Time-Weighted Difference-in-Differences in Short Panels with Common Shocks

Timo Schenk*
Erasmus School of Economics, Erasmus University Rotterdam

September 25, 2025

Abstract

This paper introduces a time-weighted difference-in-differences (TWDID) estimator for settings with few pre-treatment periods. Unlike conventional estimators, which use fixed pre-treatment weights, TWDID assigns variance-minimizing weights determined by the within-group covariance matrix of outcomes. The proposed estimator is efficient in the considered class when parallel trends hold across all periods. I introduce violations of parallel trends through common factors that have heterogeneous effects on the outcome. I show that the weights reduce the influence of the confounding factors, yielding a smaller bias than conventional DiD estimators under mild assumptions on the factors. Revisiting the impact of a cap-and-trade program on NOx emissions, TWDID yields smaller and more precise estimates than conventional approaches.

Keywords: difference-in-differences, interactive fixed effects, short panels, efficiency, bias reduction

*I thank my advisors Frank Kleibergen, Andreas Pick as well as Dmitry Arkhangelsky, Arturas Juodis, Pedro Sant'Anna, Mikkel Sølvsten, three anonymous referees and conference and seminar participants of the 2022 International Econometrics PhD Conference (Rotterdam), the Amsterdam Panel Data Workshop, ESEM, IAAE, IPDC, the NESG Meeting, and internal seminars for helpful comments and discussions. Financial support by the IAAE travel grant is highly acknowledged. Previous title: Time-Weighted Difference-in-Differences: Accounting for Common Factors in Short T Panels

1 Introduction

The presence of interactive fixed effects in the untreated potential outcomes leads to biased difference-in-difference (DID) estimates of average treatment effects. While the estimators of Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) and Chan and Kwok (2021) address this issue in large T panels, the question remains how to account for common factors in short T panels.

In this paper, I introduce a time-weighted DID (TWDID) approach to estimate the average treatment effect on the treated (ATT) of a binary treatment. While the conventional DiD estimator considers only the most recent pre-treatment period, TWDID uses a linear combination of pre-treatment periods that minimizes the variance of the resulting estimator. The corresponding pre-treatment weights are estimated from the sample covariance of outcomes. TWDID remains unbiased under the assumption that parallel trends holds across all periods, since then the ATT is identified for any set of pre-treatment weights that sum to one. The proposed estimator therefore provides substantial efficiency gains in those cases.

Next, I consider violations of parallel trends in an interactive fixed effects model. In this setup, unobserved time-varying common factors have a time-invariant but heterogeneous effect on the outcome. If those correlate with the treatment assignment, parallel trends no longer holds across all periods, but only with respect to unknown weighted averages of pre-treatment periods. This leads to a bias in the conventional DiD estimator that is proportional to the differences between the pre-treatment and post-treatment factors. Addressing the bias therefore requires pre-treatment weights that balance the confounding factors.

The main take-away of the paper is that TWDID, besides being efficient under parallel trends, typically reduces bias compared to conventional DiD. This works because

the confounding factors that generate the bias also affect the variance of the outcomes. Variance-minimizing weights therefore adjust to the dynamics of the factors without needing to estimate the factors directly. However, since the variance is also affected by noise, the bias is generally not eliminated completely. This is inherent to the setting, since latent factors can generally not be consistently recovered without stronger conditions when only a few pre-treatment periods are available.

Formally, I show that the variance-minimizing weights shrink towards noise-minimizing weights, leading to an attenuation bias that depends on the alignment between noise dynamics and factor dynamics.TWDID is asymptotically unbiased in the most favorable case in which noise dynamics and factors are perfectly aligned. I present sufficient conditions on the noise-minimizing weights under which the reduction in bias relative to conventional DiD is guaranteed. Moreover, the bias vanishes asymptotically if the factor strength increases with the sample size.

The theoretical results have implications about the empirical practice of controlling for 'pre-trends', which are discussed also in Steffens and Stuhle (2025). The TWDID estimator and the associated time weights can be jointly estimated in a linear regression of differences in outcomes on the treatment augmented with individual pre-trends. This provides useful insights for practitioners. If parallel trends holds across all periods, the pre-treatment differences in outcomes are uncorrelated with the treatment assignment. Therefore, the coefficient on the treatment still identifies the ATT when augmenting the regression. At the same time, it increases precision if the pre-trends are predictive of the changes in the outcome that are explained by the treatment. When parallel trends is violated, the pre-trends act as noisy proxies of the confounding factors. The remaining bias in TWDID is therefore similar to bias from measurement error, resulting in attenuation towards noise-minimizing time weights.

I show that the TWDID estimator and the estimated time weights are jointly asymptot-

ically normal under large N asymptotics. Standard errors of the estimated treatment effect need to account for the presence of estimated weights. This is achieved by conventional cluster-robust variance estimators applied to the joint estimating equations. Moreover, this result allows researchers to test restrictions on the variance-minimizing time weights, which serves as useful diagnostic check.

I revisit the study by Deschenes, Greenstone, and Shapiro (2017), who employ a triple-DiD design to estimate the effect of a cap-and-trade program on NOx emissions. In this setting, interactive fixed effects may arise if common shocks, such as the business cycle or weather, affect counties in a heterogeneous way. Applying TWDID leads to smaller estimated treatment effects than conventional DiD or two-way fixed effects estimators suggest, while at the same time producing narrower confidence intervals. The results suggest that part of the decline in emissions captured by conventional DiD is attributable to confounding aggregate shocks, illustrating that time-weighting can reduce this bias while improving precision.

Related literature. The TWDID estimator is closely related to the synthetic DID (SDID) estimator of Arkhangelsky et al. (2021). Both approaches use time weighting to balance the unobserved factors, but the weighting scheme differs in several ways. First, SDID estimates the time weights using only the untreated units, while TWDID estimates them from the full sample and thereby exploits information from the treated units' dynamics. Second, SDID imposes a non-negativity constraint on the time weights, providing regularization in long panels. TWDID does not impose such constraints, thus allowing for negative time weights. Finally, SDID also estimates control unit weights, which further reduces bias, but requires a large number of pre-treatment periods for consistency. When T is small, SDID may exacerbate bias compared to DiD, as suggested by Monte Carlo experiments in this paper, whereas TWDID consistently improves upon DiD in such

settings.

Other approaches to identify and estimate treatment effects in short T panels with interactive fixed effects, reviewed recently in Brown and Butts (2022), usually involve additional moment conditions. Under those, one can account for latent factors for example through forward-orthogonal differencing as in Ahn, Lee, and Schmidt (2013) or using time-invariant observable covariates with constant effect on the outcome as instruments (Callaway and Karami, 2022). Instead, TWDID exploits heterogeneity of the factor loadings.

This paper relates to other discussions around efficiency of DiD approaches when parallel trends holds across multiple periods, for example Marcus and Sant'Anna (2021) ,Roth and Sant'Anna (2023), Borusyak, Jaravel, and Spiess (2024), Harmon (2024), and Chen, Sant'Anna, and Xie (2025). Allowing violations of parallel trends, approaches by Manski and Pepper (2018) and Rambachan and Roth (2023) partially identify the ATT by restricting the way those violations evolve over time. Relative to their findings, this paper shows how the ATT can be recovered with "misspecified" efficient DiD estimators by relating the parallel trend violations to the variance in the outcomes.

This paper also relates to findings in the literature on synthetic control (SC) estimators (Abadie, Diamond, and Hainmueller, 2010; Xu, 2017; Abadie and L'hour, 2020; Ferman, 2021; Ferman and Pinto, 2021; Ben-Michael, Feller, and Rothstein, 2021). They use unit weights to balance time-invariant unobserved characteristics between treated and untreated units, which are estimated with a time-series regression over the pre-treatment periods. Consistent estimation requires a large number of pre-treatment periods, strict balancing conditions on the loadings and restrictions on the serial dependence of the errors. Similarly, TWDID requires a large number of control units and balancing conditions on the factors. Exploiting independence over the cross-section, however, TWDID estimation remains reliable when the data exhibits strong serial dependence.

When both N and T are large, one can use results from linear panel data models with

factor structures (Pesaran, 2006; Bai, 2009) to recover a long run average treatment effect. For example, Gobillon and Magnac (2016) apply the estimator of Bai (2009) to estimate the average treatment effect jointly with the factor structure. Using principal component analysis, Chan and Kwok (2021) construct factor proxies, which can then be used in a factor-augmented regression. In short T panels, however, these estimators are generally inconsistent.

The remainder of the paper is structured as follows. Section 2 introduces the interactive fixed effects model and the TWDID approach. Formal theoretical results are devoted to Section 3 Section 4 illustrates the theoretical results and finite sample properties with simulations. Section 5 contains the application and Section 6 concludes.

2 A time-weighted difference-in-differences approach

2.1 Setup

Using a panel data set for treated and untreated units, we wish to estimate the effect of a policy intervention starting in period $t = T_0$. That is, we seek to estimate the average treatment effect on the treated

$$\tau_t := ATT_t = \mathrm{E}[y_{it}(1) - y_{it}(0)|D_i = 1]$$

in the post-treatment periods $t = T_0 + 1, \ldots, T$, where $y_{it}(1), y_{it}(0)$ are the potential outcomes of unit i in period t, and $D_i \in \{0, 1\}$ indicating whether unit i is ever treated. The researcher observes $y_{it} = D_i y_{it}(1) + (1 - D_i)y_{it}(0)$ for a large number of units $i = 1, \ldots, N$ and a small number of periods $t = 1, \ldots, T$, covering at least two pre-treatment periods ($T_0 \geq 2$).

The untreated potential outcomes are generated by an interactive fixed effects model,

$$y_{it}(0) = \beta_i + \gamma_t + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \varepsilon_{it} \tag{1}$$

where β_i are unit fixed effects, γ_t are time fixed effects, $\boldsymbol{f}_t$ and $\boldsymbol{\lambda}_i$ are r-dimensional vectors of common factors and loadings, and ε_{it} is an idiosyncratic error component. Such unobserved factor structures, $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$, are present in many economic settings. In microeconomic applications, $\boldsymbol{\lambda}_i$ can be thought of a vector of unobserved, time-invariant characteristics of individual i. In contrast to the fixed effects β_i, they have a time-varying impact on the outcome y_{it} measured by $\boldsymbol{f}_t$. In macroeconomic applications, the factors $\boldsymbol{f}_t$ are unobserved common shocks (e.g. technology or weather shocks) that, in contrast to the time fixed effects γ_t, have an heterogeneous impact $\boldsymbol{\lambda}_i$ on unit i. This DGP nests important cases such as deterministic, linear trends when $f_t = t$. It also nests the two-way fixed model, when all factors are constant over time ($\boldsymbol{f}_t = f$), in which case the factor structure is absorbed by the unit fixed effects.

To simplify notation, let $\mathbb{N}_j = \{i \colon D_i = j\}$, $j = 0, 1$ denote the sets of untreated and treated units in the sample, respectively. Let $N_0 = \sum_{i=1}^{N}(1 - D_i)$ and $n_0 = \frac{N_0}{N}$ denote the number and share of untreated units, respectively. The factors are stacked into the $T \times r$ matrix $\boldsymbol{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T]'$, and similarly $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,T})'$, $\boldsymbol{y}_i(d) = (y_{i,1}(d), \ldots, y_{i,T}(d))'$ and $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,T})'$.

