Comments on “How much weak overlap can doubly-robust t-statistics handle?”

Andreas Hagemann

Stephen M. Ross School of Business
University of Michigan

UNIVERSITY OF MICHIGAN

Why is this interesting?

- Inference on ATE when treatment is as good as randomly assigned cond. on covariates when few treated or control observations in parts of covariate space

Why is this interesting?

- Inference on ATE when treatment is as good as randomly assigned cond. on covariates when few treated or control observations in parts of covariate space
- “Weak overlap” or “limited overlap” induces propensity scores close to zero or one, makes asymptotic variances of ATE estimates increase sharply

Why is this interesting?

- Inference on ATE when treatment is as good as randomly assigned cond. on covariates when few treated or control observations in parts of covariate space
- “Weak overlap” or “limited overlap” induces propensity scores close to zero or one, makes asymptotic variances of ATE estimates increase sharply
- Khan and Tamer (2010, Ecma): “weak overlap” causes slower-than-usual convergence of estimators

Why is this interesting?

- Inference on ATE when treatment is as good as randomly assigned cond. on covariates when few treated or control observations in parts of covariate space
- “Weak overlap” or “limited overlap” induces propensity scores close to zero or one, makes asymptotic variances of ATE estimates increase sharply
- Khan and Tamer (2010, Ecma): “weak overlap” causes slower-than-usual convergence of estimators
- More subtle consequence of Khan and Tamer: confidence intervals of the form “point estimate $\pm$ 1.96 $\times$ se” may be impossible

Why is this interesting?

- Inference on ATE when treatment is as good as randomly assigned cond. on covariates when few treated or control observations in parts of covariate space
- “Weak overlap” or “limited overlap” induces propensity scores close to zero or one, makes asymptotic variances of ATE estimates increase sharply
- Khan and Tamer (2010, Ecma): “weak overlap” causes slower-than-usual convergence of estimators
- More subtle consequence of Khan and Tamer: confidence intervals of the form “point estimate $\pm$ 1.96 $\times$ se” may be impossible
- Discovery of nonexistent effects under weak overlap far more likely than α

Contribution of this paper

- Confidence intervals of the form “point estimate $\pm$ 1.96 $\times$ se” are possible even with a very standard estimator

Contribution of this paper

- Confidence intervals of the form “point estimate $\pm$ 1.96 $\times$ se” are possible even with a very standard estimator
- **Key insight of Dorn (2025)**: doubly-robust (“augmented”) inverse probability weighting (AIPW) can be thresholded just right to make this work

Contribution of this paper

- Confidence intervals of the form “point estimate $\pm$ 1.96 $\times$ se” are possible even with a very standard estimator
- **Key insight of Dorn (2025)**: doubly-robust (“augmented”) inverse probability weighting (AIPW) can be thresholded just right to make this work
- If things aren’t “just right”: bias and nonstandard asymptotics if threshold too small, bias from thresholding if bias too large

Contribution of this paper

- Confidence intervals of the form “point estimate $\pm$ 1.96 $\times$ se” are possible even with a very standard estimator
- **Key insight of Dorn (2025)**: doubly-robust (“augmented”) inverse probability weighting (AIPW) can be thresholded just right to make this work
- If things aren’t “just right”: bias and nonstandard asymptotics if threshold too small, bias from thresholding if bias too large
- **Contribution of Dorn (2025)**: Characterizes when things are just right under explicit regularity conditions and gives guidance on how to choose threshold

Quick aside: why AIPW and not regression?

- Treatment D is as good as randomly assigned cond. on covariates X,

$$Y(1), Y(0) \perp\!\!\!\perp D \mid X$$

Quick aside: why AIPW and not regression?

- Treatment D is as good as randomly assigned cond. on covariates X,

$$Y(1), Y(0) \perp\!\!\!\perp D \mid X$$

- With $Y = DY(1) + (1 - D)Y(0)$ and $Y(1) - Y(0)$ not constant, regression

$$Y = \beta_0 + \beta_1 D + g(X) + U$$

  requires unrealistic conditions on U to recover ATE

Quick aside: why AIPW and not regression?

- Treatment D is as good as randomly assigned cond. on covariates X,

$$Y(1), Y(0) \perp\!\!\!\perp D \mid X$$

- With $Y = DY(1) + (1 - D)Y(0)$ and $Y(1) - Y(0)$ not constant, regression

$$Y = \beta_0 + \beta_1 D + g(X) + U$$

  requires unrealistic conditions on U to recover ATE

- AIPW recovers ATE as long as we have **overlap**

$$0 < e(X) := P(D = 1 \mid X) < 1$$

Augmented inverse propensity score weighting

- Propensity score $e(X) := P(D = 1 \mid X)$, interested in $E[Y(1) - Y(0)]$

Augmented inverse propensity score weighting

- Propensity score $e(X) := P(D = 1 \mid X)$, interested in $E[Y(1) - Y(0)]$
- Can identify $EY(1)$ ($EY(0)$ is similar) as

