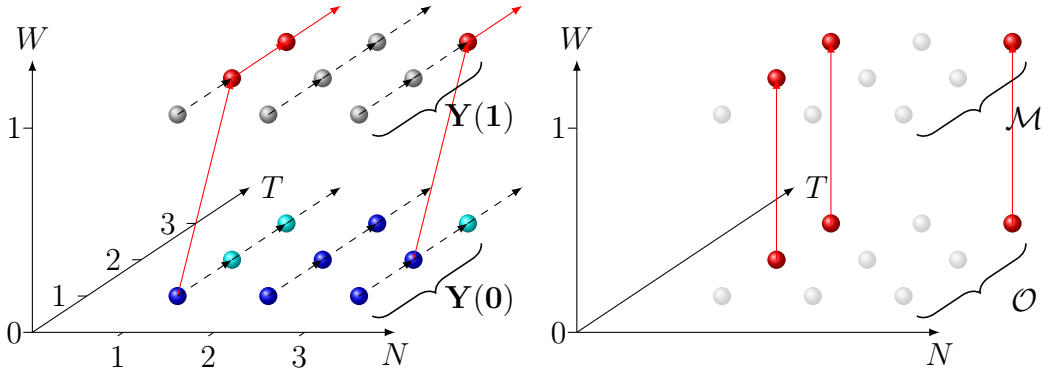
Matrix Completion Estimation in Differential Timing Settings

Tobias Schnabel

June 28, 2023



Academic Year 2022/23
Bachelor Thesis in the B.Sc. Econometrics & Operations Research program
Maastricht University School of Business and Economics
Supervisor: Prof. Martin Schumann

Abstract

Athey et al. (2021) introduce the method of Matrix Completion Estimation from the Statistics and Computer Science literatures to Panel Data Econometrics. They show that it outperforms the canonical Difference-in-Differences estimator in a variety of panel data settings, including ones with differential timing of treatment adoption. They also highlight this new method's reliance on only two identifying assumptions, SUTVA and exogeneity, which are both also needed for canonical DiD. This stands in contrast to the minimum of four additional necessary assumptions for canonical DiD. This thesis extends their work by comparing this new method not only to the canonical DiD estimator, but also to prominent methods developed in recent years to address the shortcomings of canonical DiD specifically in differential timing settings. I motivate and describe these estimators in a unified framework before comparing their performance in eight Monte Carlo simulations. My results show that the Matrix Completion Estimator shows promising potential in estimating treatment effects in such settings, but does not outperform all new DiD estimators in all eight simulations. I conclude that the Matrix Completion Estimator is preferable to canonical DiD in all settings I investigate, but not universally preferable to all recent DiD methods.

Acknowledgements

I would like to thank Martin Schumann for his close and thoughtful supervision and Guido Imbens for clarifying comments. Further, I am thankful to Florian Knäple, Michael Haas, and Kyra Pauly for proofreading. All remaining errors are mine alone.

Contents

1 Introduction **6**

2 The Rubin Causal Model **9**

2.1 Potential Outcomes and Causal Effects 9

2.2 Inference on Treatment Effects as the Ultimate Estimands 11

3 Causal Inference in Difference-in-Differences Designs **14**

3.1 The Canonical Design Estimator 14

3.2 Estimands in Richer Designs: the *Problem* 16

4 Difference-in-Differences With Differential Timing **16**

4.1 Differential Timing Setup 17

4.2 A Note on *Heterogeneity* and *Dynamics* 19

4.3 New Estimators for Staggered Designs 20

4.4 The Paradox Of Choice for The Practitioner 27

4.5 Estimands in Staggered Designs 29

5 A New Paradigm: Matrix Completion **30**

5.1 The Ultimate Estimand of Matrix Completion 30

5.2 Matrix Completion Setup 32

5.3 Connecting Matrix Completion to Difference-in-Differences 33

5.4 The Immediate Estimand: Low-Rank Matrix Approximations 36

5.5 The Matrix Completion-Nuclear Norm Minimization Estimator 40

5.6 The MC-NNM Estimator With Covariates 44

5.7 Interpreting The Ultimate Estimand 46

6 Performance Comparisons **48**

6.1 Computational Implementations 48

6.2 Data Generation and Simulation Setup 49

6.3 Estimation Process and Ultimate Estimands 52

6.4 Results 53

6.5 Limitations and Extensions 70

7 Conclusion **71**

8 References **74**

9 Main Software Packages Used **82**

A Appendix **83**

A.1 Computation of True Parameters 83

A.2 Supplementary Results 87

A.3 Correspondence 91

A.4 Thesis Proposal 93

A.5 Official Statement Of Original Bachelor Thesis 94

List of Figures

1 Unit Definition 9

2 Potential Outcomes 10

3 Unit-Specific Treatment Effect 10

4 Average Treatment Effect on the Treated 13

5 Imputing Missing Potential Outcomes 31

6 MCE Data-Generating Process, No Covariates 36

7 MCE Data-Generating Process, With Covariates 46

8 Results from Simulation 1 55

9 Results from Simulation 2 57

10 Results from Simulation 3 59

11 Results from Simulation 4 61

12 Results from Simulation 5 63

13 Results from Simulation 6 65

14 Results from Simulation 7 67

15 Results from Simulation 8 69

List of Tables

1 Estimator Comparison 28

2 Estimands in Staggered Designs 28

3 Simulation 1, Point Estimates of τ 54

4 Simulation 2, Point Estimates of τ^{ES} 56

5 Simulation 3, Point Estimates of τ 58

6 Simulation 4, Point Estimates of τ 60

7 Simulation 5, Point Estimates of τ^{ES} 62

8 Simulation 6, Point Estimates of τ^{ES} 64

9 Simulation 7, Point Estimates of τ^{ES} 66

10 Simulation 8, Point Estimates of τ^{ES} 68

11 Simulation 1, Point Estimates of τ^{ES} 87

12 Simulation 2, Point Estimates of τ 87

13 Simulation 3, Point Estimates of τ^{ES} 88

14 Simulation 4, Point Estimates of τ^{ES} 88

15 Simulation 5, Point Estimates of τ 88

16 Simulation 6, Point Estimates of τ 89

17 Simulation 7, Point Estimates of τ 89

18 Simulation 8, Point Estimates of τ 90

1 Introduction

Beginning in the 1980s, economists and econometricians revised their approach to causal inference in Applied Economics and Policy research. As we typically do not have the luxury of conducting well-controlled randomized controlled trials akin to those used in the natural sciences, more often than not, attempts at drawing causal inference on the economic effects of policies and decisions made by people and firms rely on *quasi-experimental* research designs. The most prominent methods used to implement such designs are instrumental variables, regression discontinuity, and Difference-in-Differences (DiD).

In what over time became known as the *credibility revolution*, econometricians have devoted increasing amounts of time and attention to the exact workings, necessary assumptions, and practical shortcomings of these methods (Angrist and Pischke 2010). In recent years, new discoveries about the statistical properties of the storied DiD estimator in settings where treatments are assigned to different units at different times, commonly referred to as *staggered* or *differential timing* designs, have gained substantial attention among applied microeconomists and econometricians alike. Baker, Larcker, and Wang (2022) observe that between 2000 and 2019, there were 744 papers using DiD published in the respective top five journals in the fields of Finance and Accounting, more than half of which use differential timing designs. The prevalence of DiD is even higher in the Policy literature: 10 out of 16 articles in the current 15$^{\text{th}}$ issue of *AEJ: Economic Policy* use DiD methods (American Economic Association 2023).

The econometric literature on DiD is evolving at a rapid pace. This is problematic for applied researchers, as following even the major developments in DiD best practices requires substantial time and attention on their part. Not keeping up to date with the state of the art is also hardly an option: practitioners run the risk of using *outdated* or *ill-suited* methods, such as TWFE DiD in the presence of treatment heterogeneity. In short, DiD methods lie on the top shelf of the modern econometric toolbox. They are, however, dependent on several key assumptions, which are not always explicitly testable (cf. a.o. Dette and Schumann 2020; Roth et al. 2023; Roth 2022).

In 2017, Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi published the preprint of a paper that in 2021 was published in the Journal of the American Statistical Association (Athey et al. 2021). They introduce the method of Matrix Completion Estimation (MCE) from the Machine Learning, Statistics and Computer Science literatures to Panel Data Econometrics. Among several distinct contributions, they use simulation methods to show that this new estimator estimates treatment effects more accurately than the canonical DiD estimator in a variety of panel data settings, including ones with differential timing of treatment adoption. Even more intriguingly, they also highlight this new method's reliance on only two identifying assumptions, SUTVA and exogeneity, and notably not the *parallel* or *common trends* assumption which all DiD methods hinge on.

As such, this new estimator could potentially be an all-in-one, *off-the-shelf* solution to many of the problems with DiD methods. Athey et al. (2021) do not explicitly claim this, but their discussion of the advantages of MCE does open the door to such speculation. If this new method were indeed simpler and more broadly applicable, while at the same time being as precise or more so than state-of-the-art DiD methods, it would be nothing short of a revolution in the Applied Microeconomics and Policy Evaluation fields. This thesis investigates whether such a strong interpretation of Athey et al. (2021)s' claimed advantages of Matrix Completion estimation is just a bit *too good to be true*. Athey et al. (2021) only cover [TWFE] as an afterthought, which presents a gap in the literature in understanding how, and how well, [MC-NNM] works compared to DiD methods in differential timing settings with unit- and/or time-heterogeneous treatment effects.

In a unified framework, characterized by common notation and estimands, which I introduce to articulate the mechanics and properties of each estimator, I conduct eight Monte Carlo simulations. Within these simulations, the behavior of the canonical DiD estimator is known and often problematic. I then evaluate and compare the relative performance of the Matrix Completion estimator to canonical and new DiD methods. I find that while MCE performs well across most of these simulations, it does not universally outperform all other methods. It does, however, perform better than the canonical DiD estimator. This leads me to conclude that MCE represents a good default estimator practitioners should consider instead of the canonical DiD

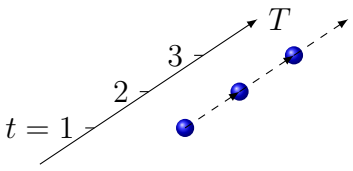
estimator, but it cannot unilaterally replace new methods tailored to certain settings. I argue that MCE should be included as a standard benchmark in applied analyses to detect potential issues with DiD estimators and alert practitioners to the possibility that their model might be misspecified.

In the section 2, I introduce the causal framework that belies both DiD and MCE. In sections 3 and 4, I synthesize the recent literature on the problems with classical DiD methods and new, alternative estimators. In section 5, I introduce the new Matrix Completion estimator, connect its motivation to that of DiD, and explain in detail how and why it works, as well as what it estimates. I describe the eight Monte Carlo simulations I carry out to assess the performance of MCE vis-à-vis recent DiD estimators and their results in section 6. Section 7 concludes.

2 The Rubin Causal Model

Empirically investigating the causal effect of an action on a unit of observation necessitates several precise definitions. The causal framework that underpins most of the popular econometric methods used in applied microeconomics is the *Rubin Causal Model* (RCM). As Imbens and Rubin (2015) explain, the central notion in this framework is that "*causality is tied to an action (or manipulation, treatment, or intervention), applied to a unit*" (p. 4). This model specifies units, that is, physical and thus observable objects, to be different when observed at different points in time. This is to allow for the possibility that time matters, which it tends to do for the subjects of the social science literature: a rock sitting on a workbench may not substantively change over time, but human preferences and (economic) behavior usually does. To correctly define causal effects, we have to be careful not to "*confuse assumptions (e.g., about similarities between different units), with definitions (e.g., of a unit, or of a causal effect)*"(p. 8). For an outcome of interest $Y_{i,t}$ with $t = 1, \ldots, T$, where i denotes the unit, we therefore have a time path like the one depicted in Figure 1 below:

Figure 1: Unit Definition

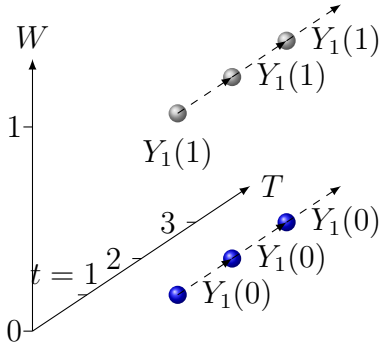


2.1 Potential Outcomes and Causal Effects

In this framework, a "*causal statement presumes that, although a unit was (at a particular point in time) subject to, or exposed to, a particular action, treatment, or regime, the same unit could have been exposed to an alternative action, treatment, or regime (at the same point in time)*" (p. 4). This modelling choice has led to the RCM's more popular moniker as the *Potential Outcomes Model*, because it allows us to associate with each action-unit pair a *potential outcome* (PO). This means that at each point in time, a unit has one out of several possible actions applied to it. In most applications, there are two actions: a *treatment* that is either applied or not applied to a unit. Popular targets of applied economic research model policy choices such as minimum

wage laws or job training programs as treatments. These policies can at any given time either be enacted or not. We can visualize the time path of a unit with two potential outcomes accordingly:

Figure 2: Potential Outcomes

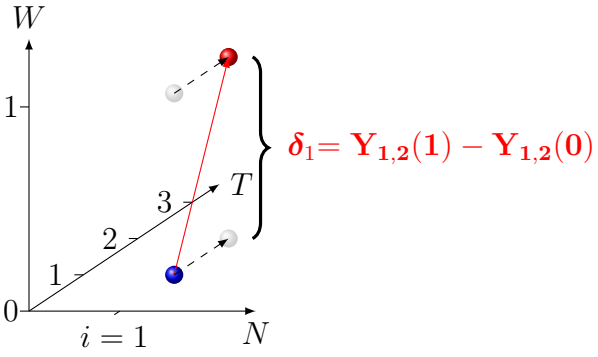


where $Y_1(0)$ indicated that a unit is *untreated* and $Y_1(1)$ that a unit is *treated*. The causal mechanism that *assigns* or *switches* a unit from being untreated to being treated, that is, from one PO to another, is commonly referred to as the *switching equation*, formalized as

$$\begin{aligned} Y_{i,t} &= W_{i,t}Y_{i,t}(1) + (1 - W_{i,t})Y_{i,t}(0) \\ &= Y_{i,t}(0) + [Y_{i,t}(1) - Y_{i,t}(0)]W_{i,t} \end{aligned}$$

where $W_{i,t} \in \{0, 1\}$ is the binary treatment indicator. This modelling approach quickly reveals the fundamental problem of causal inference in the RCM: knowing the *causal effect* δ_1 of the treatment requires us to compare two states: $Y_1(0)$ and $Y_1(1)$ at the same time. Since we cannot observe two alternate realities, this is fundamentally unobservable:

Figure 3: Unit-Specific Treatment Effect



We can, however, typically observe more than one unit, some of which are treated, others of which are not. Units that are never treated are commonly referred to as *never adopters* or *control units*, while units that do receive treatment at some point in time are referred to as *treated units*. All control units pooled together forms the *control group*, while all *treated units* treated in the *same time period* form the *treatment group*.

As Imbens and Rubin (2015) explain, the ability to observe multiple units is crucial to understanding the difference between *defining* and *estimating* a causal effect, as it enables us to draw different comparisons. While it is straightforward to define the causal effect as the difference between two potential outcomes at the same time for the same unit, this definition is hardly useful for estimation and inference, as it can never be observed. To draw inference from estimation, we have to compare *observed realizations* of PO, which necessitates multiple units as there is only ever one realized PO per unit and period. In the RCM, a before-and-after comparison of the same object or person involves two distinct units, as does the comparison of two distinct objects or persons at the same time. Such comparisons allow for *estimation* of causal effects, but are not necessary for their definition.

2.2 Inference on Treatment Effects as the Ultimate Estimands

Before we can discuss methods of estimating treatment effects, it is important to discuss two fundamental assumptions known as *exclusion restrictions* that belie all such methods based on the RCM.[1] Neither is explicitly testable, and both require a judgement with regard to their validity by the investigator. The first assumption excludes the possibility of what economists call *spillovers* and in doing so enables us to utilize multiple units for inference:

> **Assumption 1 (SUTVA) :**
> **The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different PO** (p. 10)

1. Imbens and Rubin (2015, p. 10) define this term to mean "*assumptions that rely on external, substantive, information to rule out the existence of a causal effect of a particular treatment relative to an alternative.*"

Put simply, if Assumption 1 holds, then treatments applied to one unit do not affect the outcome for another unit. Violations of this assumption can be dealt with, though methods to do so are beyond the scope of this thesis.

In a study with binary treatment in which this assumption holds, half of all PO will be unobserved as only one PO per unit and period can be observed. The estimation of causal effects requires us to predict the unobserved PO and to compare it to the observed outcome. Such predictions rely on further assumptions about the assignment (treatment) mechanism and about the possible comparisons between different units in the treatment and control groups. In most interesting settings, the units of observation have background attributes, which may be unit- or group-specific. I will call those attributes *pre-treatment variables* or *covariates*, denoted X_i per unit. Covariates can often be used to enhance the quality of our predictions of missing POs (Imbens and Rubin 2015). They do, however, also play an important role in the second exclusion restriction. Covariates could potentially affect the assignment mechanism. If units in the treatment differ substantially from those in the control group in the value of their covariates, this can affect the plausibility of our assumptions regarding the assignment mechanism. To account for this, assumptions about the assignment mechanism and its independence on POs tend to be more plausible once we condition on covariates, that is, within subpopulations that are homogeneous in covariates (Imbens and Rubin 2015).

With the importance of conditioning on covariates in mind, we arrive at the second assumption. In varying fields, this assumption is called *unconfoundedness*, *strict exogeneity*, or *selection on observables* (Imbens 2004). There is no complete consensus on whether these definitions can, in fact, be used entirely interchangeably (Imbens and Rubin 2015). In keeping with Imbens (2004), who is among the authors of Athey et al. (2021), however, I will use them interchangeably to mean the following:

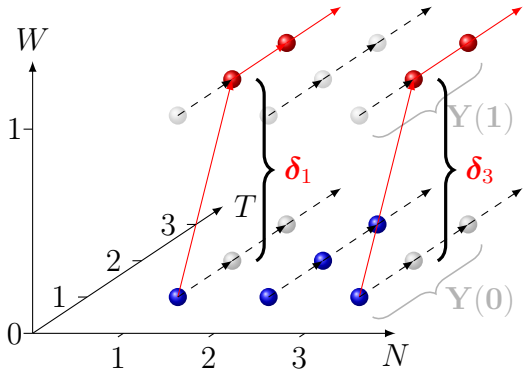
Assumption 2 (Unconfoundedness / Exogeneity) :

An assignment mechanism is unconfounded if it does not depend on the potential outcomes: $\mathbb{P}\big(\mathbf{W}|\mathbf{X},\mathbf{Y}(\mathbf{0}),\mathbf{Y}(\mathbf{1})\big)=\mathbb{P}\big(\mathbf{W}|\mathbf{X},\mathbf{Y}'(\mathbf{0}),\mathbf{Y}'(\mathbf{1})\big)\ \forall\ \mathbf{W},\mathbf{X},\mathbf{Y}(\mathbf{0}),\mathbf{Y}(\mathbf{1}),\mathbf{Y}'(\mathbf{0}),\mathbf{Y}'(\mathbf{1})$

(Imbens and Rubin 2015, p. 38)

If this assumption holds, assignment (receipt) of treatment is independent of all (both) PO when controlling for covariates. Assuming that assumptions 1 and 2 hold, we arrive at the core, or *ultimate*, interest of causal inference: causal estimands. As discussed, the unit-specific treatment effect, that is, the difference of the pair of POs, cannot be estimated. We can, however estimate this quantity averaged over the full sample: this average treatment effect (ATE) can be estimated as $\mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0)]$. One complication in this estimation is that we can only reliably estimate this ATE if the treated group is representative of the population of study. As this representativeness may not always be given, it is therefore convenient to restrict ourselves to estimating the average treatment effect *on the treated* (ATET), which averages the individual treatment effects of treated units, which in Figure 4 below corresponds to $\frac{1}{2} \times (\delta_1 + \delta_3)$.

Figure 4: Average Treatment Effect on the Treated



Formally, the ATET is defined as follows:

Definition (ATET) :

Average of individual treatment effects across treated units: $\tau = \mathbb{E}[\delta_i | W_i = 1]$

Both the ATE and ATET are similar to δ_1 in that they are unobservable: both rely on comparisons between potential outcomes for the same unit in the same period. There are several distinct popular ways of estimating the *ultimate estimand* τ through various *immediate* estimands. In recent years, popular approaches in Applied Economics have been *Difference-in-Differences* (DiD) and *Synthetic Control* (SC), or, more recently, combinations of the two (cf. Arkhangelsky et al. 2021). Given the extraordinary number of recent developments in DiD methods, this thesis focuses on that stream of what Athey et al. (2021) call the *unconfoundedness* approach.

3 Causal Inference in Difference-in-Differences Designs

The DiD method can broadly be separated into two categories: the *canonical* (sometimes also referred to as *vanilla* design, and all other cases, which I will call *richer* designs. The mechanics and properties of DiD estimators are very well understood in the canonical case, but not in all richer designs.

3.1 The Canonical Design Estimator

The canonical case is characterised by having two groups, which I will for simplicity's sake, for now, treat as if they were individual units, so we have $N = 2$, and two periods, $T = 2$. The control group is $i = 2$. One treatment group $(TREAT, i = 1)$ of units who are all treated in period $t = 2$ means that we can separate what happens *before* the treatment $(PRE, t = 1)$ from what happens *after* $(POST, t = 2)$. With outcomes $Y_{i,t}$, the DiD estimator is

$$\begin{aligned} \hat{\beta}^{2\times2} &= \left(\overline{Y}^{POST}_{TREAT} - \overline{Y}^{PRE}_{TREAT}\right) - \left(\overline{Y}^{POST}_{CONTROL} - \overline{Y}^{PRE}_{CONTROL}\right) \\ &\iff \left(\overline{Y}_{1,2}(1) - \overline{Y}_{1,1}(0)\right) - \left(\overline{Y}_{2,2}(0) - \overline{Y}_{2,1}(0)\right) \end{aligned} \tag{1}$$

This estimator, which I will call a 2×2, coincides with the Ordinary Least Squares (OLS) estimate of the linear model $Y_{i,t} = \alpha_0 + \alpha_1 \mathbb{1}_{i=1} + \alpha_2 \mathbb{1}_{t=2} + \beta\left(\mathbb{1}_{i=1}\mathbb{1}_{t=2}\right) + \epsilon_{i,t}$. Estimator (1) removes group-specific factors by taking the difference between group means. For it to produce our estimand of interest, we require four additional assumptions. The formal definition of each of the following assumption specifically relates to the canonical setup:

Assumption 3 (Common Trends) :

Differences in expected non-treatment POs conditional on covariates are unrelated to treatment status: If $Y_{i,t}(0) = \gamma_i + \delta_t + \epsilon_{i,t}$ **s.t.** $\mathbb{E}[\boldsymbol{\epsilon_{i,t}}|\boldsymbol{W_i}] = \mathbf{0}$[a], **then**

$$\mathbb{E}[\mathbf{Y_{1,2}(0)}|\mathbf{X} = \mathbf{x}] - \mathbb{E}[\mathbf{Y_{1,1}(0)}|\mathbf{X} = \mathbf{x}] = \mathbb{E}[\mathbf{Y_{2,2}(0)}|\mathbf{X} = \mathbf{x}] - \mathbb{E}[\mathbf{Y_{2,1}(0)}|\mathbf{X} = \mathbf{x}] \; \forall \; x$$

a. This implicitly encodes the assumption that the true Data-Generating Process is linear in parameters so that the linear model given here is structural or causal, which is substantially more restrictive than Assumption 2, which could also be called *Super-Exogeneity* (cf. a.o. Pearl 2000, Section 5.4.3).

Assumption 4 (No Anticipation) :

The treatment has no causal effect before its assignment, that is, units who are at some point treated do not anticipate treatment:

$$\mathbb{E}[\mathbf{Y_{i,t}(1) - Y_{i,t}(0)}|\mathbf{X = x},\ \mathbf{W_{i,t} = 1}] = 0 \ \forall\ x \textbf{ and } t\ s.t.\ W_{i,t} = 0$$

Assumption 5 (Overlap) :

Observations with value(s) $X = x$ **exist in each partition of the sample space: for some** $\epsilon > 0, \mathbb{P}(W_i = 1|X_i) < 1 - \epsilon\, a.s., \mathbb{E}[W] > 0$

Assumption 6 (Random Sampling) :

$\{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,t}, X_i, W_{i,1}, W_{i,2}, \ldots, W_{i,t}\}_{i=1}^{n}$ **is** $i.i.d$

The simplification of only analyzing two periods in the canonical design does not mean we are actually restricted to observing two periods. In the canonical design, we can pool together all post-treatment and all pre-treatment periods by including both unit- (γ_i) and time- (δ_t) fixed effects, which yields the popular twoway fixed effects (TWFE) specification (2):

$$Y_{i,t} = \gamma_i + \delta_t + \tau W_{i,t} + \epsilon_{i,t} \tag{2}$$

Estimator [TWFE] (Canonical TWFE) :

In $\hat{Y}_{i,t} = \hat{\gamma}_i + \hat{\delta}_t + \hat{\tau} W_{i,t} + e_{i,t}$ **,** $\hat{\tau}_{TWFE}$ **is the twoway fixed-effects estimator**

Under assumptions 3, 4, 5, and 6, and provided sufficiently large sample size, (1) and [TWFE] both identify the ATET τ (Goodman-Bacon 2021).