Consider the following assumptions.

Assumption 1 (No anticipation). $y_{it}(1) = y_{it}(0)$ for all $t \leq T_0$ and all $i = 1, \ldots, N$.

Assumption 2 (Potentially confounding loadings). The $\boldsymbol{\lambda}_i$ are random vectors satisfying $\mathrm{E}[\boldsymbol{\lambda}_i | D_i = d] = \boldsymbol{\mu}_\lambda^{(d)}$ with $|\boldsymbol{\mu}_\lambda^{(d)}| < \infty$ for $d = 0, 1$ and $\mathrm{Var}[\boldsymbol{\lambda}_i | D_i] = \boldsymbol{\Sigma}_{\lambda,i}$ with $\lim_{n\to\infty} \frac{1}{N_d} \sum_{i \in \mathbb{N}_d} \boldsymbol{\Sigma}_{\lambda,i} = \boldsymbol{\Sigma}_\lambda^{(d)}$ for $d = 0, 1$.

Assumption 3 (Parallel trends conditional on time-invariant unobservables). For every i, $\mathrm{E}[\boldsymbol{\varepsilon}_i | \beta_i, D_i, \boldsymbol{\lambda}_i, \boldsymbol{F}] = \boldsymbol{0}$. Moreover, $\mathrm{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' | D_i] = \boldsymbol{\Sigma}_{\varepsilon,i}$ with $\lim_{n\to\infty} \frac{1}{N_d} \sum_{i \in \mathbb{N}_d} \boldsymbol{\Sigma}_{\varepsilon,i} = \boldsymbol{\Sigma}_\varepsilon^{(d)}$ for $d = 0, 1$; and $\boldsymbol{\Sigma}_\varepsilon = n_0^{-1} \boldsymbol{\Sigma}_\varepsilon^{(0)} + (1 - n_0)^{-1} \boldsymbol{\Sigma}_\varepsilon^{(1)}$ is a positive definite $T \times T$ matrix.

Assumption 4 (Random sampling). $(\boldsymbol{y}_i(1), \boldsymbol{y}_i(0), \boldsymbol{\varepsilon}_i, \boldsymbol{\lambda}_i)$ are independent over the cross section; $n_0 \in (0,1)$ is constant as $N \to \infty$, and $T_0 \geq 2$.

Assumption 5 (Treatment effect heterogeneity). The vector of individual treatment effects $\boldsymbol{\tau}_i = (\tau_{i,T_0+1}, \ldots, \tau_{i,T})'$, $\tau_{it} = y_{it}(1) - y_{it}(0)$, satisfies $\frac{1}{N}\sum_i D_i(\boldsymbol{\tau}_i - \boldsymbol{\tau})(\boldsymbol{\varepsilon}_i', \dot{\boldsymbol{\lambda}}_i')' \xrightarrow{p} 0$ where $\dot{\boldsymbol{\lambda}}_i = \boldsymbol{\lambda}_i - \mathrm{E}[\boldsymbol{\lambda}_i | D_i]$.

Assumption 6 (Confounders do not perfectly correlate with treatment). Let $\boldsymbol{\xi}_\lambda = \boldsymbol{\mu}_\lambda^{(1)} - \boldsymbol{\mu}_\lambda^{(0)}$, $\boldsymbol{\Sigma}_\lambda = n_0^{-1}\boldsymbol{\Sigma}_\lambda^{(0)} + (1-n_0)^{-1}\boldsymbol{\Sigma}_\lambda^{(1)}$.

1. The loadings satisfy $\lim_{\zeta \to 0} \boldsymbol{\xi}_\lambda'(\boldsymbol{\Sigma}_\lambda + \zeta \boldsymbol{I})^{-1}\boldsymbol{\xi}_\lambda < \bar{w}_\lambda$ for some finite constant $\bar{w}_\lambda$.

2. The factors satisfy rank $\boldsymbol{F} = r$ with $0 < r < T_0$. Moreover, $\boldsymbol{d}'\boldsymbol{M}_{F_a}\boldsymbol{d} > \bar{w}_f$ for some constant $\bar{w}_f > 0$ and for all linear combination of post-treatment dummies $\boldsymbol{d} \in \text{span}\,\{\boldsymbol{e}_t\}_{t>T_0}$, where $\boldsymbol{F}_a = [\boldsymbol{\iota}, \boldsymbol{F}]$, $\boldsymbol{e}_t$ the t-th unit basis vector in $\mathbb{R}^T$ and $\boldsymbol{\iota} = (1, \ldots, 1)'$.

Assumption 1 ensures that we observe the untreated potential outcome of the treated units prior to the treatment. It would be violated in presence of anticipation effects, i.e. when the treatment affects the outcome before it actually starts. If sufficient pre-treatment observations are available, one can estimate the anticipation effects as it is commonly done in event-study designs.

Assumption 2 is a central characteristic of the model. It allows the loadings $\boldsymbol{\lambda}_i$ to differ systematically between treated and untreated units. The loading imbalance $\boldsymbol{\xi}_\lambda = \boldsymbol{\mu}_\lambda^{(1)} - \boldsymbol{\mu}_\lambda^{(0)}$ measures how much more (or less) the treated units are on average affected by the common factors $\boldsymbol{f}_t$. It nests the two-way fixed effects model for constant loadings $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$. In that case, the factor structure $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$ would be entirely absorbed by the time fixed effect $\tilde{\gamma}_t = \gamma_t + \boldsymbol{\lambda}'\boldsymbol{f}_t$. Moreover, it allows weak factors, as it prevents neither $\boldsymbol{\mu}_\lambda^{(j)}$ nor $\boldsymbol{\Sigma}_{\lambda,i}$ from diminishing in large samples.

Under Assumption 3, the treatment assignment is strictly exogenous once conditioned on the loadings, fixed effects and the factors. This implies parallel trends in the untreated potential outcome $y_{it}(0)$ conditional on the unobserved loadings, while allowing for heteroskedasticity and arbitrary serial dependence of the idiosyncratic errors. Lastly, Assumption 4 is a conventional restriction that ensures independence of potential outcomes, loadings, and idiosyncratic errors over the cross section. It also requires the number of treated and untreated units to grow at the same rate.

Assumption 5 imposes that within the treated units, the treatment effects are (asymptotically) uncorrelated with the idiosyncratic errors and the loadings, which is trivially satisfied when treatment effects are constant. When treatment effects are heterogeneous, it allows together with Assumption 3 to decompose the within-group covariances of outcome into three separate parts resulting from the factor structure, the idiosyncratic errors, and the treatment effect heterogeneity, respectively. While this is not strictly necessary for the main findings, it eases exposition and interpretation.

Assumption 6 imposes mild conditions on the unobserved confounders. While the basic distributional results of this paper are derived without this assumption, it is essential to guarantee bias reduction properties of the estimator. The first statement rules out multicollinearity between the loadings $\boldsymbol{\lambda}_i$ and the treatment assignment D_i, which is similar to requiring overlap in the propensity score under unconfoundedness, see for instance Imbens and Wooldridge (2009). If the loadings are correlated with D_i, it requires a non-singular covariance matrix $\boldsymbol{\Sigma}_\lambda$. However, it does not rule out constant loadings $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$. Similarly, the second statement requires sufficient variation in the pre-treatment factors, unless the factors are constant over time. It therefore excludes factors that are constant pre-treatment but vary after onset of the treatment. In that case it would be impossible to distinguish the effect from the factors from those of the treatment, even if the factors were observed. The common factors $\boldsymbol{F}$ can be viewed as realizations from some deterministic or stochas-

tic process, where the number of factors r is unknown and fixed. All results should be interpreted conditional on $\boldsymbol{F}$.

The following sections focus on the case of one treated period $t = T$, and denotes the relevant ATT by $\tau = \tau_T$. The cases of multiple treated periods and staggered adoption are discussed in Appendix A.2.

2.2 Identification and estimation with fixed time weights

This paper considers estimators based on sample mean differences $\hat{y}_t = \bar{y}_{1,t} - \bar{y}_{0,t}$ in each period, where $\bar{y}_{d,t} = \frac{1}{N_d} \sum_{i:D_i=d} y_{it}$, $d = 0, 1$ are the sample means of observed outcomes within untreated and treated units in period t. Denoting $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_T)$, the factor model (1) implies that

$$\hat{\boldsymbol{y}} = \boldsymbol{e}\tau + \boldsymbol{\iota}\boldsymbol{\xi}_\beta + \boldsymbol{F}\boldsymbol{\xi}_\lambda + \hat{\boldsymbol{u}}, \quad \sqrt{N}\hat{\boldsymbol{u}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}) \tag{2}$$

where the post-treatment dummy $\boldsymbol{e} = (0, \ldots, 0, 1)'$ is the regressor of interest, $\boldsymbol{\iota} = (1, \ldots, 1)'$ is a constant vector of ones, and $\boldsymbol{\Omega}$ the limiting covariance matrix of $\hat{\boldsymbol{y}}$. Therefore, identifying τ can be formulated as time-series regression problem where the factors $\boldsymbol{F}$ act as unobserved confounders.

This paper focuses on time-weighted DiD estimators

$$\hat{\tau}(\boldsymbol{v}) = \hat{y}_T - \sum_{t \leq T_0} v_{\text{pre},t} \hat{y}_t = \boldsymbol{v}'\hat{\boldsymbol{y}}, \quad \boldsymbol{v} \in \mathbb{V} = \{\boldsymbol{v} \in \mathbb{R}^T : [\boldsymbol{e}, \boldsymbol{\iota}]'\boldsymbol{v} = (1, 0)\}$$

These estimators compare differences in the treated period to a weighted average of pre-treatment differences, where $\boldsymbol{v}_{\text{pre}} = (v_{\text{pre},1}, \ldots, v_{\text{pre},T_0})'$ denotes the associated pre-treatment weights. It is more convenient to carry out the theoretical analysis in terms of the full vector of time weights $\boldsymbol{v} = (-\boldsymbol{v}'_{\text{pre}}, 1)'$. The set $\mathbb{V}$ restricts the weights to sum to zero and sets the post-treatment weight equal to one. Therefore, the pre-treatment weights $\boldsymbol{v}_{\text{pre}} = (-v_1, \ldots, -v_{T_0})'$ must sum to one.

Many prominent DiD estimators correspond to a particular choice of $\boldsymbol{v}$. One example is the two-way fixed effects (2wfe) estimator, which weights all pre-treatment periods equally. It corresponds to $\boldsymbol{v}_{\text{2wfe}} = (-1/T_0, \dots, -1/T_0, 1)$, An equivalent expression is $\boldsymbol{v}_{\text{2wfe}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$ with $\boldsymbol{X} = [\boldsymbol{e}, \boldsymbol{\iota}]$ and $\boldsymbol{c} = (1, 0)'$, which are the weights implied by the ordinary least-squares estimator in (2). Another example is the canonical DiD estimator, which uses the most recent pre-treatment period as 'base-period'. It corresponds to $\boldsymbol{v}_{\text{did}} = (0, \dots, 0, -1, 1)'$ or equivalently $\boldsymbol{v}_{\text{did}} = \boldsymbol{\Delta}'\boldsymbol{\Delta}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Delta}'\boldsymbol{\Delta}\boldsymbol{X})^{-1}\boldsymbol{c}$, where $\boldsymbol{\Delta}$ is the $T \times T$ matrix taking first differences, i.e. $\boldsymbol{\Delta}\boldsymbol{x} = (x_1, x_2 - x_1, \dots, x_T - x_{T-1})'$ for any vector $\boldsymbol{x} \in \mathbb{R}^T$.

