Efficient Difference-in-Differences and Event-Study Estimators

Xiaohong Chen
Yale University

Pedro H. C. Sant'Anna
Emory University

Haitian Xie
Peking University

3rd Workshop on Causal Inference and Machine Learning in Practice
August 4, 2025

Causal inference with observational data: What can we do?

- In many real-world applications, we do not have experimental data and need to rely on **observational data**.
- With observational data alone, we have no choice but to rely on **additional assumptions** for the identification of causal parameters.
- Different causal estimation and inference methods—such as those based on unconfoundedness, DiD, IV, RDD, Synthetic Control, etc.—rely on **different identification assumptions**.
- Our job as scientists is to assess the pros and cons of each method in their ability to answer the questions we (and the business) care about.

Causality with Observational Data: What can we do?

- In the last 10 years, we have seen a big jump in using ML methods to answer causal questions.
- What do people usually mean when they say that they use a **double machine learning** strategy?
- Rely on unconfoundedness/selection on observables.
- Drawback: Rule out selection on unobservables.
  We need to have data on all confounders, i.e., on everything that (jointly) affects treatment take-up and the outcomes of interest.
- In practice, many science teams question the plausibility of this condition
  (but not much is usually done to address it, especially in industry)

The appeal of Difference-in-Differences: allow for selection on unobservables

- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.

- **Advantage: Allow for selection on unobservables and for time-trends.**
  We need to assume that, absent the treatment and conditional on covariates (features), the outcome of interest would grow similarly across groups/cohorts - **a parallel trends assumption**.

- **Data Requirements:** We need data from periods before and after the treatment/intervention to use DiD (and some periods where no unit is treated).

The appeal of Difference-in-Differences: allow for selection on unobservables

- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.

- Advantage: Allow for selection on unobservables and for time-trends.
  We need to assume that, absent the treatment and conditional on covariates (features), the outcome of interest would grow similarly across groups/cohorts - a parallel trends assumption.

- Data Requirements: We need data from periods before and after the treatment/intervention to use DiD (and some periods where no unit is treated).

This talk: How to combine DiD with ML to get semiparametric efficient estimators

The appeal of Difference-in-Differences: allow for selection on unobservables

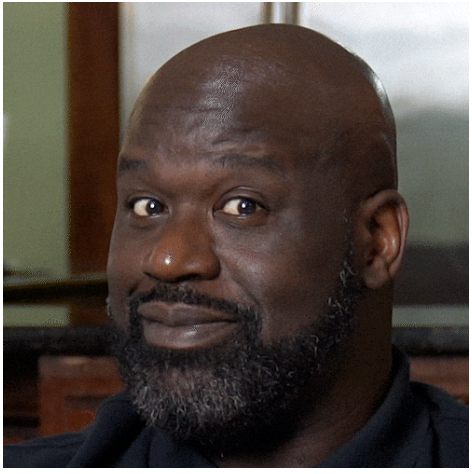
- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.

- Advantage: Allow for selection on unobservables and for time-trends.
  We need to assume that, absent the treatment and conditional on covariates (features), the outcome of interest would grow similarly across groups/cohorts - a parallel trends assumption.

- Data Requirements: We need data from periods before and after the treatment/intervention to use DiD (and some periods where no unit is treated).

This talk: How to combine DiD with ML to get semiparametric efficient estimators

Paper: Many more results, with a much greater amount of details.



Practical Takeway Messages

Our practical takeaways messages

- DiD models are usually nonparametrically overidentified:
  off-the-shelve DML using $Y = Y_{t=\text{post}} - Y_{t=\text{pre}}$ as outcome is generally not semipar. efficient
- Our EIFs are available in closed form and are automatically orthogonal moments:
  - Provide a blueprint for efficient DML estimators
- Gains in efficiency can be of first-order according to our simulations and empirical application
- Optimal to weight pre-treatments and comparison groups in a non-uniform manner
- When covariates are required for identification, the efficient weights also depend on covariate values.
  - For example, optimal weights for men may differ from those for women, as the outcomes for these two covariate groups may have heterogeneous correlations over time.