Real-life applications of DiD tend to involve data more complicated than having 2 groups over 2 periods. I will refer to any setting which deviates from the canonical design *rich*. It is well established that (1) is numerically equivalent to [TWFE] if there are two periods and the treatment is assigned to some units only in the second time period. Implementing extensions beyond the canonical setting, however, are difficult: as Goodman-Bacon (2021, footnote 5)

quotes Imai and Kim (2021, p. 8), "[it] *unfortunately, this equivalence result does not generalize to the multi-period DiD design*". This problem is at the heart of the recent advances in DiD methods:

Definition (Differential Timing) :
Differential timing, also referred to as *staggered adoption*, describes settings in which not all treated units receive treatment in the same time period t

3.2 Estimands in Richer Designs: the *Problem*

The simplest possible case shows why this soon becomes problematic. Consider a 3×3 design in which $i = 1$ is treated in period 2 (the *early* group A), $i = 2$ is the untreated (*control* group U, and $i = 3$ receives treatment in period 3 (the *late* group, B). Estimating the ATET using (2) corresponds to constructing a variance-weighted average of multiple 2×2 estimators (1): $\hat{\beta}^{2\times2}_{A-U}$, $\hat{\beta}^{2\times2}_{B-U}$, $\hat{\beta}^{2\times2}_{A-B|t<2}$, and $\hat{\beta}^{2\times2}_{B-A|t\geq1}$. This is the central result of Goodman-Bacon (2021, Theorem 1), who points out that in the presence of time-varying treatment effects, this variance-weighted average may be *severely biased* by drawing comparisons between already-treated units (these comparisons are also called *forbidden* or *dirty*, the latter of which as opposed to the *clean*, that is, unproblematic, comparisons); this issue is also empirically analyzed by Chaisemartin and D'Haultfœuille (2020), who replicate an *AER* article and find that more than of 40% of the weights are negative. This can lead to a substantial problem: with negative weights, one could plausibly have a *positive* treatment effect for every group and period, while the *overall estimated ATET might be negative*. In statistical jargon, our *estimator no longer reliably matches our estimand*. [2]

4 Difference-in-Differences With Differential Timing

To address these issues, a recent wave of research in econometric has proposed new estimators. Following Roth et al. (2023, Table 2)'s review of current work, which excludes

2. With larger panels, the sheer number of 2×2 comparisons we have to somehow weight is indicative of the complexity of the task of unbiased causal estimation: while in a 3×3 design we have 4 such comparisons, in a 4×4 design, there are already 9 separate $\hat{\tau}^{2\times2}$s.

estimators still in development, I have selected several papers that propose such new estimators. The main selection criterion used was the existence of R implementations of these new estimators.

4.1 Differential Timing Setup

Before briefly discussing the novel aspects of these estimators, however, I will first lay out an analytical setup that allows us to discuss these estimators in the same terms used so far. Table 1 at the end of this section provides an overview of the assumptions for each estimator. The notation is adapted from Roth et al. (2023, Section 3). As before, we have $i = 1, \ldots, N$ units observed over$t = 1, \ldots, T$ periods, units can receive a binary treatment in any period $t > 1$. The treatment indicator $W_{i,t}$ is binary and $G_i = \min\{t : W_{i,t} = 1\}$ records the earliest period in which unit i receives treatment. If i is untreated, then $G_i = \infty$. We now have to introduce one new assumption and adapt two existing ones to the (Differential Timing) setting.

Assumption 7 (Constant Treatment) :

Treatment is an absorbing state: $W_{i,t} = 1 \; \forall \, t \geq G_i$ **where** $G_i = \min\{t : W_{i,t} = 1\}$

Under Assumption 7, the entire path of potential outcomes of a unit can be summarised by the index of the first treatment date g: let $\mathbf{0}_s = 0 \; \forall s$ and $\mathbf{1}_s = 1 \; \forall s$ denote s-dimensional vectors of potential outcomes. Unit i's PO in t when treated in g is denoted as $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$, and as $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$ for control units. This means that each G_i forms a *treatment group* of units treated in period t. The set of all treated groups is given by $\mathcal{G} = \text{supp}(G_i) \setminus \max_i G_i$.[3] We denote the *group-specific treatment effect* as $\tau_{i,t}(g)$. To adapt Assumption 3 to differential timing, we have to impose that common trends holds for all combinations of periods and groups treated at different times. [4]

3. As detailed in Callaway and Sant'Anna (2021, Footnote 5), this set of all treated groups excludes either the control group, which is never treated, and hence has $G_i = \infty$, or, in the absence of any never-treated group, excludes the last-treated group to prevent invalid inter-group comparisons.

4. These adaptions again implicitly assume a linear Data-Generating Process with Mean Independence of the error term, cf. Assumption 3.

Assumption 8 (Common Trends, Differential Timing) :

Over time, differences in expected potential non-treatment outcomes between the group first treated in g and the control group, conditional on covariates, are unrelated to treatment status: $\forall t \neq t'$, $\forall g \neq g'$, and $\forall x$,

$$\mathbb{E}[\mathbf{Y_{i,t}}(\infty) - \mathbf{Y_{i,t'}}(\infty)|\mathbf{G_i} = \mathbf{g}, \mathbf{X} = \mathbf{x}] = \mathbb{E}[\mathbf{Y_{i,t}}(\infty) - \mathbf{Y_{i,t'}}(\infty)|\mathbf{G_i} = \mathbf{g'}, \mathbf{X} = \mathbf{x}]$$

This imposes that in the counterfactual without treatment, the expected outcomes for all adoption groups would have evolved in parallel (Roth et al. 2023). A further modification, which is required instead of Assumption 8 for certain estimators, imposes this restriction only in periods after the first treatment in any group occurs:

Assumption 9 (Common Trends, Diff. Timing, Post-Treatment Only) :

Over time, differences in expected potential non-treatment outcomes between the group first treated in g and the control group, conditional on covariates are unrelated to treatment status: $\forall t, t' \geqslant g_{min} - 1$, $g_{min} = \min \mathcal{G}$

$$\mathbb{E}[\mathbf{Y_{i,t}}(\infty) - \mathbf{Y_{i,t'}}(\infty)|\mathbf{G_i} = \mathbf{g}, \mathbf{X} = \mathbf{x}] = \mathbb{E}[\mathbf{Y_{i,t}}(\infty) - \mathbf{Y_{i,t'}}(\infty)|\mathbf{G_i} = \mathbf{g'}, \mathbf{X} = \mathbf{x}]$$

Assumption 10 (No Anticipation in Differential Timing) :

If a unit is not treated in t, its outcome is independent on when it will be treated:

$$Y_{i,t}(g) = Y_{i,t}(\infty) \; \forall i, \; \forall t < g$$

Assumption 11 (Time Invariant Treatment Effect) :

The treatment effect is constant over time if and only if for

$$\tau_{i,t}(g) = Y_{i,t}(g) - Y_{i,t}(\infty), \textbf{ we have } \tau_{i,t}(g) = \tau \; \forall \; t \geqslant g$$

Assumption 12 (Unit Homogeneous Treatment Effect) :

All units have the same treatment effect if and only if for

$$\tau_{i,t}(g) = Y_{i,t}(g) - Y_{i,t}(\infty), \textbf{ we have } \tau_{i,t}(g) = \tau \; \forall \; i$$

If *the treatment effects vary*, either across time and/or across units, then [TWFE] no longer reliably matches our estimand (Goodman-Bacon 2021).

4.2 A Note on *Heterogeneity* and *Dynamics*

The recent literature has no consensus definition of *heterogeneity*, with some authors using the term *dynamic treatment effects* to describe violations of Assumption 11, and others subsuming time heterogeneity into a more general (but typically only loosely defined) *heterogeneity*. I have made the conscious choice to define them separately, so that for each estimator I can make clear which *type*(s) of heterogeneity it is meant to address, and to avoid usage of the word *dynamic* to eliminate confusion with (Dynamic Estimation) below.

Although each of these estimators will yield numerical results even under relaxations of both Assumption 11 and 12, much of the recent work on the credibility of DiD estimates has (at least implicitly) focused on what, precisely, the *estimand* of [TWFE] represents. This depends not only on the validity of assumptions 11 and 12, but also on whether we would like to investigate treatment in an aggregate, absolute way, or in a more fine-grained, relative manner: the type of inference we draw by using (2) is commonly referred to as *static* (Chaisemartin and D'Haultfoeuille 2022):

Definition (Static Estimation) :

In (2) **(**$Y_{i,t} = \gamma_i + \delta_t + \tau W_{i,t} + \epsilon_{i,t}$**),** $W_{i,t}$ **is the absolute time of treatment**

In many applied settings, however, we are more interested in time *relative to treatment* in so-called *Event-Study designs*:

$$Y_{i,t} = \gamma_i + \delta_t + \sum_{r \neq -1} \beta_r \mathbb{1}_{R_{i,t}=r} + e_{i,t} \tag{3}$$

Definition (Dynamic Estimation) :

In (3)**,** $R_{i,t} \coloneqq t - G_i$ **is the time relative to treatment, with** $R_{i,t} = 0$ **representing the period immediately after treatment (**adopted from Roth et al. (2023, (7))[a]

a. Roth et al. (2023) actually define $R_{i,t} = t - G_i + 1$, so that the the first treated period corresponds to $R_{i,t} = 1$. The reason why I chose to make this slight change is that the R implementations discussed later on use -1 as the reference period, which corresponds to the notation used here.

As shown in a.o. Sun and Abraham (2021) and Borusyak, Jaravel, and Spiess (2023), if the treatment effect is time-varying but homogeneous across units, then β_s in (3) identifies τ_s, the treatment effect s periods after receiving treatment under assumptions 4 and 8.

4.3 New Estimators for Staggered Designs

We can now introduce the new estimators, which all aggregate heterogeneous treatment effects in varying ways to avoid these issues. The derivation of each is similar in logic: after specifying the estimand of interest (the ATET), and using Assumption 1, 2, 5, 6, 7, 10 and either 8 or 9, the new estimators "*allow inference on counterfactual outcomes for treated units using appropriately chosen "clean" control groups of untreated units, which allows us to express the ATET in terms of identified expectations*" (Roth and Sant'Anna 2023, p. 15). The papers that propose these estimators share a common motivation of outlining problems with [TWFE] in certain settings and proposing improved methods to address them. For my purposes, they differ in the immediate estimands and estimation methods used to obtain them before aggregating to the ultimate estimand.

4.3.1 De Chaisemartin & D'Haultfœuille (2020)

The estimator proposed by Chaisemartin and D'Haultfœuille (2020) is more general than the differential timing setting requires as it allows a relaxation of Assumption 7, that is, it allows for treatment to "turn off" for certain units or groups. Their estimator compares groups defined as *joiners* and *leavers*. Their notation centers on units in groups who *switch* their treatment status, of which there are in total n_S, where $n_{g,t}$ denotes the number of observations in group g at time t, and $\overline{W}_{g,t}$ denotes the average treatment status of in g at t, $n_{1,0,t}$ denotes the number of observations in g that are *treated* in period t but were *untreated* in $t-1$.[5] Under several additional assumptions on $Y(1)$ POs to accommodate this setting, Chaisemartin and D'Haultfœuille (2020, p. 2978) define estimators

5. Formally, $n_S = \sum_{(g,t):t\geqslant 2,\, W_{g,t}\neq W_{g,t-1}} n_{g,t}$ (Chaisemartin and D'Haultfœuille 2020, p. 2976)

$$\text{DID}_{+,t} = \sum_{\substack{g:\overline{W}_{g,t}=1,\\ \overline{W}_{g,t-1}=0}} \frac{n_{g,t}}{n_{1,0,t}}\Big(\overline{Y}_{g,t} - \overline{Y}_{g,t-1}\Big) - \sum_{\substack{g:\overline{W}_{g,t}=0,\\ \overline{W}_{g,t-1}=0}} \frac{n_{g,t}}{n_{0,0,t}}\Big(\overline{Y}_{g,t} - \overline{Y}_{g,t-1}\Big) \tag{4}$$

$$\text{DID}_{-,t} = \sum_{\substack{g:\overline{W}_{g,t}=1,\\ \overline{W}_{g,t-1}=1}} \frac{n_{g,t}}{n_{1,1,t}}\Big(\overline{Y}_{g,t} - \overline{Y}_{g,t-1}\Big) - \sum_{\substack{g:\overline{W}_{g,t}=0,\\ \overline{W}_{g,t-1}=1}} \frac{n_{g,t}}{n_{0,1,t}}\Big(\overline{Y}_{g,t} - \overline{Y}_{g,t-1}\Big) \tag{5}$$

$$\text{DID}_{M} = \sum_{t=2}^{T}\Big(\frac{n_{1,0,t}}{n_S}\text{DID}_{+,t} + \frac{n_{0,1,t}}{n_S}\text{DID}_{-,t}\Big) \tag{6}$$

Intuitively, (4) is a version of the basic 2×2 DiD estimator (1) that compares the average observed outcomes of groups who switch *into* treatment to those of control groups (*joiners*), (5) does the same for *leavers*, and (6) is a weighted average of the two. DID_M is an unbiased, consistent and asymptotically normal estimator of the ATET of all switching units, which in Chaisemartin and D'Haultfœuille (2020)'s general setting includes unit who leave the treatment. In the differential timing setting with assumptions 7 and 11, there are no *leavers*, so (5)$= \emptyset$ and we arrive at a static estimator:

Estimator [dCdH] (de Chaisemartin & D'Haultfœuille) :

$\hat{\tau}_{dCdH} = \sum_{t=2}^{T}\Big(\frac{n_{1,0,t}}{n_S}\textbf{DID}_{+,t}\Big)$, **DID**$_{+,t}$ **as in** (4) (Chaisemartin and D'Haultfœuille 2020)

4.3.2 Callaway & Sant'Anna (2021)

Callaway and Sant'Anna (2021) use Assumption 9 to derive a parameter called the *group-time ATET*, which estimates the ATET at time t on the treatment group treated in g: $ATET(g,t) = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(\infty)|G_i = g]$. Formally,

$$ATET(g,t) = \mathbb{E}[Y_{i,t}(g) - Y_{i,g-1}(g)|G_i = g] - \mathbb{E}[Y_{i,t}(g) - Y_{i,g-1}(g)|G_i = g'] \,\forall g' > t \tag{7}$$

$$= \mathbb{E}[Y_{i,t}(g) - Y_{i,g-1}(g)|G_i = g] - \mathbb{E}[Y_{i,t}(g) - Y_{i,g-1}(g)|G_i \in \mathcal{G}_{comp}] \forall g' > t \tag{8}$$

(7) holds $\forall\, g' > t$, so we can write $\mathcal{G}_{comp}\, s.t.\ g' > t \,\forall\, g' \in \mathcal{G}_{comp}$:, which yields (8)

Replacing expectations with their sample analogs yields, for $\mathcal{G}_{comp} = \{g' : g' > t\}$ (not-yet

treated units) [6] the nonparametric estimator

$$\widehat{ATET}(g,t) = \frac{1}{N_g} \sum_{i:G_i=g} \big[Y_{i,t} - Y_{i,g-1}\big] - \frac{1}{N_{\mathcal{G}_{comp}}} \sum_{i:G_i \in \mathcal{G}_{comp}} \big[Y_{i,t} - Y_{i,g-1}\big] \tag{9}$$

This estimator can be aggregated in many different ways detailed in Callaway and Sant'Anna (2021), one of which is the ATET for units $i \in g$ averaged across all *post-treatment periods*

$$\tau(g) = \frac{1}{T-g+1} \sum_{t=g}^{T} \widehat{ATET}(g,t) \tag{10}$$

This, finally, allows aggregation to the static overall ATET by estimating [CS] below. Callaway and Sant'Anna (2021) show asymptotic normality and consistency of (9); they also show that it avoids negative weights and lays open exactly which units are used as *clean* control groups to draw inference on unobserved POs. Their estimator is identical to [dCdH] (Chaisemartin and D'Haultfœuille 2021, p. 17):

Estimator [CS] (Callaway & Sant'Anna) :

$\hat{\tau}_{CS} = \sum_{g \in \mathcal{G}} \tau(g) \mathbb{P}\big(G_i = g | G_i \leq T\big)$ **where** $\tau(g)$ **as in** (9), (10) **with** $\mathcal{G}_{comp} = \{g' : g' > t\}$

adapted from Callaway and Sant'Anna (2021, eq. (3.7, 3.11))

The inherent flexibility of the group-time ATET (9) also allows for aggregation into an event-study type parameter that estimates the ATET ℓ periods after treatment adoption:

Estimator [CS (event-study)] (Callaway & Sant'Anna (event-study)) :

$\hat{\tau}^{ES}_{CS} = \widehat{\text{ATET}}^{w}_{\ell} = \sum_{g} w_g \, \text{ATET}(g, g+\ell)$ **where** $w_g = \mathbb{1}_{\{g+\ell \leqslant T\}} \mathbb{P}\left(G = g | G + \ell \leqslant T\right)$

and $\text{ATET}(g, g+\ell)$ **as in** (9)

adapted from Callaway and Sant'Anna (2021, eq. (3.4))

Note that [dCdH] $\hat{=}$ [CS] when the latter is specified as ATET^{w}_{0} (Chaisemartin and D'Haultfoeuille 2022, 17; Roth et al. 2023, 19).

6. Alternatively, this estimator can be specified using $\mathcal{G}_{comp} = \{\infty\}$ (never-treated units)

4.3.3 Sun & Abraham (2021)

Sun and Abraham (2021) propose an event-study estimator which uses the $ATET(g,t)$ (9) as a building block , but uses a different $\mathcal{G}_{comp}$ in the step from (8) instead of the not-yet treated units $g' : g' > t$, their estimator uses $\mathcal{G}_{comp} = \{\max_i G_i\}$, which corresponds to the never-treated control group ($\max_i G_i = \infty$), or, in the absence of such a group, to the last-to-be-treated units. In a slight change of notation, they use the relative time notation introduced above: $\ell = t - G_i$, where ℓ denotes ℓ periods *relative* to treatment. Using Assumption 7, we can (but do not have to) *bin* pre-and post-treatment observations, respectively (cf. Schmidheiny and Siegloch 2023)[7]. This approach also differs from [CS (event-study)] in how the estimation is carried out. Sun and Abraham (2021) propose a regression-based *Interaction-Weighted* estimator that is computed in three steps:

$$Y_{i,t} = \gamma_i + \delta_t + \sum_{g \notin \mathcal{G}_{comp}} \sum_{\ell \neq -1} \tau_{g,\ell} \left(\mathbb{1}_{\{G_i = g\}} W_{i,t}^{\ell} \right) + e_{i,t} \tag{11}$$

$$\tau(g) = \hat{\tau}_{g,\ell} \text{ in } (11) \tag{12}$$

First, one estimates the individual $ATET(g,t)$ using a TWFE regression (11) that interacts relative period indicators with group indicators, excluding indicators for comparison (i.e. control) groups. This yields (12), which is the DiD estimate (identical to (2)) of $ATET(g,t)$. One then estimates the weights by sample shares of each treatment group g in the relevant period(s) $\ell \in b$:

$$\mathbb{P}\{G_i = g | G_i \in [-\ell, T - \ell]\}$$

Finally, if interested in a static parameter, one can aggregate $ATET(g,t)$s to an overall ATET using (12):

7. In relative event time, differential timing necessarily creates imbalanced panels: with 2 groups in a 10-year panel, where group A is treated in $t = 3$ with 7 post-treatment lags, whereas group B is treated in $t = 6$, B will have 4 lags. In calendar time, each unit has $T = 10$, but in relative event time, the number of lags differs. Assuming constant treatment effects before and after the event time, it is therefore common to use *binning*: grouping all imbalanced leads and lags into a single lead and lag, respectively, which balances the panel.

Estimator [SA] (Sun & Abraham) :

$\hat{\tau}_{SA} = \frac{1}{b}\sum_{\ell \in b}\sum_{g} \tau(g)\mathbb{P}\big(G_i = g | G_i \in [-\ell, T-\ell]\big)$ **where** $\tau(g)$ **as in** (12)

with $\mathcal{G}_{comp} = \{\max_i G_i\}$ **, bin size b** adapted from Sun and Abraham (2021, eq. (26), (27))

We can also aggregate the group-time ATETs in the same way as in [CS (event-study)]:

Estimator [SA (event-study)] (Sun & Abraham (event-study)) :

$\hat{\tau}^{ES}_{SA} = \widehat{\text{ATET}}^{w}_{\ell} = \sum_{g} w_g \, \text{ATET}(g, g+\ell)$ **where** $w_g = \mathbb{1}_{\{g+\ell \leqslant T\}}\mathbb{P}\left(G = g | G + \ell \leqslant T\right)$

and $\text{ATET}(g, g+\ell)$ **as in** (12)

adapted from Callaway and Sant'Anna (2021, eq. (3.4)), Sun and Abraham (2021, eq. (26))

4.3.4 Marcus & Sant'Anna (2021)

Marcus and Sant'Anna (2021) propose among other things a recursive, nonparametric estimator similar to [CS] which more efficiently exploits a variation in Assumption 8: [8]

Assumption 13 (CTA, Differential Timing, not yet treated) :

Over time, differences in expected potential non-treatment outcomes between the group first treated in g and the not-yet treated groups in t are unrelated to treatment status: $\forall g, t = 2, \ldots, T \; s.t. \; t \geqslant g$,

$\mathbb{E}[\mathbf{Y_{i,t}}(\infty) - \mathbf{Y_{i,t-1}}(\infty) | \mathbf{G_i} = \mathbf{g}] = \mathbb{E}[\mathbf{Y_{i,t}}(\infty) - \mathbf{Y_{i,t-1}}(\infty) | \mathbf{G_i} = \infty]$

adapted from Marcus and Sant'Anna (2021, p. 255)

8. Including an observation-based GMM estimator usable under stronger Common Trends Asssumptions (CTA), and an interesting discussion of the robustness-efficiency tradeoff regarding how strong the variant of the CTA used in estimators

Under Assumption 13, we can write (7) as follows: $\forall\, i$ and for $2 \leqslant g \leqslant t \leqslant T$,

$$
\begin{aligned}
ATET(g,t) &= ATET_{ny}(g,t) \\
ATET_{ny}(g,t) &= \mathbb{E}[Y_{i,t} - Y_{i,g-1} | G_i = g] - \Big(\sum_{s=g}^{t} \mathbb{E}[Y_{i,t} - Y_{i,t-1} | W_s = 0, G_i \neq g] \Big) \\
\widehat{ATET}_{ny}(g,t) &= \frac{\frac{1}{n}\sum_{i=1}^{n} G_{ig}(Y_{i,t} - Y_{i,g-1})}{\frac{1}{n}\sum_{i=1}^{n} G_{i,t}} - \sum_{s=g}^{t} \Big(\frac{\frac{1}{n}\sum_{i=1}^{n}(1 - W_{i,s})(1 - G_{i,g})(Y_{i,t} - Y_{i,t-1})}{\frac{1}{n}\sum_{i=1}^{n}(1 - W_{i,s})(1 - G_{i,g})} \Big)
\end{aligned}
\tag{13}
$$

which yields (**??**), a computationally easy, $\sqrt{n}$-consistent and asymptotically well-behaved nonparametric estimator close to (9).