A time-weighted DiD estimator $\hat{\tau}(\boldsymbol{v})$ is unbiased for τ under a weighted parallel trends assumption

$$\mathrm{E}\big[y_{i,T}(0) - \sum_{t \le T_0} v_{\text{pre},t} y_{it}(0) | D_i = 1\big] = \mathrm{E}\big[y_{i,T}(0) - \sum_{t \le T_0} v_{\text{pre},t} y_{it}(0) | D_i = 0\big]$$

which restricts the expected change in $y_{it}(0)$ with respect to a particular weighted average of pre-treatment periods given by $\boldsymbol{v}$. More compactly, the condition reads as $\boldsymbol{v}'(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}) = 0$, where the vector $\boldsymbol{\mu}^{(d)} = E[\boldsymbol{y}_i(0) | D_i = d]$ contains the mean untreated potential outcome in treatment group $d = 0, 1$ stacked over all periods.

The main challenge is that parallel trends may only hold for $\boldsymbol{v}$ in an unknown subset $\mathbb{V}_0 \subset \mathbb{V}$. In the interactive fixed effects model (1), this set is $\mathbb{V}_0 = \{v \in \mathbb{V} : \ \boldsymbol{v}'\boldsymbol{F}\boldsymbol{\xi}_\lambda = 0\}$, since $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)} = \boldsymbol{\iota}\boldsymbol{\xi}_\beta + \boldsymbol{F}\boldsymbol{\xi}_\lambda$ and $\boldsymbol{v}'\boldsymbol{\iota} = 0$. In the special case of balanced loadings ($\boldsymbol{\xi}_\lambda = 0$), parallel trends hold with respect to all (combinations of) pre-treatment periods, and any choice of weights $\boldsymbol{v} \in \mathbb{V} = \mathbb{V}_0$ results in an unbiased estimator. However, in the general case, $\mathbb{V}_0$ depends on the unobserved factors $\boldsymbol{F}$. Without further restrictions on $\boldsymbol{F}$, standard choices of $\boldsymbol{v}$ lead to biased estimates of the ATT.

This paper proposes to use weights $\boldsymbol{v}^*$ that minimize the variance of $\hat{\tau}(\boldsymbol{v})$. In the setup of this paper, $\mathrm{var}[\hat{\tau}(\boldsymbol{v})] = \boldsymbol{v}'\boldsymbol{\Omega}\boldsymbol{v}/N$, and $\boldsymbol{\Omega} = \mathrm{var}[\hat{\boldsymbol{y}}] \cdot N$ can be consistently estimated

by conventional variance estimators of mean differences. The smallest variance among all time-weighted DiD estimators is obtained for

$$\boldsymbol{v}^* = \underset{\boldsymbol{v}\in\mathbb{V}}{\arg\min}\, \boldsymbol{v}'\boldsymbol{\Omega}\boldsymbol{v} = \boldsymbol{\Omega}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}$$

which I refer to as variance-minimizing weights. The Gauss-Markov Theorem tells us that $\hat{\tau}(\boldsymbol{v}^*)$ corresponds to the GLS estimator in (2) that ignores the latent factors $\boldsymbol{F}$.[1] Since $\hat{\tau}(\boldsymbol{v}^*)$ is unbiased if $\boldsymbol{\xi}_\lambda = \boldsymbol{0}$, it is efficient among time-weighted DiD estimators when parallel trends holds for all $\boldsymbol{v} \in \mathbb{V}$. For example, for i.i.d. homoskedastic errors $\hat{\boldsymbol{u}}$, the covariance matrix $\boldsymbol{\Omega} = \sigma_y^2\boldsymbol{I}$ is diagonal, $\hat{\tau}(\boldsymbol{v}^*)$ corresponds to the 2wfe estimator with equal pre-treatment weights. For random walk errors, $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}'$ with $\boldsymbol{L}$ a lower-diagonal matrix taking cumulative sums, so $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Delta}'\boldsymbol{\Delta}$. Then $\hat{\tau}(\boldsymbol{v}^*)$ corresponds to the canonical DiD estimator, known to be efficient under these error dynamics (Harmon, 2024).

In the setup of this paper, variance-minimizing weights also have favorable bias properties when $\boldsymbol{\lambda} \neq \boldsymbol{0}$. With heterogeneous loadings $\boldsymbol{\lambda}_i$, the covariance matrix $\boldsymbol{\Omega} = \boldsymbol{F}\boldsymbol{\Sigma}_\lambda\boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon$ is informative about the confounding factors. Minimizing the variance $\boldsymbol{v}'\boldsymbol{\Omega}\boldsymbol{v}$ therefore reduces the bias $b(\boldsymbol{v}) = \boldsymbol{v}'\boldsymbol{F}\boldsymbol{\xi}_\lambda$ compared to $\boldsymbol{v}_{\text{did}}$ in many cases. The remainder of this paper presents formal conditions under which such bias reduction properties can be guaranteed.

2.3 Joint estimation of time weights and the treatment effect

This section introduces the estimators $\hat{\tau}(\hat{\boldsymbol{v}})$ and $\hat{\boldsymbol{v}}$ targeting $\boldsymbol{v}^*$ and τ from a practical point of view. Formal distributional results are presented in Section 3.

The starting point is the variance estimator of mean differences $\hat{\boldsymbol{y}}$, given by

$$\widehat{\boldsymbol{\Omega}} = (1-n_0)^{-1}\widehat{\boldsymbol{\Omega}}_1 + n_0^{-1}\widehat{\boldsymbol{\Omega}}_0, \quad \widehat{\boldsymbol{\Omega}}_d = \frac{1}{N_d}\sum_{i\in\mathbb{N}_d}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_d)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_d)',\ d = 0,1$$

where $\widehat{\boldsymbol{\Omega}}_d$, $d = 0, 1$ are the sample covariance matrices of observed outcomes within untreated and treated units. The estimated weights result from plugging in $\widehat{\boldsymbol{\Omega}}$ into the general

[1] I thank Dmitry Arkhangelsky and an anonymous referee for pointing me in this direction.

expression for $\boldsymbol{v}^*$,

$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{v}'\widehat{\boldsymbol{\Omega}}\boldsymbol{v} = \widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c} \tag{3}$$

which are well-defined in short panels as long as $N_d \geq T$. The resulting time-weighted DiD (TWDID) estimator $\hat{\tau}(\hat{\boldsymbol{v}}) = \hat{\boldsymbol{v}}'\hat{\boldsymbol{y}}$ can be interpreted as the feasible GLS estimator in (2) and is the main object of interest in this paper.[2]

A useful insight is that $\hat{\tau}(\hat{\boldsymbol{v}})$ and $\hat{\boldsymbol{v}}$ can be jointly estimated from a linear regression. Substituting the conditions, the time weights can be expressed as

$$\boldsymbol{v} = \boldsymbol{v}_{\text{did}} - \boldsymbol{Q}\boldsymbol{\nu} = (-\boldsymbol{\nu}', \boldsymbol{\iota}'\boldsymbol{\nu} - 1, 1)'$$

where $\boldsymbol{\nu} \in \mathbb{R}^{T-2}$ are the weights of the first $T-2$ periods and $\boldsymbol{Q} = [\boldsymbol{I}_{T-2}, -\boldsymbol{\iota}, \boldsymbol{0}]'$ is a $T \times T-2$ matrix whose transpose takes long differences with respect to period $t = T_0$. This is without loss of generality, since the columns of $\boldsymbol{Q}$ form a basis of $\{\boldsymbol{X}'\boldsymbol{v} = \boldsymbol{0}\}$. Let $\dot{\boldsymbol{y}}_{i,\text{pre}} = \boldsymbol{Q}'\boldsymbol{y}_i = (y_{i,1} - y_{i,T_0}, \ldots, y_{i,T_0-1} - y_{i,T_0})'$ be the vector of $T-2$ unit-specific 'pre-trends'. Then $\hat{\tau}(\hat{\boldsymbol{v}})$ and $\hat{\boldsymbol{v}} = \boldsymbol{v}_{\text{did}} - \boldsymbol{Q}\hat{\boldsymbol{\nu}}$ solve the cross-sectional regression problem

$$\min_{\tau,\boldsymbol{\nu},\gamma} M_n(\tau, \boldsymbol{\nu}, \gamma), \quad M_n(\tau, \boldsymbol{\nu}, \gamma) = \sum_i p_i^{-1}(\dot{y}_{i,\text{post}} - \gamma - \tau D_i - \dot{\boldsymbol{y}}_{i,\text{pre}}'\boldsymbol{\nu})^2. \tag{4}$$

Here the difference in outcomes $\dot{y}_{i,\text{post}} = \boldsymbol{y}_i'\boldsymbol{v}_{\text{did}} = y_{i,T} - y_{i,T_0}$ is regressed on the treatment, controlling for all $T-2$ pre-trends, and weighting observations by their inverse propensity score $p_i = D_i(1-n_0)^2 + (1-D_i)n_0^2$. The canonical DiD estimator $\hat{\tau}(\boldsymbol{v}_{\text{did}})$ can be obtained from (4) when omitting the pre-trends from the regression by restricting $\boldsymbol{\nu} = \boldsymbol{0}$; similarly the 2wfe estimator $\hat{\tau}(\boldsymbol{v}_{\text{2wfe}})$ follows from restricting $\boldsymbol{\nu} = \boldsymbol{\iota}/T_0$. The regression formulation also shows that $\hat{\tau}(\hat{\boldsymbol{v}})$ and $\hat{\boldsymbol{v}}$ are invariant to the unit fixed effects β_i, since the outcomes only enter in differences. They are also invariant to time fixed effects γ_t, which are fully absorbed by the constant γ in the regression.

[2]Previous versions of this paper considered estimating the time weights only from the untreated units $\arg\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{v}'\widehat{\boldsymbol{\Omega}}_0\boldsymbol{v}$. The results continue to hold in this case when restricting heteroskedasticity, i.e. $\boldsymbol{\Sigma}_\lambda^{(0)} = \boldsymbol{\Sigma}_\lambda^{(1)}$ and $\boldsymbol{\Sigma}_\varepsilon^{(0)} = \boldsymbol{\Sigma}_\varepsilon^{(1)}$.

The coefficients of the pre-trends exactly recover the estimated weights $\hat{\boldsymbol{v}}$ defined in (3). This can be verified by tracing out γ and τ from (4). Writing $\dot{y}_{i,\text{post}} - \dot{\boldsymbol{y}}_{i,\text{pre}}'\boldsymbol{\nu} = \boldsymbol{y}_i'\boldsymbol{v}$, it follows that the regression problem for a given vector of weights is solved by $\hat{\tau}(\boldsymbol{v}) = \boldsymbol{v}'\hat{\boldsymbol{y}}$ and $\hat{\gamma}(\boldsymbol{v}) = \boldsymbol{v}'\bar{\boldsymbol{y}}_0$. The estimated weights therefore solve $\min_{v\in\mathbb{V}} M_n(\boldsymbol{v}, \hat{\gamma}(\boldsymbol{v}), \hat{\tau}(\boldsymbol{v}))$ with objective function

$$M_n(\boldsymbol{v}, \hat{\gamma}(\boldsymbol{v}), \hat{\tau}(\boldsymbol{v})) = \sum_i p_i^{-1}((\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{D_i})'\boldsymbol{v})^2 = N\boldsymbol{v}'\widehat{\boldsymbol{\Omega}}\boldsymbol{v}$$

which coincides with the definition of $\hat{\boldsymbol{v}}$.