DiD with multiple periods

DiD Framework

- "Short" panel data $\{(Y_{i,t=1}, \ldots, Y_{i,t=T}, X_i', G_i)'\}_{i=1}^n$ for n large and T finite fixed.
- Treatment is binary, may have different starting dates, and is an absorbing state.
  - Treatment does not "turn off", as we do not impose assumptions on how long its effect can last
- $Y_{i,t}(\mathbf{0_{g-1}}, \mathbf{1_{T-g+1}}) \equiv Y_{i,t}(g)$: Potential outcomes indexed by treatment starting date g.
  - Potential outcome for unit i and time t if they were treated for the first time in period g.
- G_i denotes the time unit i is first-treated, with $G_i = \infty$ if they stay untreated by time T.
  - $G_i \in \mathcal{G} \subseteq \{2, \ldots, T, \infty\}$: no unit treated in period $t = 1$
  - With single treatment date at period $t = g$, $G_i = g$ (treated units) or $G_i = \infty$ (untreated units).
  - $\mathcal{G}_{\text{trt}} = \mathcal{G} \setminus \{\infty\}$ denote the support of G among eventually treated units.
  - Let $G_g = 1\{G = g\}$.
- Due to time constraints, I will focus on the two-group setup.

Target Parameters

- The group-time average treatment effects for the treated:

$$ATT(g,t) := \mathbb{E}[Y_t(g) - Y_t(\infty)|G = g]$$

Average treatment effect at time period t of being first treated in period g compared to never being treated, among units that indeed start treatment at time g.

- $ATT(g,t)$: can track down how treatment effect varies with elapsed treatment time (aka event-time), $e = t - g$.

Maintained Assumption: Sampling, Overlap, and No-anticipation

Assumption (Maintained Assumption (M))

(i) *(S)* $\{(Y_{i,t=1}, \ldots, Y_{i,t=T}, X_i', G_i)'\}_{i=1}^n$ *is a random sample from* $(Y_{t=1}, \ldots, Y_{t=T}, X', G)'$.

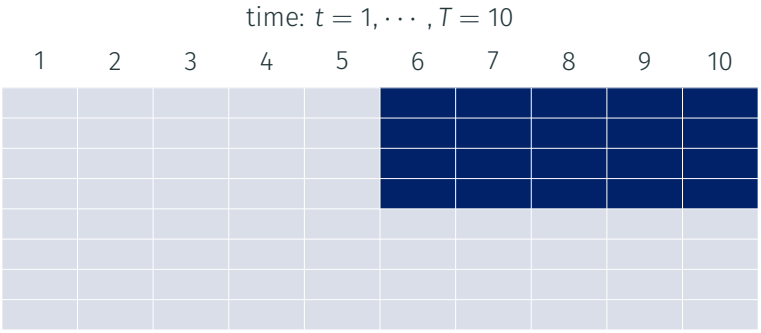
(ii) *(O) For each* $g \in \mathcal{G}$, $\mathbb{E}[G_g|X] \in (0,1)$ *almost surely (a.s.).*

(iii) *(NA) For every* $g \in \mathcal{G}_{trt}$, *and every pre-treatment periods* $t < g$, $\mathbb{E}[Y_{i,t}(g)|G = g, X] = \mathbb{E}[Y_{i,t}(\infty)|G = g, X]$ *almost surely.*

- The maintained Assumption M is not enough to identify ATT or ES type parameters.
- DiD methods impose parallel trends assumptions to identify these parameters

No variation in treatment timing

- Single treatment period at time g: $G_i = g$ (treated) or $G_i = \infty$ (untreated).
- $ES(e) = ATT(g, g + e)$.



