Time-Weighted Difference-in-Differences: Accounting for Common Factors in Short T Panels

Timo Schenk*

February 15, 2022

Preliminary work
please do not cite or redistribute without permission

Abstract

I propose a time-weighted difference-in-differences (TWDID) estimation approach that is robust against time-varying common factors in short T panels. Time weighting substantially reduces both bias and variance compared to an unweighted DID estimator through balancing the pre-treatment and post-treatment factors. To conduct valid inference on the average treatment effect, I develop a correction term that adjusts conventional standard errors for the presence of weight estimation uncertainty. Revisiting a study on the effect of a cap-and-trade program on NOx emissions, TWDID estimation reduces the standard errors of the estimated treatment effect by 10% compared to a conventional DID approach.

Keywords: Synthetic Difference-in-differences, Interactive Fixed Effects, Causal Inference, Panel Data.

1 Introduction

The presence of interactive fixed effects, for example due to time-varying common factors, leads to biased difference-in-difference (DID) estimates. 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 suggest using a time-weighted DID (TWDID) estimator. Factor imbalances between pre-treatment and post-treatment periods cause the DID estimator to be biased. This bias can be eliminated using time

*Department of Economics, University of Amsterdam (t.d.schenk@uva.nl). I am grateful for the extensive feedback from my advisors Frank Kleibergen and Andreas Pick as well as helpful discussions during the UvA econometrics seminars.

weights which give a higher weight to pre-treatment periods that are more similar to the post-treatment periods. In theory, a complete elimination of the bias requires oracle weights that perfectly balance the factors. In practice, weights that are estimated from the control unit data succeed in reducing the bias substantially.

A second effect of the factor imbalance is that it amplifies the variance of the DID estimator. By balancing the factors, time weighting reduces the variance and leads to more accurate estimates. In fact, when the number of units is large compared to the number of periods, the estimated weights converge to pseudo-true weights which minimize the variance of the estimated treatment effect. Simulations show that the amount by which the variance is reduced outweighs the additional variance caused by the weight estimation uncertainty.

As a consequence of the bias, inference based on DID estimation will be substantially oversized. The TWDID estimator reduces this problem through bias correction. However, the presence of estimated weights still leads to empirical size in excess of the nominal size when using standard covariance estimators. I propose a two-step standard error procedure that eliminates the remaining size distortions of TWDID inference. First, the weighted cluster-covariance matrix estimator (Arellano, 1987) can be used to estimate the variance of the estimated treatment effect under pseudo-true weights. For the second step, I develop a correction term that accounts for the presence of estimated weights. It uses the fact that in short T the time weights are asymptotically normal around the pseudo true weights.

Revisiting a study by Deschenes, Greenstone, and Shapiro (2017) on the effect of a cap-and-trade program on NOx emissions, I compare the practical differences between DID and TWDID estimation in a short T panel. I find evidence of time-varying common factors in the pre-treatment data. While the point estimates only differ to a small degree, TWDID reduces the standard errors of the estimated average treatment effect by 10% compared to a conventional DID approach.

The TWDID estimator is a restricted version of the Arkhangelsky et al. (2021) synthetic DID (SDID) estimator. Next to time weights, the SDID estimator also uses synthetic control unit weights to address the influence of common factors. Consistency results with estimated unit weights require a larger number of pre-treatment periods (Abadie, Diamond, and Hainmueller, 2010; Ferman, 2019; Abadie and L'hour, 2020). Therefore, several challenges occur when applying the SDID estimator in short T panels. First, the unit weight estimation causes additional variation which might outweigh any additional balancing gains. Second, it is unclear how to conduct inference in presence of estimated unit weights. Arkhangelsky et al. (2021) propose a jackknife based variance estimator which leads to conservative standard errors. However, this result requires that the weight estimation noise is negligible, which only holds for large T.

Other approaches such as Chan and Kwok (2021) and Gobillon and Magnac (2016) rely on estimating the factor structure by methods of principal components, which also requires large T. Avoiding the estimation of the whole factor structure, Pesaran (2006) finds cross-sectional averages to sufficiently proxy the factors in each period. In absence of covariates, this approach could account for one common factor when T is small. When evaluating a one-time policy intervention, however, it is not required to control for the factors in each period. Instead, it suffices to balance the differences between the average pre-treatment and post-treatment factors.

The remainder of the paper is structured as follows. Section 2 covers the Theory. Section 2.1 introduces the interactive fixed effects model and defines the TWDID estimator. Section 2.2 shows the bias and variance reduction properties. Section 2.3 covers inference. Section 3 illustrates the theoretical results with simulations. Section 4 compares TWDID and DID estimation in the study of Deschenes et al. (2017) on NOx emissions.

2 Theory

2.1 Setting

Consider a policy intervention starting in period $t = T_0$ which affects units $i = N_0+1, \ldots, N$ that are part of a large sample $i = 1, \ldots, N_0, N_0+1, \ldots, N$ observed over a small number of periods $t = 1, \ldots, T_0, T_0+1, \ldots, T$. We seek to estimate the average treatment effect in the post-treatment periods $t = T_0 + 1, \ldots, T$ on the outcome y. Let $D_i = \mathrm{I}(i > N_0)$ indicate whether unit i is ever treated and let $\mathbb{N}_j = \{i\colon D_i = j\}$, $j = 0, 1$ be the sets of untreated and treated units, respectively. Let $\kappa = \frac{N_0}{N}$ denote the share of untreated units and $\pi = \frac{T_0}{T}$ the share of pre-treatment periods.

We use the interactive fixed model (Bai, 2009)

$$y_{it} = \beta_i + \tau D_{it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \varepsilon_{it} \tag{1}$$

where $D_{it} = \mathrm{I}(i > N_0, t > T_0)$ is a dummy that is one for treated units in periods after the treatment started and zero otherwise, τ is the homogeneous treatment effect, β_i unit 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.

Unobserved factor structures $\boldsymbol{\lambda}_i' \boldsymbol{f}_t$ are present in many economic variables. Bai (2009) discusses, among others, the following examples. In microeconomics, $\boldsymbol{\lambda}_i$ 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 measured by $\boldsymbol{f}_t$ on the outcome y_{it}. In macroeconomics, the factors $\boldsymbol{f}_t$ are common shocks (e.g. technology or weather shocks) that have an heterogeneous impact $\boldsymbol{\lambda}_i$ on unit i.

I make the following assumptions.

ASSUMPTION 1 (Correlated loadings). $\mathrm{E}[\boldsymbol{\lambda}_i|D_i=1]-\mathrm{E}[\boldsymbol{\lambda}_i|D_i=0]=\boldsymbol{\xi}_\lambda<\infty$ *and* $\mathrm{var}[\boldsymbol{\lambda}_i]=\boldsymbol{\Sigma}_{\lambda,i}$ *with* $\lim_{n\to\infty}\frac{1}{N_j}\sum_{i\in\mathbb{N}_j}\boldsymbol{\Sigma}_{\lambda,i}=\boldsymbol{\Sigma}_\lambda^{(j)}$ *for* $j=0,1$, *both positive definite* $r\times r$ *matrices.*

ASSUMPTION 2 (Strict exogeneity). *For every* i, $\mathrm{E}[\boldsymbol{\varepsilon}_i|\beta_i,D_i,\boldsymbol{\lambda}_i,\boldsymbol{F}]=\mathbf{0}$. *Moreover,* $\mathrm{E}[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|D_i]=\boldsymbol{\Sigma}_{\varepsilon,i}$ *with* $\lim_{n\to\infty}\frac{1}{N_j}\sum_{i\in\mathbb{N}_j}\boldsymbol{\Sigma}_{\varepsilon,i}=\boldsymbol{\Sigma}_\varepsilon^{(j)}$ *for* $j=0,1$. *Positive definite* $T\times T$ *matrices.*

ASSUMPTION 3 (Random sampling). $(\boldsymbol{\varepsilon}_i,\boldsymbol{\lambda}_i)$ *are independent over the cross section.* $\kappa,\pi\in(0,1)$ *are both constant as* $N\to\infty$ *and* $T_0\geq 2$.

Assumption 1 is the central characteristic of the model. It allows the loadings $\boldsymbol{\lambda}_i$ to differ systematically between treated and untreated units. The loading imbalance $\boldsymbol{\xi}_\lambda$ measures how much more the treated units are on average affected by the common factors $\boldsymbol{f}_t$. It also nests the two-way fixed effects model as a special case for $\xi_\lambda=0$ and $\boldsymbol{\Sigma}_\lambda^{(1)}=\boldsymbol{\Sigma}_\lambda^{(0)}=0$, since then $\boldsymbol{\lambda}_i=\boldsymbol{\lambda}$ for all i. In that case the factor structure $\boldsymbol{\lambda}_i'\boldsymbol{f}_t$ reduces to a time fixed effect $\gamma_t=\boldsymbol{\lambda}'\boldsymbol{f}_t$.