The regression representation highlights the gains from using data-driven time weights in terms of efficiency and bias reduction from another angle. If parallel trends holds with respect to all pre-treatment periods, the pre-trends $\dot{\boldsymbol{y}}_{i,\text{pre}}$ are uncorrelated with the treatment assignment D_i. Therefore, restricting $\boldsymbol{\nu} = \mathbf{0}$ does not change the estimand, since $\tau(\boldsymbol{v}^*) = \tau$. Still, controlling for pre-trends increases precision if the (population) variance-minimizing weights $\boldsymbol{v}^* = \boldsymbol{v}_{\text{did}} - \boldsymbol{Q}\boldsymbol{\nu}^*$ differ from $\boldsymbol{v}_{\text{did}}$, i.e. if the population regression coefficient of the pre-trends $\boldsymbol{\nu}^*$ is different from zero. If parallel trends does not hold in all periods, controlling for $\dot{\boldsymbol{y}}_{i,\text{pre}}$ changes the estimand $\tau(\boldsymbol{v}^*) = \tau + b(\boldsymbol{v}^*)$. A bias term remains, because $\dot{\boldsymbol{y}}_{i,\text{pre}}$ are noisy proxies of the confounding changes in factors $\boldsymbol{Q}'\boldsymbol{F}\boldsymbol{\lambda}_i$.

Another advantage of the one-step formulation is that standard errors for $\hat{\tau}(\hat{\boldsymbol{v}})$ and $\hat{\boldsymbol{v}}$ can be easily constructed from usual cluster-robust variance estimators of the form $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{H}}^{-1}$, where $\boldsymbol{\Sigma}$ is the limiting variance of $\hat{\boldsymbol{\theta}} = (\hat{\tau}(\hat{\boldsymbol{v}}), \hat{\boldsymbol{\nu}}', \hat{\gamma})'$. This allows researches to conduct inference on the variance-minimizing weights $\boldsymbol{v}^*$. A useful diagnostic test is whether conventional DiD is efficient under parallel trends, corresponding to the restriction $\boldsymbol{v}^* = \boldsymbol{v}_{\text{did}}$.

Section 3.1 defines those formally and shows their asymptotically validity under relatively mild conditions. Extensions to multiple treated periods, staggered adoption, and including deterministic trends are discussed in Appendix A.1.

2.4 Relation to synthetic DiD and other approaches

While this paper focuses on the class of time-weighted DiD estimators in short panels, many other approaches to estimate the ATT have been studied in presence of interactive fixed effects.

A closely related approach proposed for large N large T panels is the synthetic DID estimator of Arkhangelsky et al. (2021), which weights both the pre-treatment periods and the untreated units. Specifically, the SDID estimator $\hat{\tau}_{\text{sdid}} = \hat{\tau}(\hat{\boldsymbol{v}}_{\text{sdid}}, \hat{\boldsymbol{\omega}})$ is constructed from $\hat{\tau}(\boldsymbol{v}, \boldsymbol{\omega}) = \boldsymbol{v}'(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_0(\boldsymbol{\omega}))$, where $\bar{\boldsymbol{y}}_0(\boldsymbol{\omega}) = \sum_{D_i=0} \omega_i \boldsymbol{y}_i$ the weighted average of control unit outcomes, typically referred to as synthetic control. The weights $\hat{\boldsymbol{\omega}}_{\text{sdid}}, \hat{\boldsymbol{v}}_{\text{sdid}} = (-\hat{\boldsymbol{v}}'_{\text{pre}}, 1)'$ are obtained from the penalized regressions

$$\min_{\boldsymbol{\omega} \in \Delta_{N_0}, \beta} \sum_{t \leq T_0} (\bar{y}_{1,t} - \beta - \omega' y_{0,t} - \zeta \boldsymbol{\omega}' \boldsymbol{\omega})^2, \quad \min_{\boldsymbol{v}_{\text{pre}} \in \Delta_{T_0}, \gamma} \sum_{D_i=0} (y_{i,T} - \gamma - \boldsymbol{v}'_{\text{pre}} \boldsymbol{y}_{i,\text{pre}})^2 \tag{5}$$

where $\zeta > 0$ is a penalty term in the unit weight estimation and $\Delta_k = \{\boldsymbol{x} \in \mathbb{R}^k : \ \boldsymbol{x}'\boldsymbol{\iota} = 1, \ x_i \geq 0\}$ restricts the weights to be non-negative and sum to one. The TWDID estimator $\hat{\tau}(\hat{\boldsymbol{v}})$, besides having equal control unit weights, uses more flexible time weights that differ from $\hat{\boldsymbol{v}}_{\text{sdid}}$ in two aspects. First, $\hat{\boldsymbol{v}}$ is estimated from the whole sample, while $\hat{\boldsymbol{v}}_{\text{sdid}}$ is estimated only from the untreated units. TWDID therefore takes into account additional information from the dynamics in outcomes of the treated units. As a consequence, it remains variance-minimizing under heteroskedasticity, in particular when $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_0$. Second, SDID imposes a non-negativity constraint on the time weights, which provides necessary regularization when T_0 is large. TWDID lifts this constraint as it is not required when T_0 is small.

Next, the synthetic control (SC) estimator (Abadie, Diamond, and Hainmueller, 2015; Abadie et al., 2010) in its original form is $\hat{\tau}_{\text{sc}} = \hat{\tau}(\boldsymbol{0}, \hat{\boldsymbol{\omega}})$, while Ferman and Pinto (2021) consider a demeaned version $\hat{\tau}_{\text{dsc}} = \hat{\tau}(\boldsymbol{\iota}/T_0, \hat{\boldsymbol{\omega}})$ which corrects for pre-treatment differences. TWDID, using time weights estimated from a large cross-section, can be thought of as

a "transposed" variant of the demeaned synthetic control estimator that use control unit weights estimated from a large number of pre-treatment periods.

I conclude this section with a motivation for allowing negative time weights in short T settings. Generally, researchers may find extrapolation over time more acceptable than extrapolation over the cross-section. For example, negative pre-treatment weights are required to account for unit-specific linear trend, which is common in empirical practice. When T is fixed, the unrestricted time weights consistently estimate the corresponding minimizers of the population covariance, as formally shown in the next section. Therefore, observing negative time weights carries useful information: it suggests the presence of aggregate shocks (such as linear trends) which require extrapolation.

3 Theoretical results for short panels

This section formalizes the theoretical properties of $\hat{\tau}(\hat{\boldsymbol{v}})$, focusing on the case of one treated period. The cases of multiple treated periods and staggered adoption are discussed in Appendix A.2. All proofs are in the appendix.

3.1 Baseline results under minimal assumptions on the factors

I first establish asymptotic normality of $\hat{\tau}(\hat{\boldsymbol{v}})$ around $\tau + b(\boldsymbol{v}^*)$ without restricting the factors. A corollary of this result is asymptotic efficiency of $\hat{\tau}(\hat{\boldsymbol{v}})$ in the class of time-weighted DiD estimators when parallel trends holds for all periods, in which case $b(\boldsymbol{v}^*) = 0$.

I begin by formalizing the properties of time-weighted DiD estimators with fixed weights.

LEMMA 1. *Let* $\boldsymbol{F}_a = [\boldsymbol{F}, \boldsymbol{\iota}]$, $\boldsymbol{\lambda}_{a,i} = (\boldsymbol{\lambda}_i', \beta_i)'$, $\boldsymbol{\mu}_{\lambda,a}^{(d)} = \mathrm{E}[\boldsymbol{\lambda}_{a,i} | D_i = d]$ *and* $\boldsymbol{\xi}_{\lambda,a} = \boldsymbol{\mu}_{\lambda,a}^{(1)} - \boldsymbol{\mu}_{\lambda,a}^{(0)}$. *Suppose Assumptions 1-5 and 7 hold. Then*

1. $\sqrt{N}(\hat{\boldsymbol{y}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_y)$ *as* $N \to \infty$ *with* $\boldsymbol{\mu} := \mathrm{E}[\hat{\boldsymbol{y}} | \boldsymbol{F}] = \boldsymbol{e}\tau + \boldsymbol{F}_a \boldsymbol{\xi}_{\lambda,a}$ *and limiting variance* $\boldsymbol{\Omega}_y = \boldsymbol{F}_a \boldsymbol{\Sigma}_{\lambda_a} \boldsymbol{F}_a' + \boldsymbol{\Sigma}_\varepsilon + \sigma_\tau^2 \boldsymbol{e}\boldsymbol{e}'$. *Moreover,* $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}_y$.

2. *For any matrix* $\boldsymbol{X}$ *with* rank $\boldsymbol{X} = r \leq T$ *and any positive definite matrix* $\boldsymbol{A}$ *with* rank $\boldsymbol{A} = T$ *it holds that* $\min_{\boldsymbol{v}:\ \boldsymbol{v}'\boldsymbol{X}=\boldsymbol{c}} \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} = \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}$ *at* $\boldsymbol{v}_{\min} = \boldsymbol{A}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}$ *for any* $\boldsymbol{c} \in \mathbb{R}^r$.

The first point ensures asymptotic normality of the mean differences $\hat{\boldsymbol{y}}$ with a limiting variance that can be consistently estimated. This is a natural starting point when modeling potential violations of parallel trends, see for example Rambachan and Roth (2023), and motivates the limiting regression problem in (2). An immediate consequence is asymptotic normality of $\hat{\tau}(\boldsymbol{v}) = \boldsymbol{v}'\hat{\boldsymbol{y}}$ for given weights $\boldsymbol{v}$:

$$\sqrt{N}(\hat{\tau}(\boldsymbol{v}) - \tau - b(\boldsymbol{v})) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{v}'\boldsymbol{\Omega}\boldsymbol{v} + \sigma_\tau^2)$$

where $b(\boldsymbol{v}) = \boldsymbol{v}'\boldsymbol{\mu} - \tau = \boldsymbol{v}'\boldsymbol{F}\boldsymbol{\xi}_\lambda$ is the first order bias term coming from correlations $\boldsymbol{v}'\boldsymbol{F}$ between the weights and the confounding factors over time; and correlations between the treatment D_i and the unobserved characteristics $\boldsymbol{\lambda}_i$ over the cross section, captured by $\boldsymbol{\xi}_\lambda$. The limiting variance is determined by the choice of $\boldsymbol{v}$ through the quadratic form of $\boldsymbol{\Omega} = \boldsymbol{F}\boldsymbol{\Sigma}_\lambda\boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon$, and treatment effect heterogeneity σ_τ^2 that is not affected by the choice of $\boldsymbol{v}$. Since $\boldsymbol{v}'\boldsymbol{\iota} = 0$, the estimator is (by construction) invariant to unobserved heterogeneity that is constant over time, including the unit fixed effects β_i.