Parallel Trends Assumption: All periods

Assumption (PT in all periods)

For each $t \in \{2, \ldots, T\}$,

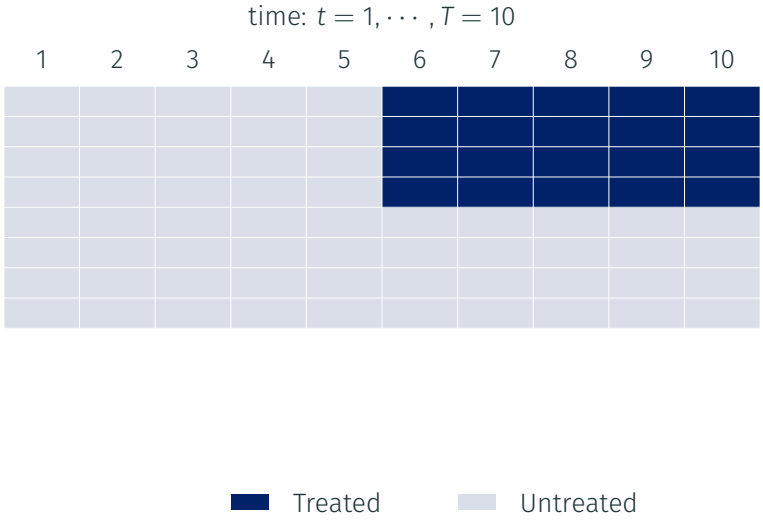
$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = g, X] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|G = \infty, X] \text{ a.s.}$$

- Impose parallel trends in all periods.
- Allows us to use any pre-treatment period as a baseline.
- Without covariates, easy to show that for any $t \geq g$ and any $t' < g$,

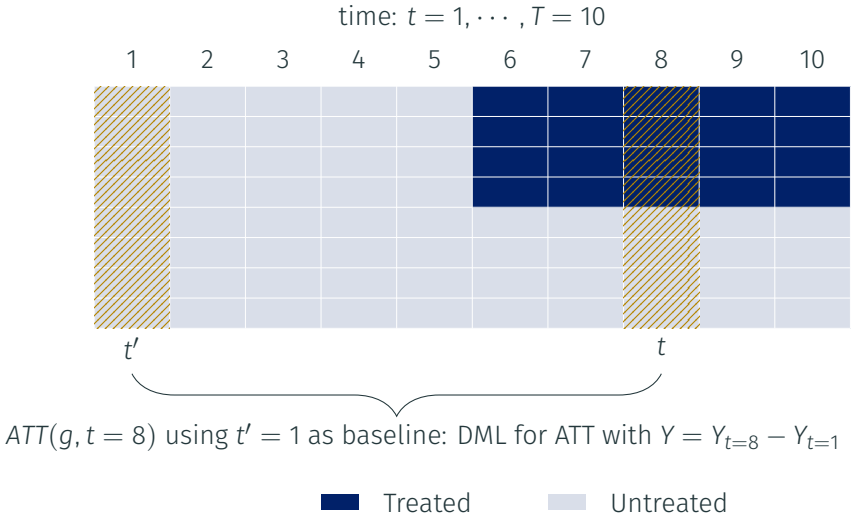
$$ATT(g,t) = \mathbb{E}[Y_t - Y_{t'}|G = g] - \mathbb{E}[Y_t - Y_{t'}|G = \infty].$$

- If we were gifted 10 more pre-treatment periods of data, we could easily use all of them to compute $ATT(g,t)$.
- How to do that efficiently, such that we maximize precision?

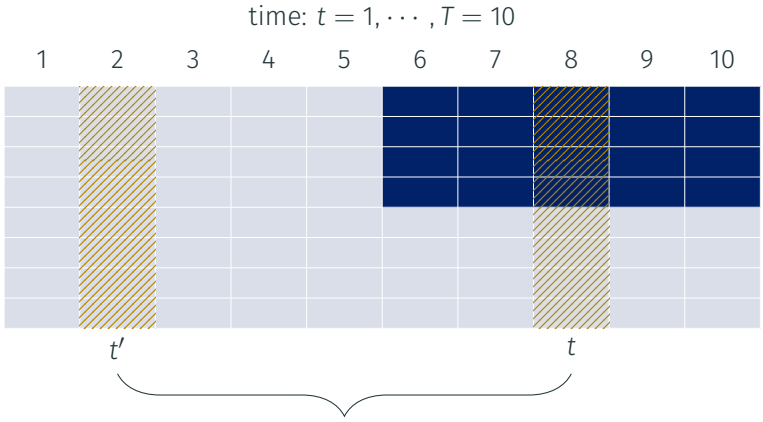
Understanding the sources of over-identification