Under Assumption 2, the treatment assignment is strictly exogeneous once conditioned on the loadings, fixed effects and the factors. Moreover, I allow for heteroskedasticity and arbitrary serial dependence of the idiosyncratic errors. The common factors $\boldsymbol{F}=(\boldsymbol{f}_1,\dots\boldsymbol{f}_T)'$ $(T\times r)$ are realizations of an arbitrary deterministic or stochastic process. The number of factors r is unknown and fixed. Without loss of generality I assume that $\boldsymbol{F}'\iota=\mathbf{0}$. All results hold conditional on $\boldsymbol{F}$.

Assumption 3 imposes independence of the error component over the cross section. It requires the number of treated and untreated units to grow at the same rate. We need at least two pre-treatment periods.

I propose a restricted version of the Arkhangelsky et al. (2021) synthetic DID estimator which uses only time weights and refer to it as the time-weighted DID estimator. It is computed in two steps.

1. Obtain a T_0 vector of time weights $\hat{\boldsymbol{v}}=(\hat{v}_1,\dots,\hat{v}_{T_0})'$ using the demeaned outcomes $\dot{y}_{i,t}=y_{i,t}-\frac{1}{N_0}\sum_{i=1}^{N_0}y_{i,t}$ of the untreated units only. Regress the average post-treatment outcome $\bar{\dot{y}}_{i,[1]}$ on the pre-treatment outcomes $\dot{\boldsymbol{y}}_{i,[0]}=(\dot{y}_{i,1},\dots,\dot{y}_{i,T_0})'$

$$\hat{\boldsymbol{v}}=\underset{\boldsymbol{v}\in\mathbb{V}}{\arg\min}\sum_{i=1}^{N_0}(\bar{\dot{y}}_{i,[1]}-\dot{\boldsymbol{y}}_{i,[0]}'\boldsymbol{v})^2 \tag{2}$$

with $\mathbb{V}=\{\boldsymbol{v}\in\mathbb{R}^{T_0}\colon v_t\geq 0,\ \sum_{t=1}^{T_0}v_t=1\}$ the set of non-negative weights that sum to one.

2. Obtain the time-weighted DID estimator $\hat{\tau}(\hat{\boldsymbol{v}})$ as solution to the weighted two-way fixed effect regression

$$\min_{\tau,\boldsymbol{\mu},\boldsymbol{\gamma}} \sum_{i=1}^{N}\sum_{t=1}^{T} v_t(y_{it} - \tau D_{it} - \mu_i - \gamma_t)^2 \tag{3}$$

The resulting estimator is

$$\hat{\tau}(\boldsymbol{v}) = \bar{\Delta}_{[1]} - \sum_{t=1}^{T_0} v_t \Delta_t \tag{4}$$

with $\Delta_t = \bar{y}_t^{(1)} - \bar{y}_t^{(0)}$, $\bar{\Delta}_{[1]} = \frac{1}{T-T_0}\sum_{t>T_0}\Delta_t$, $\bar{y}_t^{(1)} = \frac{1}{N_1}\sum_{i=N_0+1}^{N} y_{it}$ and $\bar{y}_t^{(0)} = \frac{1}{N_0}\sum_{i=1}^{N_0} y_{it}$ the treated and untreated units' average outcome, respectively.

The DID estimator is the special case of the TWDID estimator with equal weights $\bar{\boldsymbol{v}} = \frac{\iota_{T_0}}{T_0}$. The synthetic DID estimator of Arkhangelsky et al. (2021) uses both unit weights and time weights in product form $\omega_i v_t$ in (3). The treated units are weighted equally and the weights of the untreated units are non-negative and sum to one. The estimator has the form of $\hat{\tau}(\boldsymbol{v})$ with simple averages over the control units $\bar{y}_t^{(0)}$ replaced by weighted averages $\sum_{i\in\mathbb{N}_0}\omega_i y_{it}$. The control unit weights are estimated from the pre-treatment data in the same manner as the pre-treatment time weights are estimated from the control unit data. Unit weight estimation, however, requires an additional penalty term to ensure uniqueness when N_0 is large compared to T_0.

2.2 Bias and variance reduction through factor balancing

In the following part I will first consider the properties of $\hat{\tau}(\boldsymbol{v})$ under fixed weights to document the issue of factor imbalance. I then show how estimated time weights reduce the bias and the asymptotic variance of the estimated treatment effect compared to the unweighted DID approach.

Consider the treatment effect estimate $\hat{\tau}(\boldsymbol{v})$ for a given vector of weights $\boldsymbol{v}$. For equal weights $\boldsymbol{v} = \bar{\boldsymbol{v}}$ this resembles the DID estimator.

THEOREM 1. *Suppose Assumptions 1-3 hold. Then for every* $\boldsymbol{v} \in \mathbb{V}$,

1. $\mathrm{E}[\hat{\tau}(\boldsymbol{v})|\boldsymbol{F}] = \tau + b(\boldsymbol{v})$ *with bias* $b(\boldsymbol{v}) = \boldsymbol{\xi}_\lambda'\boldsymbol{\xi}_f(\boldsymbol{v})$ *and the weighted factor imbalance* $\boldsymbol{\xi}_f(\boldsymbol{v}) = \bar{\boldsymbol{f}}_{(1)} - \boldsymbol{F}'\boldsymbol{v}$,

2. $N\,\mathrm{var}[\hat{\tau}(\boldsymbol{v})|\boldsymbol{F}] = \boldsymbol{\xi}_f(\boldsymbol{v})'\boldsymbol{\Sigma}_\lambda\boldsymbol{\xi}_f(\boldsymbol{v}) + V_\varepsilon(\boldsymbol{v}) =: V_{\hat{\tau}}(\boldsymbol{v})$ *with* $\boldsymbol{\Sigma}_\lambda = \mathrm{var}[\bar{\boldsymbol{\lambda}}^{(1)} - \bar{\boldsymbol{\lambda}}^{(0)}]$, $V_\varepsilon(\boldsymbol{v}) = \mathrm{var}[\bar{\Delta}_{\varepsilon,[1]} - \sum_{t\le T_0} v_t\Delta_{\varepsilon,t}]$, $\Delta_{\varepsilon,t} = \bar{\varepsilon}_t^{(1)} - \bar{\varepsilon}_t^{(0)}$, *and*

3. $\sqrt{N}(\hat{\tau}(\boldsymbol{v}) - \tau - b(\boldsymbol{v})) \xrightarrow{d} \mathrm{N}[0, V_{\hat{\tau}}(\boldsymbol{v})]$ *as* $N \to \infty$.

The weighted factor imbalance $\boldsymbol{\xi}_f(\boldsymbol{v})$ will play an important role. It is the difference between the average post-treatment factors and the weighted average pre-treatment factors and affects the estimated treatment effect in two ways. First, the combination of a non-zero loading imbalance $\boldsymbol{\xi}_\lambda$ and factor imbalance $\boldsymbol{\xi}_f(\boldsymbol{v})$ leads to a first order bias term $b(\boldsymbol{v})$. For example, consider the case with one common factor f_t, which affects treated units are on average more than untreated units ($\xi_\lambda > 0$). If f_t is on average in the post-treatment periods higher than in the pre-treatment periods, $\hat{\tau}(\boldsymbol{v})$ will overestimate the treatment effect. Second, it increases the part of the variance resulting from variation in the loadings. This holds irrespective of whether the treatment assignment D_i correlates with the loadings $\boldsymbol{\lambda}_i$, as long as they have within group variation $\boldsymbol{\Sigma}_\lambda > 0$.

The properties of the DID estimator follow as the special case of equal weights $\boldsymbol{v} = \bar{\boldsymbol{v}}$. Researches typically refers to the common trend assumption as condition for unbiasedness. In the current setting, a trend means a non-zero factor imbalance $\boldsymbol{\xi}_f(\bar{\boldsymbol{v}})$. The trends are common if the factors affect treated and untreated units equally. Hence the DID estimator is unbiased ($b(\bar{\boldsymbol{v}}) = 0$) if either the trends are common ($\boldsymbol{\xi}_\lambda = 0$) or there are no trends ($\boldsymbol{\xi}_f(\bar{\boldsymbol{v}}) = 0$).

Now consider weights $\hat{\boldsymbol{v}}$ estimated from the control unit data as per (2). As the number of control units N_0 grows, they converge in probability to the pseudo-true time weights $\boldsymbol{v}^*$ which solve the population equivalent of (2)

$$\boldsymbol{v}^* = \arg\min_{\boldsymbol{v}\in\mathbb{V}} \left\{ \boldsymbol{\xi}_f(\boldsymbol{v})' \boldsymbol{\Sigma}_\lambda^{(0)} \boldsymbol{\xi}_f(\boldsymbol{v}) + V_\varepsilon^{(0)}(\boldsymbol{v}) \right\} \tag{5}$$

with $V_\varepsilon^{(0)}(\boldsymbol{v}) = \text{var}[\bar{\varepsilon}_{[1]}^{(0)} - \sum_{t\leq T_0} v_t \bar{\varepsilon}_t^{(0)}]$. The pseudo-true weights minimize an expression close to the variance of $\hat{\tau}(\boldsymbol{v})$ derived in Theorem 1. It is not just influenced by the factor imbalance $\boldsymbol{\xi}_f(\boldsymbol{v})$, but also by the error variance $V_\varepsilon^{(0)}(\boldsymbol{v})$. As a consequence, the pseudo-true weights do generally not balance the factors entirely. The following Theorem establishes asymptotic normality around $\boldsymbol{v}^*$.