The second point of Lemma 1 gives an explicit expression for the minimizer of a quadratic form under linear equality constraints. Applied to $\boldsymbol{X} = [\boldsymbol{e}, \boldsymbol{\iota}]$, $\boldsymbol{A} = \boldsymbol{\Omega}$ and $\boldsymbol{c} = (1, 0)'$, it yields the weights $\boldsymbol{v}^* = \boldsymbol{\Omega}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}$. They define the time-weighted DiD estimator $\hat{\tau}(\boldsymbol{v}^*)$ which has the smallest variance $\sigma_{\min}^2 = (\boldsymbol{v}^*)'\boldsymbol{\Omega}\boldsymbol{v}^* + \sigma_\tau^2$. Unbiasedness is immediate when $\boldsymbol{\xi}_\lambda = \boldsymbol{0}$. It follows that $\hat{\tau}(\boldsymbol{v}^*)$ is the efficient time-weighted DiD estimator of τ when parallel trends holds for all pre-treatment periods. In other cases, unbiasedness requires $\boldsymbol{F}'\boldsymbol{v}^* = \boldsymbol{0}$. Naturally, this is only the case under further assumptions on the factors, which are discussed in Section 3.2.

Consider now the properties of the feasible TWDID estimator $\hat{\tau}(\hat{\boldsymbol{v}})$. The estimated

weights $\hat{\boldsymbol{v}}$ obtained from $\widehat{\boldsymbol{\Omega}}$ converge in probability to $\boldsymbol{v}^*$, and therefore $\hat{\tau}(\hat{\boldsymbol{v}}) \xrightarrow{p} \tau + b(\boldsymbol{v}^*)$. However, the weight estimation uncertainty affects the limiting variance of $\hat{\tau}(\hat{\boldsymbol{v}})$, which becomes clear when decomposing

$$\hat{\tau}(\hat{\boldsymbol{v}}) = \tau + b(\boldsymbol{v}^*) + (\hat{\boldsymbol{v}} - \boldsymbol{v}^*)'\boldsymbol{F}\boldsymbol{\xi}_\lambda + \hat{\boldsymbol{u}}'\boldsymbol{v}^* + O_p(1/N)$$

with $\hat{\boldsymbol{u}} = \hat{\boldsymbol{y}} - \boldsymbol{\mu}$. The weight estimation uncertainty $\hat{\boldsymbol{v}} - \boldsymbol{v}^*$ affects the limiting distribution unless $\boldsymbol{\xi}_\lambda = \boldsymbol{0}$. This is in line with the properties of two-step estimators, discussed for example in Newey and McFadden (1994). Generally, estimation uncertainty in the first-step $\hat{\boldsymbol{v}} - \boldsymbol{v}^*$ affects the limiting distribution of the second step estimator $\hat{\tau}(\hat{\boldsymbol{v}})$ whenever the second step estimand $\tau + b(\boldsymbol{v})$ depends on the nuisance parameter $\boldsymbol{v}$. In this setup this is the case whenever $\boldsymbol{\xi}_\lambda \neq \boldsymbol{0}$.

Asymptotic normality of $\hat{\tau}(\hat{\boldsymbol{v}})$ can be established by writing $\hat{\boldsymbol{v}}$ and $\hat{\tau}(\hat{\boldsymbol{v}})$ as the solution the cross-sectional regression problem introduced in (4). This is done in the following theorem.

THEOREM 1. *Suppose Assumptions 1-5 and 7 hold. Let* $\boldsymbol{\theta} = (\tau, \boldsymbol{\nu}', \gamma)'$, $\boldsymbol{v} = \boldsymbol{v}_{\text{did}} - \boldsymbol{Q}\boldsymbol{\nu}$ *and* $\boldsymbol{z}_i = (\boldsymbol{y}_i', D_i)'$ *and* $\boldsymbol{w}_i = (D_i, \dot{\boldsymbol{y}}_{i,\text{pre}}, 1)'$. *Then* $\hat{\boldsymbol{\theta}} = (\hat{\tau}(\hat{\boldsymbol{v}}), \hat{\boldsymbol{\nu}}, \hat{\gamma})'$ *is the unique solution to* $\sum_i \boldsymbol{g}(\boldsymbol{z}_i; \boldsymbol{\theta}) = \boldsymbol{0}$ *with* $\boldsymbol{g}(\boldsymbol{z}_i; \boldsymbol{\theta}) = p_i^{-1}\boldsymbol{w}_i(\dot{y}_{i,\text{post}} - \boldsymbol{w}_i'\boldsymbol{\theta})$. *Moreover,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \boldsymbol{H}^{-1}\boldsymbol{S}\boldsymbol{H}^{-1}$$

with $\boldsymbol{\theta}^* = (\tau + b(\boldsymbol{v}^*), \boldsymbol{\nu}^*, \gamma^*)$, $\boldsymbol{H} = \lim \frac{1}{n}\sum_i E[\boldsymbol{w}_i\boldsymbol{w}_i'|\boldsymbol{F}]$ *and* $\boldsymbol{S} = \lim \frac{1}{n}\sum_i E[\boldsymbol{w}_i\boldsymbol{w}_i'(\dot{y}_{i,\text{post}} - \boldsymbol{w}_i'\boldsymbol{\theta}^*)^2|\boldsymbol{F}]$. *The sandwich estimator satisfies* $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{H}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}$.

We see that $\hat{\tau}(\hat{\boldsymbol{v}})$ is asymptotically normal around $\tau + b(\boldsymbol{v}^*)$. Appendix B.2 derives the limiting variance explicitly as

$$\sigma_{\hat{\tau}}^2 = \sigma_{\min}^2 + \boldsymbol{\xi}_\lambda'\boldsymbol{F}'\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}\boldsymbol{F}\boldsymbol{\xi}_\lambda + \boldsymbol{\xi}_\lambda'\boldsymbol{F}'\boldsymbol{\Sigma}_{\hat{\tau},\hat{\boldsymbol{v}}}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{v}}}$ is the limiting variance of $\hat{\boldsymbol{v}}$ and $\boldsymbol{\Sigma}_{\hat{\tau},\hat{\boldsymbol{v}}}$ the covariance between $\hat{\tau}(\boldsymbol{v}^*)$ and $\hat{\boldsymbol{v}}$. The take-away is that for $\boldsymbol{\xi}_\lambda \neq \boldsymbol{0}$, weight estimation uncertainty keeps the variance of $\hat{\tau}(\hat{\boldsymbol{v}})$

above the lower bound σ^2_{min} even asymptotically. A naive plug-in estimator $\hat{\boldsymbol{v}}'\widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{v}} \xrightarrow{p} \sigma^2_{\text{min}}$ therefore underestimates the variance of $\hat{\tau}(\hat{\boldsymbol{v}})$, and has to be adjusted for the weight estimation uncertainty. The last statement of Theorem 1 assures that the sandwich variance estimator $\widehat{\boldsymbol{\Sigma}}$ correctly captures all sources of uncertainty.

A corollary of Theorem 1 is asymptotic equivalence of $\hat{\tau}(\hat{\boldsymbol{v}})$ and $\hat{\tau}(\boldsymbol{v}^*)$ under 'small' violations of parallel trends, i.e. when $\sqrt{N}\boldsymbol{\xi}_\lambda \to \mathbf{0}$. Therefore, $\hat{\tau}(\hat{\boldsymbol{v}})$ is efficient in the class of time-weighted DiD estimators when parallel trends violations are asymptotically negligible.

COROLLARY 1. *Suppose Assumptions 1-5 and 7 hold. Then* $\sqrt{N}(\hat{\tau}(\hat{\boldsymbol{v}}) - \tau) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{min}})$ *as* $n \to \infty$ *with* $\sigma^2_{\text{min}} = \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}$ *the efficiency bound of time-weighted DiD estimators provided that* $\sqrt{N}\boldsymbol{\xi}_\lambda \to \mathbf{0}$.

Practically, this result promises efficiency gains in conventional event-study settings, and encourages applied researchers to use TWDID whenever they are confident that parallel trends holds across all periods. This finding is similar to results Marcus and Sant'Anna (2021), who study the optimal GMM estimator of the ATT under this assumption.

3.2 Conditions for bias reduction upon conventional DiD

I proceed with a formal discussion of the remaining bias $b(\boldsymbol{v}^*)$ that occurs when parallel trends does not hold with respect to all pre-treatment periods, i.e. $\boldsymbol{\xi}_\lambda \neq \mathbf{0}$. I derive an analytical expression of the bias $b(\boldsymbol{v}^*)$, which leads to sufficient conditions under which TWDID is guaranteed to have a smaller bias than conventional DiD.

Consider first a case in which TWDID is guaranteed to be asymptotically unbiased. Recall that the variance-minimizing weights $\boldsymbol{v}^*$ solve $\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{v}'(\boldsymbol{F}\boldsymbol{\Sigma}_\lambda\boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon)\boldsymbol{v}$. The idiosyncratic noise ε_{it} regularizes the time weights towards noise-minimizing weights $\boldsymbol{v}_\varepsilon = \arg\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{v}'\boldsymbol{\Sigma}_\varepsilon\boldsymbol{v} = \boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}$, which minimize the component of the variance

in $\hat{\tau}(\boldsymbol{v})$ attributable to idiosyncratic noise. Moreover, $\boldsymbol{v}_\varepsilon$ are the weights one would obtain in the unfavourable case in which the variance $\boldsymbol{\Omega}$ carries no information about the confounding factors. When $\boldsymbol{v}_\varepsilon$ is uncorrelated with $\boldsymbol{F}$, they coincide with the variance-minimizing weights $\boldsymbol{v}^*$. Then, TWDID is asymptotically unbiased because $b(\boldsymbol{v}^*) = \boldsymbol{v}_\varepsilon'\boldsymbol{F}\boldsymbol{\xi}_\lambda = \boldsymbol{0}$ irrespective of the loading imbalance $\boldsymbol{\xi}_\lambda$.

The condition $\boldsymbol{v}_\varepsilon'\boldsymbol{F} = \boldsymbol{0}$ imposes parallel trends with respect to the noise-minimizing weights $\boldsymbol{v}_\varepsilon$. It requires that the dynamics in the idiosyncratic errors are perfectly aligned with the factors. For example, white noise errors ($\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon^2\boldsymbol{I}$ and $\boldsymbol{v}_\varepsilon = \boldsymbol{v}_{\text{2wfe}}$) are perfectly aligned with factors satisfying $\boldsymbol{f}_T = \frac{1}{T_0}\sum_{t\leq T_0}\boldsymbol{f}_t$. In that case, TWDID is asymptotically unbiased, and Theorem 1 shows it is asymptotically equivalent to the 2wfe estimator except for additional uncertainty coming from the weight estimation.

In general, the variance-minimizing weights $\boldsymbol{v}^*$ reduce the correlation with $\boldsymbol{F}$ compared to noise-minimizing weights $\boldsymbol{v}_\varepsilon$. In case of one factor, this immediately implies that TWDID has a smaller asymptotic bias than $\hat{\tau}(\boldsymbol{v}_\varepsilon)$, i.e. $|b(\boldsymbol{v}^*)| < |b(\boldsymbol{v}_\varepsilon)|$. The following lemma formalizes this result in the general case based on an analytical expression of the bias $b(\boldsymbol{v}^*)$.