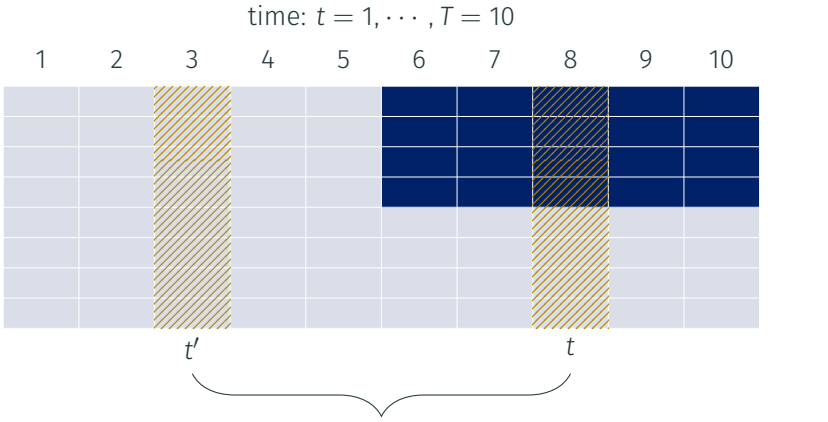
Understanding the sources of over-identification



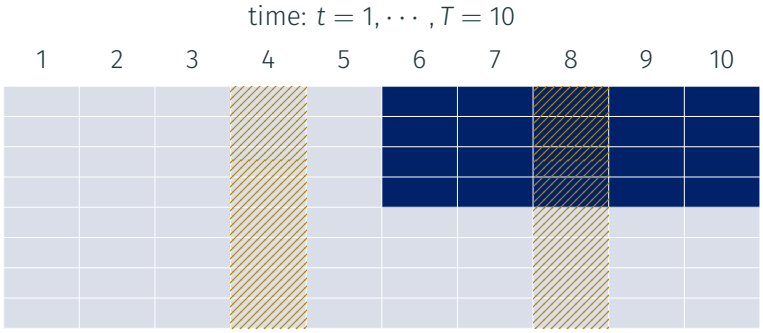
Understanding the sources of over-identification



Understanding the sources of over-identification



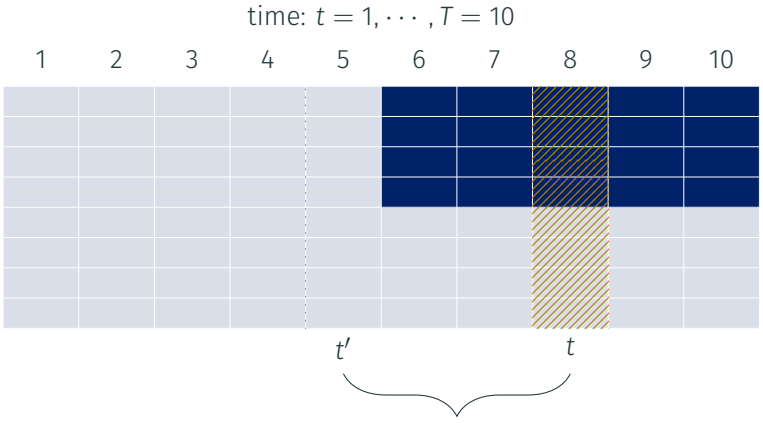
Understanding the sources of over-identification



$ATT(g, t = 8)$ using $t' = 4$ as baseline: DML for ATT with $Y = Y_{t=8} - Y_{t=4}$