THEOREM 2. *Suppose Assumptions 1-3 hold. Let T_+ be the number of strictly positive elements of $\boldsymbol{v}^*$ with $1 \leq T_+ \leq T_0$. As $N_0 \to \infty$,*

1. $\hat{\boldsymbol{v}} \xrightarrow{p} \boldsymbol{v}^*$
2. $\sqrt{N}(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) \xrightarrow{d} \text{N}[\boldsymbol{0}, \frac{1}{\kappa}\boldsymbol{\Sigma}_{\hat{v}}]$ *with* $\text{rk}\Sigma_{\hat{v}} = T_+ - 1$, $(\boldsymbol{\Sigma}_{\hat{v}})_{s,t} = 0$ *if* $\min\{v_s^*, v_t^*\} = 0$.

Appendix A.2 contains the proof.

Most importantly, the estimated weights $\hat{\boldsymbol{v}}$ and the treatment effect estimate $\hat{\tau}(\boldsymbol{v})$ converge at the same rate $\frac{1}{\sqrt{N}}$. Using estimated weights, the estimation error then becomes

$$\hat{\tau}(\hat{\boldsymbol{v}}) - \tau - b(\boldsymbol{v}^*) = Z_N(\boldsymbol{v}^*) - (\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda)'(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) + O_p(\frac{1}{N})$$

with $Z_N(\boldsymbol{v}^*) = \hat{\tau}(\boldsymbol{v}^*) - \tau - b(\boldsymbol{v}^*)$ the estimation error under pseudo-true weight and $\boldsymbol{F}_{[0]} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{T_0})'$ the $(T_0 \times r)$ matrix of pre-treatment factors. This leads us to the following result

THEOREM 3. *Suppose Assumptions 1-3 hold. Then, as* $N \to \infty$,

1. $\hat{\tau}(\hat{\boldsymbol{v}}) \xrightarrow{p} \tau + b(\boldsymbol{v}^*)$
2. $\sqrt{N}(\hat{\tau}(\hat{\boldsymbol{v}}) - \tau - b(\boldsymbol{v}^*)) \xrightarrow{d} \mathrm{N}[0, V_{\hat{\tau}(\hat{v})}]$

with

$$V_{\hat{\tau}(\hat{v})} = V_{\hat{\tau}}(\boldsymbol{v}^*) + \frac{1}{\kappa}(\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda)'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda$$

The magnitude of the bias $b(\boldsymbol{v}^*) = \boldsymbol{\xi}_\lambda'\boldsymbol{\xi}_f(\boldsymbol{v}^*)$ depends on the remaining factor imbalance under pseudo-true weights. The limit variance consists of two parts. First, $V_{\hat{\tau}}(\boldsymbol{v}^*)$ is the variance of the treatment effect estimator under fixed, pseudo-true weights $\boldsymbol{v}^*$. The second term comes from the weight estimation noise and is only present if $\boldsymbol{\xi}_\lambda \neq 0$. In this case $\hat{\tau}(\hat{\boldsymbol{v}})$ does not have oracle properties as its asymptotic distribution differs from the one of $\hat{\tau}(\boldsymbol{v}^*)$.

Comparing the DID estimator $\hat{\tau}(\bar{\boldsymbol{v}})$ to the time-weighted version $\hat{\tau}(\hat{\boldsymbol{v}})$ leads to the following conclusions. The bias of the latter is smaller if the pseudo-true weights decrease the factor imbalance compared to equal weights. This is arguably the case in relevant scenarios, although technically it is possible to make up cases in which it does not hold. Next, weighting has a two-fold effect on the relative variance

$$\frac{V_{\hat{\tau}(\hat{v})}}{V_{\hat{\tau}}(\bar{\boldsymbol{v}})} = \frac{V_{\hat{\tau}}(\boldsymbol{v}^*)}{V_{\hat{\tau}}(\bar{\boldsymbol{v}})} + \frac{\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda}{\kappa V_{\hat{\tau}}(\bar{\boldsymbol{v}})}$$

First, $\boldsymbol{v}^*$ minimizes the first term by construction. This comes at the cost of weight estimation noise, hence the second term. The following corollary provides high-level sufficient conditions to assure that TWDID in fact reduces bias and variance.

COROLLARY 1. *Suppose Assumptions 1-3 hold. If* $\boldsymbol{\Sigma}_\lambda^{(j)}, \boldsymbol{\Sigma}_\varepsilon^{(j)}$, $j = 0, 1$ *and* $\boldsymbol{F}$ *are such that*

1. $\boldsymbol{\xi}_\lambda' \left(\boldsymbol{\xi}_f(\bar{\boldsymbol{v}})\boldsymbol{\xi}_f(\bar{\boldsymbol{v}})' - \boldsymbol{\xi}_f(\boldsymbol{v}^*)\boldsymbol{\xi}_f(\boldsymbol{v}^*)'\right)\boldsymbol{\xi}_\lambda > 0$,
2. $\frac{V_{\hat{\tau}}(\boldsymbol{v}^*)}{V_{\hat{\tau}}(\bar{\boldsymbol{v}})} < 1 - \frac{\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda}{\kappa V_{\hat{\tau}}(\bar{\boldsymbol{v}})}$

then the TWDID estimator $\hat{\tau}(\hat{\boldsymbol{v}})$ *has a smaller bias and a smaller variance than the DID estimator* $\hat{\tau}(\bar{\boldsymbol{v}})$.

The Monte Carlo simulations in Section 3 show that TWDID substantially reduces bias and variance in a setting with one gaussian factor.

2.3 Inference with two-step standard errors

I propose a consistent estimator of the limit variance derived in Theorem 3

$$\widehat{V}_{\hat{\tau}(\hat{v})} = \widehat{V}_{\hat{\tau}}(\hat{\boldsymbol{v}}) + \frac{1}{\kappa}\dot{\boldsymbol{\Delta}}'_{[0]}\widehat{\boldsymbol{\Sigma}}_{\hat{v}}\dot{\boldsymbol{\Delta}}_{[0]} \tag{6}$$

which consists of two parts. The first part $\widehat{V}_{\hat{\tau}}(\hat{\boldsymbol{v}})$ is the cluster-covariance robust variance estimator of Arellano (1987) applied to the weighted data. It consistently estimates the variance under pseudo-true weights $V_{\hat{\tau}}(\boldsymbol{v}^*)$. The second part accounts for the additional variance caused by the weight estimation noise. It consist of the demeaned average pre-treatment differences $\dot{\boldsymbol{\Delta}}_{[0]} = (\Delta_1 - \bar{\Delta}_{[0]}, \ldots, \Delta_{T_0} - \bar{\Delta}_{[0]})'$ with $\bar{\Delta}_{[0]} = \frac{1}{T_0}\sum_{t\leq T_0}\Delta_t$ and a consistent estimator $\widehat{\boldsymbol{\Sigma}}_v$ of the weight variance. In the remainder of this section I will explain how to construct the different components of $\widehat{V}_{\hat{\tau}(\hat{v})}$ and associated confidence intervals in more detail.

The first part of the estimated variance, $\widehat{V}_{\hat{\tau}}(\hat{\boldsymbol{v}})$, is obtained in the following way. First estimate the time weights $\hat{\boldsymbol{v}}$. Next, weigh only the pre-treatment outcomes and call them $\tilde{y}_{it} = T_0\hat{v}_t y_{it}$ for $t \leq T_0$ and $\tilde{y}_{it} = y_{it}$ for $t > T_0$. Run a two-way fixed effects regression of the weighted outcomes $\tilde{y}_{it}$ on the treatment indicator D_{it}, which yields $\hat{\tau}(\hat{\boldsymbol{v}})$. Applying the Arellano (1987) cluster-covariance estimator[1] on the weighted data then provides $\widehat{V}_{\hat{\tau}}(\hat{\boldsymbol{v}})$. The following Theorem ensures consistency. Its details are explained in the Appendix.

THEOREM 4. *Suppose Assumptions 1-3 hold. Then for any* $\boldsymbol{v} \in \mathbb{V}$

$$\widehat{V}_{\hat{\tau}}(\boldsymbol{v}) \xrightarrow{p} V_{\hat{\tau}}(\boldsymbol{v})$$

as $N \to \infty$. *Moreover,*

$$\widehat{V}_{\hat{\tau}}(\hat{\boldsymbol{v}}) \xrightarrow{p} V_{\hat{\tau}}(\boldsymbol{v}^*)$$

with $\hat{\boldsymbol{v}}$ *the estimated weights as per* (2) *and* $\boldsymbol{v}^*$ *their probability limit as defined in* (5).