LEMMA 2. *Let* $\tilde{\boldsymbol{F}} = \boldsymbol{\Sigma}_\varepsilon^{-1/2}\boldsymbol{F}\boldsymbol{\Sigma}_\lambda^{1/2}$, $\tilde{\boldsymbol{X}} = \boldsymbol{\Sigma}_\varepsilon^{-1/2}\boldsymbol{X}$, $\tilde{\boldsymbol{v}}_\varepsilon = \boldsymbol{\Sigma}_\varepsilon^{1/2}\boldsymbol{v}_\varepsilon$ *and* $\tilde{\boldsymbol{\xi}}_\lambda = \boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{\xi}_\lambda$. *Suppose Assumptions 1-6 hold. Then* $b(\boldsymbol{v}^*) = \tilde{\boldsymbol{v}}_\varepsilon'\tilde{\boldsymbol{F}}(\boldsymbol{I}+\boldsymbol{S})^{-1}\tilde{\boldsymbol{\xi}}_\lambda$, *where* $\boldsymbol{S} = \tilde{\boldsymbol{F}}'\boldsymbol{M}_{\tilde{\boldsymbol{X}}}\tilde{\boldsymbol{F}}$ *has full rank* r *and* $\boldsymbol{M}_{\tilde{\boldsymbol{X}}} = \boldsymbol{I} - \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'$. *This implies* $b(\boldsymbol{v}^*) = 0$ *if* $\boldsymbol{v}_\varepsilon'\boldsymbol{F} = 0$ *and* $|b(\boldsymbol{v}^*)| < |b(\boldsymbol{v}_\varepsilon)|$ *if* $r = 1$.

This lemma formally shows how the bias of TWDID depends on the bias under noise-minimizing weights and the signal strength in the covariance matrix. This is best illustrated in case of one factor $\boldsymbol{f}$. The bias then simplifies to $b(\boldsymbol{v}^*) = b(\boldsymbol{v}_\varepsilon)/(1+s^2)$ with $s^2 = \tilde{\boldsymbol{f}}'\boldsymbol{M}_{\tilde{\boldsymbol{X}}}\tilde{\boldsymbol{f}}$ the signal strength in the covariance matrix and $b(\boldsymbol{v}_\varepsilon) = \boldsymbol{v}_\varepsilon'\boldsymbol{f}\xi_\lambda$ the bias under noise-minimizing weights $\boldsymbol{v}_\varepsilon$. Assumption 6 implies that $s^2 > 0$, so TWDID decreases the

magnitude of the bias relative to the estimator $\hat{\tau}(\boldsymbol{v}_\varepsilon)$ by a factor of $1+s^2$. This means that TWDID is guaranteed to have a smaller bias than the 2wfe estimator if the errors are i.i.d. over time; and a smaller bias than the canonical DiD estimator under random walk errors.

Aside specific cases of $\boldsymbol{\Sigma}_\varepsilon$, the bias is smaller than that of $\hat{\tau}(\boldsymbol{v}_{\text{did}})$ whenever $|b(\boldsymbol{v}_\varepsilon)|/|b(\boldsymbol{v}_{\text{did}})| < 1+s^2$. Clearly, TWDID is guaranteed to have a lower bias if DiD has a larger bias than $\hat{\tau}(\boldsymbol{v}_\varepsilon)$, even if there is no signal. When $|b(\boldsymbol{v}_\varepsilon)| > |b(\boldsymbol{v}_{\text{did}})| > 0$, the bias is smaller as long as the signal s^2 is sufficiently strong. TWDID therefore has larger bias only if the noise-minimizing weights point in a less favorable direction than $\boldsymbol{v}_{\text{did}}$ and the variance carries too little signal to push the weights towards reducing the bias.

3.3 Consistency under drifting factors

A key observation is that the bias $b(\boldsymbol{v}^*)$ vanishes when the factor strength increases. A natural way to formalize this is to show consistency of $\hat{\tau}(\hat{\boldsymbol{v}})$ when the factor strength increases with the sample size. I scale the factors in (1) by a signal-to-noise ratio parameter $\sigma > 0$, and consider asymptotic sequences along which $(N, \sigma^2) \to \infty$.[3]

The formal result of this section exploits that under Assumption 6 the factors are as good as observed from the covariance matrix $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{F}\boldsymbol{\Sigma}_\lambda \boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon$ when σ^2 is large. Using the Woodbury (1949) identity for matrix inversion, one can expand $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}_\varepsilon^{-1/2} \boldsymbol{M}_\sigma \boldsymbol{\Sigma}_\varepsilon^{-1/2}$, where the matrix $\boldsymbol{M}_\sigma = \boldsymbol{I} - \sigma^2 \tilde{\boldsymbol{F}}(\boldsymbol{I} + \sigma^2 \tilde{\boldsymbol{F}}'\tilde{\boldsymbol{F}})^{-1}\tilde{\boldsymbol{F}}'$ converges to the orthogonal projection on the column space of the (efficiently weighted) factors $\boldsymbol{M}_{\tilde{\boldsymbol{F}}} = \boldsymbol{I} - \tilde{\boldsymbol{F}}(\tilde{\boldsymbol{F}}'\tilde{\boldsymbol{F}})^{-1}\tilde{\boldsymbol{F}}'$. This happens "fast enough" to eliminate the correlation with the factors, in a sense that $\sigma \boldsymbol{M}_\sigma \boldsymbol{F} \to \boldsymbol{0}$.

In fact, the variance-minimizing weights $\boldsymbol{v}^*$ converge to the weights implied by the

[3] One can additionally impose weak loading imbalances $\boldsymbol{\xi}_\lambda \sigma = O(n^{-1/2})$. Then the asymptotic bias of fixed weights estimators $\sqrt{N} b(\boldsymbol{v})$ remains bounded, while TWDID is asymptotically unbiased $\sqrt{N} b(\boldsymbol{v}^*) \to 0$.

(infeasible) factor-augmented GLS estimator $\hat{\tau}(\boldsymbol{v}_0) = \boldsymbol{v}_0'\hat{\boldsymbol{y}}$. These are are given by

$$\boldsymbol{v}_0 = \boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{X}_a'(\boldsymbol{X}_a'\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{X}_a)^{-1}\boldsymbol{c}_a = \underset{\boldsymbol{v}\in\mathbb{V}_0}{\arg\min}\, \boldsymbol{v}'\boldsymbol{\Sigma}_\varepsilon\boldsymbol{v}$$

where $\boldsymbol{X}_a = [\boldsymbol{e}, \boldsymbol{\iota}, \boldsymbol{F}]$ and $\boldsymbol{c}_a = (1,0,0)'$. By construction, $\boldsymbol{v}_0$ is uncorrelated with the factors and thus eliminates the confounding factor structure without having to estimate the loadings explicitly. As a result, TWDID is consistent as $(n, \sigma^2) \to \infty$. This can also be seen from the bias expression of Lemma 2

$$b(\boldsymbol{v}^*) = \sigma\tilde{\boldsymbol{v}}_\varepsilon'\tilde{\boldsymbol{F}}(\boldsymbol{I} + \sigma^2\boldsymbol{S})^{-1}\tilde{\boldsymbol{\xi}}_\lambda = O\left(\sigma/(1+\sigma^2)\right),$$

showing that the bias diminishes for diverging factors.

Another observation is that the bias vanishes under weak factors, that is, when $\sigma^2 \to 0$. In that case, the factor structure disappears from the mean of $\hat{\boldsymbol{y}}$ and from the covariance matrix $\boldsymbol{\Omega} \to \boldsymbol{\Sigma}_\varepsilon$. The weights $\boldsymbol{v}^*$ converge to the noise-minimizing weights $\boldsymbol{v}_\varepsilon$, thus remaining well-defined. Therefore, the TWDID estimator consistently estimates τ under weak factors. While this is the case for any other time-weighted DiD estimator, TWDID achieves the smallest variance within this class. Corollary 1 reaches the same conclusion for a weak loading imbalance $\sqrt{N}\boldsymbol{\xi}_\lambda \to \boldsymbol{0}$, the difference being that then the factors still appear in the covariance matrix.

These results are summarized in the following theorem, which is the counterpart of Theorem 1 for drifting factor asymptotics.

THEOREM 2 (Drifting factor asymptotics). *Suppose Assumptions 1-7 hold. Then*

1. *(Diverging factors) If* $(N, \sigma^2) \to \infty$, *then* $\boldsymbol{v}^* \to \boldsymbol{v}_0$, $b(\boldsymbol{v}^*) \to 0$, *and* $\hat{\tau}(\hat{\boldsymbol{v}}) \xrightarrow{p} \tau$.

2. *(Weak factors) If* $N \to \infty$ *and* $\sigma^2 \to 0$, *then* $\boldsymbol{v}^* \to \boldsymbol{v}_\varepsilon$ *and* $\hat{\tau}(\hat{\boldsymbol{v}}) \xrightarrow{p} \tau$. *Moreover,* $\sqrt{N}(\hat{\tau}(\hat{\boldsymbol{v}}) - \tau) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{v}_\varepsilon'\boldsymbol{\Sigma}_\varepsilon\boldsymbol{v}_\varepsilon + \sigma_\tau^2)$ *if the rates satisfy* $N\sigma^2 \to 0$.

Similar consistency results for diverging factors have been established for synthetic control estimators in large T panels, for example in Ferman and Pinto (2021). In these

setups, asymptotically unbiased estimates also requires that the signal dominates the noise in the limit. This is typically achieved by assuming non-stationary factors while restricting the time series dependence of the idiosyncratic errors.

4 Monte Carlo Experiments

4.1 Setup

In each replication, I draw data from

$$y_{it} = \sigma_\lambda \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \varepsilon_{it}, \qquad \varepsilon_{it} = \rho \varepsilon_{i,t-1} + \sqrt{1-\rho^2}\eta_{it} \tag{6}$$

with $\lambda_{ij} = \xi_\lambda D_i + \nu_{ij}$ and $\{\nu_{ij}, \eta_{it}\}_{i,j,t}$ i.i.d. draws from the standard normal distribution. This incorporates two important parameters that I vary across different simulations. First, I vary the autocorrelation parameter $\rho \in \{0, 0.5\}$ to study the effect of persistency in the error term (while keeping $\text{var}[\varepsilon_{it}] = 1$). Second, I vary the factor strength σ_λ, including settings without factors ($\sigma_\lambda = 0$), weak factors ($\sigma_\lambda = O(1/\sqrt{N})$), and strong factors ($\sigma_\lambda = O(1)$). The loading imbalance is fixed at $\boldsymbol{\xi}_\lambda = 0.1$. That way, the factors strength and the bias are of the same order, proportional to σ_λ.

I consider up to two factors $\boldsymbol{f}_t = (f_{1t}, f_{2t})'$, both of which are being fixed across all simulation settings. The first factor comes from one draw of a persistent AR(1) process. The second factor is a deterministic, linear trend. Both factors are rescaled to have unit variance ($\sum_t f_{jt}^2 = 1$) and are plotted in Figure 1. In settings with only one factor, I set $\lambda_{i,2} = 0$.

