 1	Simulation 2
M-estimation • IPW vs MIPW • Efficiency gain in terms of both finite- and large-sample perspective	Resampling algorithm • Extension to Entropy Balancing • Performance under model misspecification

- **Data generating process**: $e(X) = \{1 + \exp(-X^T\beta)\}^{-1}, X \sim N_5(0, I)$
  - Overlap level (according to β): Strong / Moderate / Weak
  - Treatment effect: $\tau = 1$ (homogeneous)

Simulation study for implementation 1: M-estimation

- **Performance measures**: Monte-Carlo simulation of (1) standard deviation estimates and (2) Huber-White's robust standard error estimates
- **Benchmark**: ATO estimation via overlap weights (Li et al., 2018)
  - **ATO**: A causal estimand under the subpopulation for which the average treatment effect can be estimated with the smallest variance.

- **True treatment effect**: 1 (homogeneous) $\implies ATO = ATT = 1$

Estimator	IPW	MIPW	OW
Target	ATT	ATT	ATO

Results: IPW vs MIPW vs OW

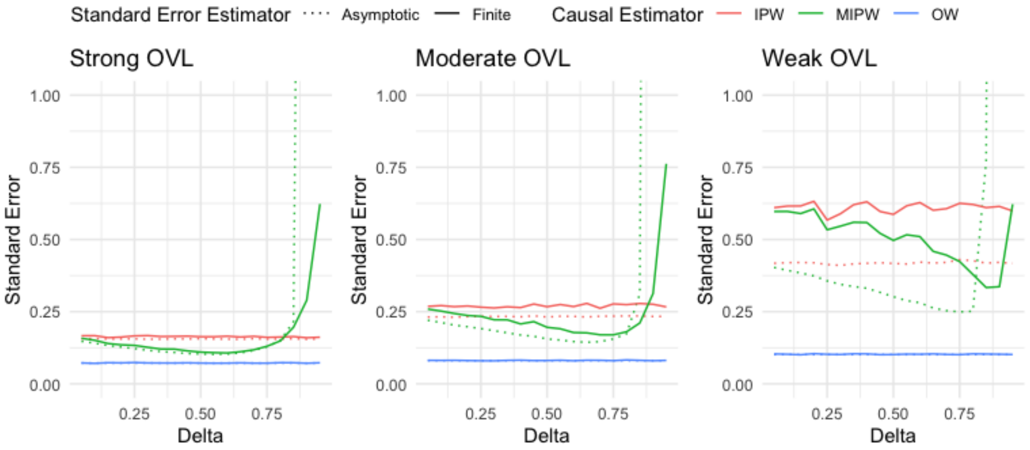


Figure 2: Monte Carlo simulation result: SD estimates (solid) and Huber-White robust SE estimates (dotted) of IPW, MIPW, OW

Simulation study for implementation 2: Resampling algorithm

- **Scenario 1**: The same weak overlap setting from previous study (true treatment effect $= 1$)
- **Scenario 2**: A modified study from Kang & Schafer (2007) to endow model misspecification but within weak overlap (true treatment effect $= 210$)
- **Extension to Entropy Balancing (EB)**: Weighting method that estimates $\frac{e}{1-e}(X_i)$ by solving a constrained optimization problem to reduce model dependence (Hainmueller, 2012).

Estimator	EB	MEB	OW
Target	ATT	ATT	ATO

Results: EB vs MEB vs OW

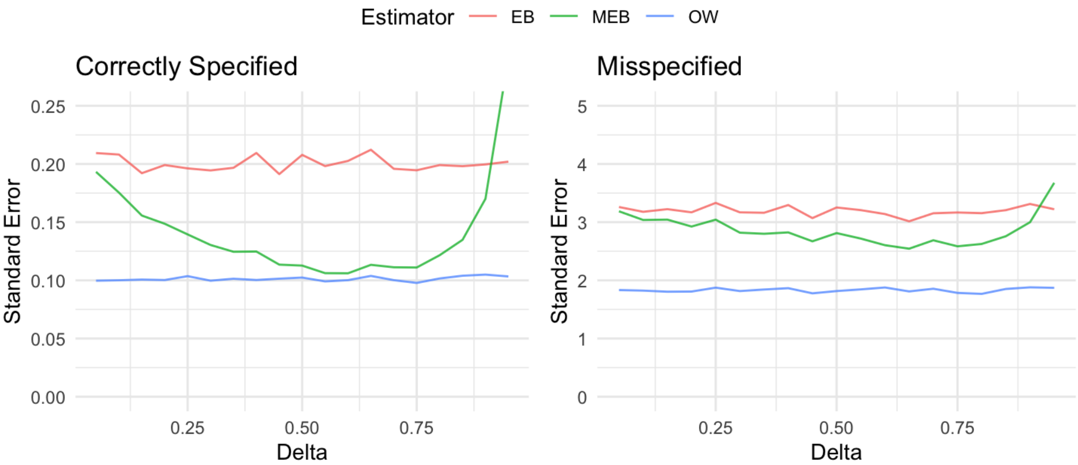


Figure 3: Monte Carlo simulation result: SD estimates of EB, MEB, OW

Results: EB vs MEB vs OW

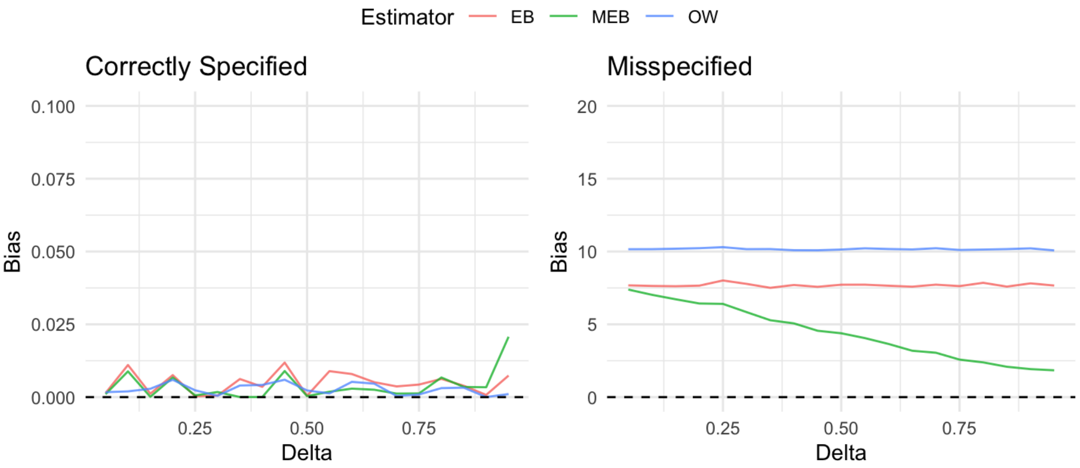


Figure 4: Monte Carlo simulation result: Finite-sample bias of EB, MEB, OW

Overview

Future Work

- **Heterogeneous mixing strategy**: What if we allow δ to vary according to the values of covariates?

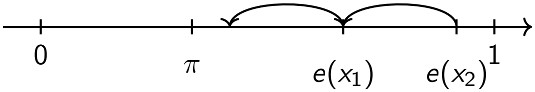


Figure 5: Homogeneous (Simple) Mixing: shrink with same ratio

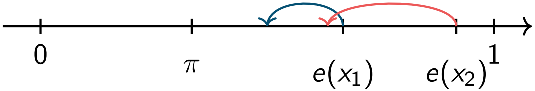


Figure 6: Heterogeneous (Advanced) Mixing: (blue) shrink less (red) shrink more

Future Work

- Primary results

Estimator	Bias	SD
IPW	0.1812	0.2656
Simple Mixing	0.1590	0.2440
Heterogeneous Mixing	0.1497	0.2232

Table 1: Advanced mixing strategy

- Other interesting topics remain, including application of mixing to matching methods.

Summary

Key Takeaways

Mixing: A simple & practical tool for handling weak overlap in causal estimation to control extremeness of inverse probability weights without additional assumptions

- **Performance**: Efficiency is enhanced without bias trade-off (even under sufficient overlap)
- **Straightforward interpretation**: No need to shift the target estimand
- **Flexibility**: Applicable to broad range of weighting methods
- **Open to further exploration**: Heterogeneous mixing strategy

Thank you!

Email: suehyunkim@snu.ac.kr
Preprint link: `https://arxiv.org/abs/2411.10801v3`