Next, I will explain why $\dot{\boldsymbol{\Delta}}'_{[0]}\widehat{\boldsymbol{\Sigma}}_{\hat{v}}\dot{\boldsymbol{\Delta}}_{[0]}$ provides an adequate estimator of $(\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda)'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda$. First, the demeaned average pre-treatment differences become

$$\dot{\boldsymbol{\Delta}}_{[0]} = \boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda + O_p(\frac{1}{\sqrt{N}})$$

implying that they consistently estimate $\boldsymbol{F}_{[0]}\boldsymbol{\xi}_\lambda$ as $N \to \infty$. The estimator of the weight variance $\widehat{\boldsymbol{\Sigma}}_v$ can be explained as follows. Let $\hat{\boldsymbol{v}}$ be the T_0 vector of estimated weights as per (2). As Theorem 3 establishes, only the non-zero weights matter for the limiting weight variance $\boldsymbol{\Sigma}_{\hat{v}}$. We can write

[1] Often referred to as clustering the standard errors on the unit (or individual) level, for instance in Bertrand, Duflo, and Mullainathan (2004)

the non-zero weights as an unrestricted least-squares estimate and thus use least-squares type of standard errors to estimate its variance.

Let $\hat{\boldsymbol{v}}_{[+]}$ be the T_+ vector which just contains the positive weights. Because the weights sum to one, I can write

$$\hat{\boldsymbol{v}}_{[+]} = \boldsymbol{e}_1 + \boldsymbol{R}\hat{\boldsymbol{v}}_{-1}$$

with $\boldsymbol{e}_1$ the T_+-dimensional unit vector, $\boldsymbol{R} = \begin{pmatrix} -\boldsymbol{\iota}'_{T_+-1} \\ \boldsymbol{I}_{T_+-1} \end{pmatrix}$ a $T_+ \times T_+ - 1$ matrix and $\hat{\boldsymbol{v}}_{-1}$ the last $T_+ - 1$ elements of $\hat{\boldsymbol{v}}_{[+]}$. The latter can be written as the unrestricted least-squares estimate

$$\hat{\boldsymbol{v}}_{-1} = (\tilde{\boldsymbol{Y}}'_{[0]}\tilde{\boldsymbol{Y}}_{[0]})^{-1}\tilde{\boldsymbol{Y}}'_{[0]}\tilde{\boldsymbol{y}}_{[1]}$$

with $\tilde{\boldsymbol{Y}}_{[0]} = \dot{\boldsymbol{Y}}^{(0)}_{[+]}\boldsymbol{R}$, $\tilde{\boldsymbol{y}}_{[1]} = \dot{\boldsymbol{y}}^{(0)}_{[1]} - \dot{\boldsymbol{Y}}^{(0)}_{[+]}\boldsymbol{e}_1$ and $\dot{\boldsymbol{Y}}^{(0)}_{[+]}$ the $N_0 \times T_+$ matrix of demeaned outcomes of the control units in the remaining pre-treatment periods. Its variance can be estimated by

$$\widehat{\boldsymbol{\Sigma}}_{\hat{v}_{-1}} = Q_N(\hat{\boldsymbol{v}})(\tilde{\boldsymbol{Y}}'_{[0]}\tilde{\boldsymbol{Y}}_{[0]})^{-1}$$

with $Q_N(\hat{\boldsymbol{v}})$ the mean sum of squared residuals of the time weight estimation. Finally, the estimator of the weight covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}_{v,[+]} = Q_N(\hat{\boldsymbol{v}})\boldsymbol{R}(\tilde{\boldsymbol{Y}}'_{[0]}\tilde{\boldsymbol{Y}}_{[0]})^{-1}\boldsymbol{R}' \tag{7}$$

which follows from $\text{var}[\hat{\boldsymbol{v}}_{[+]}] = \boldsymbol{R}\,\text{var}[\hat{\boldsymbol{v}}_{-1}]\boldsymbol{R}'$.

Inference can be based on the t-statistic following

$$\text{P}\left(\left|\frac{\hat{\tau}(\hat{\boldsymbol{v}}) - \tau_0 - b(\boldsymbol{v}^*)}{\sqrt{\widehat{V}_{\hat{\tau}(\hat{v})}/N}}\right| > q_{1-\alpha/2}\right) \xrightarrow{p} \alpha$$

with τ_0 the treatment effect under the null, α the intended size and $q_{1-\alpha/2}$ the $1-\alpha/2$ quantile of the standard normal distribution. Without any further restrictions, the resulting confidence interval

$$\left[\hat{\tau}(\hat{\boldsymbol{v}}) \pm q_{1-\alpha/2}\sqrt{\widehat{V}_{\hat{\tau}(\hat{v})}/N}\right]$$

will be centered around $\tau_0 + b(\boldsymbol{v}^*)$. Yet it is more reliable and shorter compared to an unweighted DID approach.

3 Monte Carlo Experiments

3.1 Implementation

In each Monte Carlo replication $r = 1, \ldots, R$ I generate data from

$$y^{(r)}_{it} = \tau D_{it} + \lambda^{(r)}_i f^{(r)}_t + \varepsilon^{(r)}_{it}$$

with $f_t^{(r)} \sim N[0, \sigma_f^2]$ and $\varepsilon_{it}^{(r)} \sim N[0,1]$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, all mutually independent. The loadings are $\lambda_i^{(r)} = \frac{\xi_\lambda}{\sqrt{N}} D_i + \nu_i^{(r)}$ with $\nu_i^{(r)} \sim \mathrm{N}[0,1]$. The true treatment effect is $\tau = 0$. The number of units is $N = 100$ of which half are treated ($\kappa = \frac{N_0}{N} = 0.5$) and there is one treated period $T = T_0 + 1$. I vary the number of pre-treatment periods $T_0 \in \{6, 30\}$ and the loading imbalance $\xi_\lambda \in \{2, 5\}$ along a grid of factor standard deviations $\sigma_f \in \{0, 0.1, \ldots, 2\}$. For each combination of parameters and sample size I conduct $R = 10,000$ replications. In each replication I compute the pseudo-true time weights $\boldsymbol{v}^*$ as of (5), the estimated time weights as of (2), the DID estimator $\hat{\tau}_{\text{did}} = \hat{\tau}(\bar{\boldsymbol{v}})$, the TWDID estimator $\hat{\tau}_{\text{twdid}} = \hat{\tau}(\hat{\boldsymbol{v}})$ and the corresponding variance estimators $\widehat{V}_{\text{did}}$, $\widehat{V}_{\text{twdid}}$. I compare the following Monte Carlo statistics to assess the performance of the point estimators and inference.

1. I measure the magnitude of the bias term $b(\boldsymbol{v}) = \xi_\lambda \xi_f(\boldsymbol{v})$ in terms of its standard deviation $\mathrm{sd}[b(\boldsymbol{v})]$, with $\boldsymbol{v} = \bar{\boldsymbol{v}}$ for the DID and $\boldsymbol{v} = \boldsymbol{v}^*$ for TWDID.

2. Then I look at the simulated conditional standard deviation $\mathrm{sd}[\hat{\tau}(\boldsymbol{v})|\boldsymbol{F}]$ of the point estimates. I compute it as the Monte Carlo standard deviation of the bias corrected estimator $\hat{\tau}(\boldsymbol{v}) - b(\boldsymbol{v})$.

3. Finally, I consider the feasible t-statistics of both estimation approaches

$$t_{\text{did}} = \frac{\hat{\tau}_{\text{did}} - \tau_0}{\sqrt{\widehat{V}_{\text{did}}/N}}, \qquad t_{\text{twdid}} = \frac{\hat{\tau}_{\text{twdid}} - \tau_0}{\sqrt{\widehat{V}_{\text{twdid}}/N}}$$

   I compute the true rejection frequency under $\tau_0 = 0$ using critical values from the standard normal distribution.

3.2 Results

The simulation results confirm the bias and variance reduction properties of TWDID compared to DID. The top panel of Figure 1 plots the magnitude of the bias against the strength of the factors. The bias of DID increases as the factors become stronger. The bias of TWDID is about 50% lower. This effect is stronger in a setting with a higher loading imbalance (left vs. right). The bottom panel shows the the conditional standard deviation, which is independent of the loading imbalance and instead driven by the loading variation. For sufficiently strong factors, the TWDID estimator has a substantially lower standard deviation. For weak factors, however, there is not much to be gained and the weight estimation noise leads to a slightly larger standard deviation.