I compare four estimation strategies. The first is the TWDID estimator, which imposes only the constraint that the time weights sum to one, allowing them to take negative values. The second is a restricted version of TWDID (denoted TWDID+), which in addition requires the time weights to be non-negative. For comparison with conventional approaches,

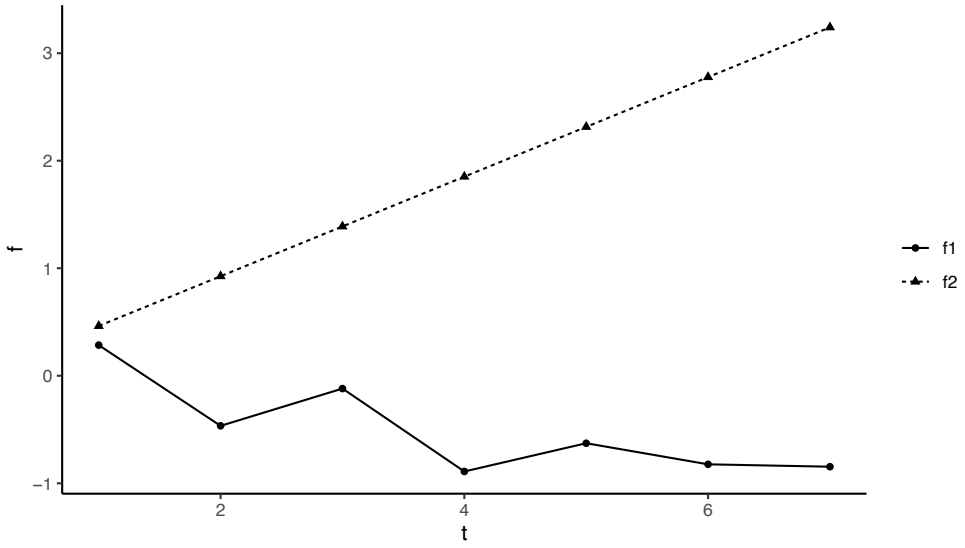

Figure 1: Factors used in the Monte Carlo simulations. The first factor is a realization of a gaussian AR(1) process with autocorrelation 0.8. The second factor is a linear trend.

I also consider the two-way fixed effects (2wfe) estimator, which assigns equal weight to all pre-treatment periods. Finally, I include the synthetic difference-in-differences (SDID) estimator, which combines non-negative time weights that sum to one with unit weights estimated from the pre-treatment periods as defined in (5).

4.2 Results

Figure 2 shows desirable properties of the time-weighting approach compared to not weighting or unit weighting. As expected, TWDID performs better than 2wfe in all settings, both in terms of bias and RMSE. Consider first the case of one factor (top rows). Here the post-treatment factor is smaller than the average pre-treatment factors (Figure 1). Because the factor affects treated units more than untreated units ($\boldsymbol{\xi}_\lambda > 0$), 2wfe has a negative bias. The magnitude of the bias is proportional to the factor strength. TWDID successfully reduces the bias and RMSE independently of the factor strength. Even in absence of factors ($\sigma_\lambda = 0$), TWDID improves upon the unbiased DiD when there is persistency in the errors due to its efficiency properties. SDID generally requires a stronger signal before it

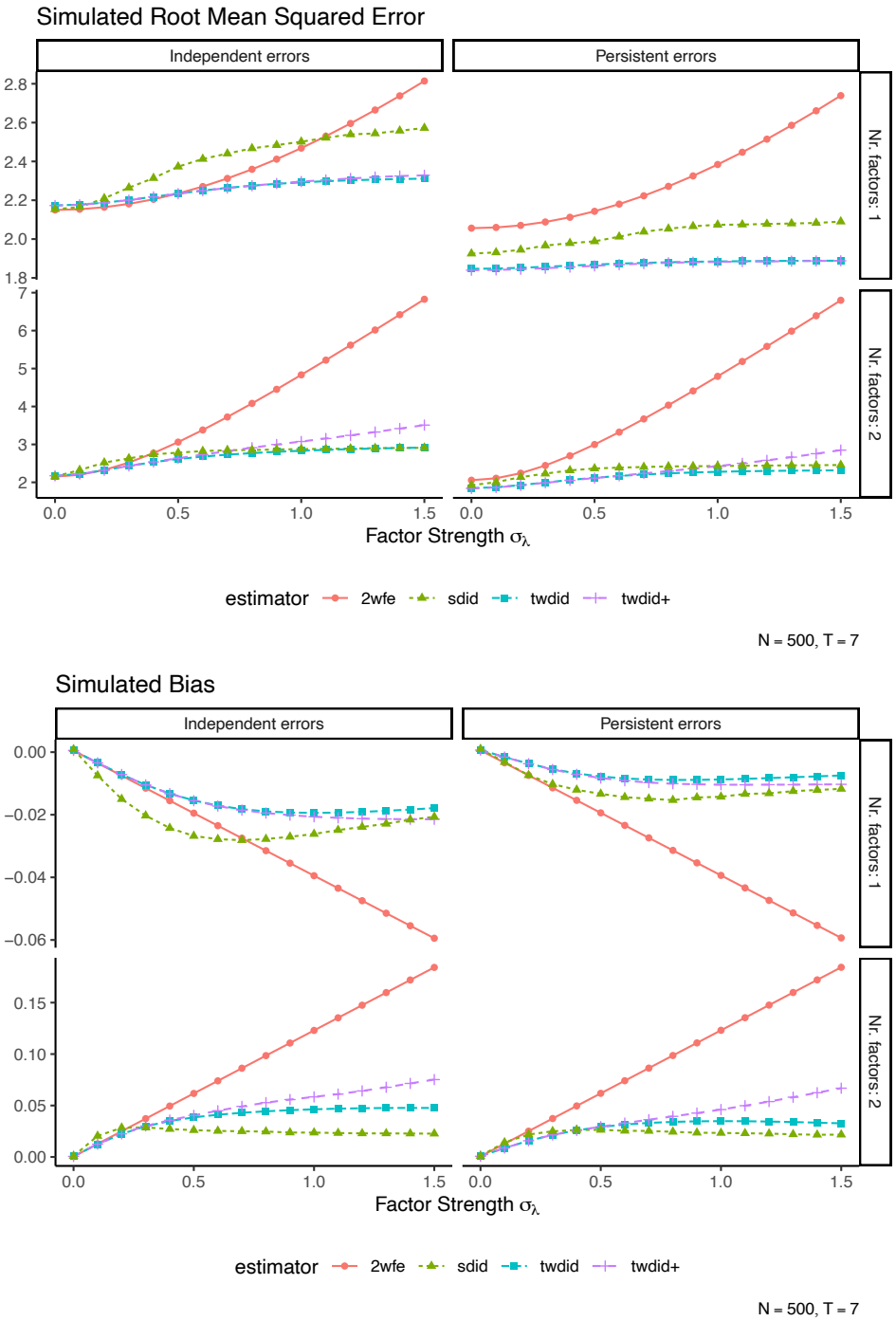

Figure 2: Simulated RMSE (top panel) and bias (bottom panel) of four estimators: two-way fixed effects (2wfe), restricted time-weighted DID (TWDID+), TWDID, and synthetic DID (SDID). The horizontal axis depicts different levels of the factor strength σ_λ. The data is generated by (6) with fixed factors.

improves upon 2wfe. Moreover, it can exacerbate bias and RMSE compared to TWDID and 2wfe when factors are weak, as the results in the single factor case suggest. With two factors, however, the comparison is more mixed. While TWDID and SDID exhibit comparable RMSEs, SDID has a lower bias. This suggests that unit weights can be successful in reducing bias further even in short T.

Overall, the simulations indicate that TWDID can reduce bias and RMSE relative to 2wfe across different levels of factor strength and error persistence. These findings highlight the advantages of TWDID in short panels. However, they should be interpreted as illustrative for the considered factor dynamics rather than taken as universal across all possible factor structures. Appendix C contains additional Monte Carlo Experiments showing desirable coverage rates, and has settings in which the factors are redrawn from a gaussian process.

5 TWDID in practice: the effect of the NOx Budget Trading Program

I revisit Deschenes et al. (2017) studying the effect of the NOx Budget Trading Program (NBP) 2003-2008 on NOx emissions. It entailed a cap and trade program to reduce NOx emissions from power plants. It was only active in the summer months May - September in the years 2003-2008 in 19 states in the US. In 2003 the program was active only in a subset of the 19 treated states. States not adjacent to the NBP states remain as non-treated states (22 in total).

Data on NOx emissions is available on county level for $N = 2539$ counties from 1997-2007. We observe $N_1 = 1,354$ counties in the treated states and $N_0 = 1,185$ in the untreated states. Per county and year we observe data for the seasons summer and winter, where summer is defined as May - September.

Specification. Consider the interactive fixed effect model

$$\tilde{y}_{ist} = \sum_{j=2004}^{2008} \tau_j^{\text{att}} D_{ist}(j) + \mu_{it} + \nu_{is} + \boldsymbol{\lambda}_i' \tilde{\boldsymbol{f}}_{st} + \tilde{\varepsilon}_{ist}$$

with $D_{ist}(j) = \mathrm{I}(i \in \mathcal{N}_1, t = j, s = 1)$ a post-treatment dummy of year j indicating whether NBP is operating in county i in season $s = 0, 1$ (winter, summer). μ_{it}, ν_{is} are county-year and county-season fixed effects, respectively. $\tilde{\boldsymbol{f}}_{st}$ are season-year specific common shocks that affect the emissions of county i with intensity $\boldsymbol{\lambda}_i$. $\tilde{\varepsilon}_{ist}$ is an idiosyncratic error term. The special case $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ resembles the additive fixed effect model that Deschenes et al. (2017) use. In that case the factor structure reduces to a season-year fixed effect.

As a preliminary step I eliminate county-year fixed effects μ_{it} by considering the difference between summer and winter observations

$$y_{it} := \tilde{y}_{i1t} - \tilde{y}_{i0t} = \sum_{j=2004}^{2008} \tau_j^{\text{att}} D_{it}(j) + \beta_i + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \varepsilon_{it}$$

with $\beta_i = \nu_{i1} - \nu_{i0}$, $\boldsymbol{f}_t = \tilde{\boldsymbol{f}}_{1t} - \tilde{\boldsymbol{f}}_{0t}$ and $\varepsilon_{it} = \tilde{\varepsilon}_{i1t} - \tilde{\varepsilon}_{i0t}$. The application matches the setting of this paper under the assumption that the program does not affect emissions in the winter months in the treated years.

Evidence against parallel trends. I first obtain evidence against $\boldsymbol{\xi}_\lambda = \mathbf{0}$ by considering how the difference in average NOx emissions $\hat{y}_t = \bar{y}_t^{(1)} - \bar{y}_t^{(0)}$ has evolved prior to the intervention. We can write

$$\hat{y}_t = \bar{\beta}^{(1)} - \bar{\beta}^{(0)} + \boldsymbol{\xi}_\lambda' \boldsymbol{f}_t + O_p(\frac{1}{\sqrt{N}}), \quad t \leq T_0$$

so $\hat{y}_t$ should be constant prior to the treatment when parallel trends holds in all periods ($\boldsymbol{\xi}_\lambda = \mathbf{0}$). However, Figure 3 does show variation of $\hat{y}_t$ in periods $t \leq T_0$. Also a formal test based on $J_{\text{pre}} = N \dot{\boldsymbol{y}}_{\text{pre}}' \widehat{\boldsymbol{\Omega}}_{\text{pre}}^{-1} \dot{\boldsymbol{y}}_{\text{pre}}$ with $\widehat{\boldsymbol{\Omega}}_{\text{pre}}/N = \widehat{\text{var}}[\dot{\boldsymbol{y}}_{\text{pre}}]$ and $\dot{\boldsymbol{y}}_{\text{pre}} = (\hat{y}_1 - \hat{y}_{T_0}, \ldots, \hat{y}_{T_0-1} - \hat{y}_{T_0})'$ rejects the null of parallel trends in all pre-treatment periods at conventional levels.