$$\psi := EY(1) = E\left[\frac{D}{e(X)}Y\right]$$

Augmented inverse propensity score weighting

- Propensity score $e(X) := P(D = 1 \mid X)$, interested in $E[Y(1) - Y(0)]$
- Can identify $EY(1)$ ($EY(0)$ is similar) as

$$\psi := EY(1) = E\left[\frac{D}{e(X)}Y\right]$$

- Sample version

$$\hat{\psi} := \sum_{i=1}^{n}\left[\frac{D_i}{\hat{e}(X_i)}Y_i\right]$$

Augmented inverse propensity score weighting

- Propensity score $e(X) := P(D = 1 \mid X)$, interested in $E[Y(1) - Y(0)]$
- Can identify $EY(1)$ ($EY(0)$ is similar) as

$$\psi := EY(1) = E\left[\frac{D}{e(X)}Y\right]$$

- Sample version

$$\hat{\psi} := \sum_{i=1}^{n}\left[\frac{D_i}{\hat{e}(X_i)}Y_i\right]$$

- Doubly-robust (augmented) IPW uses flexible estimate $\hat{\mu}(X)$ of $E[Y \mid D = 1, X]$ for

$$\hat{\psi}_{\mathrm{DR}} := \sum_{i=1}^{n}\left[\frac{D_i}{\hat{e}(X_i)}Y_i - \left(\frac{D_i}{\hat{e}(X_i)} - 1\right)\hat{\mu}(X_i)\right]$$

What if propensity score is close to zero?

- Small $e(X)$ makes identification/estimation of $\psi = EY(1) = E\frac{D}{e(X)}Y$ difficult

What if propensity score is close to zero?

- Small $e(X)$ makes identification/estimation of $\psi = EY(1) = E\frac{D}{e(X)}Y$ difficult
- Many ways of addressing this. Clipping increases small $\hat{e}(X)$ to b_n,

$$\hat{\psi}(b_n) := \sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}Y_i\right]$$

What if propensity score is close to zero?

- Small $e(X)$ makes identification/estimation of $\psi = EY(1) = E\frac{D}{e(X)}Y$ difficult
- Many ways of addressing this. Clipping increases small $\hat{e}(X)$ to b_n,

$$\hat{\psi}(b_n) := \sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}Y_i\right]$$

- **Ma and Wang (2020, JASA)**: If $P(e(X) \leq \pi) \sim \pi^\gamma$ for $\gamma > 0$ and π near zero, smaller γ lead away from normal asymptotics

What if propensity score is close to zero?

- Small $e(X)$ makes identification/estimation of $\psi = EY(1) = E\frac{D}{e(X)}Y$ difficult
- Many ways of addressing this. Clipping increases small $\hat{e}(X)$ to b_n,

$$\hat{\psi}(b_n) := \sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}Y_i\right]$$

- **Ma and Wang (2020, JASA)**: If $P(e(X) \leq \pi) \sim \pi^\gamma$ for $\gamma > 0$ and π near zero, smaller γ lead away from normal asymptotics
- **Ma and Wang**: Letting $b_n \to 0$ slowly helps but causes bias. Propose debiasing and do subsampling inference, adapts to asymptotic regime.

Back to Dorn (2025)

- Some bias reduction is built into clipped doubly-robust (augmented) IPW

$$\hat{\psi}_{\mathrm{DR}}(b_n) := \sum_{i=1}^{n} \left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} Y_i - \left(\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} - 1 \right) \hat{\mu}(X_i) \right]$$

Back to Dorn (2025)

- Some bias reduction is built into clipped doubly-robust (augmented) IPW

$$\hat{\psi}_{\mathrm{DR}}(b_n) := \sum_{i=1}^{n} \left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} Y_i - \left(\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} - 1 \right) \hat{\mu}(X_i) \right]$$

- $\hat{\mu}(X_i)$ is flexible estimate of $E[Y \mid D = 1, X]$. Second term guards against bias from estimation of $e(X_i)$.

Back to Dorn (2025)

- Some bias reduction is built into clipped doubly-robust (augmented) IPW

$$\hat{\psi}_{\mathrm{DR}}(b_n) := \sum_{i=1}^{n} \left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} Y_i - \left(\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} - 1 \right) \hat{\mu}(X_i) \right]$$

- $\hat{\mu}(X_i)$ is flexible estimate of $E[Y \mid D = 1, X]$. Second term guards against bias from estimation of $e(X_i)$.

- If we clip small $\hat{e}(X_i)$ we cause bias. Can built-in bias correction in $\hat{\psi}_{\mathrm{DR}}$ help?

Back to Dorn (2025)

- Some bias reduction is built into clipped doubly-robust (augmented) IPW

$$\hat{\psi}_{\mathrm{DR}}(b_n) := \sum_{i=1}^{n} \left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} Y_i - \left(\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} - 1 \right) \hat{\mu}(X_i) \right]$$

- $\hat{\mu}(X_i)$ is flexible estimate of $E[Y \mid D = 1, X]$. Second term guards against bias from estimation of $e(X_i)$.