Consider now inference based on the t-statistics t_{did} and t_{twdid}. Figure 2 shows the rejection frequency of the standard t-test under the null of no

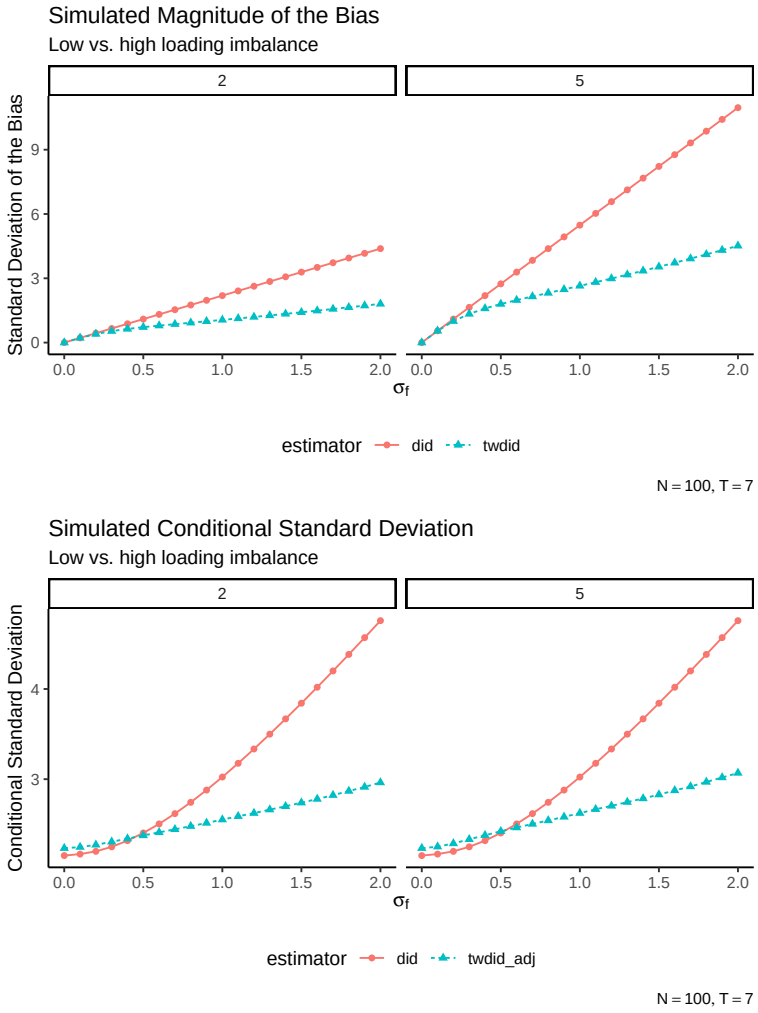

Figure 1: Magnitude of the conditional bias term $b(\boldsymbol{v}^*) = \xi_\lambda \xi_f(\boldsymbol{v}^*)$ (top) and simulated conditional standard deviation $\text{sd}[\hat{\tau}(\boldsymbol{v})]$ (bottom) of the DID and TWDID estimator, depending on the factor standard deviation σ_f.

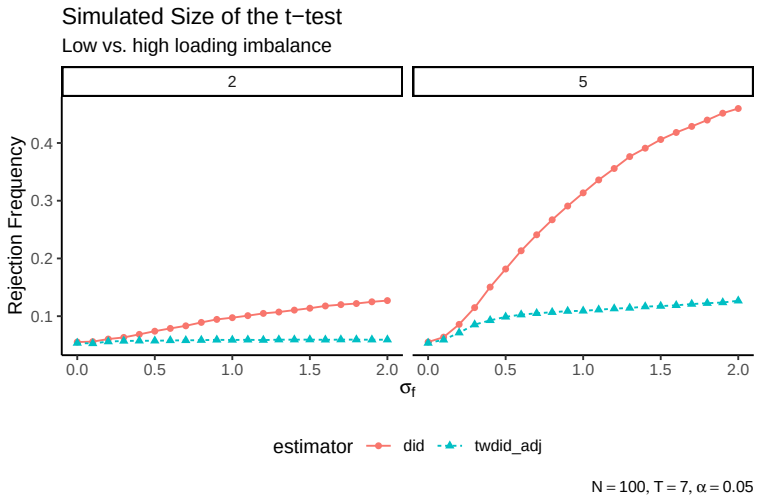

Figure 2: Simulated rejection frequency based on the feasible t-statistic t_{did} and t_{twdid} depending on the factor strength σ_f. The standard errors of the TWDID estimator have been adjusted for weight estimation uncertainty.

treatment effect $H_0 \colon \tau = 0$. If the loading imbalance is sufficiently small (left), the remaining bias of the TWDID estimator is negligible as the test remains size-correct when the factors become stronger. The larger bias of the DID estimator leads its test statistic to be distorted. For large loading imbalances (right), the bias becomes so large that both tests are distorted. Yet this distortion is less severe for TWDID than it is for DID.

4 The effect of the NOx Budget Trading Program revisited

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.

4.1 Econometric Specification

Consider the interactive fixed effect model

$$y_{ist} = \tau D_{ist} + \mu_{it} + \nu_{is} + \boldsymbol{\lambda}_i' \boldsymbol{f}_{st} + \tilde{\varepsilon}_{ist}$$

with $D_{ist} = \mathrm{I}(i \in \mathcal{N}_1, t > T_0, s = 1)$ an indicator whether NBP is operating in county i in season $s = 0, 1$ (winter, summer) of year t. μ_{it}, ν_{is} are county-year and county-season fixed effects, respectively. $\boldsymbol{f}_{st}$ 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) assume. In that case the factor structure reduces to a season-year fixed effect.

To identify τ, eliminate μ_{it} by taking the difference between summer and winter observations

$$\check{y}_{it} := y_{i1t} - y_{i0t} = \tau D_{it} + \beta_i + \boldsymbol{\lambda}_i' \check{\boldsymbol{f}}_t + \varepsilon_{it}$$

with $\beta_i = \nu_{i1} - \nu_{i0}$, $\check{\boldsymbol{f}}_t = \boldsymbol{f}_{1t} - \boldsymbol{f}_{0t}$ and $\varepsilon_{it} = \tilde{\varepsilon}_{i1t} - \tilde{\varepsilon}_{i0t}$. A key assumption hidden in this specification is that the program does not affect emissions in the winter months in the treated years. The identifying assumption is that there exists a vector of weights $\boldsymbol{v}_0$ such that

$$\frac{1}{T_1} \sum_{t > T_0} \check{\boldsymbol{f}}_t = \sum_{t \leq T_0} v_{0,t} \check{\boldsymbol{f}}_t$$

That is, the average post-treatment factor can be written as a weighted average of pre-treatment factors.

4.2 Evidence for common factors

I first obtain evidence against $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ by considering how the difference in average NOx emissions $\Delta_t = \bar{\check{y}}_t^{(1)} - \bar{\check{y}}_t^{(0)}$ has evolved prior to the intervention. We can write

$$\Delta_t = \bar{\beta}^{(1)} - \bar{\beta}^{(0)} + \boldsymbol{\xi}_\lambda' \boldsymbol{f}_t + \tau \mathrm{I}(t > T_0) + O_p(\frac{1}{\sqrt{N}})$$

where $\boldsymbol{\xi}_\lambda = 0$ under the equal loading assumption. Then, for large N, Δ_t should be constant prior to the treatment. However, Figure 3 does show variation of Δ_t in periods $t \leq T_0$.

4.3 Estimation Results

I estimate τ with a time weighted DID regression as defined in (4). Let $\hat{\tau}(\boldsymbol{v})$ be the point estimate for a given T_0 vector of time weights $\boldsymbol{v} \in \mathbb{V}$. I compare three sets of weights $\boldsymbol{v}$. First, I use equal weights $\bar{\boldsymbol{v}}$, which return the DID

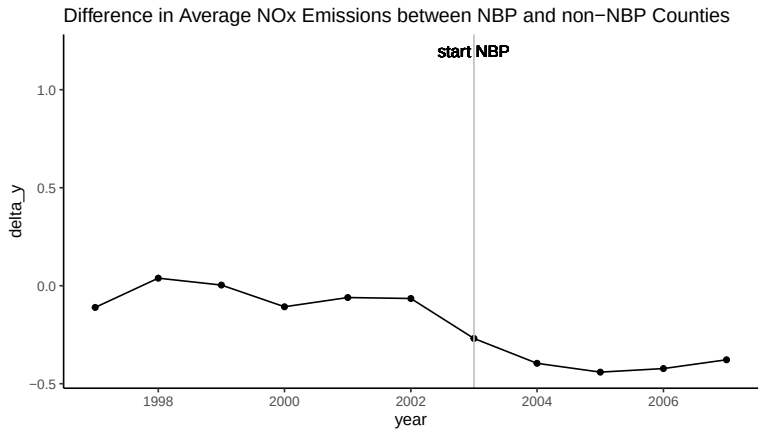

Figure 3: Difference in average NOx emissions $\bar{\hat{y}}_t^{(1)} - \bar{\hat{y}}_t^{(0)}$

estimator used by Deschenes et al. (2017). Second, for the TWDID estimator I use estimated weights $\hat{\boldsymbol{v}}$ obtained from the restricted regression (2) of pre-treatment emissions on the average post-treatment emission. Third, I consider the TWDID estimator using unrestricted weights $\hat{\boldsymbol{v}}_u$, where the only difference to $\hat{\boldsymbol{v}}$ is that I allow the weights to be non-negative. I obtain the standard errors For the DID estimator I use the cluster-covariance matrix estimator. I obtain the standard errors of the TWDID estimators with the two-step procedure discussed in Section 2.3, which accounts for the presence of estimated weigths. I omit the year 2003 from the estimation to circumvent issues related to the staggered implementation.