Treated Untreated

Understanding the sources of over-identification



$ATT(g, t = 8)$ using $t' = 5$ as baseline: DML for ATT with $Y = Y_{t=8} - Y_{t=5}$

Treated Untreated

Intuition on how to get semiparametric efficiency

- For each $1 \leq t' \leq g-1$, we fix the baseline period at t', and compute the "efficient influence function" for $ATT(g,t)$ as-if there were only 2 groups, $G = g$ and $G = \infty$, and two periods, t (post-treatment) and t' (pre-treatment)
  - Akin to compute the "DR scores" in DML language.
- Stack all the non-collinear influence functions into a vector, $\mathbb{IF}^{att(g,t)}$.
- Compute the covariance of $\mathbb{IF}^{att(g,t)}$ given covariates, $V_{gt}(X) = \text{Cov}\left(\mathbb{IF}^{att(g,t)}|X\right)$.
- Efficient Influence Function for $ATT(g,t)$ is given by

$$EIF^{att(g,t)} = \frac{1'V_{gt}(X)^{-1}}{1'V_{gt}(X)^{-1}1}\mathbb{IF}^{att(g,t)}.$$

- Next, we explore these results to obtain EIF-based estimands for $ATT(g,t)$, which serve as a blueprint for efficient estimation.

Using EIF as a blueprint for estimating ATT(g,t)

- The key is to explore that $\mathbb{E}[EIF^{att(g,t)}] = 0$ to get IF-based estimand:

$$ATT(g,t) = \mathbb{E}\left[\frac{\mathbf{1}' V_{gt}(X)^{-1}}{\mathbf{1}' V_{gt}(X)^{-1}\mathbf{1}} \theta_{g,t}(W)\right], \tag{1}$$

where $p_g(X) = \mathbb{E}[G_g|X]$, $\theta_{g,t}(W) = (\theta_{g,t,1}(W), \ldots, \theta_{g,t,g-1}(W))'$ is a $(g-1)\times 1$ column vector with

$$\theta_{g,t,t'}(W) = \frac{1}{\mathbb{P}(G=g)}\left(G_g - \frac{(1-G_g)p_g(X)}{1-p_g(X)}\right)(Y_t - Y_{t'} - \mathbb{E}[Y_t - Y_{t'}|G=\infty, X]).$$

- Here, $\theta_{g,t}(W)$ is a vector of DR DiD "integrands", each being computed pretending we were in the 2×2 DiD setup of Sant'Anna and Zhao (2020).

- Efficient DiD estimators: apply plug-in principle, or do DML.

Understanding the weights across pre-treatment periods

- The optimal way to aggregate across pre-treatment periods to learn about $ATT(g,t)$ is given by the inverse of $V_{gt}(X)$, the conditional covariance of the influence functions.
- This is OK, but it does not really help us better understand these weights.
- In the paper, we show that one can use the "easier-to-understand" matrix $V^*_{gt}(X)$ to form the optimal weights, where $dim(V^*_{gt}(X)) = dim(V_{gt}(X))$ and the (j,k)-th element being $V^*_{gt}(X)$ is given by

$$\frac{1}{p_g(X)} \text{Cov}(Y_t - Y_j, Y_t - Y_k | G = g, X) + \frac{1}{1 - p_g(X)} \text{Cov}(Y_t - Y_j, Y_t - Y_k | G = \infty, X). \tag{2}$$

Some important remarks about the efficient weights

- Efficient weights depend on the covariance of outcome changes for each treatment group;
- Efficient weights vary with the time of the ATT of interest, t.
- Efficient weights are covariate-dependent (different from GMM)
- Efficient weights are generally not constant across pre-treatment periods.
- EIF is obtained by optimally weighting the individual IFs
- Most DiD and ES estimators are not semiparametrically efficient:
  - Either only looks at a single pre-treatment t':
    DTWE, de Chaisemartin and D'Haultfœuille (2020, 2024), Callaway and Sant'Anna (2021), Sun and Abraham (2021).
  - or taking simple average over $t' < g$:
    TWFE, Wooldridge (2021), Gardner (2021),Borusyak, Jaravel and Spiess (2024), Lee and Wooldridge (2023)

An Empirical Application

What is causal effect of hospitalization on out-of-pocket medical spending?