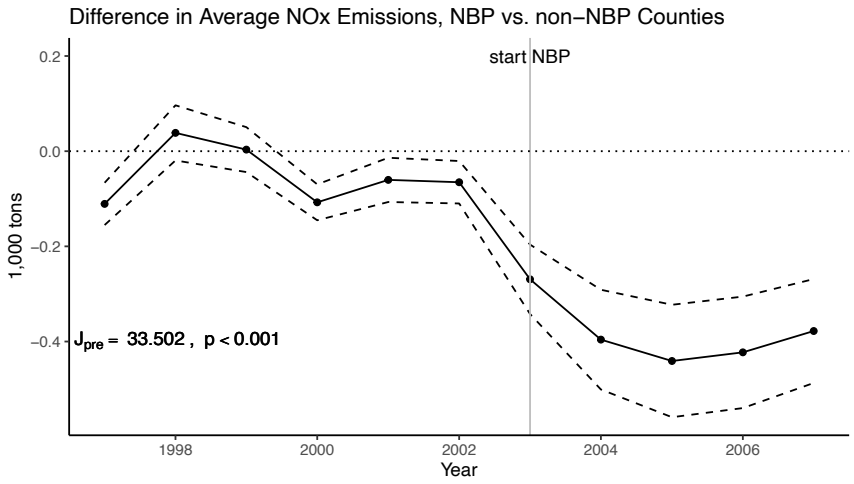

Figure 3: Difference in average NOx emissions $\bar{y}_t^{(1)} - \bar{y}_t^{(0)}$ over time, with 95% confidence band (dashed), J_{pre} the statistic testing parallel trends in all pre-treatment periods, distributed as χ^2_5 under the null.

Results. I estimate the dynamic effects of the intervention, applying the TWDID approach sequentially for each post-treatment period $j = 2004, \ldots, 2008$. For comparison I also computed the canonical DID estimator (all weight on the most recent pre-treatment period 2002) and the 2wfe estimator, which uses equal time weights. I omit the year 2003 from the analysis because not all treated states had fully implemented the program by then. For a given post-treatment period $T_1 \in \{T_0 + 1, \ldots, T\}$, I implement all estimators using the regression interpretation presented in (4). The corresponding regression equation is

$$y_{i,T_1} - y_{i,T_0} = \gamma + \tau D_i + \dot{\boldsymbol{y}}'_{i,\text{pre}}\boldsymbol{\nu} + u_i$$

with $\dot{\boldsymbol{y}}_{i,\text{pre}} = (y_{i,1} - y_{i,T_0}, \ldots, y_{i,T_0-1} - y_{i,T_0})'$ denoting the unit-specific pre-trends. In this specification, the least-squares estimate of τ is a time-weighted DiD estimate $\hat{\tau}(\boldsymbol{v}) = \boldsymbol{v}'\hat{\boldsymbol{y}}$ where the weights $\boldsymbol{v} = (-\boldsymbol{\nu}, \boldsymbol{\iota}'\boldsymbol{\nu} - 1, 1)'$ are determined by the coefficients on the pre-trends $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{T_0-1})'$. The DiD estimator corresponds to the least-square estimate of τ when restricting $\boldsymbol{\nu} = 0$, i.e. omitting the pre-trends $\dot{\boldsymbol{y}}_{i,\text{pre}}$ from the regression. Similarly, the 2wfe estimator results from restricting $\nu_t = 1/T_0$ for all t. The TWDID estimator $\hat{\tau}(\hat{\boldsymbol{v}})$ corresponds to the unrestricted regression, i.e. when $\dot{\boldsymbol{y}}_{i,\text{pre}}$ is controlled for, and observations

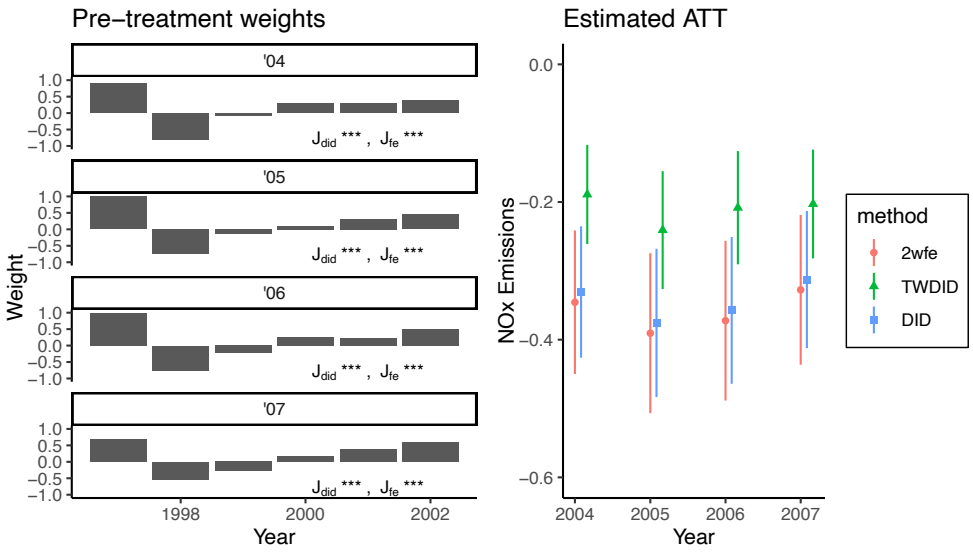

Figure 4: Left: Estimated time weights for each post-treatment period as of (7). Right: Resulting estimates of $\boldsymbol{\tau}^{\text{att}}$ and confidence intervals for both TWDID, 2wfe and DiD estimation. J^{***}_{did} and J^{***}_{2wfe} indicate that the Wald tests for $v^* = v_{\text{did}}$ and $v^* = v_{\text{2wfe}}$ reject at the 1% level, respectively.

are weighted by their inverse propensity score.

The left panel of Figure 4 shows the estimated pre-treatment weights $\hat{\boldsymbol{v}}_{\text{pre}}$ used by the TWDID estimator for a given post treatment period, while the right panel of Figure 4 shows the resulting dynamic treatment effect estimates and their 95% confidence intervals. Across all post-treatment periods, the weights are significantly different from both v_{did} and v_{2wfe} as judged by the Wald statistics testing the corresponding restrictions on ν. Because $\hat{y}_t$ significantly varies in the pre-treatment periods, the ATT estimate is sensitive to the choice of weights. While DiD and 2wfe estimation suggest effects of similar magnitude, TWDID estimation suggests, in absolute terms, smaller effect sizes. Since the weights are chosen to minimize the variance of the estimator, TWDID estimates are clearly more precise DID and 2wfe and the resulting confidence intervals are narrower.

These results can be explained by confounding aggregate shocks which affect NOx emissions differently across counties. Estimators using fixed weights, such as DID and 2wfe, are therefore sensitive to variations in the shocks before and after the start of the

program. In contrast, TWDID reduces the influence of the confounding factors on the ATT estimate under the assumption that the persistency in emissions over time is informative about those shocks. In this case, the negative pre-treatment weights suggest the presence of aggregate shocks that cannot fully be accounted for by non-negative weights used by 2wfe and DiD. Accounting for those shocks, TWDID extrapolates part of the pre-treatment decrease the difference of emissions to the post-treatment periods. The method therefore attributes smaller share of the observed decrease in emissions to the program itself, leading to lower point estimates.

6 Conclusion

This paper introduces a time-weighted difference-in-differences (TWDID) estimator for settings with few pre-treatment periods. Unlike conventional estimators, which use fixed pre-treatment weights, TWDID assigns variance-minimizing weights determined by the within-group covariance matrix of outcomes. The proposed estimator is efficient in the considered class when parallel trends hold across all periods. I introduce violations of parallel trends through common factors that have heterogeneous effects on the outcome. I show that the weights reduce the influence of the confounding factors, yielding a smaller bias than conventional DiD estimators under mild assumptions on the factors. Revisiting the impact of a cap-and-trade program on NOx emissions, TWDID yields smaller and more precise estimates than conventional approaches.

SUPPLEMENTARY MATERIAL

Online Appendix Additional results, Proof of Theorems and Lemmas

References

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, 105, 493–505.

——— (2015): "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59, 495–510.

Abadie, A. and J. L'hour (2020): "A penalized synthetic control estimator for disaggregated data," *Working Paper*.

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013): "Panel data models with multiple time-varying individual effects," *Journal of Econometrics*, 174, 1–14.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021): "Synthetic difference-in-differences," *American Economic Review*, 111, 4088–4118.

Bai, J. (2009): "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Ben-Michael, E., A. Feller, and J. Rothstein (2021): "The Augmented Synthetic Control Method," *Journal of the American Statistical Association*, 116, 1789–1803.

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting event-study designs: robust and efficient estimation," *Review of Economic Studies*, 91, 3253–3285.

Brown, N. and K. Butts (2022): "A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models," Tech. rep., Working Paper.

Callaway, B. and S. Karami (2022): "Treatment effects in interactive fixed effects models with a small number of time periods," *Journal of Econometrics*.

Callaway, B. and P. H. C. Sant'Anna (2020): "Difference-in-Differences with multiple time periods," *Journal of Econometrics*.

Chan, M. K. and S. S. Kwok (2021): "The PCDID Approach: Difference-in-Differences When Trends Are Potentially Unparallel and Stochastic," *Journal of Business & Economic Statistics*, 0, 1–18.

Chen, X., P. H. Sant'Anna, and H. Xie (2025): "Efficient Difference-in-Differences and Event Study Estimators," *arXiv preprint arXiv:2506.17729*.

Deschenes, O., M. Greenstone, and J. S. Shapiro (2017): "Defensive investments and the demand for air quality: Evidence from the NOx budget program," *American Economic Review*, 107, 2958–89.

Ferman, B. (2021): "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls," *Journal of the American Statistical Association*, 116, 1764–1772.

Ferman, B. and C. Pinto (2021): "Synthetic controls with imperfect pretreatment fit," *Quantitative Economics*, 12, 1197–1221.

Gobillon, L. and T. Magnac (2016): "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, 98, 535–551.

Harmon, N. A. (2024): "Difference-in-Differences and Efficient Estimation of Treatment Effects," *Working paper*.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5–86.

Ketz, P. (2018): "Subvector inference when the true parameter vector may be near or at the boundary," *Journal of Econometrics*, 207, 285–306.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016): "Exact post-selection inference, with application to the lasso," *The Annals of Statistics*, 44, 907–927.

Manski, C. F. and J. V. Pepper (2018): "How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions," *Review of Economics and Statistics*, 100, 232–244.

Marcus, M. and P. H. Sant'Anna (2021): "The role of parallel trends in event study settings: An application to environmental economics," *Journal of the Association of Environmental and Resource Economists*, 8, 235–275.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74, 967–1012.

Rambachan, A. and J. Roth (2023): "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 90, 2555–2591.

Roth, J. and P. H. Sant'Anna (2023): "Efficient estimation for staggered rollout designs," *Journal of Political Economy Microeconomics*, 1, 669–709.

Steffens, C. and J. Stuhle (2025): "How to control for (pre-)trends?" Tech. rep., Working Paper.

Woodbury, M. A. (1949): "The stability of out-input matrices," *Chicago, IL*, 9, 3–8.

Xu, Y. (2017): "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25, 57–76.