As before, we can aggregate the group-time ATET *building blocks* into two estimators: a static estimator of the overall ATET and, secondly, an event-study type parameter

Estimator [MS] (Marcus & Sant'Anna) :

$$
\hat{\tau}_{MS} = \frac{\sum_{g=2}^{T}\sum_{t=2}^{T} \mathbb{1}_{\{g \leqslant t\}} \mathbb{P}(G_i = g) \widehat{ATET}_{ny}(g,t)}{\sum_{g=2}^{T}\sum_{t=2}^{T} \mathbb{1}_{\{g \leqslant t\}} \mathbb{P}(G_i = g)}
$$

using (**??**), adapted from Marcus and Sant'Anna (2021, eq. (2.22),(2.19))

Estimator [MS (event-study)] (Marcus & Sant'Anna (event-study)) :

$$
\hat{\tau}^{ES}_{MS} = \sum_{g=2}^{T}\sum_{t=2}^{T} \mathbb{1}_{\{t-g+1=\ell\}} \mathbb{P}\left(G_i = g | \textbf{treated for } \geqslant \ell \textbf{ periods}\right) \mathrm{ATET}(g,t)
$$

where $\mathrm{ATET}(g,t)$ **as in** (**??**), adapted from Marcus and Sant'Anna (2021, eq. (2.23),(2.19))

4.3.5 Borusyak, Jaravel, & Spiess (2022)

Finally, Borusyak, Jaravel, and Spiess (2023) propose a regression-based approach they describe as a robust and efficient *imputation*[9] estimator with finite-sample efficiency (as opposed to the previously introduced estimators), which additionally is significantly more intuitive than

9. *At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others*" (Imbens and Rubin 2015, p. 141), this estimator is one such more explicit case.

the preceding estimators. Using Assumption 8, they construct their estimator as follows:

$$Y_{i,t}(\infty) = \gamma_i + \delta_t + \epsilon_{i,t} \tag{14}$$

$$\widehat{Y}_{i,t}(\infty) \text{ from (14)} \tag{15}$$

$$\hat{\tau}_{i,t} = Y_{i,t} - \widehat{Y}_{i,t}(\infty) \tag{16}$$

1. Fit TWFE regression (14) using observations for *not-yet treated* groups
2. Impute unrealized PO for treated units using prediction from step (15)
3. Compute unit-specific treatment effect (16)
4. Aggregate individual $\tau_{i,t}$ to ATET τ:

Estimator [BJS] (Borusyak, Jaravel, & Spiess) :

$\hat{\tau}_{BJS} = \frac{1}{|\mathcal{G}|} \sum_{i:G_i \neq \infty} \hat{\tau}_{i,t}$ **where** $\hat{\tau}_{i,t} = Y_{i,t} - \hat{Y}_{i,t}(\infty)$ **using** (15) **and** (14),
adapted from **borusyak_revisiting_2022empty citation**

The main difference between [BJS] and [CS] is that the latter "*makes all comparisons relative to the last pre-treament period, whereas* [BJS] *makes comparisons relative to the average of the pre-treatment periods*" (Roth et al. 2023, p. 18). [BJS] is likely to be more efficient than [CS] the lower heteroskedasticity and autocorrelation in $Y_{i,t}$ (Roth et al. 2023). Alternatively, at step 4 above, we can aggregate the individual $\tau_{i,t}$ into group-time ATET: as before, for each $t = 1, \ldots, T$, we define

$$\widehat{\text{ATET}}(g,t)_{BJS} = \hat{\tau}_{i,t}(g) = \frac{1}{|i : G_i = g|} \mathbb{1}_{\{i:G_i=g\}} Y_{i,t} - \widehat{Y}_{i,} \tag{17}$$

and aggregate further to estimate an event-study type parameter by using the same aggregation as [CS (event-study)]:

Estimator [BJS (event-study)] (Borusyak, Jaravel, Spiess (event-study)) :

$\hat{\tau}^{ES}_{BJS} = \text{ATET}^{w}_{\ell} = \sum_{g} w_g \, \text{ATET}(g, g+\ell)$ **where** $w_g = \mathbb{1}_{\{g+\ell \leqslant T\}} \mathbb{P}\left(G = g | G + \ell \leqslant T\right)$ **and** $\text{ATET}(g, g+\ell)$ **as in** (17)

adapted from Callaway and Sant'Anna (2021, eq. (3.4))

4.4 The Paradox Of Choice for The Practitioner

The state of the DiD literature is rapidly evolving. This presents a problem to applied researchers who use DiD methods: unless they spend substantial amounts of time following the cutting-edge advances in DiD, they run the risk of using *outdated* or *ill-suited* methods, such as TWFE DiD in the presence of treatment heterogeneity. Table 1 below shows for each of the estimators presented thus far the R packages that implement each, whether they aggregate individual 2×2s or impute unobserved POs directly.

Roth et al. (2023, Table 1) present a three-step checklist for practitioners which recommends current methods in the form of new estimators for staggered designs, including all estimators presented here, diagnostic approaches concerning the validity of the common trend and other assumptions (a.o. Dette and Schumann (2020) and Roth (2022)), and methods and research designs when relaxing the assumption of sampling a large number of clusters from an infinite super-population. Even if a practitioner chooses the *"right"* method, there is a risk of disagreement on the validity of the assumptions, especially the exact variant of Common Trend Assumption made, which has strong implications on reliability and performance of a method (Marcus and Sant'Anna 2021).

Table 1: Estimator Comparison

Estimator	R package(s)	Type	Comp. Group(s)	Main Assumptions

[CS]
[CS (event-study)]
	`did`
	`did2s`
	`staggered`
		2×2 agg.	*not-yet treated* or *never treated*	Assumption 7
				Assumption 9
				Assumption 10

[dCdH]	`DIDmultiplegt`	2×2 agg.	*not-yet treated* or *never treated*	Assumption 7 Assumption 9

[SA]
[SA (event-study)]
	`fixest`
	`did2s`
	`staggered`
		2×2 agg.	*not-yet treated* or *never treated*	Assumption 7
				Assumption 8

[MS] [MS (event-study)]	`MSA2020`	2×2 agg.	*not-yet treated*	Assumption 7 Assumption 13
[BJS] [BJS (event-study)]	`didimputation` `did2s`	imputation	*never treated*	Assumption 7 Assumption 8 Spec. of $Y(0)^{\dagger}$

† All estimators above require Assumptions 1, 2, 5, 6, and 10 in addition to the main identifying assumptions listed.

‡ *Note:* [BJS] is somewhat more flexible in the modelling of Treatment Effect Heterogeneity it allows (cf. Borusyak, Jaravel, and Spiess 2023, Assumption 3).

Table 2: Estimands in Staggered Designs

Estimator	Treatment Effect can vary over		Design is		Estimand
	Units	Time	Static	Dynamic	
[TWFE]			✓	✓†	ATET
[dCdH]	✓		✓		ATET
[CS]	✓	✓	✓		ATET
[CS (event-study)]	✓	✓		✓	ATET(g,t)§
[SA]	✓	✓	✓		ATET
[SA (event-study)]	✓	✓		✓	ATET(g,t)§
[MS]	✓	✓	✓		ATET
[MS (event-study)]	✓	✓		✓	ATET(g,t)§
[BJS]	✓	✓	✓		ATET
[BJS (event-study)]	✓	✓		✓	ATET(g,t)§

† When modified using (3)

‡ All estimators above require Assumptions 1, 2, 5, 6, and 10 in addition to the main identifying assumptions listed in Table 1.

§ I defined the *dynamic* estimators as estimating an aggregate event-study type parameter, which ultimately consists of aggregated $\widehat{\text{ATET}}(g,t)$s. The aggregation is somewhat unimportant, the actual component of the *estimand* in which they differ is the $\widehat{\text{ATET}}(g,t)$.

4.5 Estimands in Staggered Designs

The most important consequence of a choice of estimator, however, is to be very clear about what, specifically, one is trying to estimate. Unless one carefully matches the assumptions about the Data-Generating Process (DGP) with the choice of estimator, the estimator will not match the desired estimand, which can be hard to detect: imagine, for instance, using [dCdH] in a staggered design. If we can comfortably assume that treatment effects satisfy Assumption 11 and Assumption 12, there is no need to use [dCdH], as [TWFE] is unbiased and consistent. If, however, we relax Assumption 12, that is, we allow for a treatment effect that is heterogeneous across units, per the results of Goodman-Bacon (2021), we should use an estimator like [dCdH], as [TWFE] is biased. Given the lack of standardized definition of heterogeneity in the literature, however, a practitioner might easily confuse Assumption 12 for Assumption 11, that is, confuse heterogeneity across *units* with heterogeneity across *time*. In this case, however, [dCdH] does *not* estimate the ATET.[10]

Table 2 highlights the nuanced decisions and assumptions that practitioners have to make when choosing a method for their DiD analysis, such as what precise type of *heterogeneity* they have to impose, and what assumptions in Table 1 their data accommodate. All of these modelling choices, of course, presume that it is clear that DiD approaches are appropriate in the first place: it is entirely possible that a researcher has to *a priori* decide between using DiD or other approaches, such as synthetic control or fixed effects models.

10. It is not entirely clear what [dCdH] estimates in this scenario, but Chaisemartin and D'Haultfœuille (2020) and Chaisemartin and D'Haultfoeuille (2022) make clear that it does not correspond to the ATET.

5 A New Paradigm: Matrix Completion

Such an *a priori* choice is influenced by several factors, such as data availability, shape of the panel, and plausibility of key assumptions, such as Assumption 8 for DiD designs. Athey et al. (2021) introduce a method called *Matrix Completion* to the field of Econometrics [11], which they promise sidesteps many of the problems associated with differential timing settings, such as the plausibility of associated assumptions, and even the a priori choice of method as a whole. The authors also claim that this method is more accurate than [TWFE] in a variety of panel data settings, which they show using simulated data.

The focus of Athey et al. (2021), however, is not on DiD, but rather on the relative performance of Matrix Completion Estimation (MCE) compared to regularized regressions. This leaves an opening in understanding and raises the the question at the heart of this thesis: *if this new method performs so well compared to* [TWFE], *then how does it measure up against the new DiD estimators introduced in* Section 4?

Essentially, one of the key advantages of MCE that Athey et al. (2021) claim is that it is not only more accurate than existing methods, but also universally applicable while using fewer assumptions than any DiD method (or synthetic control, for that matter) requires. The following sections explain how MCE works, and investigate whether this claim, which sounds somewhat *too good* to be entirely true, also holds when comparing MCE against recent DiD methods, which are not biased in a differential timing setting.

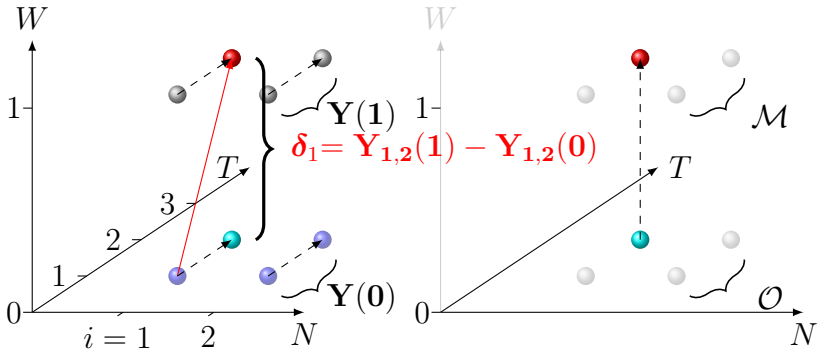
5.1 The Ultimate Estimand of Matrix Completion

The key idea underpinning MCE is simple: one way of framing the problem of causal inference under the RCM, first pointed out by Rubin (1974), is that it fundamentally is a problem of *missing data*: "*given any treatment assigned to an individual unit, the potential outcome associated with any alternate treatment is missing*" (Imbens and Rubin 2015, p. 14). If we can

11. To be clear, this method is far from new, and is very well researched in Theoretical Statistics, Computer Science, and Machine Learning, but as Economists tend not to read outside of their own discipline, and Athey et al. (2021) adapt previous MC methods to differential timing settings, it is new for my purposes.

impute certain missing PO of unit 1, visualized in cyan in Figure 5, we can immediately and directly compute the unit-specific causal effect δ_1:

Figure 5: Imputing Missing Potential Outcomes



Focusing on the (ATET) as an ultimate estimand narrows down how many such imputations we have to make: as all PO for treated units are observed, we only have to impute the missing PO for these treated units *had they never been treated* to get the ATET:

Definition (ATET over Potential Outcomes) :

Using Potential Outcome Notation, we can write the ATET as

$$\tau = \Big(\sum_{(i,t):W_{i,t}=1} [Y_{i,t}(1) - Y_{i,t}(0)] \Big) / \sum_{i,t} W_{i,t}$$

(Athey et al. 2021, p. 1717)

Like all estimators presented so far (with the exception of [TWFE]), MCE *does not directly estimate* the ATET, rather its immediate estimand is a low-rank approximation of a data matrix, which is then used to obtain the ultimate estimand. Of the estimators presented so far, it is most similar in spirit to [BJS], which also directly imputes missing POs, which are subsequently aggregated to an overall ATET.

The following sections introduce the setup in which I will frame this task, motivate the immediate estimand and estimation procedure, and characterize the estimator(s) proposed by Athey et al. (2021).

5.2 Matrix Completion Setup

To omit very frequent citation, I note that the notation and definitions in the following three sections are taken entirely from Athey et al. (2021). As before, we observe N units over T periods, where $Y_{i,t}(0)$ is the *observed* untreated PO, and unit $Y_{i,t}(1)$ is the *observed* PO if unit i has been exposed to the binary treatment. Assuming (SUTVA), for each unit in each period, we observe $(W_{i,t}, Y_{i,t})$, where the treatment indicator

$$W_{i,t} = \begin{cases} 1 & \text{if } (i,t) \in \mathcal{M}issing \\ 0 & \text{if } (i,t) \in \mathcal{O}bserved \end{cases}$$

indicates missingness of $Y_{i,t}$ in the matrix $\mathbf{Y} = \mathbf{Y(0)}$ of realized, i.e. observed, ***untreated*** outcomes. The estimator exclusively leverages information in this outcome matrix, which notably *does not include observations of treated units.* Athey et al. (2021) detail how one can incorporate those values into the estimation process, but this comes at a cost: in order to use the observed treated values, we would have to assume that the treatment effect is either constant (i.e. homogeneous), or have a *low-rank pattern* (more on this later) (Section 8.2). Assuming this, however, in a way defeats the purpose of this thesis: when treatment is (nearly) perfectly homogeneous, [TWFE] works just fine, including in differential timing settings, which negates nearly any reason to use MCE in its place. I will therefore restrict my attention to their estimator, which only uses information contained in $\mathbf{Y(0)} = \mathbf{Y}$. The matrix $\mathbf{W}$ records treatment assignment. Taken together,

$$\mathbf{Y_{N\times T}} = \begin{pmatrix} Y_{1,1} & Y_{12} & ? & \dots & Y_{1,T} \\ ? & ? & Y_{23} & \dots & ? \\ Y_{3,1} & ? & Y_{3,3} & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N,1} & ? & Y_{N,3} & \dots & ? \end{pmatrix} \text{ and } \mathbf{W}_{N\times T} = \begin{pmatrix} 0 & 0 & 1 & \dots & 0 \\ 1 & 1 & 0 & \dots & 1 \\ 0 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & 1 \end{pmatrix}$$

represent the outcome data in the form of one incomplete matrix ($\mathbf{Y}$), and one complete matrix ($\mathbf{W}$). One of the main contributions of Athey et al. (2021) is to allow for non-random patterns of

missingness in $\mathbf{Y}$, including those with time-series dependency structures, which stands in stark contrast to the origins of MCE in the Computer Science and Machine Learning literatures, where $\mathbf{Y}$ is typically assumed to be missing (completely) at random with large N, T (cf. Koltchinskii, Lounici, and Tsybakov 2011; Keshavan, Montanari, and Oh 2010).

5.3 Connecting Matrix Completion to Difference-in-Differences

Two large strands of the econometric literature, which Athey et al. (2021) call the *unconfoundedness* and *synthetic control* approaches, respectively focus on different patterns of missingness and correlational patterns in $\boldsymbol{Y}$.

Horizontal regressions focus on patterns in the time path of $Y_{i,t}$ for observed units and assume that this pattern is identical in units with missing outcomes (i.e. treated units), which means that this group of methods exploits *cross-sectional* patterns, in which the units of observation are the rows of $\boldsymbol{Y}$.

Vertical regressions, better known as *synthetic control* approaches, focus on a pattern between units in periods in which we observe all outcomes, and assume that this pattern holds for periods in which some outcomes are missing, which means that this group of methods exploits *time-series* patterns in which the units of observation are the columns of $\boldsymbol{Y}$.

$$\textit{horizontal } \mathbf{Y} = \begin{pmatrix} ✓ & ✓ & ✓ \\ \vdots & \vdots & \vdots \\ ✓ & ✓ & ✓ \\ ✓ & ✓ & ? \\ \vdots & \vdots & \vdots \\ ✓ & ✓ & ? \end{pmatrix} \text{ and } \textit{vertical } \mathbf{Y} = \begin{pmatrix} ✓ & ✓ & \dots & ✓ & ✓ & \dots & ✓ \\ ✓ & ✓ & \dots & ✓ & ✓ & \dots & ✓ \\ ✓ & ✓ & \dots & ✓ & ? & \dots & ? \end{pmatrix} \tag{18}$$

Both regressions shown in (18) rely on *block structures* of missing data, where the treated units are missing from $\boldsymbol{Y}$ in *assignment blocks* in the shape of $\begin{pmatrix} ? \\ \vdots \\ ? \end{pmatrix}$ or $\begin{pmatrix} ? & \dots & ? \end{pmatrix}$.

A third category of models that Athey et al. (2021) relate MCE to is that of *fixed effects* and *factor models* of the forms

$$Y_{i,t}(0) = \gamma_i + \delta_t + \epsilon_{i,t} \tag{19}$$

$$Y_{i,t}(0) = \sum_{r=1}^{R} u_{i,r} v_{t,r} + \epsilon_{i,t} \text{ or } \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{\epsilon} \tag{20}$$

where (19) exactly corresponds to the DGP implicitly imposed by common trends, and (20) representes a factor model with *loadings* $\boldsymbol{U}$ and *factors* $\boldsymbol{V}$, where R is the rank of $\boldsymbol{Y}$, which corresponds to the number of factors. Both types of models allow us to exploit both stable patterns over time as well as stable patterns over units. This is precisely one of the main selling points of MCE: it is very good at exploiting patterns along *all* dimensions of $\boldsymbol{Y}$. Another main contribution of the paper is to prove that, in the case of a single missing observation, unit N in period T, the proposed MC estimator, horizontal, vertical, regularized vertical *and DiD* regressions can all be characterized as using on the same objective function and only differing in the regularization and additional restrictions imposed on the parameters of the objective function (Athey et al. 2021).

Lastly, we have the differential timing setting, which unlike horizontal and vertical regressions does not follow a blocked assignment structure:

$$\mathbf{Y} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \dots & \checkmark & \text{(control unit)} \\ \checkmark & ? & ? & \dots & ? & \text{(unit in early group)} \\ \checkmark & \checkmark & ? & \dots & ? & \text{(unit in mid group)} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \checkmark & \checkmark & \checkmark & \dots & ? & \text{(unit in late group)} \end{pmatrix} \tag{21}$$

Estimators such as [TWFE] also exploit *both* cross-sectional and longitudinal correlation patters, so the main question of this thesis at its core is *whether* the new DiD estimators do this with higher accuracy than MCE, and *in which setting* which method performs better.

Although (21) looks fairly close to how one would think about observed untreated outcomes in any differential timing DiD setting, the data-generating process (DGP) is modelled very differently. A core assumption of DiD approaches, which so far was only implicitly encoded in Assumption 3 (and therefore also in 8, 9, and 11), is made explicit:

$$\boldsymbol{Y_{N\times T}} = \boldsymbol{L^*_{N\times T}} + \boldsymbol{\epsilon_{N\times T}} \quad \text{where} \quad \mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{L^*}] = \boldsymbol{0} \tag{22}$$

Definition (MCE Data-Generating Process, No Covariates) :
The complete outcome matrix $\boldsymbol{Y_{N\times T}}$ is modelled as (22)

The $\epsilon_{i,t} \in \boldsymbol{\epsilon_{N\times T}}$ can be thought of as measurement error, and are (mean) independent of $\boldsymbol{L^*}$. To illustrate what *exactly* $\boldsymbol{L^*}$ *is*, consider our earlier assumption underlying all DiD methods that the *true* DGP is linear: $Y_{i,t}(0) = \beta_i + \beta_t + e_{i,t}$ is *structural* or *causal*. Making two small modifications, we can rewrite this in matrix form:

$$Y_{i,t}(0) = \gamma_i + \delta_t + \epsilon_{i,t} \tag{23}$$

$$Y(0)_{N\times T} = \underbrace{\Gamma^i_{N\times 1}\mathbf{1}^\top_{1\times T} + \mathbf{1}^\top_{N\times 1}\Delta^t_{T\times 1}}_{} + \epsilon_{N\times T} \tag{24}$$

$$\boldsymbol{Y_{N\times T}} = \qquad \boldsymbol{L^*_{N\times T}} \qquad + \boldsymbol{\epsilon_{N\times T}} \tag{25}$$

This is one possible (albeit simple) way of expressing what the *systematic component* $\boldsymbol{L^*}$ contains. [12] In the additive case described here, DiD will consistently recover our ultimate estimand of interest. When we allow for interactions within $\boldsymbol{L^*}$ (such as in (20)), DiD is misspecified and therefore no longer matches our estimand (Bai 2009). In principle, $\boldsymbol{L^*}$ is sufficiently general to accommodate many types of DGP, the example above is merely intended to draw a connection to DiD approaches. *A priori*, we would expect MCE to perform well when there is additional information contained in the DGP that leads to correlational patterns in $\boldsymbol{Y}$ which are not fully captured by unit- and time-fixed effects. These remaining *low-rank*, that is, correlational, components of $\boldsymbol{L^*}$ will be estimated separately, in a way that is substantially

12. *Systematic component* as opposed to the *idiosyncratic* component $\boldsymbol{\epsilon}$, terminology taken from Arkhangelsky et al. (2021, Section IIA).

different from (parametric or nonparametric) DiD estimation.

To fully capture the DGP, Athey et al. (2021) make one (**!**) assumption in addition to (SUTVA):

Definition (σ-sub-Gaussian) :

A random variable X with mean $\mu = \mathbb{E}[X]$ is σ-sub-Gaussian if

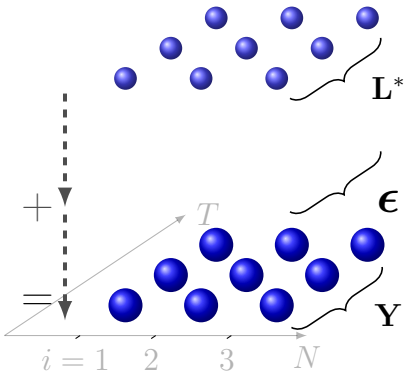
$\exists \sigma > 0$ **s.t.** $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leqslant e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall \lambda \in \mathbb{R}$ (Wainwright 2019, p. 23)

Assumption 14 (MCE Unconfoundedness / Exogeneity) :

In (22), $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{L^*}$, $\epsilon_{i,t} \in \boldsymbol{\epsilon_{N\times T}}$ are independent of each other and σ-sub-Gaussian

A σ-sub-Gaussian random variable follows a distribution with tails that decay at least as quickly as the tails of a Gaussian distribution, that is, exponentially. Athey et al. (2021) use this assumption in one of their main results, a theoretical upper bound on the estimation error of their proposed estimator. As best I can tell, it is not strictly necessary for estimation and implies Assumption 3. Figure 6 below visualizes the DGP. The *immediate* estimand of MCE is the *systematic component* $\boldsymbol{L^*}$.

Figure 6: MCE Data-Generating Process, No Covariates



5.4 The Immediate Estimand: Low-Rank Matrix Approximations

To understand why and how we want to estimate $\boldsymbol{L^*}$, it is a necessary evil to briefly delve into the realms of Linear Algebra and computer science. There are many fields, including Computer Vision, Signal Processing, and Recommendation Systems (the infamous *Netflix Problem*) which

require a solution to the same problem we are interested in: the "*recovery of a data matrix from a sampling of its entries*" (Candès and Recht 2009, p. 718). We require some additional notation to formalize this problem:

Definition (Orthogonal Projection) :

The matrix $\boldsymbol{P_{\mathcal{O}}(A)}$ is the orthogonal projection which contains all known (observed) information about $\boldsymbol{A}$, $\boldsymbol{P^{\perp}_{\mathcal{O}}(A)}$ is its orthogonal complement:

$$P_{\mathcal{O}}(A)_{i,t} = \begin{cases} A_{i,t} & \textbf{if } (i,t) \in \mathcal{O} \\ 0 & \textbf{if } (i,t) \in \mathcal{M} \end{cases} \quad \textbf{and} \quad P^{\perp}_{\mathcal{O}}(A)_{i,t} = \begin{cases} 0 & \textbf{if } (i,t) \in \mathcal{O} \\ A_{i,t} & \textbf{if } (i,t) \in \mathcal{M} \end{cases} \tag{26}$$

The matrix $\boldsymbol{Y}$ can (in principle) be recovered from $\boldsymbol{P_{\mathcal{O}}(Y)}$ if it is the *unique matrix of rank* $\leqslant r$ that is consistent with the data, that is, if $\boldsymbol{Y}$ is the *unique solution* to equation

$$\min \operatorname{rank}(\boldsymbol{L^*}) \; s.t. \; \boldsymbol{P_{\mathcal{O}}(L^*) = P_{\mathcal{O}}(Y)} \tag{27}$$

(Candes and Tao 2010, 2054)

Intuitively, a *low-rank approximation* is helpful for two reasons. First, a low-rank matrix has few linearly independent columns or rows. In the context of observed data, assuming a matrix has low rank implies that the data exhibits a certain level of regularity or structure. Second, it suggests that the observed data can be represented or compressed using fewer dimensions. This has stark computational advantages (cf. Keshavan, Montanari, and Oh 2010; Rohde and Tsybakov 2011).

Unfortunately, rank minimization is in general an NP-hard problem, which makes (27) intractable for practical purposes. Luckily for us, Candès and Recht (2009) prove that exact matrix completion is instead possible by modifying the objective function and solving the following *convex* problem:

$$\min \|\boldsymbol{L^*}\|_* \; s.t. \; \boldsymbol{P_{\mathcal{O}}(L^*) = P_{\mathcal{O}}(Y)} \tag{28}$$

Several different algorithms to approximately solve this problem are available, most of which

are efficient in terms of both space and time complexity (cf. Cai, Candès, and Shen 2010; Recht, Fazel, and Parrilo 2010).