- Dobkin, Finkelstein, Kluender and Notowidigdo (2018) study the effect of hospitalization on out-of-pocket medical spending and several other outcomes.
- They explore the variation in hospitalization time across individuals and use a DiD and ES strategy to estimate the causal effects of hospitalization.
- We follow the sample construction of Sun and Abraham (2021), which explores the publicly available survey data from the Health and Retirement Study (HRS) from the replication package of Dobkin et al. (2018):
  - Adults hospitalized at ages 50–59, excluding pregnancy-related admissions.
  - Balanced panel spanning waves 7–11 (2004–2012).
  - Final Sample: 652 individuals in 4 treatment groups
  - $G_i = 8$ (252), $G_i = 9$ (176), $G_i = 10$ (163), $G_i = \infty$ (65; re-labeled the $G_i = 11$ as "never-treated").

Parameters of Interest and Different DiD Estimators

- Parameters of Interest: Post-treatment effects $ATT(g,t)$ and aggregated event-study measures $ES(e)$, ES_{avg}.
- Estimates calculated using multiple DiD methods:
  - Our Efficient DiD estimator (EDiD)
  - Callaway and Sant'Anna (2021) and Sun and Abraham (2021) estimator using never-treated as comparison group (CS-SA)
  - Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2020) estimators using not-yet-treated s comparison group (CS-dCDH)
  - Borusyak et al. (2024), Gardner (2021), and Wooldridge (2021) imputation estimators (BJS-G-W).
- As PT is plausible in this context (Dobkin et al., 2018 and Sun and Abraham, 2021), all point estimates are expected to be similar.
- Our efficient DiD estimator is expected to deliver gains in efficiency.

Point estimates are indeed similar across estimators

Estimator	$ATT(8,8)$	$ATT(8,9)$	$ATT(8,10)$	$ATT(9,9)$	$ATT(9,10)$	$ATT(10,10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	3072	1112	1038	3063	90	2908	3024	692	1038	1585
	(806)	(637)	(817)	(690)	(641)	(894)	(486)	(471)	(816)	(521)
CS-SA	2826	825	800	3031	107	3092	2960	530	800	1430
	(1035)	(909)	(1008)	(702)	(651)	(995)	(539)	(585)	(1008)	(647)
CS-dCDH	3029	1248	800	3324	107	3092	3134	779	800	1571
	(913)	(861)	(1008)	(959)	(651)	(995)	(536)	(570)	(1008)	(566)
BJS-G-W	3029	1285	1021	3239	77	2758	3017	788	1021	1609
	(916)	(767)	(851)	(862)	(729)	(957)	(555)	(587)	(851)	(582)

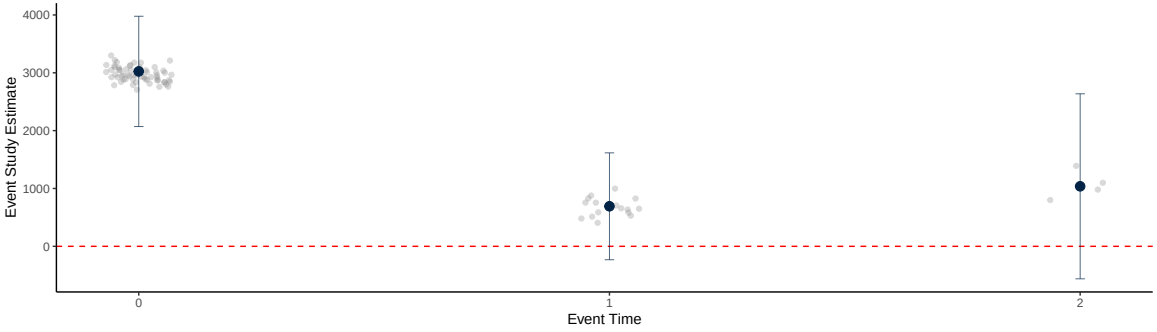
Our Efficient DiD provides substantial gains in efficiency

- Estimates of the asymptotic relative efficiency (ARE) of our proposed efficient DiD estimator with respect to the other available DiD estimators.
- Heuristically, ARE provides a relative measure of sample size needed for other DiD estimators to achieve the same precision as our efficient DiD estimator.