The estimated time weights are plotted in Figure 4. The last two pre-treatment periods receive almost all the weight, indicating that the common factors of the NOx emissions were in these two years closest to their post-treatment average. The bottom panel of Figure 4 shows the resulting treatment effect estimates and their 95% confidence intervals. All estimators suggest a significant negative effect of the NBP program on NOx emissions. Time weighting leads, in absolute terms, to a slightly lower point estimate (DID: -0.36 vs TWDID: -0.34). The standard errors of both TWDID estimates (0.050) are about 10% lower compared to DID (0.056), and the resulting confidence intervals are more narrow. This results is in line with the variance reduction property of TWDID estimation.

5 Conclusion

This paper proposes a time-weighted difference-in-difference (TWDID) estimation approach that is robust against interactive fixed effects in short T panels. It covers settings with few time periods, many cross-sectional units and sharp treatment timing. Using time weights estimated from the

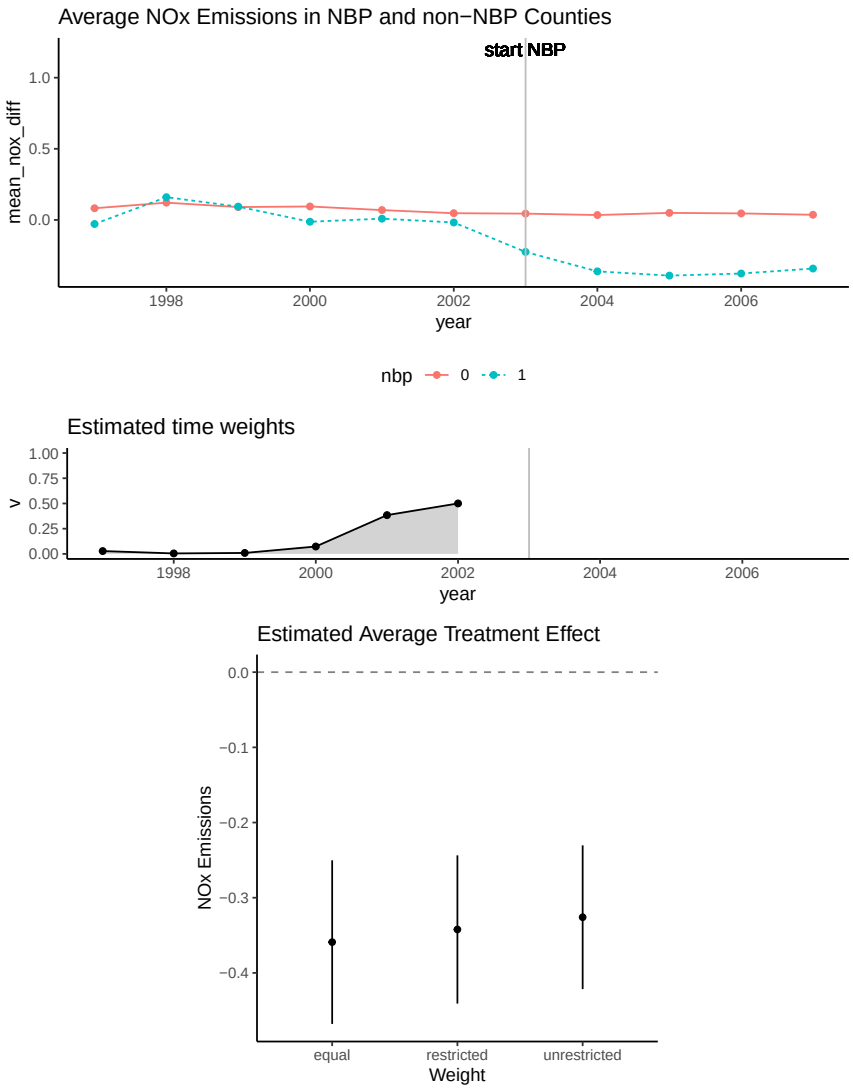

Figure 4: Top: Average NOx emissions NBP-Participating States vs. the Control States 1997 - 2008. Middle: estimated time weights as of (2). Bottom: Point estimates and confidence intervals of DID and TWDID estimation.

untreated units, the TWDID estimator offers both a lower bias and a lower variance than the unweighted DID estimator. Moreover, I show how to adjust the standard errors to cover the additional weight estimation uncertainty. I revisit a study on the effect of a cap-and-trade program on NOx emissions. Using TWDID reduces the standard errors of the estimated average treatment effect by 10% compared to a conventional DID approach.

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 Amerian Statistical Association*, 105, 493–505.

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

Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

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.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics*, 119, 249–275.

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.

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. (2019): "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls," *arXiv preprint arXiv:1906.06665*.

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

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

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

A Proofs of Theorems

A.1 Proof of Theorem 1

Again some notation. Let $\boldsymbol{b}$ be a T-vector with elements $b_t = \frac{1}{T_1} I(t > T_0) - \frac{1}{T_0} I(t \leq T_0)$. $\boldsymbol{\Upsilon}$ is a $T \times T$ diagonal matrix with the scaled time weights on the diagonal $(\boldsymbol{\Upsilon})_{t,t} = T_0 v_t I(t \leq T_0) + I(t > T_0)$. So $\boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{x} = \frac{1}{T_1}\sum_{t>T_0} - \sum_{t \leq T_0} v_t x_t$ for any T vector $\boldsymbol{x}$. Let $\bar{\boldsymbol{\nu}}_\lambda = \bar{\boldsymbol{\lambda}}^{(1)} - \bar{\boldsymbol{\lambda}}^{(0)} - \boldsymbol{\xi}_\lambda$ and

$$Z_N(\boldsymbol{v}) = \boldsymbol{b}'\boldsymbol{\Upsilon}(\bar{\boldsymbol{\varepsilon}}^{(1)} - \bar{\boldsymbol{\varepsilon}}^{(0)}) = \frac{1}{T_1}\sum_{t>T_0}(\bar{\varepsilon}_t^{(1)} - \bar{\varepsilon}_t^{(0)}) - (\bar{\boldsymbol{\varepsilon}}_{[0]}^{(1)} - \bar{\boldsymbol{\varepsilon}}_{[0]}^{(0)})'\boldsymbol{v}$$

First write

$$\hat{\tau}(\boldsymbol{v}) - \tau = \boldsymbol{\xi}_\lambda' \boldsymbol{\xi}_f(\boldsymbol{v}) + \bar{\boldsymbol{\nu}}_\lambda' \boldsymbol{\xi}_f(\boldsymbol{v}) + Z_N(\boldsymbol{v})$$

from which we arrive at the following asymptotic results.

LEMMA 1. *For any fixed* $\boldsymbol{v} \in \mathbb{V}$

1. $\sqrt{N} Z_N(\boldsymbol{v}) \xrightarrow{d} \mathrm{N}[0, V_\varepsilon(\boldsymbol{v})]$
2. $\sqrt{N}(\bar{\lambda}^{(1)} - \bar{\lambda}^{(0)} - \boldsymbol{\xi}_\lambda) \xrightarrow{d} \mathrm{N}[0, \boldsymbol{\Sigma}_\lambda]$

with $V_\varepsilon(\boldsymbol{v}) = \boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{\Sigma}_\varepsilon\boldsymbol{\Upsilon}\boldsymbol{b}$, $\boldsymbol{\Sigma}_\varepsilon = \frac{\boldsymbol{\Sigma}_\varepsilon^{(0)}}{\kappa} + \frac{\boldsymbol{\Sigma}_\varepsilon^{(1)}}{1-\kappa}$ *and* $\boldsymbol{\Sigma}_\lambda = \frac{\boldsymbol{\Sigma}_\lambda^{(0)}}{\kappa} + \frac{\boldsymbol{\Sigma}\lambda^{(1)}}{1-\kappa}$

The first two assertions of Theorem 2 follow with

$$V_{\hat{\tau}}(\boldsymbol{v}) = \boldsymbol{\xi}_f(\boldsymbol{v})'\boldsymbol{\Sigma}_\lambda\boldsymbol{\xi}_f(\boldsymbol{v}) + V_\varepsilon(\boldsymbol{v})$$

A.2 Proof of Theorem 2

The estimated time weights $\hat{\boldsymbol{v}}$ as defined in (2) can be rewritten as

$$\hat{\boldsymbol{v}} = \underset{\boldsymbol{v}\in\mathbb{V}}{\arg\min}\, Q_N(\boldsymbol{v}), \qquad Q_N(\boldsymbol{v}) = (\dot{\boldsymbol{y}}_{0,T} - \dot{\boldsymbol{Y}}_{0,p}\boldsymbol{v})'(\dot{\boldsymbol{y}}_{0,T} - \dot{\boldsymbol{Y}}_{0,p}\boldsymbol{v})$$

with $\boldsymbol{y}_{0,t} = (y_{1t}, \ldots, y_{N_0 t})$ the vector of the untreated units' outcomes in the period t, $\boldsymbol{Y}_{0,p} = [\boldsymbol{y}_{0,1}, \ldots, \boldsymbol{y}_{0,T_0}]$ the $N_0 \times T_0$ matrix of the untreated units' pre-treatment outcomes and $\mathbb{V}$ the set of T_0 vectors with non-zero elements that sum to one. We notice that $\hat{\boldsymbol{v}}$ always exists, and is unique if $\dot{\boldsymbol{Y}}_{0,p} = [\dot{\boldsymbol{y}}_{p,1}, \ldots, \dot{\boldsymbol{y}}_{p,N_0}]'$ has full column rank.