Now that we are reassured that this problem can indeed be solved, it is time to understand how the available solutions work: in (28), $\| \circ \|_*$ denotes a particular *matrix norm*. Matrix norms serve several purposes in MCE, one of which is similar to the purpose of vector norms in modern econometric methods of *shrinkage estimation*, also called *regularization*. Certain methods, such as for example *LASSO*, *shrink* estimators toward zero, in the hope of trading off unbiasedness to gain efficiency and achieve smaller mean squared error than best unbiased estimators that attain the Cramér-Rao lower bound and are therefore efficient (cf. Hansen 2022, 306; Casella and Berger 2002, 334f.). In both Linear Algebra and Econometrics, vector and matrix norms, intuitively speaking, measure *distances*:

Definition (ℓ_1 Norm) :

$\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (Golub and Van Loan 2013, p. 69)

Also known as the *Manhattan Norm*, the ℓ_1 norm is used in *LASSO* to impose sparsity on parameter estimates (James et al. 2021, Ch. 6.2).

The *Euclidean* ℓ_2 Norm, commonly used in *Ridge* regression to penalize coefficient size, is conventionally used to measure the physical distance between two points in the $\mathbb{R}^n$:

Definition (ℓ_2 Norm) :

$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ (Golub and Van Loan 2013, p. 69)

To generalize these norms to matrices, we need the inner product:

Definition (Inner Product) :

$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathbf{trace}(A^\top B)$ (Strang 2006, p. 23)

With this definition, we arrive at the first of two main building blocks of Athey et al. (2021)'s proposed estimator(s): the *Fröbenius Norm*, is the matrix equivalent o the (ℓ_2 Norm):

Definition (Fröbenius Norm) :

$\|\boldsymbol{A}\|_{\boldsymbol{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\langle A, A\rangle}$ (Golub and Van Loan 2013, p. 71)

Intuitively, this norm allows us to measure the *distance* between two matrices, for our purposes $\boldsymbol{L}^*$ and $\boldsymbol{Y}$: what for linear regression is the $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, the Fröbenius Norm is for matrices, that is, we can use it to measure how well the low-rank approximation ($\boldsymbol{L}^*$) represents the original matrix ($\boldsymbol{Y}$). [13]

It is, however, unsuitable as an objective function for our minimization problem: giving into the econometrician's temptation of simply executing the matrix completion equivalent of the regression derivation $\min SSR$ does not lead to a useful estimator:

$$\min_{L} \frac{1}{|\mathcal{O}|}\sum_{(i,t)\in\mathcal{O}}(Y_{i,t} - L_{i,t})^2 \iff \min_{L}\frac{1}{|\mathcal{O}|}\|\boldsymbol{P_{\mathcal{O}}(Y - L)}\|_F^2 \tag{29}$$

For $(i,t) \in \mathcal{M} \implies \notin \mathcal{O}$, the objective function does not depend on $L_{i,t}$ as $L_{i,t} \notin \boldsymbol{P_{\mathcal{O}}(Y - L)}$, but for $(i,t) \in \mathcal{O}$, minimization is achieved by setting $L = 0$, so this estimator would just return $\boldsymbol{Y}$ (Athey et al. 2021, p. 1721). This is where the final building block comes into play: based on the singular values, we can compute the *Nuclear Norm*:

Definition (Singular Value Decomposition) :

Any matrix $A \in \mathbb{R}^{m\times n}$ can be factored into $A = U\Sigma_{m\times n}V^\top$

where the columns of $U_{m\times m}$ are the eigenvectors of $AA^\top$, the columns of V are the eigenvectors of $A^\top A$, and the r singular values σ_i on the diagonal of $\Sigma_{n\times n}$ are the square roots of the nonzero eigenvalues of both $AA^\top$ and $A^\top A$ (Strang 2006, p. 367)

Definition (Nuclear Norm) :

$\|\boldsymbol{A}\|_* = \sum_{i=1}^{r}\sigma_i(A)$ (Golub and Van Loan 2013, p. 71)

This norm is particularly useful to us as it allows us to write (28) as the tightest possible

13. In Machine Learning parlance, this is frequently called the *loss function*.

relaxation of the NP-hard original problem (27): in jargon, this is the case because "*the nuclear ball* $\{X : \|X\|_* \leqslant 1\}$ *is the convex hull of the set of rank-one matrices with spectral norm bounded by one*"; in intelligible terms, (27) and (28) are "*formally equivalent in the sense that they have exactly the same unique solution*" (both Cai, Candès, and Shen 2010, p. 1957). [14] This insight, finally, leads us to the starting point for Athey et al. (2021)'s proposed estimator(s):

Estimator [MHT] (Mazumder, Hastie, & Tibshirani) :

$$\hat{\boldsymbol{L}}_{\boldsymbol{MHT}} = \arg\min_{L} \left[\sum_{(i,t)\in\mathcal{O}} \frac{(\boldsymbol{Y_{i,t}} - \boldsymbol{L_{i,t}})^2}{|\mathcal{O}|} + \lambda \|\boldsymbol{L}\|_* \right] \textbf{ with } \| \circ \|_*,$$

(Mazumder, Hastie, and Tibshirani 2010, (3))[a]

a. The equivalent formulation printed here is taken from the first draft of Athey et al. (2021), which can be found on *arXiv*, to facilitate consistency.

5.5 The Matrix Completion-Nuclear Norm Minimization Estimator

Following Athey et al. (2021, Section 4), I will first introduce and explain their proposed estimator in a simple setting without covariates before introducing the more complex setting with several types of covariates. Adding a Nuclear Norm *penalty term* $\lambda\|\boldsymbol{L}\|_*$ to (28) solves our earlier problems: for $(i,t) \in \mathcal{M} \implies \notin \mathcal{O}$, the objective function now *does* depend on $L_{i,t}$ as $L_{i,t} \in \|\boldsymbol{L}\|_*$, and for $(i,t) \in \mathcal{O}$, minimization is *no longer* simply achieved by setting $L = 0$, so that in this case, it no longer just returns $\boldsymbol{Y}$ and becomes useful: [MHT] matches our Matrix Completion estimand $\boldsymbol{L}^*$.

Athey et al. (2021) adapt [MHT] to increase performance. They do this by excising unit- and time-fixed effects from $\boldsymbol{L}^*$:

$$\boldsymbol{L}^* = \hat{\boldsymbol{L}} + \hat{\boldsymbol{\Gamma}}\mathbf{1}_T^\top + \mathbf{1}_N\hat{\boldsymbol{\Delta}}^\top \tag{30}$$

Although they might seem trivial, given the inconsistent usage of the term *fixed effects*, I will define them precisely: [15]

14. It is worthwhile to note that the other matrix norms one can be used to make our low-rank approximation problem tractable (cf. Negahban and Wainwright 2012; Rohde and Tsybakov 2011; Hastie et al. 2014; Keshavan, Montanari, and Oh 2010).

15. See, for example, current debates or older ones

Definition (Unit Fixed Effects) :

$\boldsymbol{\Gamma} \in \mathbb{R}^{N\times 1}$ captures the unobserved, possibly heterogeneous, time-invariant characteristics of each individual observation. It corresponds to γ_i in $Y_{i,t}(0) = \gamma_i + \delta_t + \epsilon_{i,t}$ [a]

a. cf. Assumption 3, this corresponds with Gelman (2005, Definition 1, p. 20).

Definition (Time Fixed Effects) :

$\boldsymbol{\Delta} \in \mathbb{R}^{T\times 1}$ captures the unobserved, possibly heterogeneous, period-specific effects on all observations. It corresponds to δ_t in $Y_{i,t}(0) = \gamma_i + \delta_t + \epsilon_{i,t}$ [a]

a. cf. Assumption 3, this corresponds with Gelman (2005, Definition 1, p. 20).

subtracting those two (taken together, they are typically called *twoway fixed effects* (TWFE)) components from $\boldsymbol{L}^*$ leaves us with $\hat{\boldsymbol{L}}$:

Definition (Low-Rank Systematic Component) :

$\hat{\boldsymbol{L}}$ in (30) captures the remaining low-rank systematic components of $\boldsymbol{L}^*$ after subtracting unit- and time-fixed effects.

Athey et al. (2021) note that not regularizing the TWFE is "*conceptually similar to not regularizing the intercept term in LASSO estimator, to reduce the bias created by the regularization term* [...] *including these fixed effects separately and not regularizing them greatly improves the quality of the imputations. This is partly because compared to the settings studied in the matrix completion literature, the fraction of observed values is relatively high, and so these fixed effects can be estimated accurately.*" (p. 1721). The general form of the estimator proposed by Athey, Bayati, Doudchenko, Imbens, and Khosravi hence takes the form

Estimator [MC-NNM] (ABDIK (2021)) :

$$\boldsymbol{L}^* = \hat{\boldsymbol{L}} + \hat{\boldsymbol{\Gamma}}\mathbf{1}_T^\top + \mathbf{1}_N\hat{\boldsymbol{\Delta}}^\top \text{ where } \left(\hat{\boldsymbol{L}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Delta}}\right)$$

$$\underset{\boldsymbol{L},\boldsymbol{\Gamma},\boldsymbol{\Delta}}{\arg\min} \left\{ \frac{1}{|\mathcal{O}|} \| \boldsymbol{P}_{\mathcal{O}} \left(\boldsymbol{Y} - \boldsymbol{L} - \boldsymbol{\Gamma}\mathbf{1}_T^\top - \mathbf{1}_N\boldsymbol{\Delta}^\top \right) \|_F^2 + \lambda \|\boldsymbol{L}\|_* \right\} \quad ((8))$$

Athey et al. (2021) compute this estimator using Mazumder, Hastie, and Tibshirani (2010)'s `SOFT-IMPUTE` algorithm, which for simplicity they describe without incorporating fixed

effects: given the (Singular Value Decomposition), define the shrinkage operator (31)

$$\text{shrink}_{\lambda}(A) = U\tilde{\Sigma}_{m\times n}V^{\top} \tag{31}$$

$$\sigma_i(A) = \max\left(\sigma_i(A) - \lambda,\, 0\right) \tag{32}$$

where $\tilde{\Sigma}$ is Σ with the ith singular value replaced using (32). Now define

$$\boldsymbol{L}_1(\lambda, \mathcal{O}) = \boldsymbol{P}_{\mathcal{O}}(\boldsymbol{Y})$$

$$\text{for } k = 1, 2, \ldots, \text{ define } \boldsymbol{L}_{k+1}(\lambda, \mathcal{O}) = \text{shrink}_{\frac{\lambda|\mathcal{O}|}{2}}\left\{\boldsymbol{P}_{\mathcal{O}}(\boldsymbol{Y}) + \boldsymbol{P}_{\mathcal{O}}^{\perp}\left(\boldsymbol{L}_{\boldsymbol{k}}(\boldsymbol{\lambda}, \boldsymbol{\mathcal{O}})\right)\right\}$$

$$\text{until } \{\boldsymbol{L}_k(\lambda, \mathcal{O})\}_{k\geqslant 1} \text{ converges}$$

$$\hat{\boldsymbol{L}}(\lambda, \mathcal{O}) = \lim_{k\to\infty} \boldsymbol{L}_k(\lambda, \mathcal{O}) \tag{33}$$

(33) is the limiting matrix, which corresponds to [MC-NNM]. Intuitively, the estimator is computed by iteratively regularizing the singular values of the (Orthogonal Projection) of Y and its complement. Athey et al. (2021, p. 1721) verbally describe how to incorporate TWFE, which I have synthesized with their description of the estimation process and Mazumder, Hastie, and Tibshirani (2010)'s original algorithmic representation into Algorithm 1 below.

Algorithm 1 `SOFT-IMPUTE` with TWFE *

Initialize empty list $\boldsymbol{Results}$

Choose Cross-Validation parameter $\boldsymbol{K}$, e.g. $\boldsymbol{K} = 5$

for all $\lambda_1 > \ldots > \lambda_L = 0$ **do**

  Initialize $\boldsymbol{L}_1(\lambda_l, \mathcal{O}) = \boldsymbol{P_{\mathcal{O}}(Y)}$, tolerance ϵ

  for all $k = 1, 2, \ldots,$ **do**

    $\boldsymbol{L}_{k+1}(\lambda_l, \mathcal{O}) = \text{shrink}_{\frac{\lambda_l |\mathcal{O}|}{2}} \left\{ \boldsymbol{P_{\mathcal{O}}} \left(\boldsymbol{Y - \Gamma_k 1_T^\top - 1_N \Delta_k^\top} \right) + \boldsymbol{P_{\mathcal{O}}^\perp} \left(\boldsymbol{L_k(\lambda_l, \mathcal{O})} \right) \right\}$

    $\boldsymbol{\Gamma}_{k+1} \leftarrow \frac{\partial}{\partial \boldsymbol{\Gamma}_k} \left(\boldsymbol{L}_{k+1}(\lambda_l, \mathcal{O}) \right) \overset{!}{=} 0$

    $\boldsymbol{\Delta}_{k+1} \leftarrow \frac{\partial}{\partial \boldsymbol{\Delta}_k} \left(\boldsymbol{L}_{k+1}(\lambda_l, \mathcal{O}) \right) \overset{!}{=} 0$

    if $\frac{\|\boldsymbol{L}_{k+1}(\lambda_l, \mathcal{O}) - \boldsymbol{L}_k(\lambda_l, \mathcal{O})\|_F^2}{\|\boldsymbol{L}_k(\lambda_l, \mathcal{O})\|_F^2} < \epsilon$ **then**

      $\left(\hat{\boldsymbol{L}}_{\lambda_l}, \hat{\boldsymbol{\Gamma}}_{\lambda_l}, \hat{\boldsymbol{\Delta}}_{\lambda_l} \right) \leftarrow \left(\boldsymbol{L}_{k+1}(\lambda_l, \mathcal{O}), \boldsymbol{\Gamma}_{k+1}, \boldsymbol{\Delta}_{k+1} \right)$

      break

    else

      continue

    end if

  end for

  evaluate $\textbf{MSE}_{\boldsymbol{\lambda_l}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\boldsymbol{Y}_{i,t} - \hat{\boldsymbol{L}}_{\lambda_l; i,t} \right)^2$ on $\mathcal{O} \setminus \mathcal{O}_k$, where $\mathcal{O}_k \subset \mathcal{O}$ are $\boldsymbol{K}$ randomly sampled subsets, each with cardinality*** $\lfloor |\mathcal{O}|^2 / NT \rfloor$

  return $\left(\left(\hat{\boldsymbol{L}}_{\lambda_l}, \hat{\boldsymbol{\Gamma}}_{\lambda_l}, \hat{\boldsymbol{\Delta}}_{\lambda_l} \right), \textbf{MSE}_{\boldsymbol{\lambda_l}} \right)$ tuple to $\boldsymbol{Results}$ **

end for

Notes: * Adapted from Athey et al. (2021, sections 4.2-3) and Mazumder, Hastie, and Tibshirani (2010, Alg. 1)

** Existing implementations return the entire list of results $\boldsymbol{Results}$, from which the optimal (minimal) value of λ can be selected

*** *Cardinality* refers to the fraction of observed values being kept constant when randomly selecting the K folds

Finally, we select the optimal value of the regularization parameter λ through cross-validation: for a chosen K, e.g. $K = 5$ "*random subsets* $\mathcal{O}_k \subset \mathcal{O}$ *with cardinality* $\lfloor |\mathcal{O}|^2 / NT \rfloor$. *For a sequence* $\lambda_1 > \ldots > \lambda_L = 0$ *with a sufficiently large* λ_1, *and for each subset* $\mathcal{O}_k$ *compute* $\hat{\boldsymbol{L}}(\lambda_1, \mathcal{O}_k), \ldots, \hat{\boldsymbol{L}}(\lambda_L, \mathcal{O}_k)$ *and evaluate the mean squared error (MSE) on* $\mathcal{O} \setminus \mathcal{O}_k$. *The value of* λ *which minimizes MSE is chosen*" (Athey et al. 2021, Section 4.3).

Notably, unlike the differential timing estimators presented in section 4, [MC-NNM] does not produce standard errors or confidence intervals (CI). Athey et al. (2021) briefly mention how one could go about constructing such a CI based on permutation methods from the synthetic control literature.

5.6 The MC-NNM Estimator With Covariates

This section very briefly presents the necessary extensions to accommodate covariates, following Athey et al. (2021, Section 8.1). In the setup and DGP used thus far, let

Definition (Unit-Specific Covariates) :

For unit i we observe a vector X_i and $\boldsymbol{X}_{N\times P}$ with row $i = X_i^\top$

Definition (Time-Specific Covariates) :

For period t we observe a vector Z_t and $\boldsymbol{Z}_{T\times Q}$ with row $t = Z_t^\top$

Definition (Unit-Time-Specific Covariates) :

For period t and unit i we observe $V_{i,t} \in \mathbb{R}_{j\times 1}$

Definition (Linear Terms in Covariates) :

$\tilde{\boldsymbol{X}} = [\boldsymbol{X}|\boldsymbol{I}_{N\times N}]$, $\tilde{\boldsymbol{Z}} = [\boldsymbol{Z}|\boldsymbol{I}_{T\times T}]$, and

$$\tilde{\boldsymbol{H}}^* = \begin{bmatrix} \boldsymbol{H}^*_{\boldsymbol{X},\boldsymbol{Z}} & \boldsymbol{H}^*_{\boldsymbol{X}} \\ \boldsymbol{H}^*_{\boldsymbol{Z}} & 0 \end{bmatrix} \text{ with } \boldsymbol{H}^*_{\boldsymbol{X},\boldsymbol{Z}} \in \mathbb{R}^{\boldsymbol{P}\times\boldsymbol{Q}}, \boldsymbol{H}^*_{\boldsymbol{Z}} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{Q}}, \boldsymbol{H}^*_{\boldsymbol{X}} \in \mathbb{R}^{\boldsymbol{P}\times\boldsymbol{T}}, \tilde{\boldsymbol{H}}^* \mathbb{R}^{(\boldsymbol{N}+\boldsymbol{P})\times(\boldsymbol{T}+\boldsymbol{Q})}$$

These definitions allow definition of what Athey et al. (2021, p. 1727) call their *rich* model:

$$\boldsymbol{Y} = \boldsymbol{L}^* + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{H}}^*\tilde{\boldsymbol{Z}}^\top + \boldsymbol{\Gamma}^*\boldsymbol{1}_{\boldsymbol{T}}^\top + \boldsymbol{1}_{\boldsymbol{N}}(\boldsymbol{\Delta}^*)^\top + \left[\boldsymbol{V}_{\boldsymbol{i},\boldsymbol{t}}^\top\boldsymbol{\beta}^*\right]_{\boldsymbol{i},\boldsymbol{t}} + \boldsymbol{\epsilon} \tag{34}$$

Definition (MCE Data-Generating Process) :

The complete outcome matrix $\boldsymbol{Y}_{\boldsymbol{N}\times\boldsymbol{T}}$ is modelled as (34)

The augmented estimator is now given by [MC-NNM (Covariates)] below. Athey et al. (2021, p. 1727) elaborate on how to incorporate the covariates defined above into Algorithm 1 using modifications of (31) and (32) to incorporate coordinate descent with respect to $\boldsymbol{H}$ (Mazumder, Friedman, and Hastie 2011) . They also explain how one could go about incorporating autocorrelated error structures or weighted loss functions into [MC-NNM] and [MC-NNM (Covariates)]. Given the scope of this thesis, I will restrict myself to using existing implementations of [MC-NNM] and [MC-NNM (Covariates)] in the R package `fect`.

Estimator [MC-NNM (Covariates)] (ABDIK (2021)) :

To estimate $\left(\hat{\boldsymbol{H}}^*, \hat{\boldsymbol{L}}^*, \hat{\boldsymbol{\Gamma}}^*, \hat{\boldsymbol{\Delta}}^*\right)$ in (34), we solve the convex program

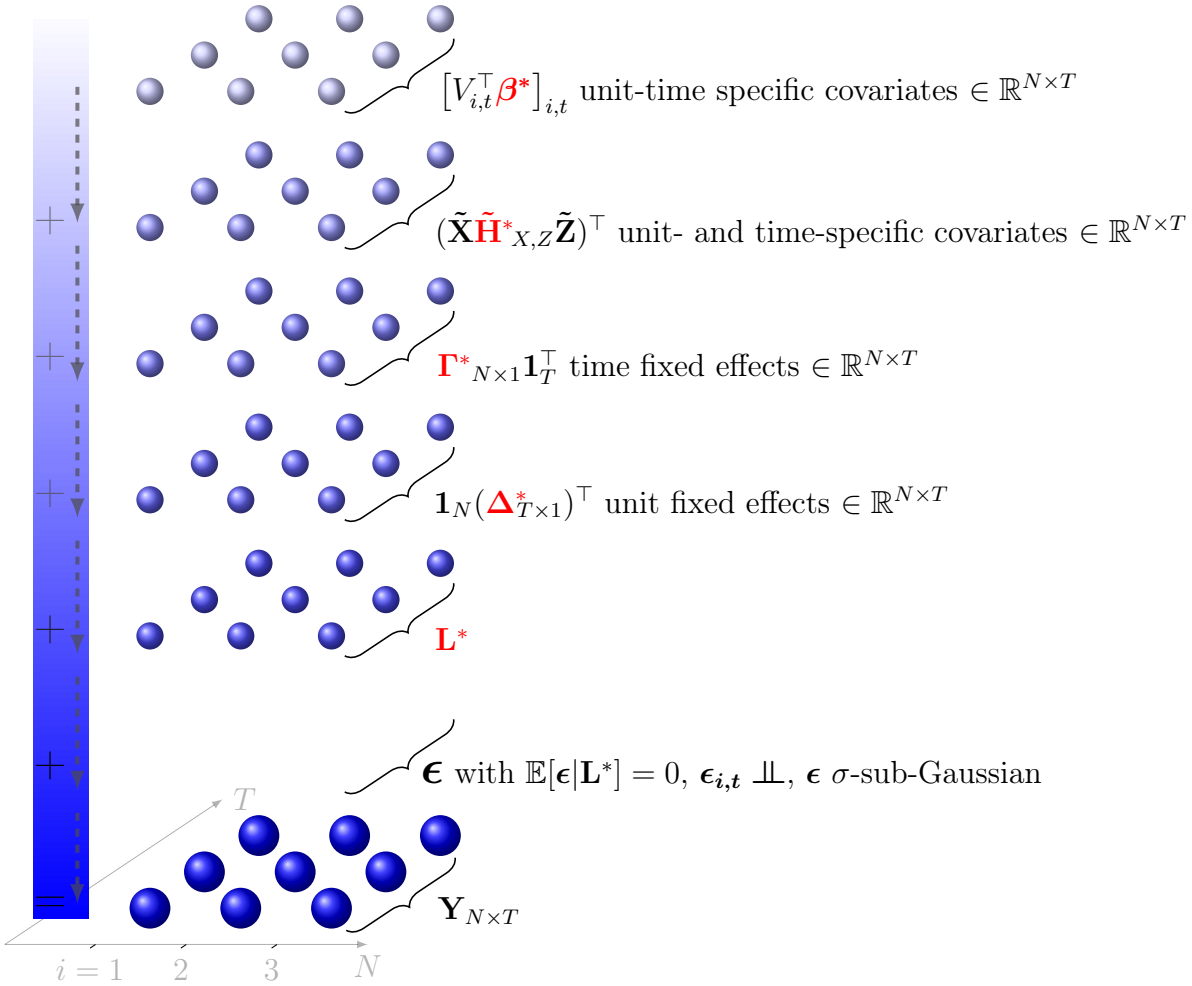
$$\min_{\boldsymbol{H},\boldsymbol{L},\boldsymbol{\Delta},\boldsymbol{\Gamma},\boldsymbol{\beta}} \left[\frac{1}{|\mathcal{O}|} \left\| \boldsymbol{P}_{\mathcal{O}} \left(\boldsymbol{Y} - \boldsymbol{L} - \tilde{\boldsymbol{X}} \boldsymbol{H}^* \tilde{\boldsymbol{Z}}^\top - \boldsymbol{\Gamma}^* \mathbf{1}_{\boldsymbol{T}}^\top - \mathbf{1}_{\boldsymbol{N}} (\boldsymbol{\Delta}^*)^\top - \left[\boldsymbol{V}_{i,t}^\top \boldsymbol{\beta}^* \right]_{i,t} \right) \right\|_{\boldsymbol{F}}^{\mathbf{2}} + \lambda_L \|\boldsymbol{L}\|_{\boldsymbol{*}} + \lambda_H \|\boldsymbol{H}\|_{\mathbf{1},\boldsymbol{\ell}} \right]$$

where $\|\boldsymbol{H}\|_{1,\ell} = \sum_{i,t} |H_{i,t}|$ denotes the element-wise (ℓ_1 Norm)

adapted from Athey et al. (2021, (18))

Figure 7 below serves as a visual aide to [MC-NNM (Covariates)].