Estimator	$ATT(8,8)$	$ATT(8,9)$	$ATT(8,10)$	$ATT(9,9)$	$ATT(9,10)$	$ATT(10,10)$	$ES(0)$	$ES(1)$	$ES(2)$	ES_{avg}
EDiD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CS-SA	1.65	2.04	1.52	1.04	1.03	1.24	1.23	1.54	1.52	1.54
CS-dCDH	1.28	1.83	1.52	1.93	1.03	1.24	1.21	1.46	1.52	1.18
BJS-G-W	1.29	1.45	1.09	1.56	1.29	1.15	1.30	1.55	1.09	1.25

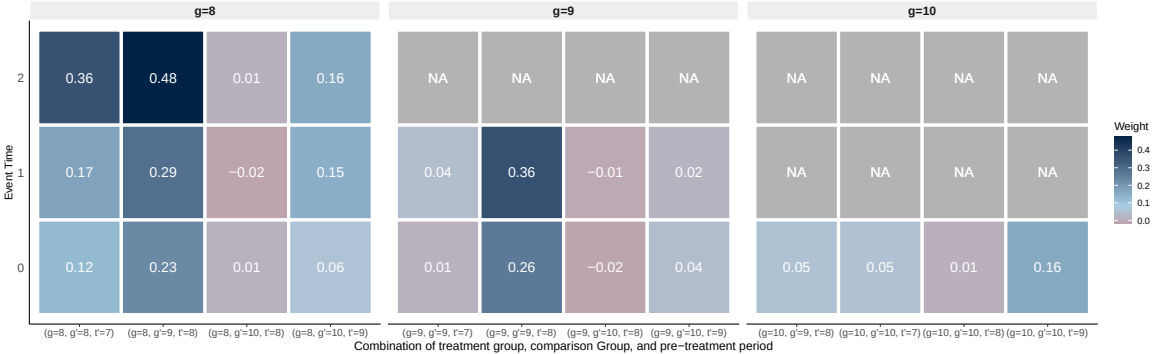
- Not possible to rank the other DiD estimators in terms of ARE in the application:
  - BJS-G-W: Efficient under homoskedasticity and no residual serial correlation (Wooldridge, 2021; Borusyak et al., 2024).

Visualizing Event-Study stability



Understanding the Efficient Weights for ATT(g,t)'s

Understanding the Efficient Weights for ES(e)'s



Conclusion

Conclusion

- DiD models are usually nonparametrically overidentified:
  off-the-shelve DML using $Y = Y_{t=\text{post}} - Y_{t=\text{pre}}$ as outcome is generally not semipar. efficient
- Using our EIF as a blueprint leads to semiparametrically efficient estimators: other implementations do not have this strong statistical guarantee
- Gains in efficiency can be of first-order according to our simulations and empirical application
- Optimal to weight pre-treatments and comparison groups in a non-uniform manner
- We will work on providing code to automate all this for you!

Thanks!

pedro.santanna@emory.edu

psantanna.com

@pedrohcgs

@pedrosantanna.bsky.social

References

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

__ and __, "Difference-in-Differences Estimators of Intertemporal Treatment Effects," *The Review of Economics and Statistics*, 2024, *Forthcoming*, 1–45.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo, "The economic consequences of hospital admissions," *American Economic Review*, 2018, *108* (2), 308–352.

Gardner, John, "Two-stage differences in differences," *Working Paper*, 2021.

Lee, Soo Jeong and Jeffrey M. Wooldridge, "A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data," *Working Paper*, 2023. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4516518.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Wooldridge, Jeffrey M, "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," *Working Paper*, 2021, pp. 1–89.