Consistency follows from two key arguments. First, the objective function uniformly convergence to its population equivalent $Q_\infty(\boldsymbol{v}) = V_f^{(0)}(\boldsymbol{v}) + V_\varepsilon(\boldsymbol{v})$

$$\sup_{\boldsymbol{v}\in\mathbb{V}} |Q_N(\boldsymbol{v}) - Q_\infty(\boldsymbol{v})| \xrightarrow{p} 0$$

Second, $\boldsymbol{v}^*$ as defined in (5) is the uniquely identifiable minimizer of $Q_\infty(\boldsymbol{v})$.

To obtain asymptotic normality, suppose first that $\boldsymbol{v}^*$ lies in the interior of $\mathbb{V}$, i.e. the non-negativity constraints are not binding in the limit. Without the non-negativity constraint we can rewrite the estimation as an unconstrained minimization and obtain an explicit solution. To do so, we transform the condition $\sum_t v_t = 1$ into $v_1 = 1 - \boldsymbol{\iota}'_{T_0-1}\boldsymbol{v}_{-1}$ with $\boldsymbol{v}_{-1} = (v_2, \ldots, v_{T_0})$, so $\boldsymbol{v} = \boldsymbol{e}_1 + \boldsymbol{R}\boldsymbol{v}_{-1}$ with $\boldsymbol{R} = \begin{pmatrix} -\boldsymbol{\iota}'_{T_0-1} \\ \boldsymbol{I}_{T_0-1} \end{pmatrix}$. The minimization problem then becomes

$$\min_{\boldsymbol{v}_{-1} \in \mathbb{R}^{T_0-1}} (\tilde{\boldsymbol{y}}_{(1)} - \tilde{\boldsymbol{Y}}_p \boldsymbol{v}_{-1})'(\tilde{\boldsymbol{y}}_{(1)} - \tilde{\boldsymbol{Y}}_p \boldsymbol{v}_{-1})$$

with $\tilde{\boldsymbol{y}}_{(1)} = \bar{\boldsymbol{y}}_{(1)} - \boldsymbol{y}_1$ and $\tilde{\boldsymbol{Y}}_p = \boldsymbol{Y}_p \boldsymbol{R}$. The solution is

$$\hat{\boldsymbol{v}}_{-1} = (\tilde{\boldsymbol{Y}}'_p \tilde{\boldsymbol{Y}}_p)^{-1} \tilde{\boldsymbol{Y}}'_p \tilde{\boldsymbol{y}}_{(1)}$$

and the corresponding T_0 vector of time weights will be $\hat{\boldsymbol{v}} = (1 - \sum_{t=2}^{T_0} \hat{v}_t, \hat{\boldsymbol{v}}'_{-1})'$. That is a least-squares regression with the first regressor subtracted. By the same argument we can write $\boldsymbol{v}^* = \boldsymbol{e}_1 + \boldsymbol{R}\boldsymbol{v}^*_{-1}$. The idea is to first show asymptotic normality for the last $T_0 - 1$ weights,

$$\sqrt{N_0}(\hat{\boldsymbol{v}}_{-1} - \boldsymbol{v}^*_{-1}) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_{\hat{v}_{-1}}]$$

which is easy to obtain because $\hat{\boldsymbol{v}}_{-1}$ is an unrestricted least squares estimator. Because of the sum-to-1 condition it will immediately follow for the full vector

$$\sqrt{N_0}(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) = \sqrt{N_0}\boldsymbol{R}(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_{\hat{v}}]$$

with $\boldsymbol{\Sigma}_{\hat{v}} = \boldsymbol{R}\boldsymbol{\Sigma}_{\hat{v}_{-1}}\boldsymbol{R}'$. Note that both $\boldsymbol{\Sigma}_{\hat{v}_{-1}}$ and $\boldsymbol{\Sigma}_{\hat{v}}$ have rank $T_0 - 1$. In absence of heteroskedasticity we suggest using the simple least squares variance estimator

$$\widehat{\boldsymbol{\Sigma}}_{\hat{v}_{-1}} = Q_N(\hat{\boldsymbol{v}})(\boldsymbol{R}'\boldsymbol{Y}'_p\boldsymbol{Y}_p\boldsymbol{R})^{-1}$$

.

Suppose now that $\boldsymbol{v}^*$ lies on the boundary of $\mathbb{V}$. That is, at least one element is exactly zero and $\boldsymbol{v}^*$ is sparse. In general, asymptotic normality of extremum estimators can break down near the boundary of the parameter space (Ketz, 2018). In this special case, however, the part of $\hat{\boldsymbol{v}}$ belonging to the non-zero elements of $\boldsymbol{v}^*$ will still be asymptotically normal.

Formally, suppose without loss of generality that the first k elements $(0 < k < T_0 - 1)$ of $\boldsymbol{v}^*$ are zero and write $\boldsymbol{v} = (\boldsymbol{v}'_k, \boldsymbol{v}'_{-k})'$.

LEMMA 2. *I conjecture:* $\hat{\boldsymbol{v}}_k = O_p(N^{-q})$ *with* $q > 0.5$

The idea is that $\hat{\boldsymbol{v}}_k$ converges to zero so quickly that the asymptotic distribution is the same as if we had just set $\hat{\boldsymbol{v}}_k = 0$ in the first place. But this means that the first k periods are irrelevant and we have reduced the exercise to $T_0 - k$ pre-treatment periods, for which the pseudo-true weights $\boldsymbol{v}^*_{-k}$ are not sparse. Then we can apply the results from the previous paragraph to get asymptotic normality for the non-zero elements.

A.3 Proof of Theorem 3

Let $\boldsymbol{G}_N = (\bar{\boldsymbol{\nu}}_\lambda' \boldsymbol{\xi}_f(\boldsymbol{v}^*) + Z_N(\boldsymbol{v}^*), (\hat{\boldsymbol{v}} - \boldsymbol{v}^*)')'$.

LEMMA 3. *1.* $(\boldsymbol{F}_{(0)}' \bar{\boldsymbol{\nu}}_\lambda)'(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) = O_p(\frac{1}{N})$

2. $(\bar{\boldsymbol{\varepsilon}}_{(0)}^{(1)} - \bar{\boldsymbol{\varepsilon}}_{(0)}^{(0)})'(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) = O_p(\frac{1}{N})$

3. $\sqrt{N} G_N \xrightarrow{d} \mathrm{N}[\boldsymbol{0}, \boldsymbol{\Sigma}_G]$
with

$$\boldsymbol{\Sigma}_G = \begin{pmatrix} V_{\hat{\tau}} & \frac{1}{\sqrt{\kappa}} \boldsymbol{C}' \\ \frac{1}{\sqrt{\kappa}} \boldsymbol{C} & \frac{1}{\kappa} \boldsymbol{\Sigma}_{\hat{v}} \end{pmatrix}$$

Adding and subtracting $\hat{\tau}(\boldsymbol{v}^*)$ we get

$$\hat{\tau}(\hat{\boldsymbol{v}}) - \tau - b(\boldsymbol{v}^*) = (1, -(\boldsymbol{F}_{(0)} \xi_\lambda)') \boldsymbol{G}_N + O_p(\frac{1}{N})$$

and thus assertions 1 and 2 of Theorem 3 follow with

$$V_{\hat{\tau}(\hat{v})} = V_{\hat{\tau}} + \frac{1}{\kappa} (\boldsymbol{F}_{(0)} \xi_\lambda)' \boldsymbol{\Sigma}_{\hat{v}} \boldsymbol{F}_{(0)} \xi_\lambda - \frac{2}{\sqrt{\kappa}} (\boldsymbol{F}_{(0)} \xi_\lambda)' \boldsymbol{C}$$

Consistency of $\widehat{V}_{\hat{\tau}(\hat{v})}$ follows from the fact that $\widehat{V}_{\hat{\tau}}(\hat{\boldsymbol{v}}) \xrightarrow{p} V_{\hat{\tau}}(\boldsymbol{v}^*)$ and $\dot{\boldsymbol{\Delta}}_{(0)} \xrightarrow{p} \boldsymbol{F}_{(0)} \xi_\lambda$. So far $\widehat{V}_{\hat{\tau}(\hat{v})}$ ignores the covariance term $\boldsymbol{C}$. In simulations $\boldsymbol{C}$ is negligible as long as $\hat{\boldsymbol{v}}$ is estimated using demeaned observations.

A.4 Proof of Theorem 4

Let $\dot{\boldsymbol{y}}_i = \boldsymbol{y}_i - \frac{1}{N} \sum_{i=1}^N \boldsymbol{y}_i$ denote the T-vector of cross-sectionally demeaned outcomes, $\ddot{\boldsymbol{y}}_i = \boldsymbol{M}_T \dot{\boldsymbol{y}}_i$ is the vector of double demeaned outcomes with $\boldsymbol{M}_T = \boldsymbol{I}_T - \frac{\iota_T \iota_T'}{T}$.

I now derive probability limit of $\widehat{V}(\boldsymbol{v})$ for some fixed $\boldsymbol{v} \in \mathbb{V}$. First note that the denominator reduces to $\frac{1}{N} \sum_{i=1}^N \ddot{\boldsymbol{D}}_i' \ddot{\boldsymbol{D}}_i = T k_N k_T (1 - k_N)(1 - k_T) =: Q_N$. Hence we only need to consider the numerator.

$$\begin{aligned} \hat{\ddot{\boldsymbol{u}}}_{i,\boldsymbol{v}} &= \ddot{y}_{i,\boldsymbol{v}} - \ddot{\boldsymbol{D}}_i \hat{\tau}(\boldsymbol{v}) \\ \ddot{y}_{i,\boldsymbol{v}} &= \boldsymbol{M}_T \boldsymbol{\Upsilon} \dot{\boldsymbol{y}}_i = \tau \boldsymbol{M}_T \boldsymbol{\Upsilon} \dot{\boldsymbol{D}}_i + \ddot{\boldsymbol{u}}_{i,\boldsymbol{v}} \\ \ddot{\boldsymbol{u}}_{i,\boldsymbol{v}} &= \boldsymbol{M}_T (\boldsymbol{F}_{\boldsymbol{v}} \dot{\boldsymbol{\lambda}}_i + \dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}) \end{aligned}$$

with $\boldsymbol{F}_{\boldsymbol{v}} = \boldsymbol{\Upsilon} \boldsymbol{F}$ and $\dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}} = \boldsymbol{\Upsilon} \dot{\boldsymbol{\varepsilon}}_i$. Note that $\boldsymbol{\Upsilon} \dot{\boldsymbol{D}}_i = \dot{\boldsymbol{D}}_i$ because $(\boldsymbol{\Upsilon})_{t,t} \neq 1$

only if $D_{it} = 0$. Substitute $\hat{\ddot{\boldsymbol{u}}}_{i,\boldsymbol{v}} = \ddot{\boldsymbol{u}}_{i,\boldsymbol{v}} - (\hat{\tau} - \tau)\ddot{\boldsymbol{D}}_i$ to get

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\hat{\ddot{\boldsymbol{u}}}_{i,\boldsymbol{v}}\hat{\ddot{\boldsymbol{u}}}_{i,\boldsymbol{v}}'\ddot{\boldsymbol{D}}_i = \frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{u}}_{i,\boldsymbol{v}}\ddot{\boldsymbol{u}}_{i,\boldsymbol{v}}'\ddot{\boldsymbol{D}}_i \\ - 2(\hat{\tau}(\boldsymbol{v}) - \tau)\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{u}}_{i,\boldsymbol{v}}\ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{D}}_i \\ + (\hat{\tau}(\boldsymbol{v}) - \tau)^2\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{D}}_i\ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{D}}_i\end{aligned}$$

The following lemma ensures that the latter two terms converge to zero given consistency of $\hat{\tau}(\boldsymbol{v})$.

LEMMA 4. *We have these convergence results as* $N \to \infty$

- $\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{u}}_{i,\boldsymbol{v}}\ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{D}}_i \xrightarrow{p} c_1 < \infty$
- $\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{D}}_i\ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{D}}_i \xrightarrow{p} c_2 < \infty$
- $\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\boldsymbol{F}_{\boldsymbol{v}}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}'\ddot{\boldsymbol{D}}_i \xrightarrow{p} 0$

Use $\ddot{\boldsymbol{D}}_i'\ddot{\boldsymbol{u}}_{i,\boldsymbol{v}} = \dot{\boldsymbol{D}}_i'\boldsymbol{M}_T'\boldsymbol{M}_T\dot{\boldsymbol{u}}_{i,\boldsymbol{v}} = \ddot{\boldsymbol{D}}_i'\dot{\boldsymbol{u}}_{i,\boldsymbol{v}}$, $\dot{\boldsymbol{u}}_{i,\boldsymbol{v}} = \boldsymbol{F}_{\boldsymbol{v}}\dot{\boldsymbol{\lambda}}_i + \dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}$ to get

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\dot{\boldsymbol{u}}_{i,\boldsymbol{v}}\dot{\boldsymbol{u}}_{i,\boldsymbol{v}}'\ddot{\boldsymbol{D}}_i = \frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\boldsymbol{\Upsilon}\boldsymbol{F}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\lambda}}_i'\boldsymbol{F}'\boldsymbol{\Upsilon}\ddot{\boldsymbol{D}}_i + \frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}\dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}'\ddot{\boldsymbol{D}}_i \\ + \frac{2}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\boldsymbol{F}_{\boldsymbol{v}}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}'\ddot{\boldsymbol{D}}_i\end{aligned} \tag{8}$$

where the last term converges to zero by independence of $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}_i$. Consider the first part of (8).

Then we can write $\ddot{\boldsymbol{D}}_i = Q_N\boldsymbol{b}(\frac{1}{1-k_N}I(i > N_0) - \frac{1}{k_N}I(i \leq N_0))$. Also, $\boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{F} = \bar{\boldsymbol{f}}_{(1)} - \boldsymbol{F}_p'\boldsymbol{v}$. Let $\Sigma_\lambda^{(0)} := \frac{1}{N_0}\sum_{i=1}^{N_0}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\lambda}}_i'$ and $\Sigma_\lambda^{(1)} := \frac{1}{N_1}\sum_{i=N_0+1}^{N}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\lambda}}_i'$. Then, all together we get

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\boldsymbol{\Upsilon}\boldsymbol{F}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\lambda}}_i'\boldsymbol{F}'\boldsymbol{\Upsilon}\ddot{\boldsymbol{D}}_i &= \frac{Q_N^2}{Nk_N^2}\sum_{i=1}^{N_0} \boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{F}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\lambda}}_i'\boldsymbol{F}'\boldsymbol{\Upsilon}\boldsymbol{b} \\ &\quad + \frac{Q_N^2}{N(1-k_N)^2}\sum_{i=N_0}^{N} \boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{F}\dot{\boldsymbol{\lambda}}_i\dot{\boldsymbol{\lambda}}_i'\boldsymbol{F}'\boldsymbol{\Upsilon}\boldsymbol{b} \\ &= Q_N^2\left[(\bar{\boldsymbol{f}}_{(1)} - \boldsymbol{F}_p'\boldsymbol{v})'\left(\frac{\Sigma_\lambda^{(0)}}{k_N} + \frac{\Sigma_\lambda^{(1)}}{1-k_N}\right)(\bar{\boldsymbol{f}}_{(1)} - \boldsymbol{F}_p'\boldsymbol{v})\right]\end{aligned}$$

Consider now the second part of (8). First, observe that

$$\frac{\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i'\dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}}{Q_N} = \boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{\Delta}_\varepsilon = \frac{1}{T_1}\sum_{t>T_0}\Delta_{\varepsilon,t} - \sum_{t\leq T_0} v_t\Delta_{\varepsilon,t}$$

with $\boldsymbol{\Delta}_\varepsilon = \bar{\boldsymbol{\varepsilon}}^{(1)} - \bar{\boldsymbol{\varepsilon}}^{(0)}$. We should therefore have that

$$\frac{\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i' \dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}} \dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}' \ddot{\boldsymbol{D}}_i}{Q_N^2} \xrightarrow{p} \frac{\operatorname{var}[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \ddot{\boldsymbol{D}}_i' \dot{\boldsymbol{\varepsilon}}_{i,\boldsymbol{v}}]}{Q_N^2} = N \ \operatorname{var}[\boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{\Delta}_\varepsilon].$$

Moreover, by independence of the units we have $N \operatorname{var}[\boldsymbol{\Delta}_\varepsilon] = N \operatorname{var}[\bar{\boldsymbol{\varepsilon}}^{(1)}] + N \operatorname{var}[\bar{\boldsymbol{\varepsilon}}^{(0)}] = \frac{\boldsymbol{\Omega}}{k_N(1-k_N)}$ and thus

$$V_\varepsilon(\boldsymbol{v}) = \frac{\boldsymbol{b}'\boldsymbol{\Upsilon}\boldsymbol{\Omega}\boldsymbol{\Upsilon}\boldsymbol{b}}{k_N(1-k_N)} = \frac{\bar{\sigma}_{(0),\boldsymbol{v}} + \bar{\sigma}_{(1)} - 2\bar{\sigma}_{(0,1),\boldsymbol{v}}}{k_N(1-k_N)}$$

with $\bar{\sigma}_{(0),\boldsymbol{v}} = \boldsymbol{v}'\boldsymbol{\Omega}_{(0)}\boldsymbol{v}$, $\bar{\sigma}_{(1)} = \frac{1}{T_1^2}\boldsymbol{\iota}_{T_1}'\boldsymbol{\Omega}_{(1)}\boldsymbol{\iota}_{T_1}$ and $\bar{\sigma}_{(0,1),\boldsymbol{v}} = \frac{1}{T_1}\boldsymbol{v}'\boldsymbol{\Omega}_{(0,1)}\boldsymbol{\iota}_{T_1}$.