Figure 7: MCE Data-Generating Process, With Covariates



5.7 Interpreting The Ultimate Estimand

One assumption underlying [MC-NNM] which I have not explicitly discussed is that of a time-invariant treatment effect (Assumption 11). Athey et al. (2021, 1717) write that "*[i]mplicit in our setup is that we rule out dynamic effects and make the stable-unit-treatment-value assumption (Imbens and Rubin 2015)* [...] *Cases where such assumptions are restrictive include those analyzed in the dynamic treatment regime literature* ([...] Hernán and Robins (2021)). *In the case where units are only exposed to the treatment in the last period this issue is not material.*" This, at first sight, seems to make the comparison of [MC-NNM] to all estimators presented in section 4, with the exception of [dCdH] pointless. After all, what makes these estimators attractive is the fact that they accommodate time-varying treatment effects.

All hope is, however, not lost: "*in the case with staggered adoption violations of the*

no-dynamics assumption simply changes the interpretation of the estimand, but does not in general invalidate a causal interpretation" (Athey et al. 2021, 1717). This phrasing necessarily raises the question of what the interpretation of the ultimate estimand is in settings where the treatment effect varies over time (as well as across units, which [MC-NNM] already accommodates).

Revisiting [BJS (event-study)] gives some helpful intuition here. When we can aggregate imputed POs into group-time ATETs identical to the immediate estimands of [CS (event-study)], [MS (event-study)], and [SA (event-study)], and even further aggregate those into overall *ultimate* parameters such as the (ATET over Potential Outcomes) and/or event-study type parameters, then that is exactly what we would be doing by applying [MC-NNM] in a setting where Assumption 11 is violated.

To ensure that this intuition is correct, I asked the corresponding author of Athey et al. (2021), Guido Imbens, who kindly took the time to confirm: "*if there are dynamic effects, you would be estimating some complicated average of the dynamic effects, along the lines of* [CS]. *You could focus on particular weighted averages like* (Liu, Wang, and Xu 2022). *It is not obvious what the most natural thing would be to focus on*" (Appendix A). I have chosen to focus on a time-average of group-time ATETs, which is easily comparable between estimators and is explained in detail in subsection 6.3.

Determining the exact statistical properties and precise interpretation of this estimand is squarely outside the scope of this thesis and its author's skill set. I will therefore trust that a sensible interpretation of this ultimate estimand exists and compare the estimators presented so far in multiple different ways in a simulation-based setup inspired by Baker, Larcker, and Wang (2022) and Liu, Wang, and Xu (2022).

6 Performance Comparisons

This section analyzes the performance of the estimators discussed so far in Monte Carlo simulations based on several different Data-Generating Processes (cf. Cameron and Trivedi 2005, Section 7.7.1). All code is designed with reproducibility in mind and can be found (along with replication instructions) on GitHub. The analysis was carried out in `RStudio` version `2023.6.0.421` using `R` v. `4.3.1`. The estimators in section 4 are implemented using the packages shown in Table 1 (`did2s` v. `1.0.2`; `DIDmultiplegt` v. `0.1.0`; `didimputation` v. `0.3.0`). . Unfortunately, `MarcusSantAnna2020` v. `0.1.1`; has not been updated in the last three years and is therefore not functional. I was therefore unable to implement either [MS] or [MS (event-study)]. Similarly, `DDImultiplegt` produces many errors and was removed from `CRAN` during the time I implemented these comparisons. It produces estimates of the ATET, but neither standard errors nor event-study-type estimates. I therefore omit results for event-study type parameter estimation for [dCdH].

6.1 Computational Implementations

To implement [MC-NNM], there are three packages available: `MCPanel`, maintained by two authors of Athey et al. (2021), and `gsynth` and `fect`, both maintained by two authors of Liu, Wang, and Xu (2022). The former package is not documented and was last updated in 2017. The latter packages are straightforward to use, and achieve lower RMSE in a small comparison that can be found in the `MC-Implementation-Comparison` file here. To resolve why this might be the case, I contacted Yiqing Xu, who is one of the maintainers of `fect`. He confirmed that the implementations of `MCPanel`, `gsynth` and `fect` were identical except for minor details (Appendix A). `fect` has one advantage over `gsynth` for my purposes, as it includes an option to obtain period-specific ATET estimates. This option, and the fact that `fect` is substantially easier to use than `MCPanel` and provides bootstrapped standard errors for [MC-NNM], led me to exclusively rely on `fect` v. `1.0.0` to implement [MC-NNM (Covariates)] and [MC-NNM]. References for all packages can be found in section 9.

6.2 Data Generation and Simulation Setup

I simulate outcome data using a total of eight different Data-Generating Processes (DGPs). These DGPs are adapted from and extend the work of Baker, Larcker, and Wang (2022), who use real financial data as a baseline, from which they then draw simulated unit- ($\widetilde{\gamma}_i$) and time- ($\widetilde{\delta}_t$) fixed effects. This idea in turn is based on Bertrand, Duflo, and Mullainathan (2004), who use MC simulations to assess the trustworthiness of DiD estimates, an idea similarly adopted by Arkhangelsky et al. (2021), who include [MC-NNM] as a benchmark estimator. The latter authors also report bootstrapped *placebo* standard errors for [MC-NNM], a practice which they do not describe in the paper, and for which I have not found a corresponding article to in the context of [MC-NNM], but which I will nonetheless adopt given that two of the authors of Arkhangelsky et al. (2021) (Susan Athey and Guido Imbens) also authored the original Athey et al. (2021).

The first six of these eight DGPs are in a sense *tailor-made* for DiD- and DiD-adjacent estimators. If [MC-NNM] performs well in these settings, it could be a convincing off-the-shelf alternative to the estimators presented in section 4. In the following, let $\mathbb{1}$ denote dummy variables, t a variable that stores the period, t_{Gi} a variable that stores the period in which group i is first treated, and G_{Gi} a dummy variable storing membership of an observation in each treatment group. The DGPs are modelled as follows:

$$\widetilde{Y}_{i,t}(0) = \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \tag{35}$$

$$\widetilde{\gamma}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5), \quad \widetilde{\delta}_t \overset{i.i.d}{\sim} \mathcal{N}(0, 0.5), \quad \widetilde{\epsilon}_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0, 0.5), \quad \widetilde{\tau}^{\sigma}_{\mu} \overset{i.i.d}{\sim} \mathcal{N}(0, x) \tag{36}$$

The untreated (observed and potential) outcome is given by (35), and is composed of unit-and time-fixed effects with additive mean-zero noise. Parameterizations for the fixed effects and error are given in (36), and are identical throughout all simulations. This underlying setup ideally suits DiD methods: a quick comparison with Assumption 3 shows that this linear model guarantees parallel trends, and therefore satisfies the central identifying assumption of [TWFE] as well as all estimators presented in section 4. The simulations differ only in how the treatment effect(s) $\widetilde{\tau}^{\sigma}_{\mu}$ are assigned and parameterized. In the remainder of this section, let $\widetilde{Y}_{i,t}$ denote the

simulated observed outcomes (both treated and untreated).

All simulation functions are coded to handle near-arbitrary N, T, and treatment timings, but for this thesis I chose the following parameters: all DGPs are simulated for $N = 1,000$ units over $T = 100$ periods, where the units are randomly divided into (nearly perfectly) equally-sized groups. Treatment is assigned to groups using $\widetilde{\tau}_{\mu}^{\sigma}$ with a group-specific μ, which ensures that in each run of each simulation, the true treatment effect varies as given by the distribution of $\widetilde{\tau}_{\mu}^{\sigma}$ in (36). The group-specific standard deviation σ varies between the DGP, but is identical for all groups in each DGP. [16] One group remains untreated throughout the panel and serves as a control group, it therefore has $\widetilde{\tau}_{\mu}^{\sigma} = 0$. In DGPs 1 and 2, this is the group for which $\mathbb{1}_{treat} = 0$, in the remaining DGPs, group 5 serves as never-treated control.

6.2.1 Simulation 1

$$\widetilde{Y}_{i,t}^{1} = \widetilde{\tau}_{1}^{0.2} \mathbb{1}_{treat} \mathbb{1}_{post} + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \tag{37}$$

(37) generates data that features a single treatment group treated in period 50 with a treatment that satisfies Assumption 11 (time invariance) and Assumption 12 (homogeneity across units). In this setting, [TWFE] identifies the (ATET).

6.2.2 Simulation 2

$$\widetilde{Y}_{i,t}^{2} = \widetilde{\tau}_{0.05}^{0.025} \mathbb{1}_{treat} \mathbb{1}_{post}(t - post - 1) + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \tag{38}$$

(38) generates data with a unit-homogeneous but time-varying treatment effect, which increases by 5% every period after initial treatment in period 50. This setting satisfies the assumptions of the event-study type estimators ([CS (event-study)], [SA (event-study)], [MS (event-study)], and [BJS (event-study)]).

16. The variation in σ between DGPs only serves to ensure that the range of $Y_{i,t}$ values drawn is sufficiently tight to be plotted.

6.2.3 Simulation 3

$$\widetilde{Y}^3_{i,t} = \widetilde{\tau}^{0.2}_1\big(G_{G1}\mathbb{1}_{t\geqslant G1} + G_{G2}\mathbb{1}_{t\geqslant G2} + G_{G3}\mathbb{1}_{t\geqslant G3} + G_{G4}\mathbb{1}_{t\geqslant G4}\big) + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \tag{39}$$

(39) generates data with a unit-homogeneous and time-invariant treatment effect that is assigned to 4 distinct treatment groups (treatment in periods 20, 40 , 60, and 80), that is, it corresponds to a differential timing design. Unlike Baker, Larcker, and Wang (2022), one group stays untreated throughout the entire panel (group 5). All estimators in section 4 can accommodate settings in which there is no untreated group, but [MC-NNM] cannot. The reason for this lies in the theoretical upper bound for the estimation error and is explained in Athey et al. (2021, Remark 6.1) . This setting satisfies the identifying assumptions of [dCdH] and all other static estimators ([CS], [SA], [MS], and [BJS]).

6.2.4 Simulation 4

$$\widetilde{Y}^4_{i,t} = \left(\widetilde{\tau}^{0.2}_{1.5}G_{G1}\mathbb{1}_{t\geqslant G1} + \widetilde{\tau}^{0.2}_{1}G_{G2}\mathbb{1}_{t\geqslant G2} + \widetilde{\tau}^{0.2}_{0.5}G_{G3}\mathbb{1}_{t\geqslant G3} + \widetilde{\tau}^{0.2}_{0.25}G_{G4}\mathbb{1}_{t\geqslant G4}\right) + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \tag{40}$$

(40) generates data with a time-invariant but unit-heterogeneous treatment effect with treatment groups 20, 40, 60, and 80. This setting satisfies the assumptions of [dCdH] and all other *static* estimators.

6.2.5 Simulation 5

$$\begin{aligned}\widetilde{Y}^5_{i,t} = \widetilde{\tau}^{0.025}_{0.05}&\left(G_{G1}\mathbb{1}_{t\geqslant G1} + G_{G2}\mathbb{1}_{t\geqslant G2} + G_{G3}\mathbb{1}_{t\geqslant G3} + G_{G4}\mathbb{1}_{t\geqslant G4}\right) \\ &\times \left(t - t_{G1}G_{G1} - t_{G2}G_{G2} - t_{G3}G_{G3} - t_{G4}G_{G4} - t_{G5}G_{G5}\right) + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t}\end{aligned} \tag{41}$$

(41) generates data with a unit-homogeneous but time-varying treatment effect with treatment groups 20, 40, 60, and 80. This setting again satisfies the assumptions of the *dynamic* estimators.

6.2.6 Simulation 6

$$\widetilde{Y}_{i,t}^{6} = \left(\widetilde{\tau}_{0.05}^{0.025} G_{G1}\mathbb{1}_{t\geqslant G1} + \widetilde{\tau}_{0.075}^{0.025} G_{G2}\mathbb{1}_{t\geqslant G2} + \widetilde{\tau}_{0.1}^{0.025} G_{G3}\mathbb{1}_{t\geqslant G3} + \widetilde{\tau}_{0.2}^{0.025} G_{G4}\mathbb{1}_{t\geqslant G4}\right) \times \left(t - t_{G1}G_{G1} - t_{G2}G_{G2} - t_{G3}G_{G3} - t_{G4}G_{G4} - t_{G5}G_{G5}\right) + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \tag{42}$$

(42) generates data with a unit-heterogeneous *and* time-varying treatment effect with treatment groups 20, 40, 60, and 80. This setting satisfies the assumptions of the *dynamic* estimators.

6.2.7 Simulations 7 and 8

$$\widetilde{Y}_{i,t}^{7} = \left(\widetilde{\tau}_{0.05}^{0.025} G_{G1}\mathbb{1}_{t\geqslant G1} + \widetilde{\tau}_{0.075}^{0.025} G_{G2}\mathbb{1}_{t\geqslant G2} + \widetilde{\tau}_{0.1}^{0.025} G_{G3}\mathbb{1}_{t\geqslant G3} + \widetilde{\tau}_{0.2}^{0.025} G_{G4}\mathbb{1}_{t\geqslant G4}\right) \times \left(t - t_{G1}G_{G1} - t_{G2}G_{G2} - t_{G3}G_{G3} - t_{G4}G_{G4} - t_{G5}G_{G5}\right) + \underbrace{\left(0.005 \times group \times t + \widetilde{X}_{i,t}\right)}_{\text{Nuisance Variable}} + \widetilde{\gamma}_i + \widetilde{\delta}_t + \widetilde{\epsilon}_{i,t} \text{ where } \widetilde{X}_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0, 0.02) \tag{43}$$

Finally, (43) generates data with unit-heterogeneous *and* time-varying treatment effect with treatment groups 20, 40, 60, and 80. Additionally, it includes a nuisance term that breaks (unconditional) parallel trends. The difference between simulations 7 and 8 lies only in whether the estimation includes the nuisance term as a covariate: simulation 7 estimates with the covariate; in this setting, all DiD-adjacent methods rely on conditional parallel trends assumptions. Simulation 8 estimates without the covariate, thereby violating the respective central identifying assumptions of all methods except for [MC-NNM].

6.3 Estimation Process and Ultimate Estimands

Each of the eight simulations is carried out separately. I simulate 500 iterations of each DGP. On each iteration, I store the true ATET τ as defined earlier and the true event-study

estimate, which I define as the time-average of period-specific ATETs (τ_t), formally defined as:

$$\tau^{ES} = \frac{1}{T}\sum_{t=1}^{T}\tau_t = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sum \mathbb{1}_{\{W_{i,t}=1\}}}\sum_{i=1}^{N}\left(\tau_{i,t} \times \mathbb{1}_{\{W_{i,t}=1\}}\right)\right) \tag{44}$$

where $\mathbb{1}_{\{W_{i,t}=1\}}$ is an indicator that equals 1 if unit i is under treatment in period t. I describe how the true effects are computed in my simulations in appendix A.1.

From the DiD perspective, this parameter corresponds to estimating the event-study regression $Y_{i,t} = \gamma_i + \delta_t + \beta_r \sum_{r \neq -1} \tau_r \mathbb{1}_{R_{i,t}=r} + e_{i,t}$ introduced earlier. For simulations 1 and 2, where there is only one treatment group, this is given by $\frac{1}{100}\sum_{r=-50}^{50}\beta_r$. For simulations 3 to 8, with treatment groups 20, 40, 60, and 80, this parameter is estimated as $\frac{1}{100}\sum_{r=-20}^{80}\beta_r$.

Estimator [TWFE] (event-study) (Canonical TWFE (event-study)) :

In $Y_{i,t} = \gamma_i + \delta_t + \beta_r \sum_{r \neq -1} \tau_r \mathbb{1}_{R_{i,t}=r} + e_{i,t}$ **,** $\hat{\tau}^{ES}_{TWFE} = \frac{1}{T}\sum_{r:\ |r|=T}\beta_r$ **is the twoway fixed-effects estimator corresponding to (44), where the** r **reflect the last** 100 **relative periods**

(44) represents a one-number estimate of the ATET when Assumption 11 is violated, that is, when the treatment effect varies over time. To make this number comparable across estimators, I wrote wrapper functions around the package implementations of [TWFE], [CS (event-study)], [SA (event-study)], [BJS (event-study)] to take the mean of all calendar-period estimates for each estimator. In DGPs 1, 3, 4, and 5, this event-study-type parameter is not particularly interesting (but nonetheless reported in the appendix). In DGPs 2, 5, 6, 7, and 8, however, where the true treatment effect increases over time, the estimates of the *static* τ are not of interest, as they do not identify the (ATET) (Roth et al. 2023). In these settings, I use the event-study estimates to asses the accuracy of each estimator.

6.4 Results

After storing the true values, I estimate each *static* ATET using [MC-NNM],[TWFE], [CS], [SA], [dCdH], and [BJS] and each *dynamic* event-study parameter using [MC-NNM], [TWFE] (event-study), [CS (event-study)], [SA (event-study)], and [BJS (event-study)]. [dCdH] in principle

supports event-study estimates, but the package's functionality in its current version does not produce results, so they are omitted. After the 500 simulation iterations are complete, results are stored and the processed further. Given the high computational cost of simulating $8 \times 500 = 4,000$ iterations of $5 \times 2 + 1 = 11$ estimators, in simulations 1 to 7, I use only 100 bootstrap iterations, $k = 2$ folds for cross-validation of the $num_{lam} = 4$ candidate regularization parameters λ for the estimation of [MC-NNM] described in Algorithm 1. As simulation 8 produces data that is more challenging for all estimators, I use 100 bootstrap iterations, $k = 4$ folds for cross-validation of the $num_{lam} = 6$ candidate regularization parameters λ for the estimation of [MC-NNM].

On a 10-core Apple *M1 Pro* CPU, simulating 500 iterations of one DGP and obtaining the 11 estimates for each takes between 4 and 8 hours, with simulation 1 having the lowest, and simulation 8 the highest execution time. For the sake of brevity, for each simulation, I will only contrast the performance of [MC-NNM] to that of the DiD (-adjacent) estimator we would *a priori* expect to perform well in each scenario.

6.4.1 Simulation 1

Figure 8 below shows the raw data from one draw and distribution of point estimates from the full simulation of (37). Numerical results are displayed in Table 3 below and in Table 11 in the Appendix. As expected, [TWFE] (denotes as DiD in the tables) identifies τ very well.

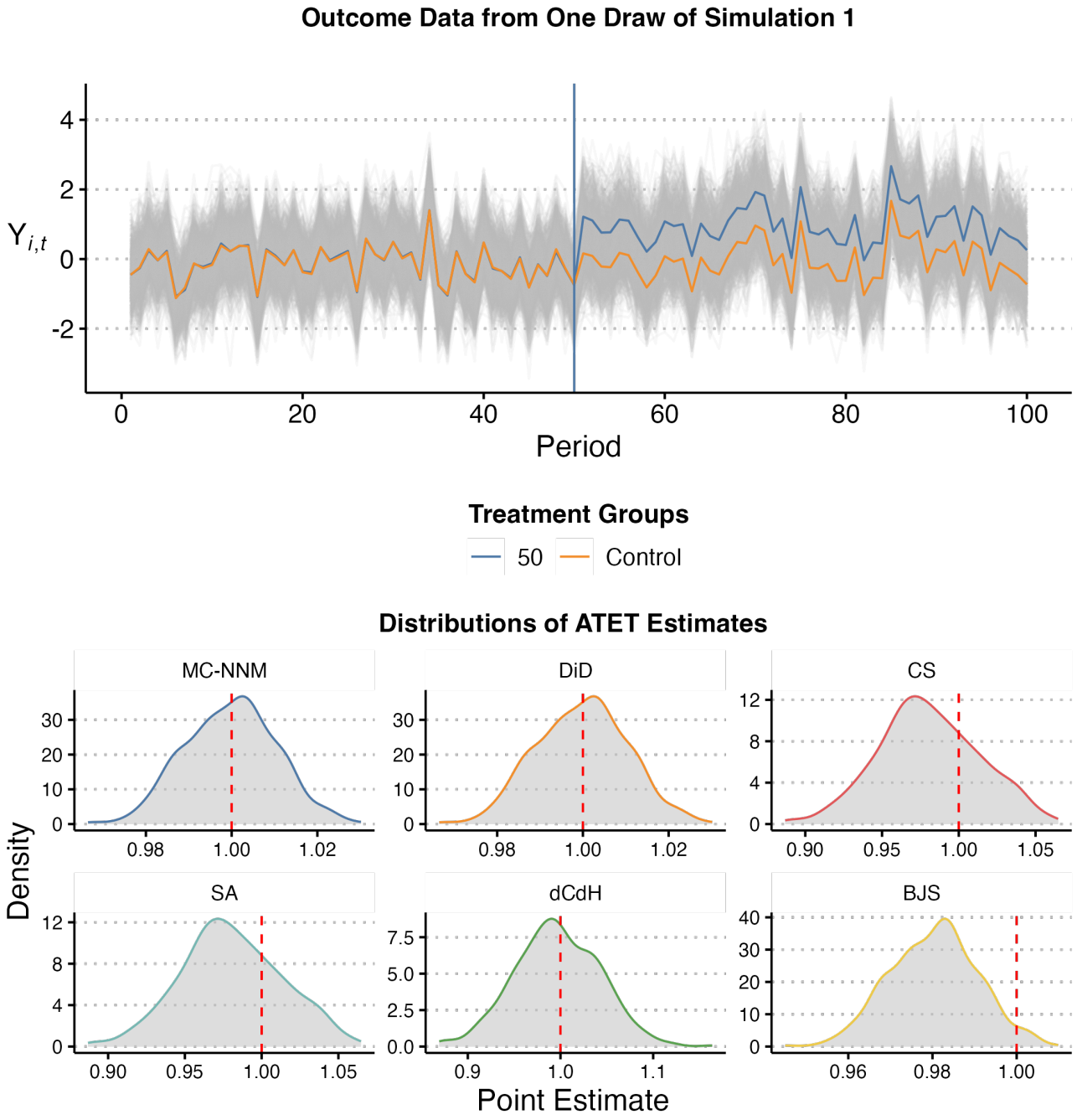
Table 3: Simulation 1, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**1.00**	**1.00**	**1.00**	**0.000**	**-**	**-**	**-**
MC-NNM	0.97	1.00	1.03	0.011	0.01	0.00	0.01
DiD	0.97	1.00	1.03	0.011	0.01	0.00	0.01
CS	0.89	0.98	1.06	0.033	0.03	-0.02	0.04
SA	0.89	0.98	1.06	0.033	0.03	-0.02	0.04
dCdH	0.87	1.00	1.16	0.046	-	0.00	0.05
BJS	0.95	0.98	1.01	0.010	0.01	-0.02	0.02

Results obtained from 500 iterations

Interestingly, [MC-NNM] achieves identical bias, RMSE, mean standard error (Mean SE), and range of point estimates.

Figure 8: Results from Simulation 1



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the ATET as defined on p. 12. Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.2 Simulation 2

Figure 9 below shows the raw data from one draw and distribution of point estimates from the full simulation of (38). Numerical results are displayed in Table 4 below and in Table 12 in the Appendix.

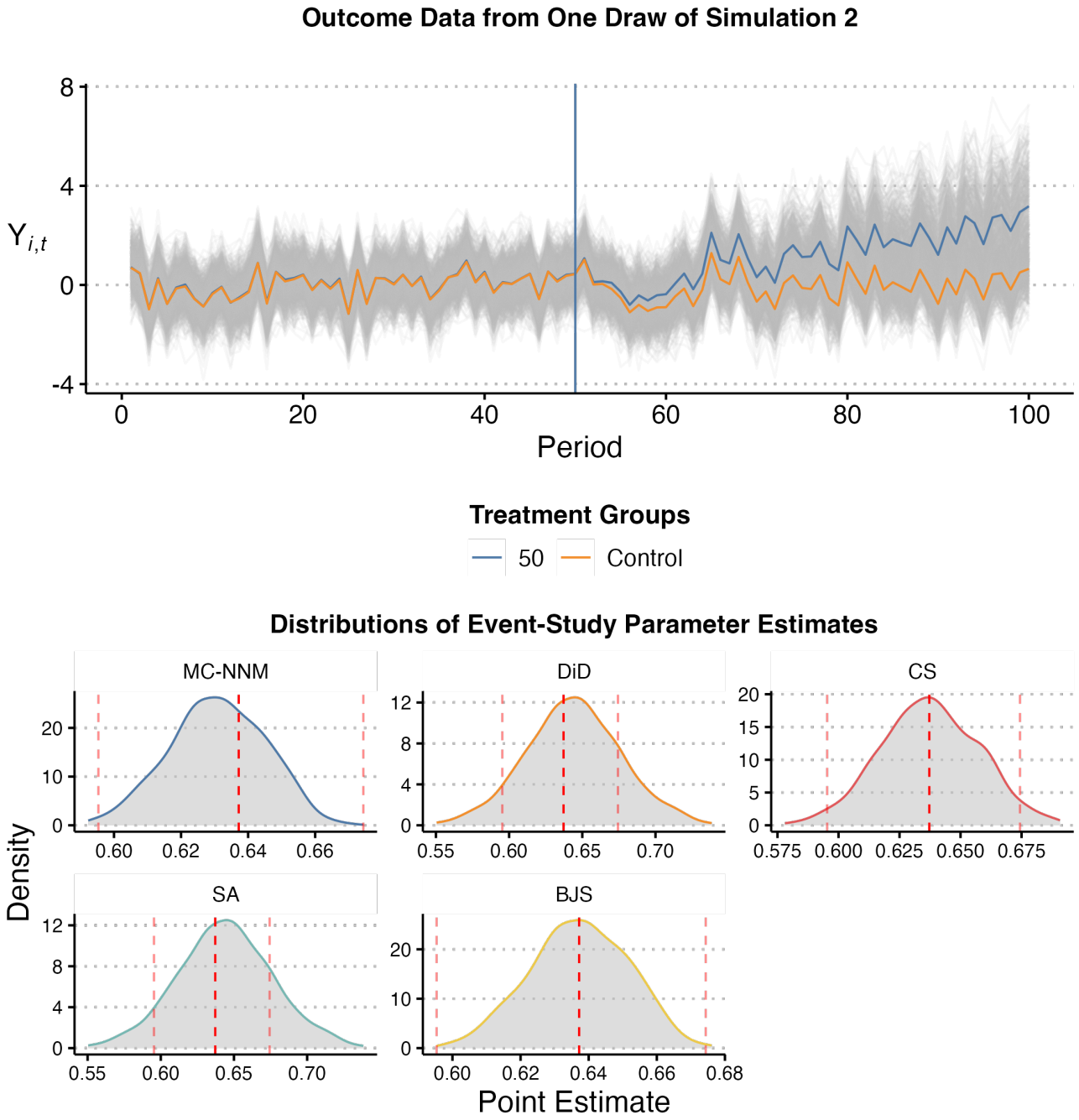
Table 4: Simulation 2, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.60**	**0.64**	**0.67**	**0.014**	**-**	**-**	**-**
MC-NNM	0.59	0.63	0.67	0.014	0.02	-0.01	0.02
DiD	0.55	0.64	0.74	0.032	0.05	0.01	0.03
CS	0.58	0.64	0.69	0.020	0.03	0.00	0.02
SA	0.55	0.64	0.74	0.032	0.05	0.01	0.03
BJS	0.60	0.64	0.68	0.014	0.02	0.00	0.01

Results obtained from 500 iterations

As this simulation has a time-varying but unit-homogeneous treatment effect, we would expect [TWFE] (event-study) to identify τ^{ES} well, which it does. [MC-NNM] performs only marginally worse than the best-performing estimator in this simulation, [BJS (event-study)], and together with [BJS (event-study)] has the lowest mean SE even with only 100 bootstrap iterations.