- If we clip small $\hat{e}(X_i)$ we cause bias. Can built-in bias correction in $\hat{\psi}_{\mathrm{DR}}$ help?

- **Dorn**: b_n needs to be "big" to restore normality (standard confidence intervals) but "small" to not change estimand. Such b_n exist!

Main result

$$\hat{\psi}_{\mathrm{DR}}(b_n) := \sum_{i=1}^{n} \left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} Y_i - \left(\frac{D_i}{\max\{\hat{e}(X_i), b_n\}} - 1 \right) \hat{\mu}(X_i) \right]$$

Informal statement of Dorn's main result

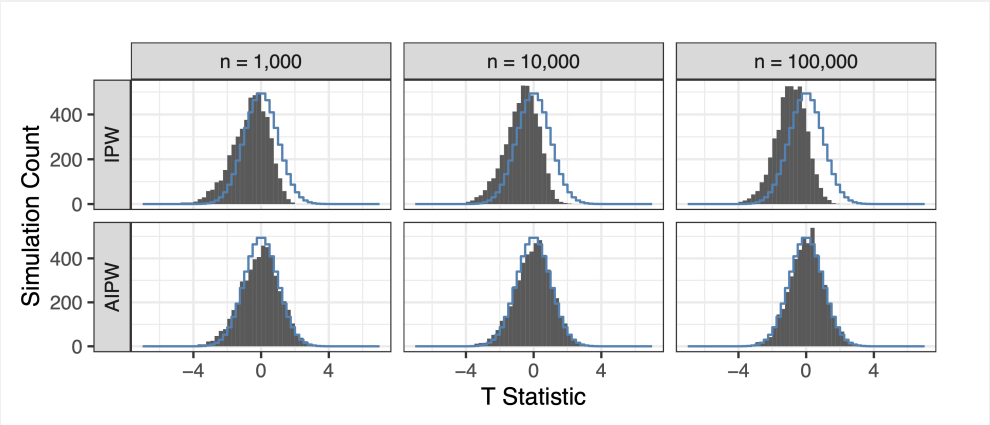
If $e(X)$ and $E[Y \mid T = 1, X]$ are smooth enough and even if overlap is weak (but not too weak), then there are local polynomial estimators of $e(X)$ and $E[Y \mid D = 1, X]$, and a clipping rate b_n that depend on smoothness (but not on the precise nature of overlap) such that

$$(\hat{\psi}_{\mathrm{DR}}(b_n) \pm 1.96 \times \mathrm{se})$$

is a 95% confidence interval as $n \to \infty$.

Formal statement and proof are highly nontrivial

Simulation inspired by Ma and Wang (2020, JASA)



Comments, questions, personal opinions

- Robustness? Good estimates of all quantities are essential, reliance on good guesses of smoothness of at least two functions, tail behavior, and b_n.

Comments, questions, personal opinions

- Robustness? Good estimates of all quantities are essential, reliance on good guesses of smoothness of at least two functions, tail behavior, and b_n.
- Symmetry? Not clear that the same conditions need to apply to $EY(0)$.

Comments, questions, personal opinions

- Robustness? Good estimates of all quantities are essential, reliance on good guesses of smoothness of at least two functions, tail behavior, and b_n.
- Symmetry? Not clear that the same conditions need to apply to $EY(0)$.
- Would normalizing the weights to sum to 1 help stabilize the estimate? Such as,

$$\sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}Y_i\right] \Big/ \sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}\right]$$

Comments, questions, personal opinions

- Robustness? Good estimates of all quantities are essential, reliance on good guesses of smoothness of at least two functions, tail behavior, and b_n.
- Symmetry? Not clear that the same conditions need to apply to $EY(0)$.
- Would normalizing the weights to sum to 1 help stabilize the estimate? Such as,

$$\sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}Y_i\right] \Big/ \sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}\right]$$

- Not specific to this paper but: why are we going after and what are we hoping to learn from populations with almost no control group?

Comments, questions, personal opinions

- Robustness? Good estimates of all quantities are essential, reliance on good guesses of smoothness of at least two functions, tail behavior, and b_n.
- Symmetry? Not clear that the same conditions need to apply to $EY(0)$.
- Would normalizing the weights to sum to 1 help stabilize the estimate? Such as,

$$\sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}Y_i\right] \Big/ \sum_{i=1}^{n}\left[\frac{D_i}{\max\{\hat{e}(X_i), b_n\}}\right]$$

- Not specific to this paper but: why are we going after and what are we hoping to learn from populations with almost no control group?
- Not specific to this paper but: large sample solution to a small sample problem.

Concluding remarks

- Strong theoretical results with a clear view towards empirical practice
- Immediately implementable through rules of thumb for bandwidth backed by theoretical results
- Thoughtful, well motivated, and exceptionally well written paper