Figure 9: Results from Simulation 2



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the event-study parameter defined in (53). Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.3 Simulation 3

Figure 10 below shows the raw data from one draw and distribution of point estimates from the full simulation of (39). Numerical results are displayed in Table 5 below and in Table 13 in the Appendix.

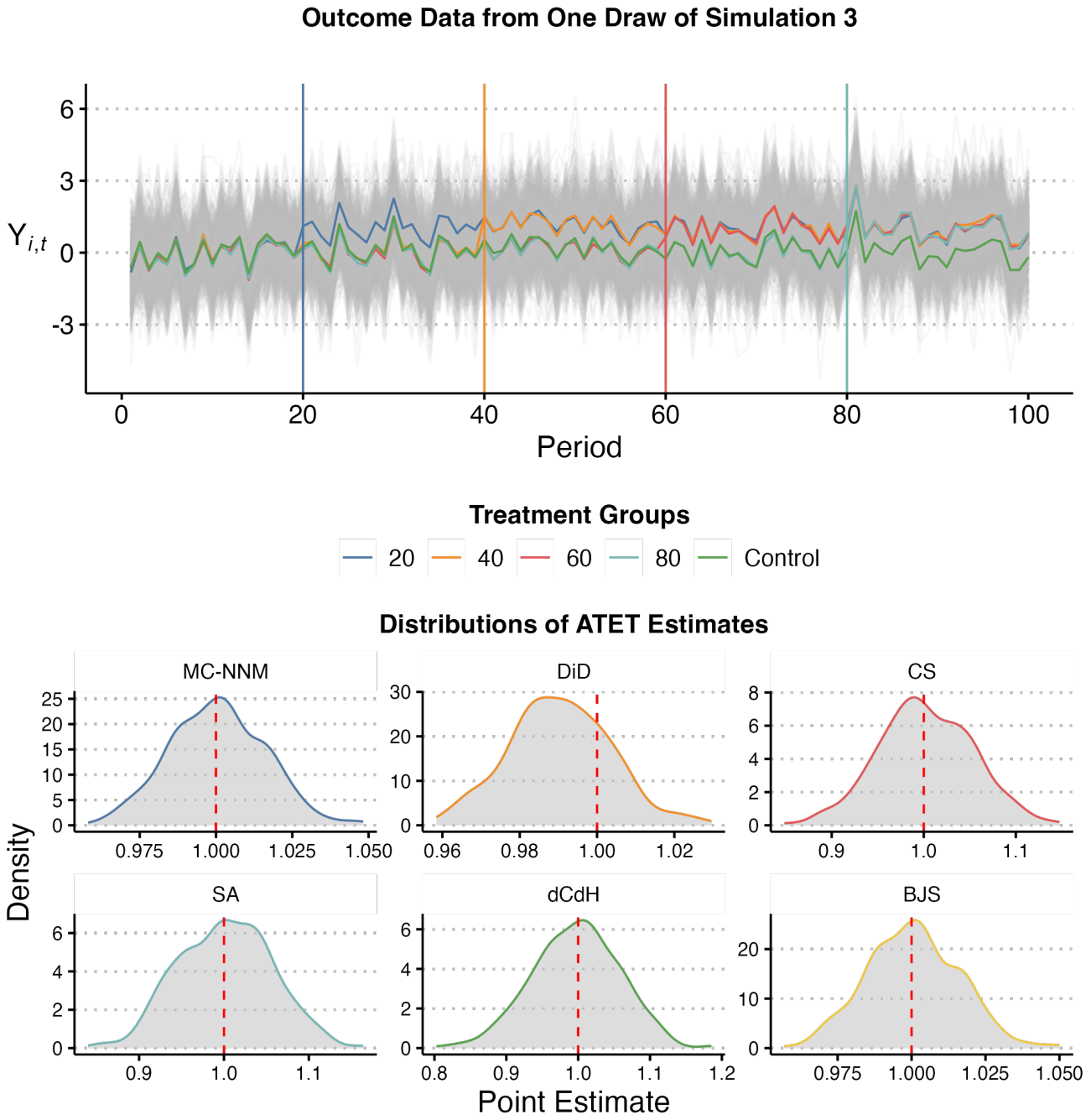
Table 5: Simulation 3, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**1.00**	**1.00**	**1.00**	**0.000**	**-**	**-**	**-**
MC-NNM	0.96	1.00	1.05	0.016	0.02	0.00	0.02
DiD	0.96	0.99	1.03	0.013	0.01	-0.01	0.02
CS	0.85	1.00	1.15	0.051	0.05	0.00	0.05
SA	0.84	1.00	1.16	0.055	0.06	0.00	0.05
dCdH	0.80	1.00	1.19	0.061	-	0.00	0.06
BJS	0.96	1.00	1.05	0.016	0.02	0.00	0.02

Results obtained from 500 iterations

In this differential timing setting with a unit-homogeneous and time-invariant treatment effect, all estimators identify τ well, although some estimators, such as [dCdH] and [SA] show a relatively wide standard deviation of their point estimates. [MC-NNM] and [BJS (event-study)] perform best.

Figure 10: Results from Simulation 3



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the ATET as defined on p. 12. Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.4 Simulation 4

Figure 11 below shows the raw data from one draw and distribution of point estimates from the full simulation of (40). Numerical results are displayed in Table 6 below and in Table 14 in the Appendix.

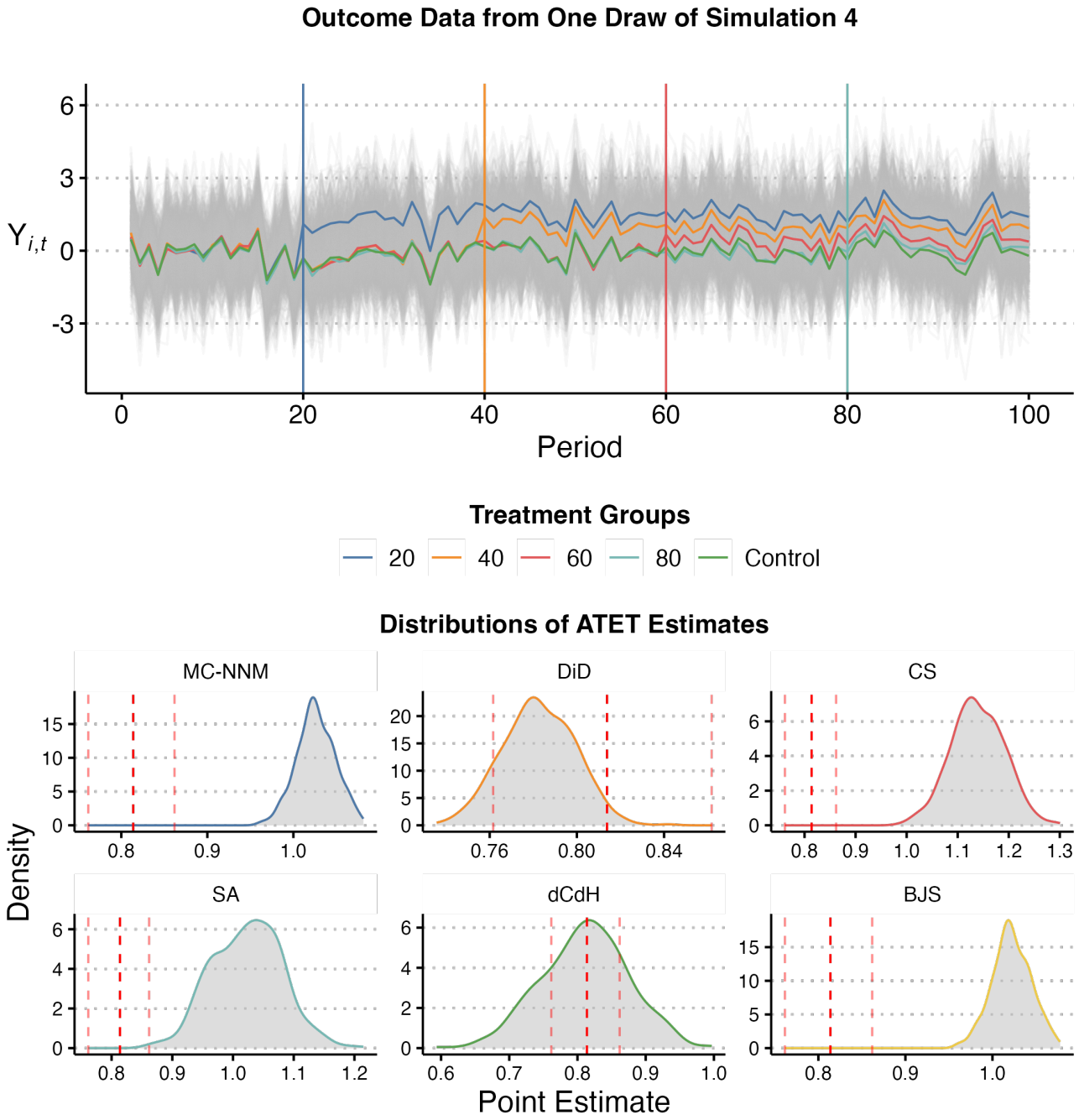
Table 6: Simulation 4, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.76**	**0.81**	**0.86**	**0.017**	**-**	**-**	**-**
MC-NNM	0.96	1.03	1.08	0.022	0.02	0.21	0.21
DiD	0.74	0.78	0.84	0.017	0.01	-0.03	0.04
CS	0.99	1.14	1.30	0.052	0.05	0.33	0.33
SA	0.86	1.02	1.21	0.058	0.06	0.21	0.22
dCdH	0.59	0.81	1.00	0.064	-	0.00	0.06
BJS	0.96	1.02	1.08	0.022	0.02	0.21	0.21

Results obtained from 500 iterations

In this differential timing setting with unit-heterogeneous but time-invariant treatment effects, [dCdH] identifies τ the best. [MC-NNM] and [BJS] perform equally, outperform [CS] and [SA], but underperform [TWFE]. This is not surprising: when treatment effects vary across units but not over time, [TWFE] still identifies τ, as shown in a.o. Goodman-Bacon (2021).

Figure 11: Results from Simulation 4



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the ATET as defined on p. 12. Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.5 Simulation 5

Figure 12 below shows the raw data from one draw and distribution of point estimates from the full simulation of (41). Numerical results are displayed in Table 7 below and in Table 15 in the Appendix.

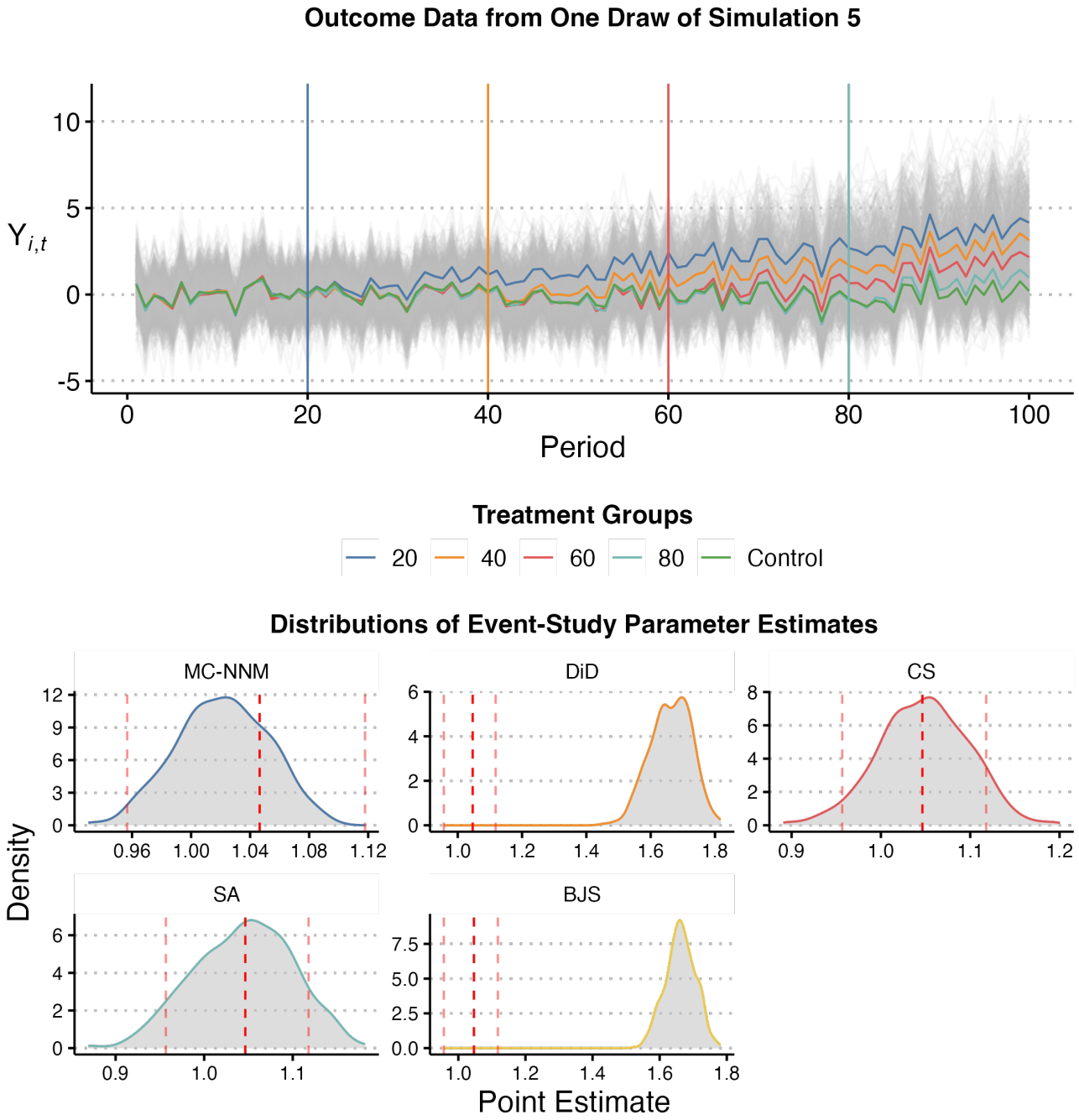
Table 7: Simulation 5, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.96**	**1.05**	**1.12**	**0.029**	**-**	**-**	**-**
MC-NNM	0.93	1.02	1.10	0.031	0.07	-0.03	0.04
DiD	1.45	1.66	1.82	0.063	0.09	0.62	0.62
CS	0.89	1.05	1.20	0.050	0.08	0.00	0.05
SA	0.87	1.04	1.18	0.055	0.11	0.00	0.06
BJS	1.53	1.66	1.78	0.044	0.07	0.62	0.62

Results obtained from 500 iterations

In this differential timing setting with unit-homogeneous but time-varying treatment effects, [MC-NNM] identifies τ^{ES} the best in terms of mean SE and RMSE. [CS (event-study)] and [SA (event-study)] outperform [MC-NNM] in terms of bias. All three outperform [BJS (event-study)], which in turn narrowly beats out [TWFE] (event-study). This is again not surprising: when treatment effects do not vary across units but do vary over time, [TWFE] identifies neither τ nor τ^{ES}, as shown in a.o. Goodman-Bacon (2021) and Roth et al. (2023).

Figure 12: Results from Simulation 5



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the event-study parameter defined in (53). Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.6 Simulation 6

Figure 13 below shows the raw data from one draw and distribution of point estimates from the full simulation of (42). Numerical results are displayed in Table 8 below and in Table 16 in the Appendix.

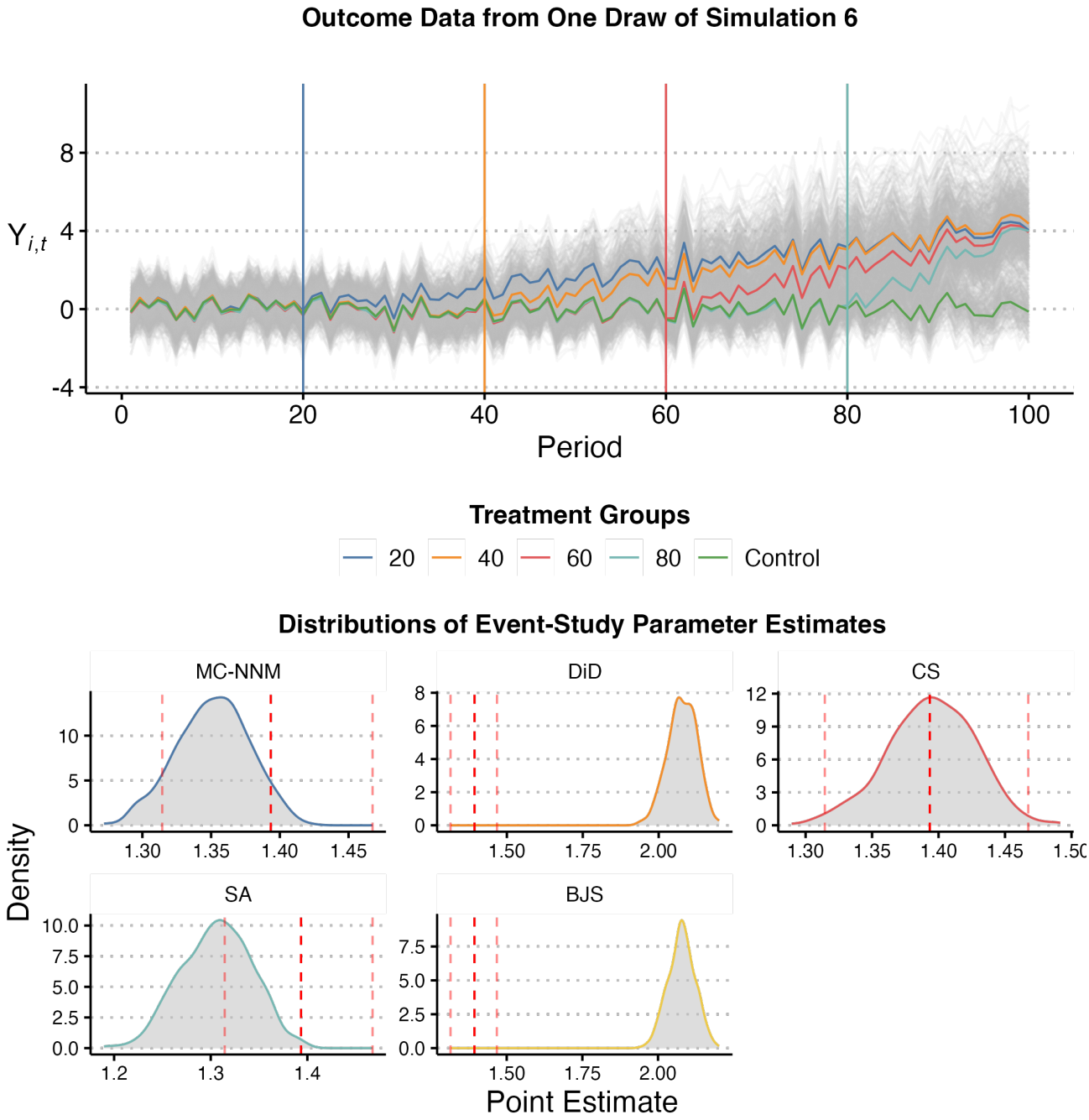
Table 8: Simulation 6, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**1.31**	**1.39**	**1.47**	**0.026**	**-**	**-**	**-**
MC-NNM	1.27	1.35	1.42	0.026	0.04	-0.04	0.05
DiD	1.93	2.08	2.20	0.048	0.06	0.68	0.69
CS	1.29	1.39	1.49	0.033	0.05	0.00	0.03
SA	1.19	1.31	1.40	0.036	0.06	-0.09	0.09
BJS	1.94	2.08	2.20	0.044	0.06	0.68	0.69

Results obtained from 500 iterations

In this differential timing setting with unit-heterogeneous *and* time-varying treatment effects, [MC-NNM] identifies τ^{ES} second-best in terms of bias and RMSE. [CS (event-study)] and [SA (event-study)] outperforms all other estimators in these two metrics. Together, [MC-NNM] and [CS (event-study)] dominate all other estimators. [TWFE] (event-study) is severely biased, which is perfectly consistent with theoretical, simulation-based, and empirical results in Callaway and Sant'Anna (2021), Sun and Abraham (2021), Chaisemartin and D'Haultfœuille (2020), and Borusyak, Jaravel, and Spiess (2023).

Figure 13: Results from Simulation 6



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the event-study parameter defined in (53). Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.7 Simulation 7

Figure 14 below shows the raw data from one draw and distribution of point estimates from the full simulation of (43). Numerical results are displayed in Table 9 below and in Table 17 in the Appendix.

Table 9: Simulation 7, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**1.26**	**1.39**	**1.53**	**0.045**	**-**	**-**	**-**
MC-NNM	1.21	1.35	1.49	0.046	0.06	-0.04	0.06
DiD	-10.05	-9.83	-9.59	0.081	0.10	-11.22	11.22
CS	-8.50	-3.36	1.74	1.875	2.68	-4.75	5.11
SA	-0.64	1.31	3.24	0.652	1.44	-0.09	0.66
BJS	1.86	2.08	2.32	0.080	0.09	0.69	0.69

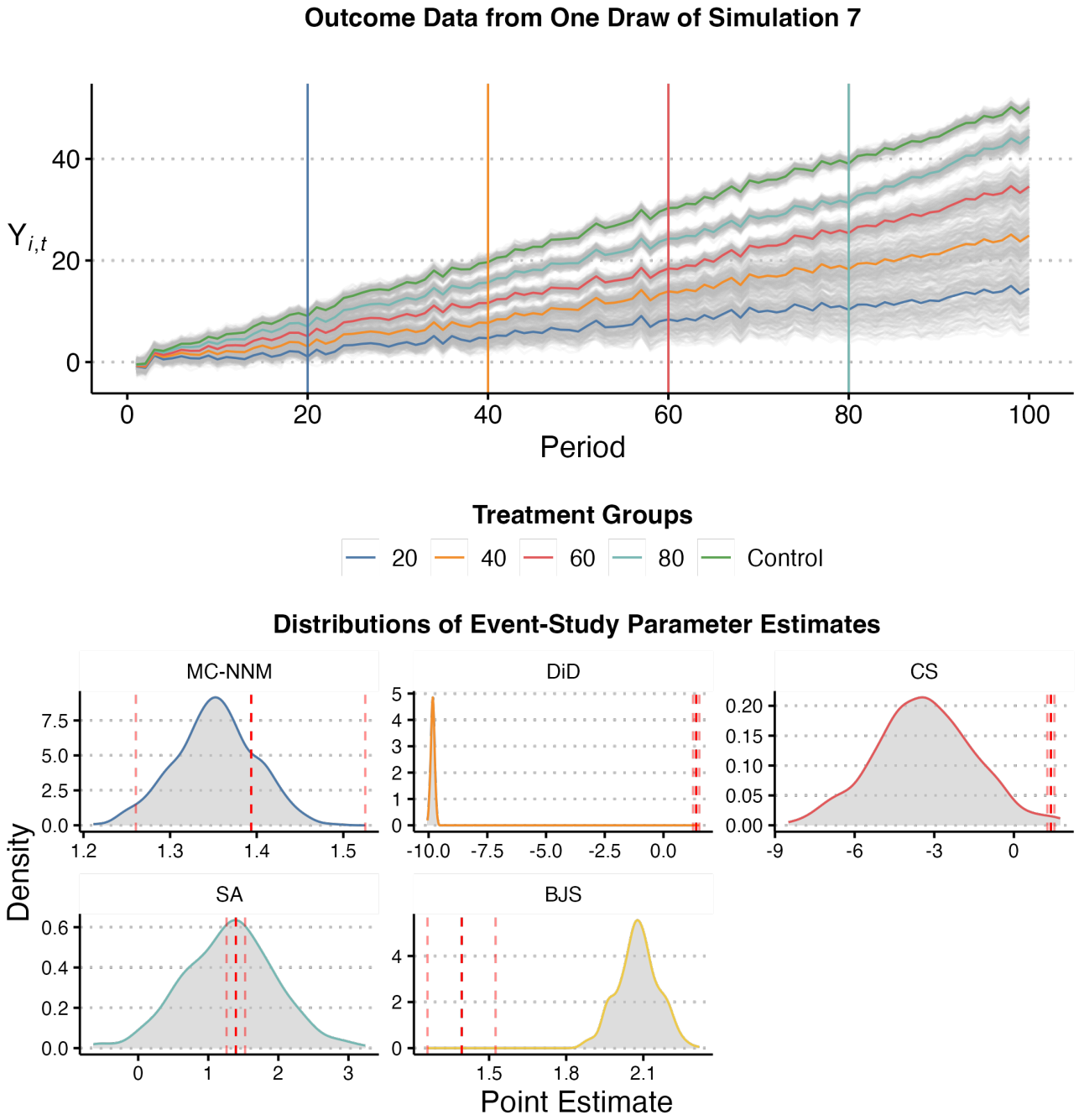
Results obtained from 500 iterations

Finally, in this setting, which is similar to simulation 6 in that it features a differential timing setting with unit-heterogeneous *and* time-varying treatment effects, but additionally includes a nuisance term described in (43). To incorporate this covariate into the estimation of [TWFE] (event-study), [CS (event-study)], [SA (event-study)], and [BJS (event-study)], I use the corresponding extensions with covariates, which are sufficiently straightforward not to merit their detailed description, and can be found in the corresponding papers.

In this complex setting, [TWFE] (event-study) should hopelessly fail. Indeed, it severely misestimates τ^{ES}. We do, however, see a drastic outperformance by [MC-NNM (Covariates)]. It estimates τ^{ES} remarkably well by every metric. The second-best-performing estimator is [SA (event-study)]. These results lend support to the central advantage Athey et al. (2021) claim for [MC-NNM]: precise treatment effect estimation in many different panel data settings, all while using fewer assumptions than any DiD-adjacent method.

Interestingly, as shown in Table 17, [dCdH] identifies the *static* τ with zero bias and negligible RMSE. This is intriguing, as Chaisemartin and D'Haultfœuille (2020) explicitly rule out time-varying treatment effects in the assumptions they use to construct [dCdH].

Figure 14: Results from Simulation 7



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the event-study parameter defined in (53). Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.4.8 Simulation 8

Figure 15 below shows the raw data from one draw and distribution of point estimates from the full simulation of (43). Numerical results are displayed in Table 10 below and in Table 18 in the Appendix.

Table 10: Simulation 8, Point Estimates of τ^{ES}

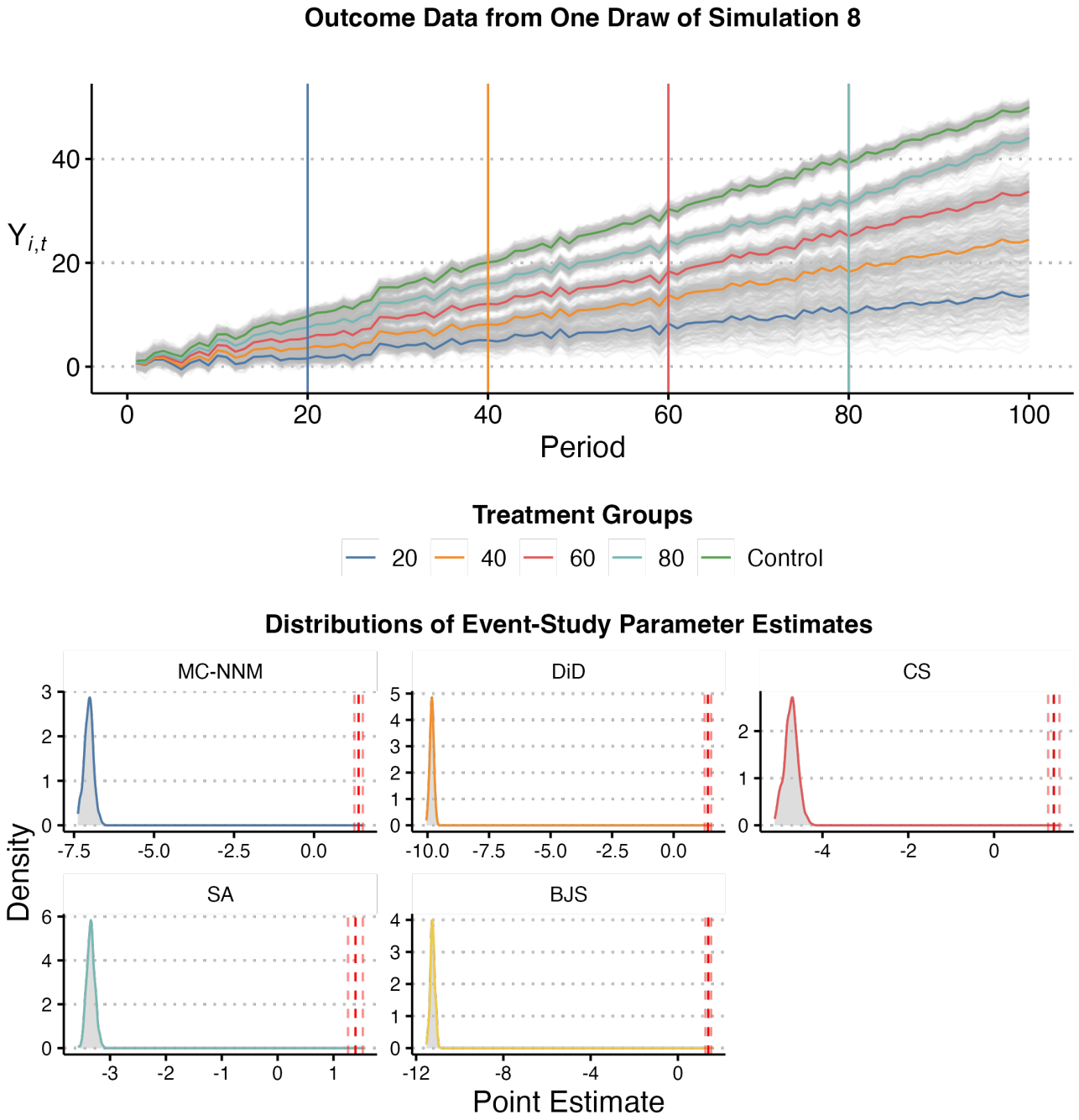
Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**1.26**	**1.39**	**1.53**	**0.045**	**-**	**-**	**-**
MC-NNM	-7.37	-7.04	-6.58	0.142	0.15	-8.43	8.43
DiD	-10.05	-9.83	-9.59	0.081	0.10	-11.22	11.22
CS	-5.11	-4.73	-4.24	0.153	0.18	-6.12	6.13
SA	-3.57	-3.34	-3.12	0.072	0.08	-4.74	4.74
BJS	-11.53	-11.24	-10.92	0.102	0.11	-12.64	12.64

Results obtained from 500 iterations

This final setting is identical to simulation 7 in that it features a differential timing setting with unit-heterogeneous *and* time-varying treatment effects, but additionally includes a nuisance term described in (43). The only difference to simulation 7 is that τ^{ES} is estimated without using covariates, which means that the key identifying assumption of all estimators except for [MC-NNM] is violated.

In this complex setting, [TWFE] (event-study) should perform even worse than in simulation 7. Indeed, it severely misestimates τ^{ES}. So do, however, all other estimators. [MC-NNM] does substantially better than [TWFE] (event-study) and [BJS (event-study)], but is severeley outperformed by [CS (event-study)] and [SA (event-study)] by all metrics. This is somewhat counterintuitive and deals a heavy blow to the speculation that [MC-NNM] can replace DiD-based estimators when common trends are not plausible. I cannot explain why [CS (event-study)] and [SA (event-study)] estimate τ^{ES} comparably well even though their central identifying assumption is clearly violated.

Figure 15: Results from Simulation 8



Upper Panel shows all observatins from one draw of the simulation with group means. Lower panel shows for each estimator described in sections 4.3 and 5.5/6 densities of point estimates of the event-study parameter defined in (53). Vertical Red Lines indicate minimum, mean, and maximum of true parameter value.

6.5 Limitations and Extensions

The results presented in subsection 6.4 provide a first indication that [MC-NNM] is a plausible alternative to DiD methods in all eight simulation settings. They are, however, by no means definitive or exhaustive. There are several reasons for this, which also serve as starting points for interesting extensions of this thesis.

First, it is plausible, though by no means necessary, that estimation precision of [MC-NNM] in all simulations would have benefited from using more bootstrap iterations to compute standard errors, more folds and more candidate λs to perform cross-validation of the regularization parameter(s). Unfortunately, given the already high computational complexity of the simulations, doing would very likely have resulted in an execution time for all simulations in stark excess of the 2-3 days per run on the high-end consumer CPU I used. Researchers with access to the type of high-performance computing (HPC) clusters used to obtain the results in Athey et al. (a.o. 2021), Arkhangelsky et al. (2021), and Liu, Wang, and Xu (2022) could relatively easily adapt my code to implement such an extension.

Second, the leading current implementation of [MC-NNM] provided by `fect()` only incorporates one of the four practical extension to [MC-NNM] proposed by Athey et al. (2021, section 8), the inclusion of covariates. The other three extensions would allow for the inclusion of treated outcome data in the estimation (under the assumption of a time-invariant or low-rank-pattern treatment effect), the incorporation of time-series dependencies in the error term to model autocorrelated panel data, and the use of a weighted loss function similar to propensity scoring. It is entirely unclear whether any or all of these extensions would improve the performance of [MC-NNM] on both simulated and empirical data, and until `fect`'s maintainers or others show practitioners the mercy of implementing these extension in the `C++` code carrying out the estimation, we will not find out.

Third, it would be interesting to carry out further simulations similar to simulation 8, with DGPs that are not linear in parameters. Doing so would provide additional evidence on the relative performance of [MC-NNM] in settings where all other estimators are not applicable from

both an identification and a design-based perspective (Athey and Imbens 2022). In a way, my choice of additive, linear DGPs in this thesis might have been *biased* in favor of DiD-adjacent methods, and a performance comparison based on a *factor* DGP such as the one described in (20) might have yielded results more favorable to [MC-NNM].

Fourth, a straightforward extension using only the generalized methods I have already programmed would be to repeat some or all of the simulations while systematically varying the shape of the panel (that is, comparing the performance for varying N and T, or N/T ratios similar to the comparisons between [MC-NNM] and elastic-net regularized regressions drawn in Athey et al. (2021)), the number of treatment groups, and the fraction of never-treated control groups in the panel. The computational time needed to do this in combination with the deadline for this thesis approaching prevented me from implementing these comparisons.

Lastly, although it might be somewhat *cliché* to say, we simply need more research. This thesis is, to the best of my knowledge, as of today the only work comparing [MC-NNM] to methods from the recent DiD literature (although Liu, Wang, and Xu (2022, 2) do compare it to the *fixed effects counterfactual estimator, of which* [canonical] *DiD is a special case* in staggered settings). Additional theoretical work on *which exact estimand* [MC-NNM] identifies in settings with time-varying treatment effects would also be helpful. The publication of a `STATA` version of `fect` roughly two months ago satisfies the necessary (and possibly sufficient) condition for applied economists to begin experimenting with [MC-NNM].

7 Conclusion

Three months ago, I proposed the topic for this thesis with three concrete goals: to "*review and synthesize the contributions made by* Athey et al. (2021)", to "*examine whether recent advances in DID methods can be incorporated into the MCE framework*", and to "*compare the accuracy of the new MC estimator with that of recent DiD methods in data regimes in which canonical DiD assumptions are not met*" (cf. Appendix A.4).

As Athey et al. (2021) discuss [TWFE] only as an afterthought, there is a gap in the literature in understanding how, and how well, [MC-NNM] works compared to DiD methods in differential timing settings with unit- and/or time-heterogeneous treatment effects. This thesis provides a framework to think about the similarities and differences between these two fundamentally different methods using compatible notation, unified motivation, and identical ultimate estimands. Within this framework, I have provided brief summaries of the most relevant DiD-adjacent estimators introduced in recent years, and analyzed the relative performance of each of these methods in a variety of different panel data settings with increasing complexity and degree of real-world applicability.

While my simulations cover a variety of different settings, they are not exhaustive. Equipped with sufficient computational resources and time, each estimator could have been better fine-tuned towards each Data-Generating Process. [SA]'s performance, for instance, can depend on whether and how one decides to bin pre-and post-treatment observations. The computation of standard errors for [CS] could be switched to using the bootstrap rather than analytical. The bug-laden implementation of `didmultiplegt()` could be fixed to provide both uncertainty- and event-study-type point estimates.

In my opinion, my results show encouraging signs of [MC-NNM]'s performance in a variety of data configurations, exactly as claimed by Athey et al. (2021). As my results show, [MC-NNM] performs *well* (or *decently*, depending on one's standards) in all simulations, especially compared to [TWFE], but it is outperformed by application-specific estimators such as the one proposed by Callaway and Sant'Anna (2021) in simulation 6.

I think a sensible conclusion from these results is that [MC-NNM] is preferable to [TWFE] as a baseline estimator, full stop. The former relies on two identifying assumptions, the latter on at least six. [MC-NNM] performs equally well as [TWFE] in the simulation most favorable to [TWFE], slightly underperforms it in simulation 4, where [TWFE] identifies τ by construction, and outperforms it on every metric in all others. [MC-NNM] requires fewer implicit (for example linearity) and explicit (such as time-invariant treatment effects) modelling choices and produces sensible estimates with arbitrary heterogeneity across time and units, which [TWFE], as a deluge

of work over the recent years has shown, does not. Even if one is willing to impose a true model that is linear in parameters, [MC-NNM] is preferable. The only instance I can think of in which this does not necessarily hold is if one has reason to believe that the σ-sub-Gaussianity portion of Assumption 14 to be violated, in which case the error term has very high kurtosis, which would also be problematic in any linear model, such as the one imposed by [TWFE]. In short, I believe that [MC-NNM] should be included in addition to [TWFE] as a benchmark estimator in differential timing settings. If the estimates of [MC-NNM] and [TWFE] or [TWFE] (event-study) differ materially, as my results show, the canonical estimator is likely to be inadequate. [MC-NNM] is a very new method, so it will not immediately replace DiD, nor should it, as simple linear models are easy to use and interpret. It does, however, serve as an indicator that treatment effects might be heterogeneous and/or time-varying, and that therefore [TWFE] should not be used. Such a diagnostic function improves upon the practice of simply assuming that treatment effects are homogeneous and/or time-invariant.

I do not, however, want to tout Matrix Completion as a universal silver bullet to solve all the problems that come with Difference-in-Differences based methods. Neither do I want to convey that Athey et al. (2021) claim their method to be superior to new DiD methods in differential timing settings, they do not. Their claim of matrix completion's universal applicability is narrower than that, and mainly pertains to the fact that [MC-NNM] can handle nearly any shape of data matrix and treatment timing configuration, but it can easily lead one to speculate, as I did when I first read their paper, that [MC-NNM] might be the solution to the many problems practitioners face when using DiD detailed in Section 4.

Such a far-reaching interpretation of their arguments, as my results show, is wishful thinking reminiscent of the saying "*if something seems too good to be true, it probably is*". Clean identification of causal treatment effects in messy data is hard. I hope to have shown that while [MC-NNM] shows promising signs of doing this task well in many cases, it cannot single-handedly supplant existing methods just yet.

8 References

American Economic Association. 2023. "American Economic Journal: Economic Policy." *American Economic Journal: Economic Policy* 15 (1). https://www.aeaweb.org/issues/710.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *The Journal of Economic Perspectives* 24 (2): 3–30. ISSN: 0895-3309, accessed June 24, 2023. https://www.jstor.org/stable/25703496.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences" [in en]. *American Economic Review* 111, no. 12 (December): 4088–4118. ISSN: 0002-8282, accessed May 1, 2023. https://doi.org/10.1257/aer.20190159. https://pubs.aeaweb.org/doi/10.1257/aer.20190159.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." Appendix at https://arxiv.org/pdf/1710.10251.pdf, *Journal of the American Statistical Association* 116, no. 536 (October): 1716–1730. ISSN: 0162-1459, accessed May 1, 2023. https://doi.org/10.1080/01621459.2021.1891924. https://doi.org/10.1080/01621459.2021.1891924.

Athey, Susan, and Guido W. Imbens. 2022. "Design-based analysis in Difference-In-Differences settings with staggered adoption" [in en]. *Journal of Econometrics,* Annals Issue in Honor of Gary Chamberlain, 226, no. 1 (January): 62–79. ISSN: 0304-4076, accessed June 5, 2023. https://doi.org/10.1016/j.jeconom.2020.10.012. https://www.sciencedirect.com/science/article/pii/S0304407621000488.

Bai, Jushan. 2009. "Panel Data Models With Interactive Fixed Effects." *Econometrica* 77 (4): 1229–1279. ISSN: 0012-9682. https://doi.org/10.3982/ECTA6135.

Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" [In en]. *Journal of Financial Economics* 144, no. 2 (May): 370–395. ISSN: 0304-405X, accessed June 6, 2023. https://doi.org/10.1016/j.jfineco.2022.01.004. https://www.sciencedirect.com/science/article/pii/S0304405X22000204.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" Publisher: Oxford University Press / USA, *Quarterly Journal of Economics* 119, no. 1 (February): 249–275. ISSN: 00335533, accessed June 10, 2023. https://doi.org/10.1162/003355304772839588. https://mu.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=12336618&site=ehost-live&scope=site.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2023. *Revisiting Event Study Designs: Robust and Efficient Estimation* [in en]. ArXiv:2108.12419 [econ], April. Accessed June 23, 2023. http://arxiv.org/abs/2108.12419.

Cai, Jian-Feng, Emmanuel J. Candès, and Zuowei Shen. 2010. "A Singular Value Thresholding Algorithm for Matrix Completion." Publisher: Society for Industrial and Applied Mathematics, *SIAM Journal on Optimization* 20, no. 4 (January): 1956–1982. ISSN: 1052-6234, accessed June 2, 2023. https://doi.org/10.1137/080738970. https://epubs.siam.org/doi/10.1137/080738970.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230. ISSN: 0304-4076. https://doi.org/10.1016/j.jeconom.2020.12.001.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications* [in en]. Google-Books-ID: TdlKAgAAQBAJ. Cambridge University Press, May. ISBN: 978-1-139-44486-6.

Candes, Emmanuel J., and Terence Tao. 2010. "The Power of Convex Relaxation: Near-Optimal Matrix Completion." Conference Name: IEEE Transactions on Information Theory, *IEEE Transactions on Information Theory* 56, no. 5 (May): 2053–2080. ISSN: 1557-9654. https://doi.org/10.1109/TIT.2010.2044061.

Candès, Emmanuel J., and Benjamin Recht. 2009. "Exact Matrix Completion via Convex Optimization" [in en]. *Foundations of Computational Mathematics* 9, no. 6 (December): 717–772. ISSN: 1615-3383, accessed April 24, 2023. https://doi.org/10.1007/s10208-009-9045-5. https://doi.org/10.1007/s10208-009-9045-5.

Casella, George, and Roger L. Berger. 2002. *Statistical inference* [in eng]. 2nd ed. Duxbury advanced series in statistics and decision sciences. OCLC: 46538638. Australia: Thomson Learning. ISBN: 978-0-534-24312-8, accessed June 2, 2023. http://catdir.loc.gov/catdir/enhancements/fy1302/2001025794-t.html.

Chaisemartin, Clément de, and Xavier D'Haultfœuille. 2021. *Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey* [in en]. SSRN Scholarly Paper. Rochester, NY, December. Accessed May 29, 2023. https://doi.org/10.2139/ssrn.3980758. https://papers.ssrn.com/abstract=3980758.

———. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects" [in en]. *American Economic Review* 110, no. 9 (September): 2964–2996. ISSN: 0002-8282, accessed March 12, 2023. https://doi.org/10.1257/aer.20181169. https://www.aeaweb.org/articles?id=10.1257/aer.20181169.

Chaisemartin, Clément de, and Xavier D'Haultfoeuille. 2022. *Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey.* Working Paper, January. Accessed April 24, 2023. https://doi.org/10.3386/w29691. https://www.nber.org/papers/w29691.

Dette, Holger, and Martin Schumann. 2020. "Difference-in-Differences Estimation under Non-Parallel Trends" [in en]. *Working Paper,* https://www.ruhr-uni-bochum.de/imperia/md/content/mathematik3/publications/dette_schumann2020.pdf.

Gelman, Andrew. 2005. "Analysis of Variance: Why It Is More Important than Ever." Publisher: Institute of Mathematical Statistics, *The Annals of Statistics* 33 (1): 1–31. ISSN: 0090-5364, accessed June 4, 2023. https://www.jstor.org/stable/3448650.

Golub, Gene H., and Charles F. Van Loan. 2013. *Matrix computations* [in eng]. 4. ed. Johns Hopkins studies in the mathematical sciences. OCLC: 835290390. Baltimore: Johns Hopkins Univ Press. ISBN: 978-1-4214-0794-4, accessed June 2, 2023. http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=025701078&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing" [in en]. *Journal of Econometrics,* Themed Issue: Treatment Effect 1, 225, no. 2 (December): 254–277. ISSN: 0304-4076, accessed March 12, 2023. https://doi.org/10.1016/j.jeconom.2021.03.014. https://www.sciencedirect.com/science/article/pii/S0304407621001445.

Hansen, Bruce. 2022. *Econometrics* [in en]. Google-Books-ID: UipXEAAAQBAJ. Princeton University Press, June. ISBN: 978-0-691-23615-5.

Hastie, Trevor, Rahul Mazumder, Jason Lee, and Reza Zadeh. 2014. *Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares.* ArXiv:1410.2596 [stat], October. Accessed April 24, 2023. https://doi.org/10.48550/arXiv.1410.2596. http://arxiv.org/abs/1410.2596.

Hernán, Miguel, and James M. Robins. 2021. *Causal inference* [in eng]. Chapman & Hall/CRC monographs on statistics & applied probability. OCLC: 1009571686. Boca Raton: Chapman & Hall/CRC. ISBN: 978-1-4200-7616-5.

Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data" [in en]. Publisher: Cambridge University Press, *Political Analysis* 29, no. 3 (July): 405–415. ISSN: 1047-1987, 1476-4989, accessed April 24, 2023. https://doi.org/10.1017/pan.2020.33. https://www.cambridge.org/core/journals/political-analysis/article/abs/on-the-use-of-twoway-fixed-effects-regression-models-for-causal-inference-with-panel-data/F10006D0210407C5F9C7CAC1EEE3EF0D.

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics,* https://scholar.harvard.edu/imbens/publications/nonparametric-estimation-average-treatment-effects-under-exogeneity-review.

Imbens, Guido, and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: an introduction* [in Englisch]. New York: Cambridge University Press. ISBN: 978-0-521-88588-1, accessed April 25, 2023. http://catdir.loc.gov/catdir/enhancements/fy1513/2014020988-t.html.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An introduction to statistical learning: with applications in R* [in Englisch]. Second edition. Springer texts in statistics. New York: Springer. ISBN: 978-1-07-161418-1 978-1-07-161417-4, accessed January 27, 2023. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2985424.

Keshavan, Raghunandan H., Andrea Montanari, and Sewoong Oh. 2010. "Matrix completion from a few entries." *IRE Professional Group on Information Theory* 56, no. 6 (June): 2980–2998. ISSN: 0018-9448, accessed April 24, 2023. https://doi.org/10.1109/TIT.2010.2046205. http://www.scopus.com/inward/record.url?scp=77956897560&partnerID=8YFLogxK.

Koltchinskii, Vladimir, Karim Lounici, and Alexandre B. Tsybakov. 2011. "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion." Publisher: Institute of Mathematical Statistics, *The Annals of Statistics* 39 (5): 2302–2329. ISSN: 0090-5364, accessed April 24, 2023. https://www.jstor.org/stable/41713579.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2022. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data" [in en]. *American Journal of Political Science* n/a, no. n/a (August). ISSN: 1540-5907, accessed June 6, 2023. https://doi.org/10.1111/ajps.12723. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12723.

Marcus, Michelle, and Pedro H. C. Sant'Anna. 2021. "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics." Publisher: The University of Chicago Press, *Journal of the Association of Environmental and Resource Economists* 8, no. 2 (March): 235–275. ISSN: 2333-5955, accessed May 30, 2023. https://doi.org/10.1086/711509. https://www.journals.uchicago.edu/doi/full/10.1086/711509.

Mazumder, Rahul, Jerome H. Friedman, and Trevor Hastie. 2011. "SparseNet: Coordinate Descent With Nonconvex Penalties." *Journal of the American Statistical Association* 106, no. 495 (September): 1125–1138. ISSN: 0162-1459, accessed June 3, 2023. https://doi.org/10.1198/jasa.2011.tm09738. https://doi.org/10.1198/jasa.2011.tm09738.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. 2010. "Spectral Regularization Algorithms for Learning Large Incomplete Matrices." *Journal of Machine Learning Research* 11 (80): 2287–2322. ISSN: 1533-7928, accessed April 24, 2023. http://jmlr.org/papers/v11/mazumder10a.html.

Negahban, Sahand, and Martin J. Wainwright. 2012. "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise." *Journal of Machine Learning Research* 13 (53): 1665–1697. ISSN: 1533-7928, accessed April 24, 2023. http://jmlr.org/papers/v13/negahban12a.html.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference* [in Englisch]. Cambridge, U.K. ; Cambridge University Press. ISBN: 978-1-139-64936-0 978-0-511-80316-1, accessed April 25, 2023. https://doi.org/10.1017/CBO9780511803161.

Recht, Benjamin, Maryam Fazel, and Pablo A. Parrilo. 2010. "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization." ArXiv:0706.4138 [math, stat], *SIAM Review* 52, no. 3 (January): 471–501. ISSN: 0036-1445, 1095-7200, accessed June 2, 2023. https://doi.org/10.1137/070697835. http://arxiv.org/abs/0706.4138.

Rohde, Angelika, and Alexandre B. Tsybakov. 2011. "Estimation of high-dimensional low-rank matrices." ArXiv:0912.5338 [math, stat], *The Annals of Statistics* 39, no. 2 (April). ISSN: 0090-5364, accessed April 24, 2023. https://doi.org/10.1214/10-AOS860. http://arxiv.org/abs/0912.5338.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends" [in en]. *American Economic Review: Insights* 4, no. 3 (September): 305–322. Accessed May 28, 2023. https://doi.org/10.1257/aeri.20210236. https://www.aeaweb.org/articles?id=10.1257/aeri.20210236.

Roth, Jonathan, and Pedro H. C. Sant'Anna. 2023. *Efficient Estimation for Staggered Rollout Designs.* ArXiv:2102.01291 [econ, math, stat], February. Accessed April 24, 2023. https://doi.org/10.48550/arXiv.2102.01291. http://arxiv.org/abs/2102.01291.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. *What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature.* ArXiv:2201.01194 [econ, stat], January. Accessed February 24, 2023. https://doi.org/10.48550/arXiv.2201.01194. http://arxiv.org/abs/2201.01194.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5): 688–701. ISSN: 0022-0663. https://doi.org/10.1037/h0037350.

Schmidheiny, Kurt, and Sebastian Siegloch. 2023. "On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization" [in en]. *Journal*

of Applied Econometrics n/a, no. n/a (February). ISSN: 1099-1255, accessed May 30, 2023. https://doi.org/10.1002/jae.2971. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2971.

Strang, Gilbert. 2006. *Linear Algebra and Its Applications, 4th Edition* [in Englisch]. 4th ed. Belmont, CA: Cengage Learning, January. ISBN: 978-0-03-010567-8.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects" [in en]. *Journal of Econometrics,* Themed Issue: Treatment Effect 1, 225, no. 2 (December): 175–199. ISSN: 0304-4076, accessed April 24, 2023. https://doi.org/10.1016/j.jeconom.2020.09.006. https://www.sciencedirect.com/science/article/pii/S030440762030378X.

Wainwright, Martin J. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. ISBN: 978-1-108-49802-9, accessed June 2, 2023. https://doi.org/10.1017/9781108627771. https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E.

9 Main Software Packages Used

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2017. *MCPanel: Matrix Completion algorithms for Causal Panel Data Models.* R package version 0.0.

Butts, Kyle. 2021a. *did2s: Two-Stage Difference-in-Differences Following Gardner (2021).* https://github.com/kylebutts/did2s/.

———. 2021b. *didimputation: Difference-in-Differences estimator from Borusyak, Jaravel, and Spiess (2021).* https://github.com/kylebutts/didimputation.

Liu, Licheng, Ziyi Liu, Ye Wang, and Yiqing Xu. 2022. *fect: Fixed Effects Counterfactuals.* R package version 1.0.0. https://CRAN.R-project.org/package=fect.

Posit team. 2023. *RStudio: Integrated Development Environment for R.* Boston, MA: Posit Software, PBC. http://www.posit.co/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sant'Anna, Pedro H. C., and Michelle Marcus. 2020. *MarcusSantAnna2020: Unconditional Event-study analysis with variation in Treatment Timing.* R package version 0.1.1.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. *Welcome to the tidyverse.* 43. https://doi.org/10.21105/joss.01686.

Xu, Yiqing, and Licheng Liu. 2021. *gsynth: Generalized Synthetic Control Method.* R package version 1.2.1. https://CRAN.R-project.org/package=gsynth.

Zhang, Shuo, and Clément de Chaisemartin. 2020. *DIDmultiplegt: Estimation in DID with Multiple Groups and Periods.* R package version 0.1.0. https://CRAN.R-project.org/package=DIDmultiplegt.

A Appendix

A.1 Computation of True Parameters

True values for the ATET and the event-study parameter in each iteration are obtained as follows: in each iteration, I call function `est_true()` with the data draw of that iteration and the corresponding iteration number. This method itself invokes `get_true_values()` (line 13) on the drawn data, which extracts several attributes of the data: it checks a logical flag that all DGP simulation functions set to see whether the data was generated with a time-invariant or time-varying treatment effect. Function `dgp_7_sim()` corresponding to (43) on line 59 below is the most complex simulation function. As shown on line 76, the average treatment effect for each group differs (lines 77-80 correspond to the μ values in (43)). For DGPs 2, 5, 6, and 7, `y` is constructed as shown in line 101, that is, using the cumulative treatment effect stored in the `cum.t.eff` variable (DGP 7 is the only DGP that includes the nuisance term). `cum.t.eff` itself is constructed as the cumulative sum of each treatment effect (`t.eff`, 0 for control units). For these DGP, the flag `use_cum_te` is set to `TRUE` (line 102). For DGPs 1, 3, and 4, the simulation function uses a different line 101 to construct y: `y = unit_fe + period_fe + t.eff + error`. In those DGPs, the flag `use_cum_te` is set to `FALSE`.

The function `get_true_values()` regroups the data by period, filters out control groups, selects for each period the treated units, computes for each period the average of the (cumulative) treatment effect, sums those values up over all periods and divides by the number of periods in the panel (lines 32-42). This provides for the time-invariant and time-varying treatment effects DGPs the true value of the event-study parameter. Line 52 computes the true ATET as the mean of the group-specific μ values set on lines 76-82. These two true parameters are returned and stored.

```
1 ### Function to extract true ATET and cumulative ATET from data in each iteration
2 est_true <- function(data, iteration = 0) {
3   vals = get_true_values(data)
4   weighted_est <- sum(vals$est * vals$share) / sum(vals$share)
5   ce = mean(vals$cum_est)
```

```
6
7     out = list(est = weighted_est, se = 0, cum_est = ce,
8                cum_se = 0, iter = iteration, estimator = "TRUE")
9
10    return(out)
11  }
12  ### Function to extract true ATET and cumulative ATET from simulated data
13  get_true_values <- function(data){
14    units = get_num_units(data)
15    periods = get_num_periods(data)
16    treat_times <- sort(get_treat_times(data))
17    tgroups <- treat_times[-length(treat_times)]
18
19    # check whether DGP uses t.eff or cum.t.eff
20    use_cum_effect = as.logical(data$use_cum_te[1])
21
22    # get number of treated units
23    numtreated = data %>% filter(group %in% tgroups) %>%  get_num_units()
24
25    # get share of each treatment group
26    group_share <- data %>%
27      filter(group %in% tgroups) %>%
28      group_by(group) %>%
29      summarise(share = n() / numtreated)
30
31    # calculate true dynamic treatment effect
32    if (use_cum_effect) {
33      ce = data %>%  group_by(period) %>%
34        filter(group %in% tgroups) %>%
35        filter(treat == 1) %>%
36        summarise(mce = mean(cum.t.eff)) %>%
37        summarise(cum_est = sum(mce) / periods)
38    } else {
39      ce = data %>%  group_by(period) %>%
40        filter(group %in% tgroups) %>%
```

```
41        filter(treat == 1) %>%
42        summarise(mce = mean(t.eff)) %>%
43        summarise(cum_est = sum(mce) / periods)
44    }
45
46    # calculate true static treatment effect
47    group_data <- data %>%
48      filter(group %in% tgroups) %>%
49      left_join(group_share, by = "group") %>%
50      # left_join(ce, by = "period") %>%
51      group_by(group) %>%
52      summarise(est = mean(avg.te),
53                cum_est = mean(as.numeric(ce)), share = mean(share))
54
55    return(group_data)
56  }
57  ## DGP 7  Multiple Treatment Groups, Time-Varying Heterogeneous TE,
58  # CONDITIONAL PARALLEL TRENDS
59  dgp_7_sim <- function(nobs = 1000,
60                        nperiods = 100,
61                        nobsgroups = 50,
62                        treatgroups = c(nperiods/5, 2*(nperiods/5), 3*(nperiods/5),
                          ↪  4*(nperiods/5))) {
63
64    # Unit Fixed Effects
65    unit <- tibble(
66      unit = 1:nobs,
67      obsgroup = sample(1:nobsgroups, nobs, replace = T),
68      unit_fe = rnorm(nobs, 0, 0.5),
69      group = case_when(
70        obsgroup %in% 1:(nobsgroups%/%5) ~ treatgroups[1],
71        obsgroup %in% ((nobsgroups%/%5) + 1):(2*nobsgroups%/%5) ~ treatgroups[2],
72        obsgroup %in% ((2*nobsgroups%/%5) + 1):(3*nobsgroups%/%5) ~ treatgroups[3],
73        obsgroup %in% ((3*nobsgroups%/%5) + 1):(4*nobsgroups%/%5) ~ treatgroups[4],
74        obsgroup %in% ((4*nobsgroups%/%5) + 1):nobsgroups ~ nperiods
```

```
75      ),
76      avg.te = case_when(
77        group == treatgroups[1] ~ .05,
78        group == treatgroups[2] ~ .075,
79        group == treatgroups[3] ~ .1,
80        group == treatgroups[4] ~ .2,
81        TRUE ~ 0
82      )) %>%
83      rowwise() %>%
84      mutate(te = rnorm(1, avg.te, .045)) %>%
85      ungroup()
86
87    # generate Time FE
88    period <- tibble(
89      period = 1:nperiods,
90      period_fe = rnorm(nperiods, 0, 0.5)
91    )
92
93    # interact unit, period, and nuisance parameter that breaks common trends
94    crossing(unit, period) %>%
95      mutate(
96        nuisance = 0.005 * period * group + rnorm(n(), 0, 0.02),
97        error = rnorm(n(), 0, 0.5),
98        treat = ifelse(period >= group, 1, 0),
99        t.eff = ifelse(treat == 1, te, 0),
100       cum.t.eff = ave(t.eff, unit, FUN = cumsum),
101       y = unit_fe + period_fe + cum.t.eff + error + nuisance,
102       use_cum_te = T,
103       use_cov = T) %>%
104     # change column order
105     select(unit, period, obsgroup, te, group, treat, cum.t.eff, nuisance, everything())
106
107 }
```

A.2 Supplementary Results

A.2.1 Simulation 1

Table 11: Simulation 1, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.49**	**0.50**	**0.51**	**0.004**	**-**	**-**	**-**
MC-NNM	0.48	0.49	0.51	0.005	0.02	-0.01	0.01
DiD	0.42	0.51	0.59	0.032	0.05	0.01	0.03
CS	0.45	0.50	0.54	0.017	0.02	0.00	0.02
SA	0.42	0.51	0.59	0.032	0.05	0.01	0.03
BJS	0.48	0.50	0.52	0.005	0.02	0.00	0.01

Results obtained from 500 iterations

[MC-NNM] identifies τ^{ES} more precisely than [TWFE] (event-study). Both perform marginally worse than [BJS (event-study)]. Back to Table 3

A.2.2 Simulation 2

Table 12: Simulation 2, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.05**	**0.05**	**0.05**	**0.000**	**-**	**-**	**-**
MC-NNM	1.20	1.27	1.35	0.029	0.03	1.22	1.22
DiD	1.20	1.27	1.35	0.029	0.01	1.22	1.22
CS	1.13	1.25	1.35	0.039	0.04	1.20	1.20
SA	1.13	1.25	1.35	0.039	0.04	1.20	1.20
dCdH	-0.10	0.05	0.19	0.042	-	0.00	0.04
BJS	1.17	1.25	1.33	0.028	0.03	1.20	1.20

Results obtained from 500 iterations

The only estimator that identifies τ well is [dCdH]. Back to Table 4

A.2.3 Simulation 3

All estimators with the exception of [SA (event-study)] identify τ^{ES} well. [MC-NNM] performs second-best, after [BJS (event-study)]. Back to Table 5

Table 13: Simulation 3, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.79**	**0.81**	**0.83**	**0.007**	**-**	**-**	**-**
MC-NNM	0.76	0.80	0.84	0.013	0.06	-0.01	0.02
DiD	0.66	0.81	0.95	0.049	0.08	0.00	0.05
CS	0.69	0.81	0.93	0.041	0.08	0.00	0.04
SA	0.37	0.51	0.65	0.050	0.10	-0.30	0.31
BJS	0.77	0.81	0.86	0.015	0.05	0.00	0.01

Results obtained from 500 iterations

Table 14: Simulation 4, Point Estimates of τ^{ES}

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.89**	**0.92**	**0.95**	**0.010**	**-**	**-**	**-**
MC-NNM	0.87	0.91	0.95	0.015	0.06	-0.01	0.02
DiD	0.67	0.82	0.97	0.050	0.08	-0.10	0.11
CS	0.80	0.92	1.05	0.042	0.08	0.00	0.04
SA	0.44	0.58	0.72	0.050	0.10	-0.34	0.35
BJS	0.87	0.92	0.97	0.017	0.05	0.00	0.02

Results obtained from 500 iterations

A.2.4 Simulation 4

[MC-NNM] and [BJS (event-study)] identify τ^{ES} equally well, although [BJS (event-study)] is marginally more efficient. Back to Table 6

A.2.5 Simulation 5

Table 15: Simulation 5, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.05**	**0.05**	**0.05**	**0.000**	**-**	**-**	**-**
MC-NNM	1.39	1.53	1.62	0.041	0.04	1.48	1.48
DiD	0.39	0.48	0.56	0.029	0.01	0.43	0.43
CS	1.10	1.29	1.48	0.062	0.06	1.24	1.25
SA	1.37	1.55	1.73	0.066	0.06	1.50	1.50
dCdH	-0.14	0.05	0.24	0.060	-	0.00	0.06
BJS	1.41	1.55	1.65	0.041	0.04	1.50	1.50

Results obtained from 500 iterations

Here again, the only estimator that identifies τ well is [dCdH]. Back to Table 7

A.2.6 Simulation 6

Table 16: Simulation 6, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.10**	**0.11**	**0.11**	**0.002**	**-**	**-**	**-**
MC-NNM	2.02	2.12	2.22	0.035	0.03	2.01	2.01
DiD	1.10	1.18	1.28	0.031	0.01	1.08	1.08
CS	1.59	1.72	1.84	0.041	0.04	1.62	1.62
SA	2.03	2.16	2.28	0.044	0.04	2.05	2.05
dCdH	0.01	0.10	0.20	0.030	-	0.00	0.03
BJS	2.06	2.16	2.26	0.035	0.03	2.05	2.05

Results obtained from 500 iterations

Yet again, the only estimator that identifies τ well is [dCdH]. Back to Table 8

A.2.7 Simulation 7

Table 17: Simulation 7, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.10**	**0.11**	**0.11**	**0.002**	**-**	**-**	**-**
MC-NNM	1.94	2.12	2.31	0.065	0.06	2.01	2.01
DiD	0.58	0.72	0.86	0.044	0.01	0.62	0.62
CS	-10.49	-4.14	2.15	2.315	2.32	-4.25	4.84
SA	-2.17	2.16	6.42	1.434	1.45	2.05	2.50
dCdH	0.00	0.11	0.20	0.033	-	0.00	0.03
BJS	1.98	2.16	2.35	0.065	0.06	2.05	2.05

Results obtained from 500 iterations

Yet again, the only estimator that identifies τ well is [dCdH]. **Back to Table 9**

A.2.8 Simulation 8

Table 18: Simulation 8, Point Estimates of τ

Estimator	Min	Mean	Max	SD	Mean SE	Bias	RMSE
TRUE	**0.10**	**0.11**	**0.11**	**0.002**	**-**	**-**	**-**
MC-NNM	-10.68	-10.25	-9.68	0.175	0.16	-10.36	10.36
DiD	-3.85	-3.33	-2.68	0.205	0.04	-3.44	3.44
CS	-6.31	-5.84	-5.24	0.189	0.19	-5.95	5.95
SA	-8.63	-8.10	-7.46	0.209	0.07	-8.20	8.20
dCdH	-0.17	-0.07	0.02	0.033	-	-0.17	0.18
BJS	-10.78	-10.35	-9.78	0.177	0.10	-10.46	10.46

Results obtained from 500 iterations

Yet again, the only estimator that identifies τ well is [dCdH]. **Back to Table 10**

A.3 Correspondence

From: **Guido Imbens** imbens@stanford.edu
Subject: Re: Question regarding MC-NNM in dynamic treatment regimes
Date: 6. June 2023 at 20:41
To: Schnabel, Tobias (Stud. SBE) t.schnabel@student.maastrichtuniversity.nl

yes, if there are dynamic effects, you would be estimating some complicated average of the dynamic effects, along the lines of callaway-san'tanna.

You could focus on particular weighted averages like the Xu paper. It is not obvious what the most natural thing would be to focus on.

guido

On Tue, Jun 6, 2023 at 9:09 AM Schnabel, Tobias (Stud. SBE) <t.schnabel@student.maastrichtuniversity.nl> wrote:

> Dear Prof. Imbens,
>
> I hope this email finds you well. I am an undergraduate student in Econometrics at Universiteit Maastricht and am writing my thesis on your Matrix Completion paper, for which I have one question I would very much appreciate your input on. Please forgive me for simply emailing you out of the blue for this, but your email is listed as the corresponding author, so I figured I would just try:
>
> On p. 1717 in Section 2 (highlighted PDF attached for convenience), you write that "*Also, **in the case with staggered adoption violations of the no-dynamics assumption simply changes the interpretation of the estimand**, but does not in general invalidate a causal interpretation*".
>
> My question specifically is: does this changed interpretation correspond to the **group-time average effect** (Callaway & San't Anna) a.k.a "**cohort-specific ATET**" (Sun and Abraham). Is this the changed interpretation which you were referring to, and if not, what estimand does MC-NNM yield in a staggered dynamic treatment setting?
>
> Context:
> Essentially, my thesis is about comparing MC-NNM to new estimators specifically intended for heterogeneous (and possibly dynamic) treatment effect estimation. I first wanted to compare the accuracy using the overall ATET, but realized upon re-reading the quoted sentence from your paper, that this might be a nonsensical (or unfair to MC-NNM) comparison. Would individual group-time ATET be better suited here, possibly estimated using the separate methods and aggregated the same way into an ATET? I did find a Political Science paper which in my reading equivocates MC-NNM results under dynamic treatment to group-time ATTs, but wanted to make sure I understood your paper correctly:
>
> fect: Fixed Effects Counterfactual Estimators
> yiqingxu.org
>
> Please forgive me in case the answer to my question can be found in the paper or another obvious source.
> Thank you very much in advance, best wishes and greetings from Limburg,
>
> Tobias Schnabel

--
Guido Imbens
https://stanford.zoom.us/my/imbens, passcode 19630903

Note: responses may be delayed
Press inquiries: gsbmediarelations@stanford.edu

Back to subsection 5.7

From: **Yiqing Xu** yiqingxu@stanford.edu
Subject: Re: Question regarding fect / gsynth packages
Date: 6. June 2023 at 22:55
To: Schnabel, Tobias (Stud. SBE) t.schnabel@student.maastrichtuniversity.nl

Hi Tobias,

If I remember correctly, the small differences come from covariates and hyper-parameter tuning.

Ziy, please correct me if I'm wrong.

Bets,
Yiqing

On Tue, Jun 6, 2023 at 1:34 PM Schnabel, Tobias (Stud. SBE) <t.schnabel@student.maastrichtuniversity.nl> wrote:

> Dear Prof. Xu,
>
> thank you for providing *gsynth* and *fect*, they are immensely helpful to anyone interested in implementing matrix completion! I am currently writing my undergraduate thesis on a comparison between Matrix Completion and Staggered DiD, and your packages have been a lifesaver.
>
> I do have two small questions, which I hope I did not overlook the answer to, and hope that you don't mind my asking:
>
> 1. Is there **any meaningful difference between your implementation of the MC-NNM estimator and Susan Athey's MCPanel package**? Given the lack of documentation on *MCPanel*, and the last commit on it being 6 years ago, I have tried to use very basic comparisons using your *simdata* dataset, which has yielded slightly different results between your implementation (lower MSPE) and Prof. Athey's. I did see a tweet by Scott Cunningham saying that *MCPanel* was "folded into" *gsynth*, but was hoping to make sure.
>
> 2. Between fect and gsynth, which one would you recommend I use (for matrix completion) in terms of long-term compatibility moving forward? I saw you reply to a GitHub issue 2 years ago saying that you planned on merging them, but again was just hoping to make sure.
>
> Thank you in advance, best regards,
>
> Tobias Schnabel

--
Yiqing Xu

Assistant Professor
Department of Political Science
Stanford University
https://yiqingxu.org/

Back to subsection 6.1

A.4 Thesis Proposal

Matrix Completion Estimation in Varying Panel Data Settings

as discussed with Prof. Martin Schumann

In Athey et al. (2021), the authors synthesize the literature on Causal Inference in Panel Data Econometrics. They characterize the common goal of prominent approaches in this field as estimating average (causal) treatment effects by imputing missing potential outcomes. They show that two popular frameworks for causal inference, the unconfoundedness and synthetic control approaches, can be viewed as matrix completion (MC) estimators. They then introduce a new MC estimator that they claim shows improved accuracy compared to existing methods in certain settings. MC estimation (MCE) is a technique that so far has been used extensively in the Computer Science and Statistics literatures, but has not been widely adopted in Econometrics. The empirical part of their paper is focused on comparing the accuracy of their new estimator to several other methods, including a Difference-in-Differences (DID) approach.

Since the initial publication of Athey et al. (2021), there has been a deluge of advancements in the DID literature. In the characterization of Roth et al. (2023), substantial parts of this frontier work focus on assessing and reducing estimation bias when canonical assumptions of DID are relaxed (cf. Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfoeuille, 2018; de Chaisemartin & D'Haultfœuille, 2020; Goodman-Bacon, 2021). There is currently no literature that incorporates these advances into the MCE framework.

The main goals of this proposed thesis are threefold:

1. To review and synthesize the contributions to Panel Data Econometrics made in Athey et al. (2021),
2. To examine whether recent advances in DID methods can be incorporated into the MCE framework, and
3. To compare the accuracy of Athey et al. (2021)'s new MC estimator with that of recent DID methods in data regimes in which canonical DID assumptions are not met.

References

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix Completion Methods for Causal Panel Data Models [Appendix at https://arxiv.org/pdf/1710.10251.pdf]. *Journal of the American Statistical Association*, *116*(536), 1716–1730. https://doi.org/10.1080/01621459.2021.1891924

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods [Publisher: Elsevier]. *Journal of Econometrics*, *225*(2), 200–230.

de Chaisemartin, C., & D'Haultfoeuille, X. (2018). Fuzzy Differences-in-Differences. *The Review of Economic Studies*, *85*(2), 999–1028. https://doi.org/10.1093/restud/rdx049

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*(9), 2964–2996. https://doi.org/10.1257/aer.20181169

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature [arXiv:2201.01194 [econ, stat]]. https://doi.org/10.48550/arXiv.2201.01194

Back to section 7

A.5 Official Statement Of Original Bachelor Thesis

By signing this statement, I hereby acknowledge the submitted thesis, entitled:

Matrix Completion Estimation in Differential Timing Settings

to be produced independently by me, without external help. Wherever I paraphrase or cite literally, a reference to the original source (journal, book, report, internet, etc.) is given. By signing this statement, I explicitly declare that I am aware of the fraud sanctions as stated in the Education and Examination Regulations (EERs) of the SBE.

Place: Maastricht

Date: June 28, 2023

First and last name: Tobias Schnabel

Study programme: B.Sc. Econometrics & Operations Research

EBT Code: EBT0003

ID number: i6255807

Signature: