Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels

Nicholas Brown* and Kyle Butts†

June 15, 2024

We study the estimation and inference of dynamic average treatment effect parameters when parallel trends hold conditional on interactive fixed effects. We study the case of staggered treatment, where units enter into treatment at different time periods. Our proposed generalized method of moments (GMM) estimator consists of two parts: first, we estimate the unobserved time effects by applying the fixed-T consistent quasi-differencing estimator of Ahn et al. (2013) to the never-treated group. Second, we estimate the interactive fixed effects for treated groups post-treatment to recover their unobserved counterfactual outcomes. We subtract this quantity from the observed outcomes and average over group membership to achieve our estimator of the Average Treatment Effect on the Treated (ATT). We also demonstrate the robustness of two-way fixed effects to certain parallel trends violations and describe how to test for its consistency. We investigate the effect of Walmart openings on local economic conditions and demonstrate that our methods ameliorate pre-trend violations commonly found in the literature.

JEL Classification Number: C13, C21, C23, C26

Keywords: factor model, panel treatment effect, causal inference, fixed-T

*Queen's University, Economics Department (n.brown@queensu.ca)

†University of Colorado Boulder, Economics Department (kyle.butts@colorado.edu)

1 — Introduction

Difference-in-differences estimators are one of the most popular causal inference tools and computationally simple but rely on strong parallel-trends assumptions. In many empirical settings, treatment is assigned non-randomly based on trends in economic variables, rendering this method unusable. For example, in urban economics, place-based policies target places with worsening labor markets (Neumark and Simpson, 2015), new apartments are built in appreciating neighborhoods (Asquith et al., 2021; Pennington, 2021), and firms open new stores in growing economies (Basker, 2005; Neumark et al., 2008). Estimation of treatment effects in this setting is confounded by the pre-existing economic trends. It is common, though, that the causes of these trends are due to larger economic forces and not location-specific shocks. Continuing our examples, the national decline of manufacturing caused targeted manufacturing hubs to decline, consumer tastes for walkable neighborhoods caused certain existing neighborhoods to become increasingly demanded, and national macroeconomic changes benefited certain counties.

A recent but growing literature models these kind of parallel trends deviations using interactive fixed effects. While interactive fixed effects relax the parallel trends assumptions relied on by difference-in-differences, these estimators require long panels, which are often impractical because of (i) lack of data, (ii) strong assumptions like serially uncorrelated outcomes and homogeneous treatment effects, or (iii) the presence of structural breaks, e.g. recessions or structural changes to the macroeconomy that render previous time periods uninformative about the current economy. This paper proposes a treatment effect estimator under the more general interactive fixed effect model that is robust to certain violations of parallel trends while remaining consistent in short panels and under heterogeneous treatment effects.

We model untreated potential outcomes $y_{it}(\infty)$ with interactive fixed effects

$$y_{it}(\infty) = \boldsymbol{f}_t'\boldsymbol{\gamma}_i + u_{it}, \tag{1}$$

where $\boldsymbol{F}_t$ is a $p \times 1$ vector of unobservable factors, $\boldsymbol{\gamma}_i$ is a $p \times 1$ vector of unobservable factor loadings, and $\mathbb{E}[u_{it}] = 0$ for all (i, t).[1] We can view the factors $\boldsymbol{F}_t$ as macroeconomic shocks

1. We follow Callaway and Sant'Anna (2021) and define the state of not receiving treatment in the sample as '∞'. This is useful in settings with staggered treatment timing where potential outcomes are denoted by the period where

with factor loadings $\boldsymbol{\gamma}_i$ denoting a unit's exposure to the shocks. Another possibility lets the $\boldsymbol{\gamma}_i$ represent time-invariant characteristics with a marginal effect on the outcome $\boldsymbol{F}_t$ that changes over time.[2] Note that this model nests the standard two-way error model when $\boldsymbol{F}_t' = (\lambda_t, 1)$ and $\boldsymbol{\gamma}_i' = (1, \mu_i)$; that is, $\boldsymbol{F}_t'\boldsymbol{\gamma}_i = \lambda_t + \mu_i$. The interactive structure allows for more general patterns of unobserved heterogeneity. Importantly, we allow for treatment to be correlated with a unit's exposure to macroeconomic shocks via their factor loadings $\boldsymbol{\gamma}_i$.

For a concrete example, our empirical application focuses on estimating the effect of Walmart store openings on county-level employment. Estimation of a standard two-way fixed effect event-study model suggests that Walmart opened stores in counties that had higher retail employment growth prior to the opening (e.g. Neumark et al. (2008)). In Figure 2, we present an event-study graph and overlay a line of best fit on the pre-treatment estimates. That the line is positive sloping and the estimates are different from zero at the 5% level suggests that estimated positive impacts are due to pre-existing trends rather than the effect of Walmart per se. However, there seems to be a discrete jump when the Walmart opened. The goal then is to remove these pre-existing trends to isolate the treatment effect. It is plausible to assume that during their period of mass expansion, Walmart selected appealing locations based on their local demographic background and national economic trends, while ignoring transitory local economic shocks. Our framework allows this type of selection mechanism and effectively 'controls' for these pre-existing trends.

Our main treatment effect estimator only requires fixed-T consistent estimates of the column space of $\boldsymbol{F}_t$. Using the estimated factors, we compute a matrix that projects the pre-treatment outcomes onto the estimated post-treatment factors, imputing the untreated potential outcome for treated units. Averaging over the difference between the post-treatment observed outcomes and the estimated untreated potential outcomes gives a consistent estimator of average treatment effects. In specifications that include the two-way error model, we show how to explicitly remove the additive fixed effects with a double-demeaning transformation that maintains the common factor structure across treated groups and the never-treated group.

There are two major benefits of our general identification argument. First, fixed-T consistent

a unit start treatments.

2. Ahn et al. (2013) suggest a wage equation where $\boldsymbol{\gamma}_i$ are unobserved worker characteristics of an individual and $\boldsymbol{F}_t$ are their time-varying prices or returns to those characteristics. See Bai (2009) for a collection of economic examples that justify the inclusion of a factor structure.

estimation of $\boldsymbol{F}_t$ is possible through a variety of approaches, most popularly quasi-differencing (Ahn et al., 2001, 2013; Callaway and Karami, 2023) and cross-sectional averages (Westerlund et al., 2019; Juodis and Sarafidis, 2022a,b; Brown et al., 2023). These techniques allow the user to tailor their factor estimator to the specific data and problem under consideration, including how many pre-treatment time periods are available. Our identification result provides a recipe for using any consistent estimator of the factors to estimate treatment effects, opening up the large factor-model literature for causal inference methods. Second, our imputation method allows researchers to graph the estimated counterfactual untreated potential outcomes and the observed outcomes for treated units as a visual check for the parallel trends assumption, similar to a synthetic control plot.

We derive asymptotic properties of an imputation estimator with factor proxies that contain the true unobserved factors in their column space. The resulting estimator takes the form of a generalized method of moments (GMM) estimator, which allows estimation and inference via common statistical software. We implement the estimator using the quasi-long-differencing factor estimator of Ahn et al. (2013) because it is consistent when the number of pre-treatment time periods is small. One advantage of this estimator is that we can form statistical tests for the consistency of the two-way fixed effects (TWFE) estimator.

Relation to Literature

Current estimators that allow for selection based on a factor model either require (i) the number of time periods available is large, e.g. synthetic control (Abadie, 2021), factor-model imputation via principal components (Gobillon and Magnac, 2016; Xu, 2017; Bai and Ng, 2021), and the matrix completion method (Athey et al., 2021; Fernández-Val et al., 2021); or (ii) that an individual's error term u_{it} is uncorrelated over time (Feng, 2020; Imbens et al., 2021).[3] Both of these restrictions are non-realistic in many applied microeconomic data sets where the number of time periods is much smaller than the number of units and serial correlation of shocks is expected. Further, large-T estimators often place restrictions on the dynamic heterogeneity of treatment. Our method requires neither large T nor error term restrictions.

3. Imbens et al. (2021) allow correlation within the post- and pre-treatment sets of the idiosyncratic errors, but assume independence between the two sets. This assumption is still strong in a static modeling context.

A recent set of papers has proposed 'imputation' in the two-way fixed effects setting (e.g. Borusyak et al., 2024; Gardner, 2021; Wooldridge, 2021).[4] While these estimators are consistent in fixed-T settings, these approaches only allow for level fixed effects and preclude interactions like in equation (1). Borusyak et al. (2024) allow a structure similar to equation (1) but requires the the factors $\boldsymbol{F}_t$ be observed. We generalize these techniques by proposing an estimator that imputes the untreated potential outcomes under the more general (1) with unobserved interactive effects.

Our work also contributes to an emerging literature on adjusting for parallel trends violations in short panels. Freyaldenhoven et al. (2019) propose a similar instrumental variable type estimator in the presence of time-varying confounders. Their results rely importantly on homogeneous treatment effects. Their simulations show that heterogeneous treatment effects bias their estimates severely, while our estimator allows for arbitrary time heterogeneity. The most similar paper to our current approach is Callaway and Karami (2023), who also allow for heterogeneous effects in short panels. They prove identification using a similar strategy to QLD and instrumental variables and derive asymptotic normality assuming the number of time periods is fixed. They require time-invariant instruments whose effects on the outcome are constant over time. Their instruments are valid for the QLD estimator in our application, but we also allow for time-varying covariates as instruments. They do not provide a general identification scheme like ours and so their results do not readily extend to other estimators like principal components or common correlated effects.

The rest of the paper is divided into the following sections: Section 2 describes the theory behind our methods and presents identification results of the group-specific dynamic treatment effect parameters. Section 3 provides the main asymptotic theory for a particular QLD estimator. We also discuss practical concerns for practitioners. We include a small Monte Carlo experiment in Section 4 to examine the finite-sample performance of our estimator. Finally, Section 5 contains our application and Section 6 leaves with some concluding remarks.

4. The imputation procedure has been proposed in various settings in causal inference, see Imbens and Rubin (2015).

2 — Model and Identification

We assume a panel data set with units $i = 1, \ldots, N$ and periods $t = 1, \ldots, T$. Treatment turns on in different periods for units in different groups; we denote these groups by the period they start treatment. For each unit, we define G_i to be unit i's group with possible values $\{g_1, \ldots, g_G\} \equiv \mathcal{G} \subseteq \{2, \ldots, T\}$. We follow Callaway and Sant'Anna (2021) and denote $G_i = \infty$ for units that never receive treatment. We assume that $0 < P(G_i = g) < 1$ for all $g \in \mathcal{G} \cup \{\infty\}$, so that the number of individuals in each group and the never-treated group grow with N. Treated potential outcomes are a function of group-timing. We denote $y_{it}(g)$ as the treated potential outcome for unit at i at time t if they were treated at time g. For treatment indicators, we define the vector of treatment statuses $\boldsymbol{d}_i = (d_{i1}, ..., d_{iT})$ where $d_{it} = \mathbf{1}(t \geq G_i)$ and the indicator $D_{ig} = \mathbf{1}(G_i = g)$ if unit i is a member of group g. Let $T_0 = \min_j\{g_j\} - 1$ be the last period before the earliest treatment adoption.

Following Callaway and Sant'Anna (2021), we aim to estimate group-time average treatment effects on the treated:

$$\text{ATT}(g, t) = \tau_{gt} \equiv \mathbb{E}(y_{it}(g) \mid G_i = g) - \mathbb{E}(y_{it}(\infty) \mid G_i = g) \tag{2}$$

These quantities represent the average effect of treatment at time t for units that start treatment in period g for $t \geq g$. Once these quantities are obtained, it is trivial to estimate other aggregations, including averaging over all post-treatment observations to estimate an overall ATT, and averaging over (i, t) where $t - G_i = \ell$ to estimate event-study estimands ATT^{ℓ}'s. We discuss these and other extensions from Callaway and Sant'Anna (2021) in Section 3.

We now state our main identifying assumptions.

Assumption 1 (Sampling). The random vectors $\{(\boldsymbol{d}_i, \boldsymbol{\gamma}_i, \boldsymbol{u}_i)\}$ are randomly sampled from an infinite population and have finite moments up to the fourth order. ■

Assumption 2 (Untreated potential outcomes). The untreated potential outcomes take the form

$$y_{it}(\infty) = \boldsymbol{F}_t'\boldsymbol{\gamma}_i + u_{it}$$

where $\mathbb{E}(u_{it} \mid \boldsymbol{d}_i, \boldsymbol{\gamma}_i) = 0$ for $t = 1, ..., T$. ■

Assumption 3 (No anticipation). For all units i and groups $g \in \mathcal{G}$, $y_{it} = y_{it}(\infty)$ for $t < g$. ■

Assumption 2 imposes a factor-model for the untreated potential outcomes. We discuss the inclusion of covariates and the subsequent relaxation of assumption 2 in the Appendix. We allow for heterogeneous and dynamic treatment effects of any form, i.e. $y_{it}(g) = \tau_{igt} + y_{it}(\infty)$. We also allow arbitrary serial correlation among the idiosyncratic errors.[5] We assume the common factors $\boldsymbol{F}_t$ are nonrandom parameters and the number of factors p is fixed in the asymptotic analysis.

Assumption 2 is more general than the standard difference-in-differences parallel trend assumption since we include the factor structure in our potential outcome model. In particular, it assumes that the error term is uncorrelated with treatment status *after* controlling for the factor loadings. Treatment can still be correlated with contemporaneous shocks so long as the shocks, but not necessarily the exposure to them, are 'common' across the sample. For example, our identification strategy is valid if workers select into a job training program based on their exposure (or adaptability) to macroeconomic productivity shocks. Importantly, we put no restrictions on the time series dependence of $\boldsymbol{u}_i$ and allow for arbitrary heteroskedasticity.

The two-way error model cannot generally accommodate differential exposure.[6] In the more general factor model and Assumption 2, changes in untreated potential outcomes are given by

$$\mathbb{E}(y_{it}(\infty) - y_{it-1}(\infty) \mid G_i = g) = \lambda_t + (\boldsymbol{F}_t - \boldsymbol{F}_{t-1})' \, \mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g)$$

Unless either (i) the factor loadings have the same mean across treatment groups, $\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) = \mathbb{E}[\boldsymbol{\gamma}_i]$, or (ii) the factors are time-invariant, then the standard parallel trends assumption that the group g and the never-treated group follow common trends would not hold. If either of the two cases hold for all g and t, the two-way error model is correctly specified.[7] However, these are knife-edge cases which are not the focus of the paper. Our Assumption 2 allows for the factor loadings to be correlated with treatment timing and opens up treatment effect estimation for a much broader set of empirical questions.

The key econometric challenge lies in that we do not observe $y_{it}(\infty)$ whenever $d_{it} = 1$. Our goal is to consistently estimate $\mathbb{E}(y_{it}(\infty) \mid G_i = g)$ under equation (1) to consistently estimate

5. This condition may need to be strengthened for inference when $T \to \infty$.
6. The following derivation is also shown in Callaway and Karami (2023), but we are repeating it here for exposition.
7. We explicitly prove this result later.

group-time average treatment effects. Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024) implicitly rely on this insight in studying the two-way error model.

Prior attempts at estimating average treatment effects in a factor-model setting focus on finding conditions that allow for estimation of $\boldsymbol{\gamma}_i$ and $\boldsymbol{F}_t$ jointly as in Gobillon and Magnac (2016), Xu (2017), and Bai and Ng (2021), or a generalized version of a factor model as in Feng (2020) and Arkhangelsky et al. (2021). These techniques require the number of pre-treatment periods to grow to infinity and often place restrictions on both the dynamics of the treatment effects' distribution and the serial dependence among the idiosyncratic errors. Instead, we pursue identification noting that

$$\mathbb{E}(y_{it}(\infty) \mid G_i = g) = \boldsymbol{F}_t' \, \mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) \tag{3}$$

Therefore, we only need to estimate the *average* of the factor loadings among a treatment group, which we can always do even with a small number of post-treatment time periods. We can then accommodate either a large or small number of pre-treatment periods and allow for estimation using a broad range of known strategies.

2.1. *ATT*(g, t) *Identification*

We begin by describing the intuition behind our identification result. Consider a unit subject to treatment at time g. Define $\boldsymbol{y}_{i,t<g}$ and $\boldsymbol{y}_{i,t\geq g}$ as respectively the first $(g-1)$ and last $(T-g+1)$ outcomes for unit i, or the 'pre-treatment' and 'post-treatment' outcomes. Define $\boldsymbol{F}$ to be the matrix of factor shocks with rows given by $\boldsymbol{F}_t$. We similarly define $\boldsymbol{F}_{t<g}$ and $\boldsymbol{F}_{t\geq g}$ as the first and last rows of matrix $\boldsymbol{F}$. Equation (3) implies

$$\mathbb{E}(\boldsymbol{y}_{i,t<g}(\infty) \mid \boldsymbol{G}_i = g) = \boldsymbol{F}_{t<g} \, \mathbb{E}(\boldsymbol{\gamma}_i \mid \boldsymbol{G}_i = g) \tag{4}$$

If the factors were observed, we could consistently estimate the mean values of the p-vector of average factor loadings for treated group $G_i = g$ via ordinary least squares. More formally, if $\text{Rank}(\boldsymbol{F}_{t<g}) = p$, the coefficient from the population regression of $\mathbb{E}(y_{i,t<g}(\infty) \mid G_i = g)$ on $\boldsymbol{F}_{t<g}$ is $\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g)$. Equation (3) also gives us

$$\mathbb{E}(\boldsymbol{y}_{i,t\geq g}(\infty) \mid \boldsymbol{G}_i = g) = \boldsymbol{F}_{t\geq g} \, \mathbb{E}(\boldsymbol{\gamma}_i \mid \boldsymbol{G}_i = g) \tag{5}$$

for the post-treated outcomes. Because we assume $\boldsymbol{F}$ is known (for now), we can predict $\mathbb{E}(\boldsymbol{y}_{i,t} \mid G_i = g)$ for $t \geq g$ by multiplying $\boldsymbol{F}_t$ by the OLS estimate from the prior infeasible regression. We then obtain $\mathbb{E}(y_{it}(\infty) \mid G_i = g)$ for the post-treatment outcomes, which we can subtract from y_{it} and average over the respective sample to obtain an estimate of $\text{ATT}(g, t)$.

We now define a useful matrix function for a more formal derivation of our main result. Given matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ that are respectively $n \times k$ and $m \times k$, suppose $\text{Rank}(\boldsymbol{X}_0) = k$. We define the *imputation matrix*

$$\boldsymbol{P}(\boldsymbol{X}_1, \boldsymbol{X}_0) \equiv \boldsymbol{X}_1(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0' \tag{6}$$

This matrix takes a similar form to a projection matrix but "imputes" the fitted values from regressing on $\boldsymbol{X}_0$ onto a different matrix $\boldsymbol{X}_1$. The next theorem provides our main identification result:

Theorem 1. Suppose $\boldsymbol{F}$ is known and $\text{Rank}(\boldsymbol{F}_{t \leq T_0}) = p$. Under Assumptions 1, 2, and 3 for all $g \in \mathcal{G}$,

$$\text{ATT}(g, t) = \mathbb{E}(y_{it} - \boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g) \tag{7}$$

for $t \geq g$.

Moreover, let $\boldsymbol{F}^*$ be a full rank $T \times m$ matrix where $m < T_0$ and $\boldsymbol{F} \in \text{col}(\boldsymbol{F}^*)$, the column space of $\boldsymbol{F}^*$. Then the imputation matrix is invariant to $\boldsymbol{F}^*$

$$\boldsymbol{P}(\boldsymbol{F}_t^{*'}, \boldsymbol{F}_{t<g}^*)\boldsymbol{F}_{t<g}\boldsymbol{\gamma}_i = \boldsymbol{F}_t'\boldsymbol{\gamma}_i \tag{8}$$

■

All proofs are contained in the Appendix. Theorem 1 shows that we can identify the ATTs if we know the factor matrix. The second part of the theorem suggests that any rotation of the true factor matrix, $\boldsymbol{F}$, can be used in the imputation matrix. This result is important because it is well understood that $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are not separately identified (Bai, 2009; Ahn et al., 2013; Xu, 2017). All of the estimators discussed so far can at best approximate the column space of the factors because both $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are unobserved. The second part of the theorem shows that our identification scheme allows for this class of estimators. To see how, note that $\boldsymbol{F} \in \text{col}(\boldsymbol{F}^*)$ implies the existence

of a $m \times p$ matrix $\boldsymbol{A}$ such that $\boldsymbol{F}^* \boldsymbol{A} = \boldsymbol{F}$. Thus

$$\begin{aligned}
\boldsymbol{F}_t^{*\prime} \left(\boldsymbol{F}_{t<g}^{*\prime} \boldsymbol{F}_{t<g}^{*} \right)^{-1} \boldsymbol{F}_{t<g}^{*\prime} \boldsymbol{F}_{t<g} &= \boldsymbol{F}_t^{*\prime} \left(\boldsymbol{F}_{t<g}^{*\prime} \boldsymbol{F}_{t<g}^{*} \right)^{-1} \boldsymbol{F}_{t<g}^{*\prime} \boldsymbol{F}_{t<g}^{*} \boldsymbol{A} \\
&= \boldsymbol{F}_t^{*\prime} \boldsymbol{A} \\
&= \boldsymbol{F}_t'
\end{aligned}$$

We only require $m \geq p$ for this result because we only need $\boldsymbol{F}$ to be in the column space of $\boldsymbol{F}^*$.

We view Theorem 1 as an extension of earlier treatment effect identification results to the factor model. The identification argument in equation (7) is similar to the bridge function argument of Imbens et al. (2021), but does not require restrictions on the time series dependence of the outcomes because we put structure on the non-parallel trending (i.e. factor model). It can also be seen as an extension of the imputation arguments from Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024) who study the additive error model. In fact, Gardner (2021) and Borusyak et al. (2024) implicitly use the imputation matrix but with known factors.

Theorem 1 shows we can apply these conclusions to any estimator that achieves fixed-T consistency by asymptotically spanning the factor space. The most popular classes of these estimators examples include the common correlated effects (Pesaran, 2006) and quasi-differencing (Ahn et al., 2013). Westerlund (2020) shows that principal components can also fit in this class, so long as strong restrictions are met. As long as the column space of the factors are consistently estimated using the control sample, dynamic ATTs are identified as in Theorem 1, regardless of the normalization used for estimation.

To present a general framework for the estimation of the factors, we formally present the identifying assumptions needed for factor space estimators. We assume the factor estimator can be defined in terms of a finite-dimensional moment function that itself is a function of a finite-dimensional set of parameters: $\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})$. The '$\infty$' subscript denotes the fact that $\boldsymbol{g}$ is only a function of the never-treated group because this group can identify the entire factor space. Here, $\boldsymbol{\theta}$ denotes a set of parameters that can control for the entirety of the factor space; we give examples of $\boldsymbol{\theta}$ below. We again stress that while our identification argument can extend to settings where T grows, our current asymptotic results are only appropriate for fixed-T applications because the number of parameters and moment conditions is finite.

Assumption 4. There exists a unique $q \times 1$ vector of parameters $\boldsymbol{\theta}$ and a $T \times m$ function $\boldsymbol{F}(\boldsymbol{\theta})$ such that the following conditions hold:

(i) For some full-rank matrix $\boldsymbol{A}$, $\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{A} = \boldsymbol{F}$ where $\text{Rank}(\boldsymbol{F}(\boldsymbol{\theta})) = m < T_0$

(ii) There is a $s \times 1$ vector of moment functions $\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})$ such that

$$\mathbb{E}(\boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \mid G_i = \infty) = \boldsymbol{0} \tag{9}$$

(iii) Let $\boldsymbol{D}_\infty = \mathbb{E}(\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \mid G_i = \infty)$. Then $\text{Rank}(\boldsymbol{D}_\infty) = q$.

(iv) $\mathbb{E}(\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})' \mid G_i = \infty)$ is positive definite.

Assumption 4 is written with fixed-T estimation and inference in mind. As mentioned before, accommodating principal components estimation requires additional restrictions, as well as a large time series in the pre-treatment periods. However, the general identification result is the same and our estimator is still valid for estimating dynamic effects in the post-treatment period. Part (i) implies that the estimated factors can be reduced to a finite dimension of estimable parameters. The matrix $\boldsymbol{A}$ is the full rank linear rotation that turns $\boldsymbol{F}(\boldsymbol{\theta})$ into $\boldsymbol{F}$. For the example estimators expressed above, $\boldsymbol{F}(\boldsymbol{\theta})$ asymptotically spans the unknown factors, $\boldsymbol{F}$. Parts (ii)-(iv) imply the parameters $\boldsymbol{\theta}$ are identified and consistently estimable and are standard in GMM applications. The parameters $\boldsymbol{\theta}$ themselves are often the result of an underlying normalization like in Ahn et al. (2001, 2013), Juodis and Sarafidis (2022a,b), and Callaway and Karami (2023). Sometimes they are population moments estimated by cross-sectional averages like in Westerlund et al. (2019) and Brown et al. (2023); e.g. $\boldsymbol{\theta} = \text{vec}(\mathbb{E}[\boldsymbol{X}_i])$.[8]

2.2. Quasi-Long-Differencing

A leading example of a set of moment equations for factor-space estimation is the quasi-long differencing (QLD) estimator of Ahn et al. (2013). They propose a QLD transformation given by

$$\boldsymbol{H}(\boldsymbol{\theta}) = (\boldsymbol{I}_{T-p}, \boldsymbol{\Theta}) \tag{10}$$

8. The common correlated effects model requires stronger conditions for consistency, like the existence of time-varying covariates that are linear in the same common factors $\boldsymbol{F}$. Assumption 4 would require that the expected value of the data $[\boldsymbol{y}_i, \boldsymbol{X}_i]$ has rank p so that their cross-sectional averages are consistent for the factor space. See our companion paper Brown et al. (2023) for details.

where $\boldsymbol{\Theta}$ is a $(T-p)\times p$ matrix of unrestricted parameters and $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\theta})$[9]. They normalize the factors as

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Theta} \\ -\boldsymbol{I}_p \end{pmatrix} \tag{11}$$

so that $\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{F}\boldsymbol{\gamma}_i = \mathbf{0}$ by construction. We modify their proposed moment conditions to use just the never-treated group:

$$\mathbb{E}(g_{i\infty}(\boldsymbol{\theta}) \mid G_i = \infty) = \mathbb{E}(\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{y}_i \otimes \boldsymbol{w}_i \mid G_i = \infty) = \mathbf{0} \tag{12}$$

where $\boldsymbol{w}_i$ is a vector of instruments that are exogenous with respect to the idiosyncratic error in Assumption 2 but correlated with $\boldsymbol{\gamma}_i$; see the Appendix for a discussion of the identifying conditions. Essentially, we require the instruments to be strictly exogenous with the defactored errors, but correlate strongly with the factor loadings $\boldsymbol{\gamma}_i$. We also require there be at least as many instruments as factors. We discuss the choice of instruments $\boldsymbol{w}_i$ in more practical terms in Section 5.

While both approaches are valid in the first stage of our setting, we use the Ahn et al. (2013) estimator because it is more general than Callaway and Karami (2023). For one, they allow for a larger set of instruments. One identification strategy proposed by Callaway and Karami (2023) requires time-invariant covariates whose effects on y_{it} are independent of time, meaning the researcher must decide which of the time-invariant observables have constant effects on the outcome. Ahn et al. (2013) can allow for arbitrary time effects on covariates while still using those covariates as instruments. Ahn et al. (2013) also give a road map to estimation based on weakly exogenous covariates that allows for dynamic modeling. This aspect of the estimator is left for future research.

Because the QLD matrix is a function of p, we need a consistent estimator of the number of common factors. Ahn et al. (2013) propose a number of consistent estimators of p. First, they

9. We reuse the "$\boldsymbol{\theta}$" notation throughout the remainder of the text.

consider the usual Hansen-Sargan over-identifying test restriction. Let

$$J(\boldsymbol{\theta}) = \left(N_\infty^{-1} \sum_{i=1}^{N} D_{i\infty} \boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \right)' \left(N_\infty^{-1} \sum_{i=1}^{N} D_{i\infty} \boldsymbol{g}_{i\infty}(\tilde{\boldsymbol{\theta}}) \boldsymbol{g}_{i\infty}(\tilde{\boldsymbol{\theta}}) \right)^{-1} \left(N_\infty^{-1} \sum_{i=1}^{N} D_{i\infty} \boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \right) \tag{13}$$

where $\tilde{\boldsymbol{\theta}}$ is an initial consistent estimator of $\boldsymbol{\theta}$ for the given p being tested. Letting $\widehat{\boldsymbol{\theta}}$ minimize the statistic above, Ahn et al. (2013) show that $N_\infty J(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{(T-p)(q-p)}$ where q is the number of instruments in $\boldsymbol{w}_i$, which we can see implicitly must be at least as large as p. One can start at $p = 0$ then continue to increase p until rejection[10]. See Section 3 of Ahn et al. (2013) for further discussion. There is currently no formal derivation of the limiting properties of $\widehat{\boldsymbol{\theta}}$ when the practitioner overestimates p. However, a number of simulation studies (Ahn et al., 2013; Breitung and Hansen, 2021; Brown, 2023) have shown that QLD estimators still have favorable finite-sample properties when p is overestimated. We also demonstrate via simulations in Section 4 that our estimator performs well when p is estimated in this manner.

2.3. Two-Way Error Model

We now demonstrate how to explicitly nest the standard two-way error model. While this structure is a special case of the factor model studied above, we consider the special case for two main reasons. First, eliminating the additive effects saves degrees of freedom to estimate the factor model and provides efficiency by reducing the burden on the first-stage factor estimator. Second, a thorough study of the additive model will provide insight into the link between TWFE estimation and more complicated and computationally involved factor model estimation. It will also allow us to show when TWFE estimation is consistent in the presence of interactive fixed effects.

We first note that care must be taken when eliminating additive effects so that the overall factor structure is preserved. The methods in Borusyak et al. (2024), Gardner (2021), and Wooldridge (2021) that estimate the additive effects using the full untreated sample (never-treated and not-yet-treated observations) will not maintain a common factor structure between the untreated and treated groups. For example, consider the first order conditions from the regression of $(1 - d_{it})y_{it}$ on unit and time effects. The estimators for the unit effect of a unit treated at time g and a

10. One must choose a rejection level b_{N_∞} such that $b_{N_\infty} \to 0$ and $-\ln(b_{N_\infty})/N_\infty \to 0$ as $N_\infty \to \infty$. See Cragg and Donald (1997).

never-treated unit respectively satisfy

$$\sum_{t=1}^{g-1}(y_{it} - \widehat{\lambda}_t - \widehat{\mu}_i) = 0 \tag{14}$$

$$\sum_{t=1}^{T}(y_{it} - \widehat{\lambda}_t - \widehat{\mu}_i) = 0 \tag{15}$$

The control sample will remove more time averages than in every treated sample, meaning the factors are demeaned using different subsamplesamples. As such, the transformed factors are not equal across groups and we cannot use the control sample to estimate the factors for the treated samples.

We first define the following averages for the purpose of removing the additive effects:

$$\overline{y}_{\infty,t} = \frac{1}{N_\infty}\sum_{i=1}^{N} D_{i\infty} y_{it} \tag{16}$$

$$\overline{y}_{i,t\leq T_0} = \frac{1}{T_0}\sum_{t=1}^{T_0} y_{it} \tag{17}$$

$$\overline{y}_{\infty,t<T_0} = \frac{1}{N_\infty T_0}\sum_{i=1}^{N}\sum_{t=1}^{T_0} D_{i\infty} y_{it} \tag{18}$$

where $\overline{y}_{\infty,t}$ is the cross-sectional averages of the never-treated units for period t, $\overline{y}_{i,t\leq T_0}$ is the time-averages of unit i before any group is treated, and $\overline{y}_{\infty,t<T_0}$ is the total average of the never-treated units before any group is treated.

We then perform all estimation on the residuals $\tilde{y}_{it} \equiv y_{it} - \overline{y}_{\infty,t} - \overline{y}_{i,t<T_0} + \overline{y}_{\infty,t<T_0}$. These residuals are reminiscent of the usual TWFE residuals, except we carefully select this transformation to accomplish two things. First, this transformation leaves the treatment dummy variables unaffected to prevent problems with negative weighting when aggregating heterogeneous treatment effects (Goodman-Bacon, 2021; Borusyak et al., 2024). Second, it preserves a common factor structure for all units and time periods[11]. The TWFE imputation estimator of Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024) would not share this property because they estimate μ_i and λ_t based on the full sample $d_{it} = 0$, while we use a specific subsample.

11. Such a transformation should not be used when considering the common correlated effects estimator because it would violate the CCE rank condition. See Brown et al. (2023).

This result is summarized in the following lemma:

Lemma 1. $\mathbb{E}(\tilde{y}_{it} \mid G_i = g) = \mathbb{E}\big(d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_{\infty}) \mid G_i = g\big)$ for $t = 1, ..., T$ and $g \in \mathcal{G} \cup \{\infty\}$ where $\overline{\boldsymbol{F}}_{t<T_0}$ is the average of $\boldsymbol{F}_t$ in the pre-treatment periods and $\overline{\boldsymbol{\gamma}}_{\infty}$ is the average of $\boldsymbol{\gamma}_i$ among the control units. ■

Lemma 1 demonstrates how to explicitly nest the two-way error model model while allowing for a general common factor structure. Since we are not interested in inference on the factors themselves, this form will suffice for the imputation process. The transformed outcomes take the form

$$\tilde{y}_{it} = d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_{\infty}) + \tilde{u}_{it}. \tag{19}$$

For ease of exposition, we rewrite the above equation as:

$$\tilde{y}_{it} = d_{it}\tau_{it} + \tilde{\boldsymbol{F}}_t'\tilde{\boldsymbol{\gamma}}_i + \tilde{u}_{it}. \tag{20}$$

Lemma 1 has the added benefit of showing us when the ATTs are identified by our TWFE transformation alone.

Corollary 1. Under Assumptions 1-3, $\text{ATT}(g, t)$ is identified by the fixed effects imputation transformation if $\mathbb{E}(\gamma_i \mid G_i = g) = \mathbb{E}[\gamma_i]$ for all $g \in \mathcal{G} \cup \{\infty\}$. ■

This result is an immediate consequence of Assumptions 1 – 3 as $\mathbb{E}(\boldsymbol{\gamma}_j \mid G_i = g) = \mathbb{E}[\boldsymbol{\gamma}_i]$ for $j \neq i$ under random sampling. Corollary 1 tells us that TWFE imputation is sufficient to estimate the ATTs, even when a non-trivial factor structure exists, so long as the average factor loadings do not differ systemically with treatment status. Asymptotic normality of our imputation procedure under a two-way error model is studied in the Appendix. We also provide simple tests for mean independence of the factor loadings in Remark 4, i.e. consistency of the TWFE estimator. However, if the researcher believes a TWFE estimator is sufficient, they should use one of the other techniques mentioned above. Our method sacrifices potential efficiency by not using all observations to eliminate the additive effects in order to allow for additional interactive effects.

3 — Estimation and Inference

This section considers estimation of the group-time average treatment effects. A major benefit of our approach is the simplicity of inference while allowing for a large number of possible estimation techniques in the first stage. Our moment conditions lead to a simple GMM estimator for which inference is standard and can be computed via routine packages in common statistical software. Further, we can use the moment conditions to test the fundamental features of the model.

3.1. Asymptotic Normality

Equations (7) and (9) provide us with the necessary moment conditions to estimate the ATTs. We collect them here in their unconditional form:

$$\mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty}=1)}\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\right]=\mathbf{0}$$
$$\mathbb{E}[\boldsymbol{g}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})]=\mathbb{E}\left[\frac{D_{ig_G}}{\mathbb{P}(D_{ig_G}=1)}\left(\boldsymbol{y}_{i,t\geq g_G}-\boldsymbol{P}(\boldsymbol{F}_{t\geq g_G}(\boldsymbol{\theta}),\boldsymbol{F}_{t<g_G}(\boldsymbol{\theta}))\boldsymbol{y}_{i,t<g_G}-\boldsymbol{\tau}_{g_G}\right)\right]=\mathbf{0}$$
$$\vdots$$
$$\mathbb{E}[\boldsymbol{g}_{i1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})]=\mathbb{E}\left[\frac{D_{ig_1}}{\mathbb{P}(D_{ig_1}=1)}\left(\boldsymbol{y}_{i,t\geq g_1}-\boldsymbol{P}(\boldsymbol{F}_{t\geq g_1}(\boldsymbol{\theta}),\boldsymbol{F}_{t<g_1}(\boldsymbol{\theta}))\boldsymbol{y}_{i,t<g_1}-\boldsymbol{\tau}_{g_1}\right)\right]=\mathbf{0}$$

where $\boldsymbol{\tau}_g=(\tau_{gg},...,\tau_{gT})'$ is the vector of post-treatment treatment effects. We stack these over g as $\boldsymbol{\tau}=(\boldsymbol{\tau}_{g_1}',...,\boldsymbol{\tau}_{g_G}')'$. The first set of moment conditions identify the factor space by Assumption 4 and the remaining moments identify the τ_{gt} via our imputation method. Implementation requires replacing $\mathbb{P}(D_{ig}=1)$ with its sample counterpart N_g/N. This setting can also accommodate cases as in Hahn et al. (2018) where the factor structure is estimated nonparametrically in the first stage but the parametric estimator in the second stage is still $O_p(N^{-1/2})$. We leave this case for future study.

We need one final regularity condition to implement the asymptotically efficient GMM estimator:

Assumption 5. $\mathbb{E}[\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g)\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g)']$ is positive definite for each $g\in\mathcal{G}$. ■

We collect the moment functions into the vector $\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})=(\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})',\boldsymbol{g}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})',...,\boldsymbol{g}_{ig_1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})')'$. In an abuse of notation, we assume $\boldsymbol{g}_{i\infty}$ is the moment function from equation (9) but scaled by

$D_{i\infty}/P(D_{i\infty}=1)$. We define $\boldsymbol{\Delta} = \mathbb{E}[\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})']$ which is positive definite by Assumptions 4 and 5. Then our GMM estimator $(\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')'$ solves

$$\min_{\boldsymbol{\theta},\boldsymbol{\tau}} \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right)' \widehat{\boldsymbol{\Delta}}^{-1} \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right) \tag{21}$$

where $\widehat{\boldsymbol{\Delta}} \xrightarrow{p} \boldsymbol{\Delta}$ uses an initial consistent estimator of $(\boldsymbol{\theta}', \boldsymbol{\tau}')'$.

Theorem 2. Under Assumptions 1-5, $\sqrt{N}\big((\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')' - (\boldsymbol{\theta}', \boldsymbol{\tau}')'\big)$ is jointly asymptotically normal as $N \to \infty$ and

$$\begin{aligned}
\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &\xrightarrow{d} N\left(\mathbf{0}, \left(\boldsymbol{D}_{\infty}'\boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty}\right)^{-1}\right) \\
\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_G} - \boldsymbol{\tau}_{g_G}) &\xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Delta}_{g_G} + \boldsymbol{D}_{g_G}\left(\boldsymbol{D}_{\infty}'\boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty}\right)^{-1}\boldsymbol{D}_{g_G}'\right) \\
&\vdots \\
\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_1} - \boldsymbol{\tau}_{g_1}) &\xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Delta}_{g_1} + \boldsymbol{D}_{g_1}\left(\boldsymbol{D}_{\infty}'\boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty}\right)^{-1}\boldsymbol{D}_{g_1}'\right)
\end{aligned}$$

where $\boldsymbol{D}_g$ is the gradient of group g's moment function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}_g$ is the variance of group g's moment function. Further, the asymptotic covariance between $\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_h} - \boldsymbol{\tau}_{g_h})$ and $\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_k} - \boldsymbol{\tau}_{g_k})$ is given by $\boldsymbol{D}_{g_h}(\boldsymbol{D}_{\infty}'\boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty})^{-1}\boldsymbol{D}_{g_k}'$. We derive the functional forms of the various matrices for the QLD estimator in the Appendix[12]. ■

Valid inference is easy to obtain because we use a GMM framework. Analytic standard errors are computed and reported by most routine statistical packages implementing GMM estimation. We achieve the desired $\sqrt{N}$-convergence rate because we assume a $\sqrt{N}$-convergent estimator of the factor proxies. Examples include quasi-differencing, common correlated effects, and even principal components, though the latter also requires T to go to infinity for the asymptotic results to hold. Because we have proved asymptotic normality, one can also use the usual nonparametric panel bootstrap. We derive an asymptotically linear representation of the ATT estimates in the Appendix that also allow for the multiplier bootstrap and uniform inference as in Callaway and Karami (2023).

12. We also include in the appendix a derivation of the asymptotic variance when QLD is used to estimate the factors.

The asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$ generally depends on the estimation of $\boldsymbol{\theta}$ in the first stage by the term $D_g(D_\infty' \Delta_\infty^{-1} D_\infty)^{-1} D_g'$. We can see directly from Theorem 2 that a smaller Avar$(\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))$ leads to a smaller Avar$(\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g))$ (in the matrix sense) when $\boldsymbol{D}_g$ has full rank. This result also suggests that more efficient estimation of the factors is an important avenue of future work and demonstrates why our general identification result is so powerful: we can use different estimators of the factors if we believe we can achieve substantial efficiency gains.

Estimation of $\boldsymbol{\tau}_g$ is not dependent on the first stage estimation of $\boldsymbol{\theta}$ when $\boldsymbol{D}_g = \mathbf{0}$. A sufficient condition for this equality occurs when the transformed factor loadings for group g center about zero. The fixed-T common correlated effects analysis of Westerlund et al. (2019) implies such a condition. We may also think this condition holds in certain applications where the factor model is relevant. For example, suppose γ_i is exposure to an information shock f_t such that $\gamma_i \in [0, 1]$ with probability one. If non-institutional investors of a given asset do not have access to privately held limited information, we would expect $\gamma_i \approx 0$ for units in said group. When the gradient $\boldsymbol{D}_g = \mathbf{0}$ for a given g, the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$ is just $\boldsymbol{\Delta}_g$. This quantity is simple to estimate via a nonparametric variance estimator. Let

$$\widehat{\boldsymbol{\Delta}}_g = \frac{1}{N_g - 1} \sum_{i=1}^{N} D_{ig} \left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{g_G} \right) \left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{g_G} \right)' \tag{22}$$

where $\widehat{\boldsymbol{\Delta}}_{ig} = \boldsymbol{y}_{i,t \geq g} - \boldsymbol{P}(\boldsymbol{F}_{t \geq g}(\widehat{\boldsymbol{\theta}}), \boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))\boldsymbol{y}_{i,t<g}$. This estimator is sufficient to generate valid standard errors whenever $\boldsymbol{D}_g = \mathbf{0}$.

Theorem 3. Under Assumptions 1-5, $\widehat{\boldsymbol{\Delta}}_g^{-1} \xrightarrow{p} \boldsymbol{\Delta}_g^{-1}$.

3.2. Extensions

We conclude this section with a few extensions of our estimator to highlight the flexibility of our approach.

Remark 1 (Limited Anticipation)**.** We can relax the limited anticipation assumption by simply redefining the last pre-treatment period as $q_g - 1$ and incorporate the additional $g - q_g$ periods into the moment conditions, so long as there are still enough pre-treatment periods to construct the imputation matrix. Then $\boldsymbol{\tau}_g$ is a $T - q_g + 1$ vector that makes treatment anticipation a testable

hypothesis:

$$H_0 : \tau_{g,q_g} = ... = \tau_{g,g-1} = 0 \tag{23}$$

■

Remark 2 (Other Aggregate Treatment Effects)**.** Our estimation method can handle other aggregations of $y_{it} - \hat{y}_{it}(\infty)$. For example, one could aggregate over all post-treatment (i, t) to estimate an overall ATT or over event-time indicators to estimate aggregate event-study estimates.[13] Researchers can perform heterogeneity analyses by aggregating for units with different values of X_i like gender, race, or age to estimate a conditional ATT. All one needs to do to estimate such aggregate effects is to correctly specify the unconditional treatment effect moment conditions. If there are *a priori* restrictions on treatment effects as in Borusyak et al. (2024), these can be imposed on the moment conditions as well.

Remark 3 (Assessing Model Fit)**.** Our key identifying assumption is that after subtracting off (the estimated) factor model, there is no remaining confounders in the post-treatment periods that are correlated with treatment (a generalized parallel-trends assumption). As is common in the difference-in-differences literature, we can assess the plausibility of this assumption using pre-treatment "placebo" effects. This is done by extending the projection matrix into the pre-treatment periods $\boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g})$. Under the no anticipation assumption,

$$\mathbb{E}((\boldsymbol{I}_g - \boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g}))\, \boldsymbol{y}_{i,t\leq g} \mid G_i = g) = \boldsymbol{0} \tag{24}$$

so that the properly standardized vector of pre-treatment residuals is asymptotically normal and centered at 0. While this is not a formal test, pre-treatment estimates are typically presented to readers to help assess the plausibility of the identifying assumption.

The synthetic control literature provides an alternative procedure that plots the raw outcome data for the treated unit and the synthetic control prediction. Readers then can visually inspect the model fit and see if they believe the synthetic control makes a good counterfactual estimator (Abadie, 2021). Our proposed estimator can be used to produce estimates for $y_{it}(\infty)$ in all periods

13. Alternatively, we allow for aggregation of ATT(g, t) estimates as in Callaway and Sant'Anna (2021) by deriving the influence function in the Appendix.

for the treated observations:

$$\hat{y}_{it}(\infty) = \boldsymbol{P}(\boldsymbol{F}_t, \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} + \overline{y}_{\infty,t} + \overline{y}_{i,t<T_0} - \overline{y}_{\infty,t<T_0} \tag{25}$$

where the first term on the right-hand side imputes $\hat{y}_{it}(\infty)$ and the last three terms in the sum 'undo' the within-transformation[14]. In the pre-treatment periods, our estimates $\hat{y}_{it}(\infty)$ should be approximately equal to the observed y_{it} under our assumptions. Similar to synthetic control estimators, comparing the imputed values to the true value can validate the 'fit' of our model. However, since we have many treated units, doing so unit by unit is not practical. There are two complementary ways to aggregate treated units that will prove useful.

First, one can aggregate over a group and plot the average of y_{it} and the average of $\hat{y}_{it}(\infty)$ separately for each group $g \in \mathcal{G}$. This will create a set of 'synthetic-control' like plots. To produce an 'overall' plot, the observed outcome y_{it} and the estimated untreated potential outcome $\hat{y}_{it}(\infty)$ should be 'recentered' to event-time, i.e. reindex time to $e = t - G_i$, so that treatment is centered at event-time 0. Then y_{ie} and $\hat{y}_{ie}(\infty)$ can be aggregated for each value of event-time e. We produce such a plot in our empirical example. ■

Remark 4 (TWFE Specification Testing)**.** This paper is motivated by the fact that the two-way error model's generality is suspicious in practice. Therefore, we think a test of the two-way error structure versus a more complicated interactive effects model is of practical importance. Ahn et al. (2013) discuss consistent estimation of p. Their tests have a new interpretation under this null hypothesis when testing for p on the residuals $\tilde{y}_{it}$. If Assumption 1 and 2 hold with $\boldsymbol{F}_t'\boldsymbol{\gamma}_i = \mathbf{0}$ almost surely for all (i, t), then $p = 0$.

If the null hypothesis is true, the more computationally burdensome QLD procedure is unnecessary for estimating the ATTs.[15] Even if the two-way error model is unrepresentative of the factor structure, Corollary 1 shows that mean independence of the factor loadings with respect to treatment timing is sufficient for consistency of TWFE. See the Appendix for an additional test of the equality of the factor loadings' conditional means. ■

14. Leave this part out if you do not remove the additive effects by hand.

15. Even if TWFE is consistent, it is not necessarily more efficient than our procedure. See Section 4 for example.

4 — Simulations

We present a brief simulation study to compare our estimator to alternatives in the literature. We specifically study the version of our estimator using the quasi-long differencing estimator from Ahn et al. (2013) as the first stage. Since the focus of this paper is to propose a set of fixed-T estimators while the majority of estimators are large-T based, we will present simulations with $T_0 = 4$ and $T_0 = 10$ and three post-periods. We draw $N = 200$ observations, which is a relatively small number for a nonlinear estimation problem.

We consider the setting with two factors: the first being a unit-specific linear time-trend, $f_{1t} = t$, and the second, f_{2t}, being generated by an AR(1) process with autocorrelation coefficient of 0.75. The factors are kept constant across simulation draws to match our sampling assumptions. For each unit we generate a unit 'fixed effect' as iid with $\mu_i \sim N(0, 1)$ and the two factor loadings to be correlated with the unit fixed effects by drawing individually from $\gamma_i \sim N(\mu_i, 1)$.

We generate untreated potential outcomes either following a general factor model (1) or a two-way fixed effects model. When we use the two-way fixed effects model, we use μ_i as the unit fixed effect and f_{2t} as the time fixed effect. Depending on the simulation, we either draw the error term to be white-noise $N(0, 1)$ or from an AR(1) process with autocorrelation coefficient of 0.75. When the error term is not autocorrelated for individuals, many of the large-T factor model estimators are expected to perform well.

We assign treatment in two ways depending on the simulation. First, we assign treatment completely randomly with probability of treatment at 50% for all units. This mechanism implies that the factor loadings have the same mean (and distribution) for treated and control units. In this case, corollary 1 shows that parallel trends holds even when a factor model generates data. We label this setting as 'parallel trends' in the tables. Alternatively, we generate treatment with probability increasing in the first factor loading such that parallel trends fail (since treated units are more exposed to the time-trend in f_t).[16] Treatment effects grow over time being $1, 2,$ and 3 in each post-treatment period.

We model a researcher as having available 'proxies' for the underlying factor loadings by

16. In particular, we rescale $\tilde{\gamma}_i = \frac{\gamma_i - \min_i \gamma_i}{\max_i \gamma_i - \min_i \gamma_i}$ to be between 0 and 1. Then we rescale $\tilde{\gamma}_i$ to have mean 0.5 so that the probability of treatment stays at 50%.

generating a covariate $w_{i,r} = \gamma_{i,r} + \xi_{i,r}$ where $\xi_{i,r}$ is white-noise measurement error for $r = 1, 2$. The covariates $\boldsymbol{w}_i$ will be used as covariates in some TWFE specifications and as our instrument for our factor-model estimation. In the baseline simulation, we consider the case where $\xi_{i,r} \sim N(0, 1)$. In a set of simulations, we vary the level of noise to see how the instrument strength affects estimates.

We estimate event-study treatment effects using a set of estimators. First, we estimate the two-way error model without covariates using the imputation estimator proposed by Borusyak et al. (2024) and Gardner (2021).[17] Secondly, as is sometimes done by applied researchers, we augment the two-way error model by including a noisy measure of the factor loadings:

$$y_{it}(0) = \mu_i + \lambda_t + \boldsymbol{w}_i \beta_t + u_{it} \tag{26}$$

where β_t allows for trends to vary based on $\boldsymbol{w}_i$.[18] In the case where $\boldsymbol{w}_i = \boldsymbol{\gamma}_i$, i.e. the factors are observable, this model is correctly specified. However, when $\text{Var}(\xi_i) > 0$, i.e. the covariates are noisy measures for the underlying factor loadings, model (26) will only partially absorb the factor model. We compare this method to our estimator using the QLD transformation of Ahn et al. (2013) to estimate the factors.[19] The covariates $\boldsymbol{w}_i$ are our instruments in the first stage to estimate the QLD parameters. See Section 2.2.

Additionally, we consider a set of alternative estimators that are designed for factor models but typically requiring large panels. We use the synthetic control estimator (aggregated across treated units) following Abadie et al. (2010); Ben-Michael et al. (2021); the generalized synthetic control method that is an imputation style estimator that estimates the factor model using principal components (Xu, 2017); and the augmented synthetic control method (Ben-Michael et al., 2021).

In the simulations, we consider a few 'oracle'-style estimators to inform us on the importance of different estimation steps on the underlying variance of the estimators. First, we consider factor imputation where $\boldsymbol{F}$ is known which completely removes noise from the first-stage estimation. Second, when p is needed (generalized synthetic control and the QLD estimator), we consider an

17. We use the R package `did2s` (Butts and Gardner, 2022) for estimation.

18. This model is a specific parametric form of conditional parallel trends that models conditional trends as being linear in covariates.

19. We use the `Optim.jl` package for GMM estimation (Mogensen and Riseth, 2018).

Table 1—Simulation with $T_0 = 4$

	DGP 1		DGP 2		DGP 3		DGP 4	
Estimator	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
TWFE	0.00	0.02	-0.01	0.40	2.35	5.77	3.84	15.53
TWFE with $\boldsymbol{W}_i\beta_t$	0.00	0.03	-0.00	0.13	0.86	0.88	1.08	1.45
Synthetic Control	-0.00	0.04	0.00	0.23	0.11	0.07	0.13	0.07
Augmented Synthetic Control	-0.00	0.04	0.00	0.23	0.11	0.07	0.13	0.07
Generalized Synth (p known)	0.00	0.03	0.01	0.92	-0.93	1.18	-0.04	0.04
Generalized Synth (p estimated)	-0.28	0.09	-0.01	0.39	-1.15	1.52	-0.04	0.04
Factor Imputation ($\boldsymbol{F}$ known)	-0.32	0.15	0.01	0.14	-0.00	0.02	0.00	0.02
QLD (p known)	-0.00	0.41	0.02	0.16	-0.00	0.03	0.00	0.03
QLD (p estimated)	-0.03	0.06	0.02	0.16	-0.00	0.03	0.00	0.03
Model	TWFE		Factor		Factor		Factor	
Parallel Trends	✓		✓					
AR(1) Error Term							✓	

Notes. This table presents a set of simulations with 5000 iterations. Each row in a panel consists of a treatment effect estimator as described in the text. There are 4 different data-generating processes as described in the main text with key details listed in the second portion of the table. The columns present the average bias and mean-squared error of the estimate of $\hat{\tau}^0$ relative to the true effect of 1.

oracle that knows p but not $\boldsymbol{F}$.

Results are presented in table 1 for $T_0 = 4$ and 2 for $T_0 = 10$. For each data-generating process, we present the average bias for the estimate as well as the mean-squared error for the first post-period treatment effect. The second panel of the table describes the data generating processes.

The first data-generating process is where the untreated potential outcome is generated two-way fixed effects model holds.[20] In this case, most estimators work well except for the factor imputation where $\boldsymbol{F}$ is known and the generalized synthetic control estimator. The former is expected since $\boldsymbol{F}$ is not part of the model. In either case, performing the within-transformation

20. We do not include the $\boldsymbol{F}$ known estimates in this case since the underlying model is not a factor model.

Table 2 — Simulation with $T_0 = 10$

	DGP 1		DGP 2		DGP 3		DGP 4	
Estimator	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
TWFE	-0.00	0.02	-0.03	1.59	6.74	47.37	6.53	44.33
TWFE with $\boldsymbol{W}_i\beta_t$	-0.00	0.02	-0.00	0.56	2.33	6.25	2.34	6.25
Synthetic Control	-0.00	0.03	-0.00	0.05	0.01	0.05	0.04	0.05
Augmented Synthetic Control	-0.00	0.03	-0.00	0.05	0.01	0.05	0.04	0.05
Generalized Synth (p known)	-0.01	0.02	0.01	0.12	-0.28	0.16	-0.21	0.13
Generalized Synth (p estimated)	-0.01	0.02	0.00	0.06	-0.01	0.04	0.07	0.07
Factor Imputation ($\boldsymbol{F}$ known)	-0.32	0.12	-0.00	0.01	-0.00	0.01	-0.00	0.02
QLD (p known)	0.01	0.22	-0.00	0.01	-0.00	0.02	-0.00	0.03
QLD (p estimated)	0.02	0.03	-0.00	0.01	-0.00	0.02	-0.00	0.03
Model	TWFE		Factor		Factor		Factor	
Parallel Trends	✓		✓					
AR(1) Error Term							✓	

Notes. This table presents a set of simulations with 5000 iterations. Each row in a panel consists of a treatment effect estimator as described in the text. There are 4 different data-generating processes as described in the main text with key details listed in the second portion of the table. The columns present the average bias and mean-squared error of the estimate of $\hat{\tau}^0$ relative to the true effect of 1.

described above fixes the problem.

It is worth noting that the more robust factor imputation pays an efficiency cost with larger mean-squared error. However, this flips in the second data-generating process where outcomes are generated under a factor model but with parallel trends holding. In this case, all estimators are still unbiased but the factor imputation estimator is the more efficient because it removes the factor structure and leaves a white noise error.

Next we turn to where parallel trends does not hold in data-generating processes 3 and 4. In the third DGP, the error term is a white-noise error and most of the factor-model based estimators work well (curiously not when p is known for the generalized synthetic control estimator). However, in the fourth DGP, the large-T estimators start producing biased estimates with larger mean-squared errors even with a moderately large 10 pre-periods. Last, comparing the general estimation procedure to the two oracle estimators, it appears most of the efficiency cost comes from estimation of $\boldsymbol{F}$ and not from the estimation of p.

Our first set of simulations show that including $\boldsymbol{w}_i\beta_t$ in the TWFE model does remove some bias, but the estimates still perform worse than our imputation procedure due to w_i being a noisy measure. To highlight the problems with noisy proxies for factor loadings, figure 1 presents a set of simulation results using the 3rd data-generating process where the covariates $\boldsymbol{w}_i$ has different amount of noise added in. In particular, we choose different values of $\text{Var}(\xi_{i,r})$ to have different signal-to-noise measures. The signal-to-noise definition is

$$\text{signal to noise ratio} = \frac{\text{Var}(\gamma_{i,r})}{\text{Var}(\gamma_{i,r}) + \text{Var}(\xi_{i,r})} \tag{27}$$

For each signal to noise ratio, we estimate the TWFE imputation estimator with covariates and the factor model imputation estimator. Figure 1 presents the results of estimates for τ_8. At one extreme, where the signal to noise ratio is approximately 0, i.e. ξ_i is white noise, the estimated bias for the TWFE imputation estimator is the same as the TWFE imputation estimator that does not include covariates. At the other extreme, where the signal to noise ratio is approximately 1, i.e. $w_i = \gamma_i$, the bias is completely removed. Except in settings with very large amounts of noise, the factor model imputation estimator remains unbiased and precise. This experiment echos the results of Kejriwal et al. (2024). However, we note that our results are still generous to

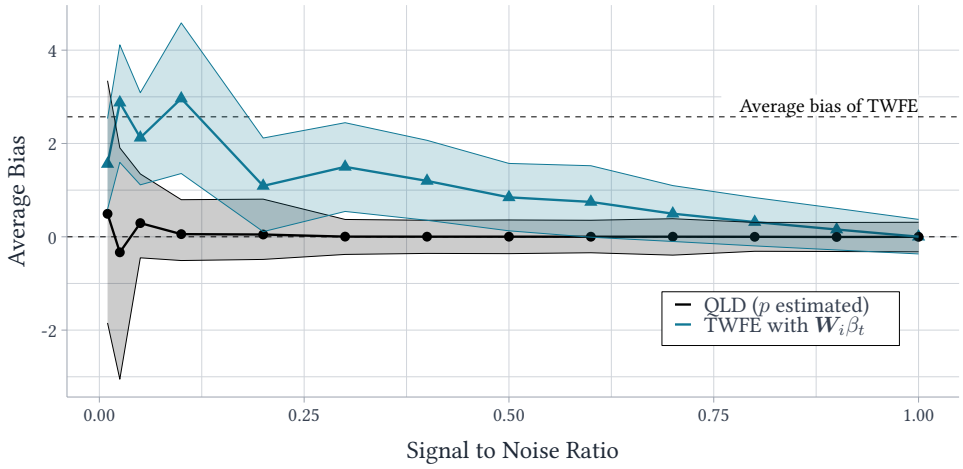


Figure 1 — Bias of TWFE Imputation with Covariates

Notes. This figure plots the average and empirical 95% confidence intervals for treatment effect estimates in of $\hat{\tau}^0$ relative to the true effect of 1. We estimate the TWFE imputation estimator that includes $w_i\beta_t$ linearly in the model and our the factor imputation we propose using w_i instead as an instrument. We vary the signal to noise ratios of w_i to make it a better or worse measure for the factor loading. For each signal to noise ratio, we run 5000 simulations for each value of the signal to noise ratio.

estimators that use such noisy measure because we generate w_i as an unbiased estimator of γ_i. The instrument requirement for QLD estimation does not require unbiased estimation of $\boldsymbol{\gamma}_i$ for identification of the normalized parameters.

5 — Application

We revisit the literature on estimating local labor market effects of Walmart store openings (Basker, 2005; Neumark et al., 2008; Volpe and Boland, 2022). The primary identification concern is that Walmart targets where to open stores based on local economic trajectories (Neumark et al., 2008). For instance, if Walmart targeted areas with positive underlying economic fundamentals in anticipation of their growing consumptive expenditures, then the non-treated counties would fail to be a valid counterfactual group for difference-in-differences. Indeed, we observe significant differences in employment trends for treated and untreated counties in our data. Volpe and Boland (2022) point to conflicting results on retail employment with two leading papers finding effects of opposite signs. Employing different instrumental variable strategies, Basker (2005) finds positive effects on retail employment while Neumark et al. (2008) finds negative effects.

We construct a dataset following the description in Basker (2005). In particular, we use the County Business Patterns dataset from 1964 and 1977-1999, subsetting to counties that (i) had more than 1500 employees overall in 1964 and (ii) had non-negative aggregate employment growth between 1964 and 1977.[21] We use a geocoded data set of Walmart openings from Arcidiacono et al. (2020) to construct our treatment variable. Our treatment dummy is equal to one if the county has any Walmart in that year and our group variable denotes the year of entrance for the *first* Walmart in the county. [22] We drop any county that was treated with $g \leq T_0 = 1985$ so that we we have 9 pre-periods to use when estimating the factor model. Our remaining sample consists of 1274 counties (about 500 fewer than the sample used in Basker (2005) since we drop units treated between 1977 and 1985). We estimate impacts on retail and wholesale employment.[23] Walmart is a vertically integrated business, so we expect Walmart to compete in the retail and wholesale sectors (Basker, 2005).

First, we estimate the two-way fixed effect imputation estimator proposed by Borusyak et al. (2024) and estimate event-study effects on (log) retail and wholesale employment. In particular, we use the following model

$$\log(y_{it}) = \mu_i + \lambda_t + \sum_{\ell=-22}^{13} \tau^\ell d_{it}^\ell + u_{it} \tag{28}$$

where i denotes county, t denotes year, y_{it} is either retail or wholesale employment, and $d_{it}^\ell = 1(t - g_i = \ell)$ are indicator variables denoting event-time. Results of the event-study estimates are presented in panel (a) of figure 2 and figure 3.

For both retail and wholesale employment, counties receiving Walmarts had faster employment growth relative to the control counties, emphasizing our concern over endogenous opening decisions. In the spirit of Freyaldenhoven et al. (Forthcoming) and Rambachan and Roth (2023), we draw the line of best fit for the 15 most-recent pre-treatment estimates ($\hat{\tau}^\ell$ for $-15 \leq \ell < 0$) and extend it into the post-treatment estimates. For both retail and wholesale employment, the

21. We use the 1977-1999 dataset with imputed values from Eckert et al. (2021).

22. For our sample 82.4% of our counties receive ≤ 1 Walmart and another 10.4% receive two Walmarts in the sample, alleviating some concerns of making the treatment binary.

23. Retail employment corresponds with NAICS 2-digit codes 44 and 45 and wholesale employment corresponds to NAICS 2-digit code 42.

pre-trend lines would suggest that a large portion of the estimated effect is a continuation of already existing trends. However, there still appears to be positive effects on retail employment (if the pre-trend violations were indeed linear in the post-treatment period).

We use the QLD estimator of Ahn et al. (2013) to estimate the factors as described in Section 2.2. For this factor estimator, we need a set of instruments that satisfy the two standard instrument requirements: relevancy and exclusion. Intuitively, the relevancy restriction requires that the instruments are correlated with the full vector of factor-loadings. That is, the instruments should be selected as 'proxies' for the kinds of economic factor-loadings that the researcher is concerned of. The exclusion restriction requires that the instrument values are uncorrelated with location-specific idiosyncratic shocks. For this reason, we use baseline covariate values as instruments to avoid shocks to the covariates that are correlated with shocks to the outcome variable.

We select instruments that we suspect are driven by the general macroeconomic trends that cause differential retail employment growth in the 1980s and 1990s. For example, retail employment is likely driven by consumptive expenditures, which in turn are reflective of local labor market trends. Therefore, we use instruments that we think proxy for characteristics that determine local labor market trends. Specifically, we use the 1980 baseline values of the following variables as instruments: share of population employed in manufacturing, shares of population below and above the poverty line, shares of population employed in the private-sector and by the government, and shares of population with high-school and college degrees.[24] Note that instead of estimating $\text{ATT}(g,t)$, we estimate ATT^{ℓ} pooling across (i,t) with $\ell = t - g_i$ as described after Theorem 2.

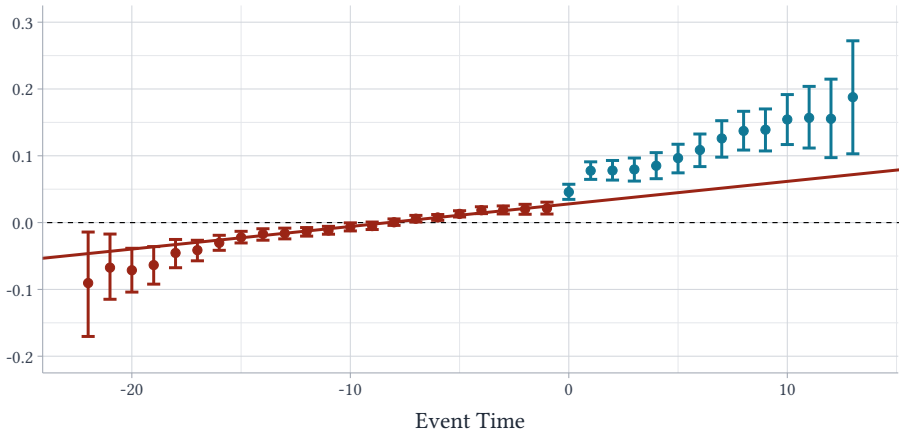
The results of our estimator are presented in panel (b) of figure 2 and figure 3.[25] For retail employment, there is basically no pre-trend violations with the pre-treatment point estimates centered on zero. After removing the pre-existing economic trends, the point estimates are smaller than estimated by the two-way error model with an estimated effect on employment of around 6% on average in the post-treatment periods. Evaluated at the median baseline retail employment of 1417 employees, this would imply an increase in about 85 jobs, which is in line with the estimates

24. All of these values are obtained from 1980 Census Tables accessed from Manson (2020).

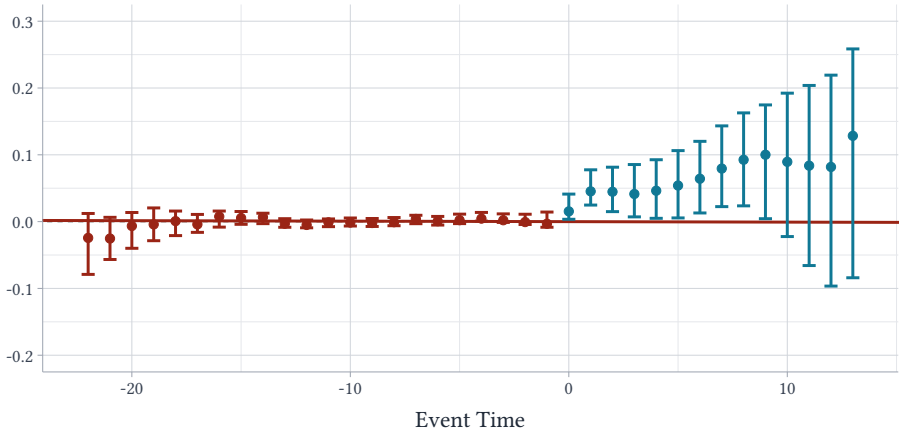
25. We carry out the test to determine the correct number of factors p following the discussion in Ahn et al. (2013). For retail, the p-value of the over-identification test were as follows: p = 0 with a p-value of 1.56e-5; p = 1 with a p-value of 0.001; p = 2 with a p-value of 0.133. Since $p = 2$ is the first value where we fail to reject the null at a 10% level, we set $p = 2$. Similarly, we selected $p = 1$ for wholesale since the p-values were: p = 0 with a p-value of 0.049; and p = 1 with a p-value of 0.40.

Figure 2 — Effect of Walmart on County log Retail Employment

(a) TWFE Imputation Estimator

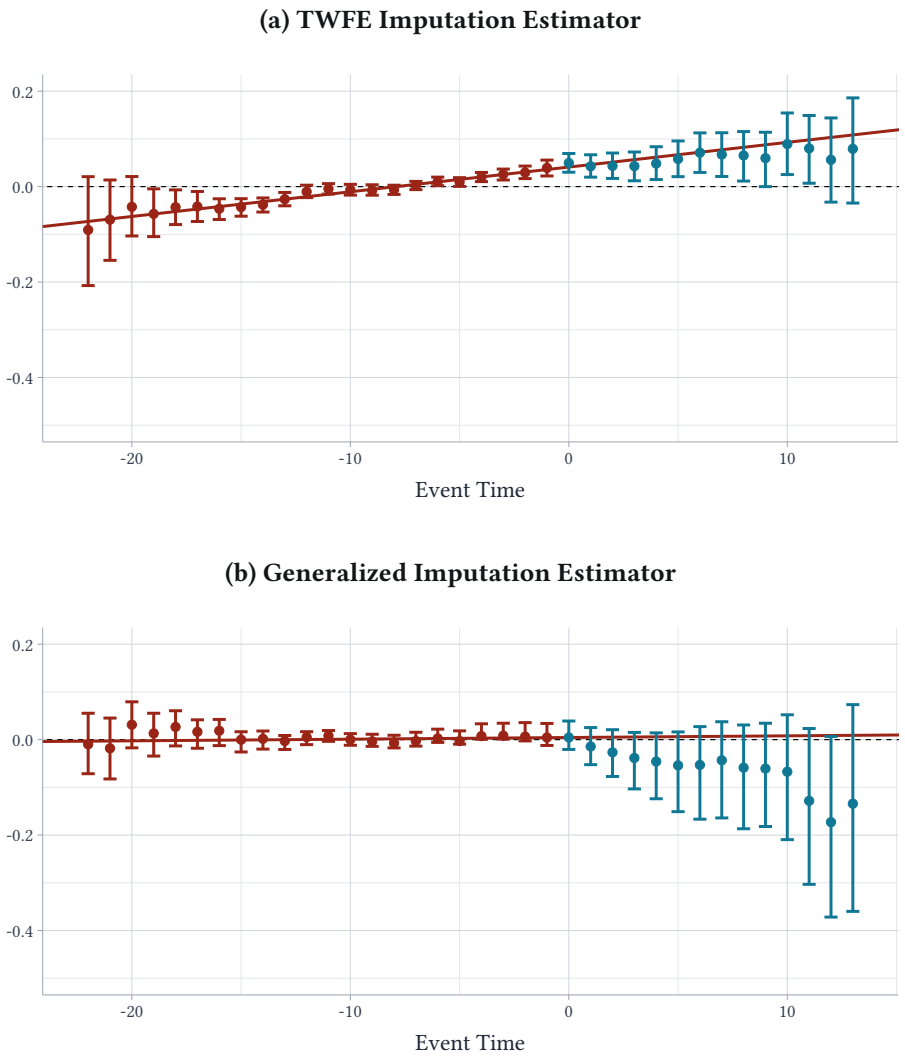


(b) Generalized Imputation Estimator



Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log retail employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (2024). Panel (b) estimates effects using the generalized imputation estimator we propose in Section 3 with $p = 2$ and using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree. The red lines correspond to a linear estimate of pre-treatment point estimates for event time -15 to -1 and is extended into the post-treatment periods.

Figure 3 — Effect of Walmart on County log wholesale Employment



Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log wholesale employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (2024). Panel (b) estimates effects using the generalized imputation estimator we propose in Section 3 with $p = 1$ and using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree. The red lines correspond to a linear estimate of pre-treatment point estimates for event time -15 to -1 and is extended into the post-treatment periods.

of Basker (2005) and Stapp (2014) who use alternative instrumental variables strategies. It is important to note that post-treatment estimates are noisier than the TWFE estimates largely due to estimating the factor proxies in the first stage. This problem is at its worst for the furthest event-times due to very few counties being averaged over in the last few bins. We view this as a worthy trade-off since the point estimates are much less likely to be biased.

Turning to wholesale employment, we see a similar story with our estimator removing most of the pre-trend violations. In this case, however, the estimated effects flip signs with an estimated effect of around -6%, although they are not statistically significant at the 5% level. Evaluated at the 1977 median wholesale employment of 410, this suggests a decrease of about 25 jobs, which is similar to what Basker (2005) finds. Overall, we find effects very much in line with those reported in Basker (2005).

Our estimator allows for any root-N consistent estimator of the factor's column space to be 'plugged-in' and used for estimation of treatment effects. To show the versatility of the method, we use three different factor estimators in figure 4. First, we use our original quasi-differencing estimator from figure 2. Second, we use the common correlated effects (CCE) estimator originally proposed in Pesaran (2006). This estimator uses a set of covariates, $\boldsymbol{X}$, which are generated by the same factors, $\boldsymbol{F}$, as the outcome variable:

$$X_{it} = \boldsymbol{\alpha}_i' \boldsymbol{F}_t + \nu_{it}. \tag{29}$$

Under this assumption, the cross-sectional averages of X (averaged over the never-treated group) consistently span the column space of $\boldsymbol{F}$. In our application, we use log employment for the manufacturing, construction, agriculture, and healthcare 2-digit NAICS codes. The choice of these covariates is plausible if the same sort of national shocks that affect retail employment also affect these other sectors. We more formally analyze this estimator in Brown et al. (2023), which derives the asymptotic distribution of the estimates. One advantage of this factor estimator is that it allows decomposition of treatment effects into direct effects and mediated effects that operate through the covariates, X_{it}.

Last, we use the principal components estimator originally proposed in Bai (2009). This estimator uses the eigenvectors of the matrix $\boldsymbol{Y}\boldsymbol{Y}'$ with the p largest eigenvalues as estimates for

$\boldsymbol{F}$.[26] The advantage of this estimator is that no instrument or additional covariates are required. However this comes at the cost of requiring long panels, which may be infeasible to assume in our application.

The results of each estimator are presented in figure 4. All three estimators are effective at removing underlying trends that the treated counties experienced. Moreover, the estimated effects are similar between estimators suggesting that all three are doing a good job at estimating the underlying factors. This figure highlights the broad applicability of our identification results, allowing the factor estimator of choice to be tailored to the research context at hand. In panel (b), we use $\log$ wholesale employment as an outcome. The CCE and the quasi-differencing estimators produce very similar results, while the principal components estimator suggests positive growth in employment outcomes in later years. Corresponding confidence intervals are very large, suggesting that these results are too noisy to draw any meaningful conclusions. This could be due to wholesale employment being too auto-correlated for the factor estimates to be consistent, or because we do not have a large enough time series to get a meaningful asymptotic approximation of the factors.

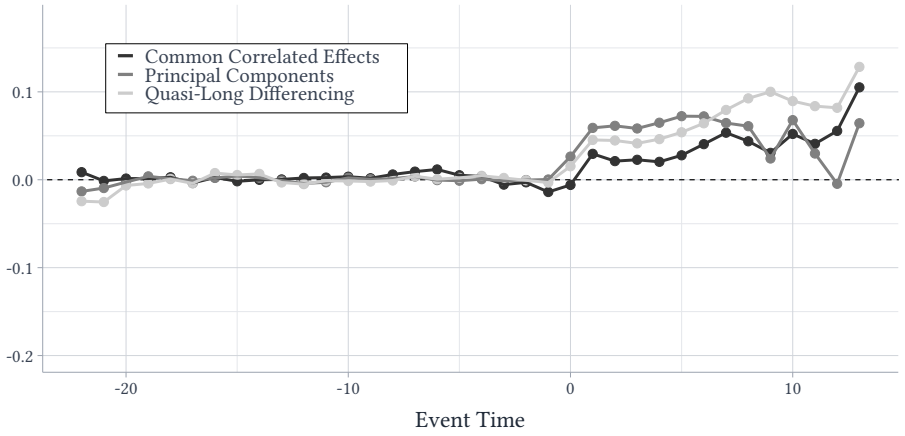
As we discuss above, one reason the synthetic control literature is increasingly popular is that it allows researchers to transparently plot the counterfactual estimates of $y(0)$ for the treated unit. For this reason, we plot the observed $\tilde{y}_{it}$ and the imputed $\hat{\tilde{y}}_{it}(0)$ for (log) retail and wholesale employment in figure 5. In pre-treatment ($\ell < 0$), the imputed estimate, our 'synthetic control' follows closely with the observed $\tilde{y}_{it}$ giving us confidence in our ability to approximate the factor structure. In the post-periods, we see the observed counties and the imputed untreated version of the counties pulling apart. The gap between the two are our estimated treatment effects.

As discussed above, a common approach in empirical work is to include a set of time-invariant covariates interacted with time-period-specific coefficients, $w_i\beta_t$, to model some forms of non-parallel trends (e.g. Abadie (2005); Sant'Anna and Zhao (2020)). Mirroring our simulations, we rerun our TWFE model using our quasi-long differencing instruments as w_i. Intuitively, these worked well as instruments since we think they are likely to be correlated with the true underlying factor loadings. Figure 6 presents the results. Including these variables in our TWFE model fails

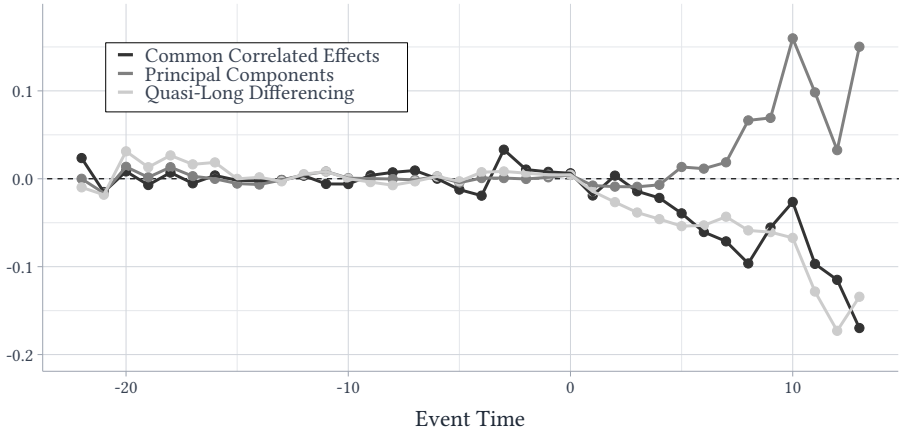
26. This imputation estimator is proposed by Xu (2017) in the context of large panels. The author uses an alternative identification strategy that fails to work in short panels.

Figure 4 — Generalized Imputation Estimator for Effect of Walmart on County Employment with Different Factor Estimators

(a) log **Retail Employment**



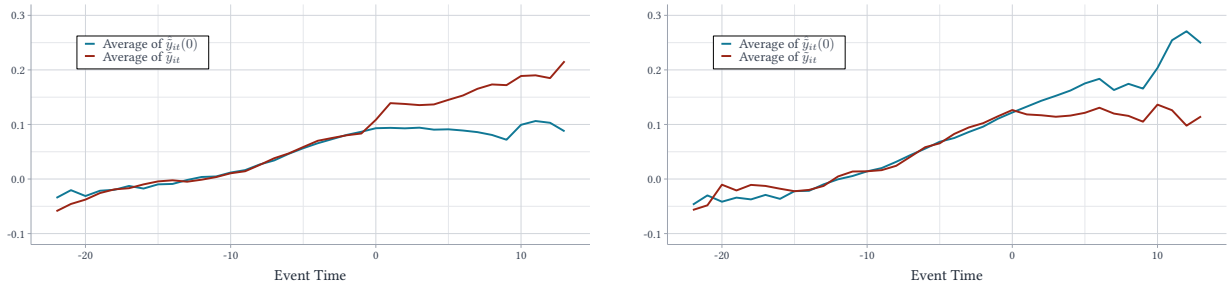
(b) log **wholesale Employment**



Notes. This figure presents estimated treatment effects of Walmart entry on county-level log retail employment using the generalized imputation procedure proposed in section 2.1. The factor estimation procedures include the principal components estimator proposed in Bai (2009), the common correlated effects estimator proposed in Pesaran (2006), and the quasi-differencing estimator proposed in Ahn et al. (2013). Details of the estimation procedures appear in the text.

Figure 5 — Synthetic Control Style Plot of the Effect of Walmart on County Employment

(a) log **Retail Employment** **(b)** log **wholesale Employment**



Notes. This figure plots the observed $\tilde{y}_{it}$ and the imputed $\hat{\tilde{y}}_{it}(0)$ for treated units averaged over event time $\ell = t - g_i$. We impute within-transformed potential outcome using the generalized imputation estimator we propose in Section 3 using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree.

to absorb the non-parallel trends we think are present in the estimates. This is in line with the intuition that w_i is a correlated but noisy measures of the underlying factor-loadings and causes attenuation bias in estimates of β_t that fail to absorb the non-parallel trends.

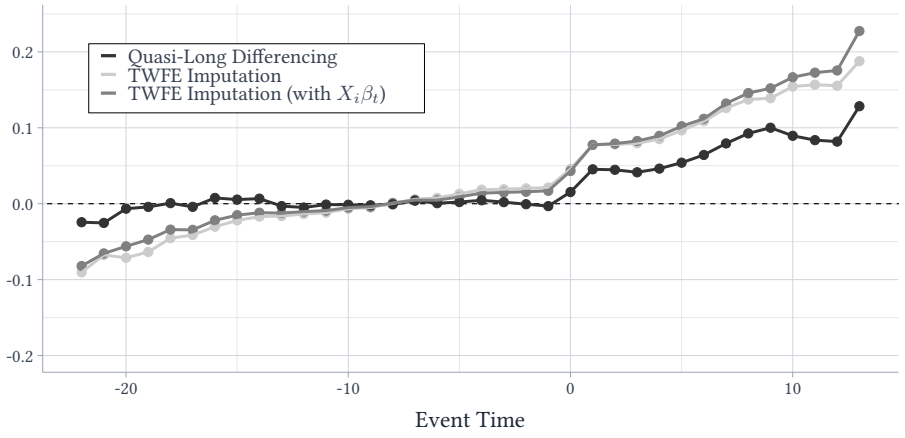
To highlight the importance of the uncertainty from estimation of the factors in the first stage, we recreate confidence intervals from our generalized imputation estimator with the QLD first stage using the nonparametric standard errors that are derived in Theorem 3. Results are given in figure 7. The standard errors on point estimates are far smaller, with estimates becoming strongly significant in wholesale employment. In this empirical application, this is likely because there are relatively few never-treated counties and hence estimation of $\boldsymbol{F}$ is imprecise. This result shows an important step for future research in finding more efficient estimates of the factors. For instance, we consider the common correlated effects estimator in a follow-up paper. The CCE model generally implies that the nonparametric standard errors are valid when there is a common factor model for time-varying covariates.

6 — Conclusions

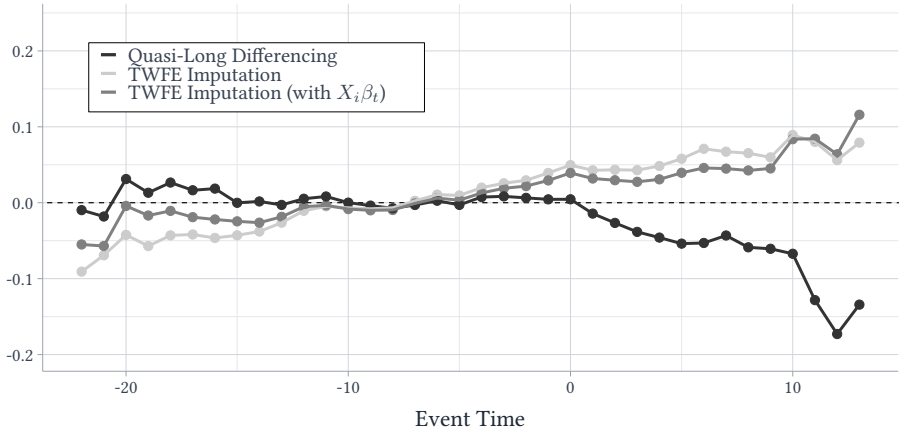
We consider identification and inference of functions of heterogeneous treatment effects in a linear panel data model. We show how to relax the usual parallel trends assumption by introducing a

Figure 6 — Time-interacted covariates in TWFE model

(a) log **Retail Employment**

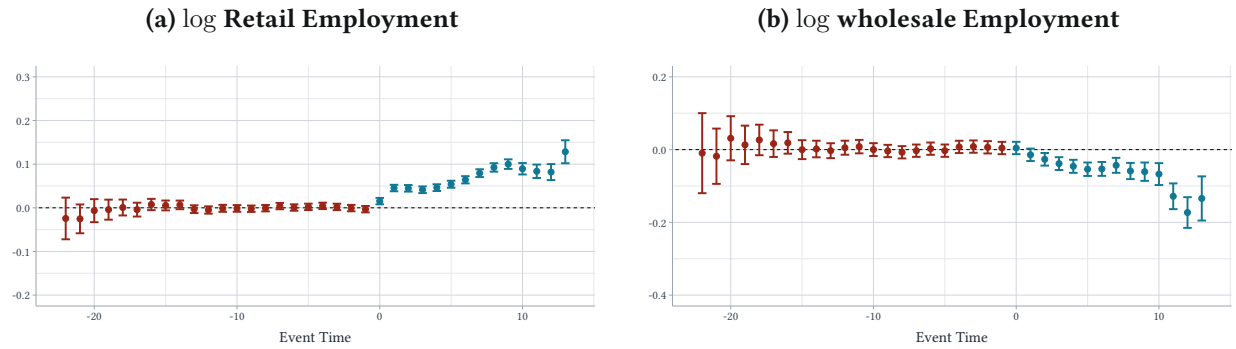


(b) log **wholesale Employment**



Notes. This figure reproduces estimates from figures 2 and 3 and additionally plots estimates modifying the TWFE model to include a set of time-invariant covariates interacted with time-specific coefficients, $w_i\beta_t$ (see previous note).

Figure 7 — Generalized Imputation Estimator for Effect of Walmart on County Employment with Naive Standard Errors



Notes. This figure recreates estimates from panel (b) of figure 2 and figure 3 with confidence intervals formed ignoring the uncertainty deriving from first-stage estimates of θ.

linear factor model in the error. Our main identification result shows that a consistent estimator of the unobserved factors is all that one needs to estimate the dynamic treatment effect coefficients. This result is general and can be implemented by a number of modern interactive fixed effects estimators, such as quasi-long-differencing, internally generated instruments, common correlated effects, or principal components, allowing for both large and small numbers of pre-treatment time periods. While we specifically consider the quasi-long-differencing estimator of Ahn et al. (2013), further work should demonstrate both theoretical and finite-sample properties of these various estimators of the factors and how they affect to ATT estimation, especially for larger time series. The GMM imputation framework should also be examined in the context of unbalanced panels as in Rai (2023).

While a factor model nests the usual two-way error structure, we explicitly model the level fixed effects in addition to the factors. This setting allows us to provide useful tests for the consistency of the TWFE estimator. We also show that one must remove the unit and time fixed effects in a particular way so as to preserve the common factor structure in all time periods for all individuals. We provide such a transformation and prove a novel identification result for TWFE imputation estimators of ATTs.

We implement the QLD estimator of Ahn et al. (2013) in a study of the local impact of Walmart openings. We demonstrate findings consistent with the IV estimation strategy of Basker (2005). Our estimator is shown to remove pre-trends that bias the usual TWFE estimates. Similar results are found using common correlated effects in the first stage. A principal components estimator is

also explored, but performs suspiciously for the given problem. The QLD identification scheme can also allow sequentially exogenous outcomes like those generated by dynamic models. We leave this possibility for future study.

Acknowledgments

We would like to thank Stephane Bonhomme, Brantly Callaway, Brian Cadena, Peter Hull, Jeffrey Wooldridge, and Taylor Jaworski as well as seminar participants from the University of Georgia, Queen's University, the 2022 Midwest Econometrics Group, the CU Boulder Econometrics Brown-bag, the 2023 CEA Annual Meeting, and the 2023 IAAE Annual Conference for their insightful questions and comments. We also thank the Associate Editor for their careful review of our paper, as well as the thoughtful reports from two anonymous referees. All errors are our own.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." *The review of economic studies* 72 (1): 1–19.

Abadie, Alberto. 2021. "Using synthetic controls: Feasibility, data requirements, and methodological aspects." *Journal of Economic Literature* 59 (2): 391–425. 10.1257/jel.20191450.

Abadie, Alberto, Alexis Diamond, and Jens 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 (490): 493–505.

Abadir, Karim M., and Jan R. Magnus. 2005. *Matrix Algebra.* Volume 1. Cambridge University Press, . 10.1017/cbo9780511810800.

Ahn, Seung C, Young H Lee, and Peter Schmidt. 2013. "Panel data models with multiple time-varying individual effects." *Journal of econometrics* 174 (1): 1–14. 10.1016/j.jeconom.2012.12.002.

Ahn, Seung Chan, Young Hoon Lee, and Peter Schmidt. 2001. "GMM estimation of linear panel data models with time-varying individual effects." *Journal of Econometrics* 101 (2): 219–255. 10.1016/s0304-4076(00)00083-x.

Arcidiacono, Peter, Paul B Ellickson, Carl F Mela, and John D Singleton. 2020. "The competitive effects of entry: Evidence from supercenter expansion." *American Economic Journal: Applied Economics* 12 (3): 175–206. 10.2139/ssrn.3045492.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. 2021. "Synthetic difference-in-differences." *American Economic Review* 111 (12): 4088–4118. 10.1257/aer.20190159.

Asquith, Brian J, Evan Mast, and Davin Reed. 2021. "Local effects of large new apartment buildings in low-income areas." *Review of Economics and Statistics* 1–46.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2021. "Matrix completion methods for causal panel data models." *Journal of the American Statistical Association* 116 (536): 1716–1730. 10.1080/01621459.2021.1891924.

Bai, Jushan. 2009. "Panel data models with interactive fixed effects." *Econometrica* 77 (4): 1229–1279. 10.3982/ecta6135.

Bai, Jushan, and Serena Ng. 2021. "Matrix completion, counterfactuals, and factor analysis of missing data." *Journal of the American Statistical Association* 116 (536): 1746–1763.

Basker, Emek. 2005. "Job Creation or Destruction? Labor Market Effects of Wal-Mart Expansion." *Review of Economics and Statistics* 87 (1): 174–183. 10.1162/0034653053327568.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. 2021. "The augmented synthetic control method." *Journal of the American Statistical Association* 116 (536): 1789–1803.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event Study Designs: Robust and Efficient Estimation." 10.47004/wp.cem.2022.1122, Review of Economic Studies.

Breitung, Jörg, and Philipp Hansen. 2021. "Alternative estimation approaches for the factor augmented panel data model with small T." *Empirical Economics* 60 327–351. 10.1007/s00181-020-01948-7.

Brown, Nicholas. 2023. "Moment-based Estimation of Linear Panel Data Models with Factor-augmented Errors." Working Paper.

Brown, Nicholas, Kyle Butts, and Joakim Westerlund. 2023. "Simple Difference-in-Differences Estimation in Fixed-T Panels."

Brown, Nicholas L., Peter Schmidt, and Jeffrey M. Wooldridge. 2023. "Simple Alternatives to the Common Correlated Effects Model." 10.13140/RG.2.2.12655.76969/1.

Butts, Kyle, and John Gardner. 2022. "did2s: Two-Stage Difference-in-Differences." *R Journal* 14 (3): . 10.32614/rj-2022-048.

Callaway, Brantly, and Sonia Karami. 2023. "Treatment effects in interactive fixed effects models with a small number of time periods." *Journal of Econometrics* 233 (1): 184–208. 10.1016/j.jeconom.2022.02.001.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230. 10.1016/j.jeconom.2020.12.001.

Chan, Marc K, and Simon S Kwok. 2022. "The PCDID approach: difference-in-differences when trends are potentially unparallel and stochastic." *Journal of Business & Economic Statistics* 40 (3): 1216–1233. 10.1080/07350015.2021.1914636.

Cragg, John G, and Stephen G Donald. 1997. "Inferring the rank of a matrix." *Journal of econometrics* 76 (1-2): 223–250.

Eckert, Fabian, Teresa C. Fort, Peter K. Schott, and Natalie J. Yang. 2021. "Imputing Missing Values in the US Census Bureau's County Business Patterns."Technical report, National Bureau of Economic Research. 10.3386/w26632.

Feng, Yingjie. 2020. "Causal inference in possibly nonlinear factor models." *arXiv preprint arXiv:2008.13651*.

Fernández-Val, Iván, Hugo Freeman, and Martin Weidner. 2021. "Low-rank approximations of nonseparable panel models." *The Econometrics Journal* 24 (2): C40–C77.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. Forthcoming. "Visualization, identification, and estimation in the linear panel event-study design." 10.3386/w29170.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. "Pre-Event Trends in the Panel Event-Study Design." *American Economic Review* 109 (9): 3307–3338. 10.1257/aer.20180609.

Gardner, John. 2021. "Two-Stage Difference-in-Differences."

Gobillon, Laurent, and Thierry Magnac. 2016. "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls." *Review of Economics and Statistics* 98 (3): 535–551. 10.1162/REST_a_00537.

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

Hahn, Jinyong, Zhipeng Liao, and Geert Ridder. 2018. "Nonparametric two-step sieve M estimation and inference." *Econometric Theory* 34 (6): 1281–1324. 10.1017/s0266466618000014.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 1029–1054. 10.2307/1912775.

Imbens, Guido, Nathan Kallus, and Xiaojie Mao. 2021. "Controlling for Unmeasured Confounding in Panel Data Using Minimal Bridge Functions: From Two-Way Fixed Effects to Factor Models."

Imbens, Guido W, and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Juodis, Artūras, and Vasilis Sarafidis. 2022b. "An incidental parameters free inference approach for panels with common shocks." *Journal of Econometrics* 229 (1): 19–54. 10.1016/j.jeconom.2021.03.011.

Juodis, Artūras, and Vasilis Sarafidis. 2022a. "A Linear Estimator for Factor-Augmented Fixed-T Panels With Endogenous Regressors." *Journal of Business & Economic Statistics* 40 (1): 1–15. 10.1080/07350015.2020.1766469.

Kejriwal, Mohitosh, Xiaoxiao Li, Linh Nguyen, and Evan Totty. 2024. "The efficacy of ability proxies for estimating the returns to schooling: A factor model-based evaluation." *Journal of Applied Econometrics* 39 (1): 3–21.

Manson, Steven M. 2020. "IPUMS national historical geographic information system: Version 15.0."

Mogensen, P, and A Riseth. 2018. "Optim: A mathematical optimization package for Julia." *Journal of Open Source Software* 3 (24): , https://joss.theoj.org/papers/10.21105/joss.00615.

Neumark, David, and Helen Simpson. 2015. "Place-based policies." In *Handbook of regional and urban economics*, Volume 5. 1197–1287, Elsevier.

Neumark, David, Junfu Zhang, and Stephen Ciccarella. 2008. "The effects of Wal-Mart on local labor markets." *Journal of Urban Economics* 63 (2): 405–430. 10.1016/j.jue.2007.07.004.

Pennington, Kate. 2021. "Does building new housing cause displacement?: the supply and demand effects of construction in San Francisco." 10.2139/ssrn.3867764.

Pesaran, M Hashem. 2006. "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica* 74 (4): 967–1012.

Rai, Bhavna. 2023. "Efficient estimation with missing data and endogeneity." *Econometric Reviews* 42 (2): 220–239. 10.1080/07474938.2023.2178089.

Rambachan, Ashesh, and Jonathan Roth. 2023. "A more credible approach to parallel trends." *Review of Economic Studies* rdad018.

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

Stapp, Jacob. 2014. "The Walmart Effect: Labor Market Implications in Rural and Urban Counties." *SS-AAEA Journal of Agricultural Economics* 2014 (318-2016-9525): , https://ideas.repec.org/a/ags/ssaaea/232737.html.

Volpe, Richard, and Michael A Boland. 2022. "The Economic Impacts of Walmart Supercenters." *Annual Review of Resource Economics* 14 43–62. 10.1146/annurev-resource-111820-032827.

Westerlund, Joakim. 2020. "A cross-section average-based principal components approach for fixed-T panels." *Journal of Applied Econometrics* 35 (6): 776–785. 10.1002/jae.2786.

Westerlund, Joakim, Yana Petrova, and Milda Norkutė. 2019. "CCE in fixed-T panels." *Journal of Applied Econometrics* 34 746–761. 10.1002/jae.2707.

Windmeijer, Frank. 2005. "A finite sample correction for the variance of linear efficient two-step GMM estimators." *Journal of econometrics* 126 (1): 25–51.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* MIT press.

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906345.

Xu, Yiqing. 2017. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25 (1): 57–76.

Appendix for "**Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels**"

A — Proofs

Proof of Theorem 1

Let $t \geq g$ for the given group g.

$$\mathbb{E}(y_{it} - \boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g) = \mathbb{E}(y_{it}(1) \mid G_i = g) - \mathbb{E}(\boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g)$$

We use the fact that

$$\begin{aligned}
\mathbb{E}(\boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} \mid G_i = g) &= \mathbb{E}\big(\boldsymbol{F}_t'(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \mid G_i = g\big) \\
&= \mathbb{E}\big(\boldsymbol{F}_t'(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}'\big[\boldsymbol{F}_{t<g}\boldsymbol{\gamma}_i + u_{i,t<g}\big] \mid G_i = g\big) \\
&= \mathbb{E}\big(\boldsymbol{F}_t'\boldsymbol{\gamma}_i + \boldsymbol{F}_t'(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}'u_{i,t<g} \mid G_i = g\big) \\
&= \mathbb{E}(y_{it}(\infty) \mid G_i = g)
\end{aligned}$$

The second equality hold by Assumption 2 and the fact that $y_{i,t<g} = y_{i,t<g}(0)$. The final equality holds by Assumption 2.

For the second part of the theorem, note that from the column span condition, there exists a $m \times p$ matrix $\boldsymbol{A}$ such that

$$\boldsymbol{F}^* \boldsymbol{A} = \boldsymbol{F} \tag{A1}$$

$\boldsymbol{A}$ defines the linear combinations of the columns of $\boldsymbol{F}^*$ that span the columns of $\boldsymbol{F}$. Thus

$\boldsymbol{F}_t^{*\prime}\boldsymbol{A} = \boldsymbol{F}_t'$. We then have

$$\begin{aligned}\boldsymbol{F}_t^{*\prime}(\boldsymbol{F}_{t<g}^{*\prime}\boldsymbol{F}_{t<g}^{*\prime})^{-1}\boldsymbol{F}_{t<g}^{*\prime}\boldsymbol{F}_{t<g}\boldsymbol{\gamma}_i &= \boldsymbol{F}_t^{*\prime}(\boldsymbol{F}_{t<g}^{*\prime}\boldsymbol{F}_{t<g}^{*})^{-1}\boldsymbol{F}_{t<g}^{*\prime}\boldsymbol{F}_{t<g}^{*\prime}\boldsymbol{A}\boldsymbol{\gamma}_i \\ &= \boldsymbol{F}_t^{*\prime}\boldsymbol{A}\boldsymbol{\gamma}_i \\ &= \boldsymbol{F}_t^{*\prime}\boldsymbol{\gamma}_i\end{aligned}$$

If $m = p$ so that $\boldsymbol{F}$ also has full column rank, we can make the stronger statement that the imputation matrices of $\boldsymbol{F}$ and $\boldsymbol{F}^*$ are equal:

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{F}_{t\geq g}, \boldsymbol{F}_{t<g}) &= \boldsymbol{F}_{t\geq g}(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})^{-1}\boldsymbol{F}_{t<g}' \\ &= \boldsymbol{F}_{t\geq g}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{F}_{t<g}' \\ &= \boldsymbol{F}_{t\geq g}^{*\prime}(\boldsymbol{F}_{t<g}^{*\prime}\boldsymbol{F}_{t<g}^{*})^{-1}\boldsymbol{F}_{t<g}^{*\prime} \\ &= \boldsymbol{P}(\boldsymbol{F}_{t\geq g}^{*}, \boldsymbol{F}_{t<g}^{*})\end{aligned}$$

where the second equality holds because $\boldsymbol{A}$ and $(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g})$ are full rank.

□

Proof of Lemma 1

We first derive the averages defined in Section 2.2 in terms of the potential outcome framework:

$$\begin{aligned}\overline{y}_{\infty,t} &= \frac{1}{N_\infty}\sum_{i=1}^{N} D_{i\infty}y_{it} = \overline{\mu}_\infty + \lambda_t + \boldsymbol{F}_t\overline{\boldsymbol{\gamma}}_\infty + \overline{u}_{t,\infty} \\ \overline{y}_{i,t\leq T_0} &= \frac{1}{T_0}\sum_{t=1}^{T_0} y_{it} = \mu_i + \overline{\lambda}_{t<T_0} + \overline{\boldsymbol{F}}_{t<T_0}\boldsymbol{\gamma}_i + \overline{u}_{i,t<T_0} \\ \overline{y}_{\infty,t<T_0} &= \frac{1}{N_\infty T_0}\sum_{i=1}^{N}\sum_{t=1}^{T_0} D_{i\infty}y_{it} = \overline{\mu}_\infty + \overline{\lambda}_{t<T_0} + \overline{\boldsymbol{F}}_{t<T_0}\overline{\boldsymbol{\gamma}}_\infty + \overline{u}_{\infty,t<T_0}\end{aligned}$$

where $\overline{\mu}_\infty$ and $\overline{\boldsymbol{\gamma}}_\infty$ are the averages of the never-treated individuals' heterogeneity and $\overline{\boldsymbol{F}}_{t<T_0}$ and $\overline{\lambda}_{t<T_0}$ are the averages of the time effects before anyone is treated. The error averages have the same interpretation as the outcome averages.

The definition of τ_{it} is the difference between treated and untreated potential outcomes for

unit i at time t, so for any (i,t), $y_{it} = d_{it}y_{it}(1) + (1-d_{it})y_{it}(\infty) = d_{it}\tau_{it} + y_{it}(\infty)$. Then

$$\begin{aligned}\tilde{y}_{it} &= d_{it}\tau_{it} + \boldsymbol{F}_t'\boldsymbol{\gamma}_i - \overline{\boldsymbol{F}}_{t<T_0}'\boldsymbol{\gamma}_i - \boldsymbol{F}_t'\overline{\boldsymbol{\gamma}}_\infty + \overline{\boldsymbol{F}}_{t<T_0}\overline{\boldsymbol{\gamma}}_\infty + u_{it} - \overline{u}_{t,\infty} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0} \\ &= d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_\infty) + u_{it} - \overline{u}_{t,\infty} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0}\end{aligned}$$

Taking expectation conditional on $G_i = g$ gives $\mathbb{E}(u_{it} - \overline{u}_{i,t<T_0} \mid G_i = g) = 0$ by Assumption 2 and $\mathbb{E}(\overline{u}_{\infty,t<T_0} - \overline{u}_{t,\infty} \mid G_i = g) = \mathbb{E}[\overline{u}_{\infty,t<T_0} - \overline{u}_{t,\infty}] = 0$ by random sampling and iterated expectations.

□

Proof of Theorem 2

We can appeal to standard large sample GMM theory as in Hansen (1982) due to the types of first-stage factor estimators we consider. We do not consider true "fixed effects" estimators where the number of parameters grows with the sample size. The IV and cross-sectional averages approaches are based on eliminating the factors (which are fixed in the asymptotic analysis) by reducing them to a smaller set of parameters. For example, while the CCE estimator can be implemented as a pooled regression where unit dummies are interacted with cross-sectional averages, the estimator itself takes a form similar to the within transformation in the linear fixed effects model. In fact, we prove asymptotic unbiasedness of dynamic ATT estimators using the CCE estimator in the first stage (Brown et al., 2023)[27].

Consider the QLD estimator of Ahn et al. (2013). They study the linear model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \tag{A2}$$

They jointly estimate the QLD parameters $\boldsymbol{\theta}$ along with the conditional response parameters $\boldsymbol{\beta}$ using the moment conditions

$$\mathbb{E}[\boldsymbol{H}(\boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \otimes \boldsymbol{w}_i] = \boldsymbol{0} \tag{A3}$$

27. We consider CCE in a separate paper because the additional modeling assumptions allow for stronger results than those considered in this paper.

They show that the estimator is well-behaved and does not suffer from asymptotic bias. As described in Windmeijer (2005), the most likely source of finite-sample bias comes from estimating the optimal weight matrix. The appendix of Ahn et al. (2013) describes a continuous updating estimator (CUE) based on their moment conditions, which may have less finite-sample bias than the optimal two-step estimator. However, we may also sacrifice efficiency in large samples if their assumed covariance structure is incorrect.

We now derive the asymptotic variance of the full estimator under a general first-step estimator of the factors. Note that $\boldsymbol{g}_{i\infty}(\boldsymbol{\theta}) \otimes \boldsymbol{g}_{ig}(\boldsymbol{\theta}, \boldsymbol{\tau}_g) = \mathbf{0}$ (from the D_{ig} terms) and $\boldsymbol{g}_{ih}(\boldsymbol{\theta}, \boldsymbol{\tau}_h) \otimes \boldsymbol{g}_{ik}(\boldsymbol{\theta}, \boldsymbol{\tau}_k) = \mathbf{0}$ almost surely uniformly over the parameter space for all $g \in \mathcal{G}$ and $h \neq k$. The covariance matrix of these moment functions, which we denote as $\boldsymbol{\Delta}$, is a block diagonal matrix.

$$\boldsymbol{\Delta} = \begin{pmatrix} \mathbb{E}[\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})'] & \mathbf{0} & 0 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbb{E}[\boldsymbol{g}_{ig_G}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_G})\boldsymbol{g}_{ig_G}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_G})'] & 0 & \dots & \mathbf{0} \\ \vdots & & \ddots & & \\ \mathbf{0} & \mathbf{0} & 0 & \dots & \mathbb{E}[\boldsymbol{g}_{ig_1}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_1})\boldsymbol{g}_{ig_1}(\boldsymbol{\theta}, \boldsymbol{\tau})'] \end{pmatrix}$$

We write the individual blocks as $\boldsymbol{\Delta}_g$ for $g \in \mathcal{G} \cup \{\infty\}$. The gradient is also simple to compute because all of the moments are linear in the treatment effects. We define the overall gradient $\boldsymbol{D}$ and show it is a lower triangular matrix which we write in terms of its constituent blocks:

$$\boldsymbol{D} = \begin{pmatrix} \mathbb{E}[\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_{i\infty}(\boldsymbol{\theta})] & \mathbf{0} & 0 & \dots & \mathbf{0} \\ \mathbb{E}[\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_{ig_G}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_G})] & -\boldsymbol{I}_{T-g_G+1} & 0 & \dots & \mathbf{0} \\ \vdots & & \ddots & & \\ \mathbb{E}[\nabla_{\boldsymbol{\theta}} \boldsymbol{g}_{ig_1}(\boldsymbol{\theta}, \boldsymbol{\tau}_{g_1})] & \mathbf{0} & 0 & \dots & -\boldsymbol{I}_{T-g_1+1} \end{pmatrix}$$

where we write the blocks in the first column as $\boldsymbol{D}_g$ for $g \in \mathcal{G} \cup \{\infty\}$. The diagonal is made up of negative identity matrices because $\mathbb{E}\left[\frac{D_{ig_h}}{\mathbb{P}(D_{ig_h}=1)}\right] = 1$.

Given we use the optimal weight matrix, the overall asymptotic variance is given by $(\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D})^{-1}$.

$\boldsymbol{\Delta}$ is a block diagonal matrix so its inverse is trivial to compute. First, we have

$$\boldsymbol{\Delta}^{-1}\boldsymbol{D} = \begin{pmatrix} \boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{\Delta}_{g_G}^{-1}\boldsymbol{D}_{g_G} & -\boldsymbol{\Delta}_{g_G}^{-1} & \dots & \mathbf{0} \\ \vdots & & \ddots & \\ \boldsymbol{\Delta}_{g_1}^{-1}\boldsymbol{D}_{g_1} & \mathbf{0} & \dots & -\boldsymbol{\Delta}_{g_1}^{-1} \end{pmatrix}$$

The transpose of the gradient matrix is

$$\boldsymbol{D}' = \begin{pmatrix} \boldsymbol{D}'_{\infty} & \boldsymbol{D}'_{g_G} & \dots & \boldsymbol{D}'_{g_1} \\ \mathbf{0} & -\boldsymbol{I}_{T-g_G+1} & \dots & \mathbf{0} \\ \vdots & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & -\boldsymbol{I}_{T-g_1+1} \end{pmatrix}$$

so that we get

$$\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D} = \begin{pmatrix} \sum_{g\in\mathcal{G}\cup\{\infty\}} \boldsymbol{D}'_g\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g & -\boldsymbol{D}'_{g_G}\boldsymbol{\Delta}_{g_G}^{-1} & \dots & -\boldsymbol{D}'_{g_1}\boldsymbol{\Delta}_{g_G}^{-1} \\ -\boldsymbol{\Delta}_{g_G}^{-1}\boldsymbol{D}_{g_G} & \boldsymbol{\Delta}_{g_G}^{-1} & \dots & \mathbf{0} \\ \vdots & & \ddots & \\ -\boldsymbol{\Delta}_{g_1}^{-1}\boldsymbol{D}_{g_1} & \mathbf{0} & \dots & \boldsymbol{\Delta}_{g_1}^{-1} \end{pmatrix}$$

We write this matrix as

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$$

where $\boldsymbol{A} = \sum_{g\in\mathcal{G}\cup\{\infty\}} \boldsymbol{D}'_g\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g$ and $\boldsymbol{D} = \text{diag}\{\boldsymbol{\Delta}_g^{-1}\}_{g\in\mathcal{G}}$. We then apply Exercise 5.16 of

Abadir and Magnus (2005) to get the final inverse. The top left corner of the inverse is $\boldsymbol{F}^{-1}$ where

$$\begin{aligned}(\boldsymbol{F})^{-1} &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} \\ &= \left(\sum_{g\in\mathcal{G}\cup\{\infty\}} \boldsymbol{D}_g'\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g - \left(\sum_{g\in\mathcal{G}} \boldsymbol{D}_g'\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g\right)\right)^{-1} \\ &= (\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1} \\ &= \text{Avar}(\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))\end{aligned}$$

The rest of the first column of matrices takes the form

$$\begin{aligned}-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{F}^{-1} &= \begin{pmatrix}\boldsymbol{D}_{g_G} \\ \vdots \\ \boldsymbol{D}_{g_1}\end{pmatrix}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1} \\ &= \begin{pmatrix}\boldsymbol{D}_{g_G}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1} \\ \vdots \\ \boldsymbol{D}_{g_1}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1}\end{pmatrix}\end{aligned}$$

and the rest of the first row is $-\boldsymbol{F}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = (-\boldsymbol{D}^{-1}\boldsymbol{B}'\boldsymbol{F}^{-1})' = (-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{F}^{-1})'$.

Finally, the bottom-right block, which also gives the asymptotic covariance matrix of the ATT estimators, is

$$\boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{F}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} = \boldsymbol{D}^{-1} + \begin{pmatrix}\boldsymbol{D}_{g_G}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1}\boldsymbol{D}_{g_G}' & \dots & \boldsymbol{D}_{g_G}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1}\boldsymbol{D}_{g_1}' \\ & \ddots & \\ \boldsymbol{D}_{g_1}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1}\boldsymbol{D}_{g_G}' & \dots & \boldsymbol{D}_{g_1}(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1}\boldsymbol{D}_{g_1}'\end{pmatrix}$$

The g'th diagonal elements of the resulting matrix is $\boldsymbol{\Delta}_g + \boldsymbol{D}_g(\boldsymbol{D}_\infty'\boldsymbol{\Delta}_\infty^{-1}\boldsymbol{D}_\infty)^{-1}\boldsymbol{D}_g'$.

□

Proof of Theorem 3

We derive the limiting theory by multiplying $\widehat{\boldsymbol{\Delta}}_g$ by $(N_g-1)/N_g$ which produces the same limit as $N \to \infty$. We write

$$\frac{N_g-1}{N_g}\widehat{\boldsymbol{\Delta}}_g = \frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\widehat{\boldsymbol{\Delta}}_{ig}\widehat{\boldsymbol{\Delta}}_{ig}' - \widehat{\boldsymbol{\tau}}_g\widehat{\boldsymbol{\tau}}_g'$$

We already know that $\widehat{\boldsymbol{\tau}}_g \xrightarrow{p} \boldsymbol{\tau}_g$ by Theorem 3.1. Note that

$$\begin{aligned}\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\widehat{\boldsymbol{\Delta}}_{ig}\widehat{\boldsymbol{\Delta}}_{ig}' = &\left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t\geq g}\boldsymbol{y}_{i,t\geq g}'\right) - \left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t\geq g}\boldsymbol{y}_{i,t<g}'\right)\boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))' \\ &- \boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))\left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t<g}\boldsymbol{y}_{i,t\geq g}'\right) \\ &- \boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))\left(\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\boldsymbol{y}_{i,t<g}\boldsymbol{y}_{i,t\geq g}'\right)\boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))'\end{aligned}$$

Given $\boldsymbol{P}(\boldsymbol{F}_{t\geq g}(\widehat{\boldsymbol{\theta}}),\boldsymbol{F}_{t<g}(\widehat{\boldsymbol{\theta}}))$ is equal to its infeasible counterpart $\boldsymbol{P}(\boldsymbol{F}_{t\geq g},\boldsymbol{F}_{t<g})$ plus a $O_p(N^{-1/2})$ term, Assumption 1 and the weak law of large numbers imply

$$\frac{1}{N_g}\sum_{i=1}^{N} D_{ig}\widehat{\boldsymbol{\Delta}}_{ig}\widehat{\boldsymbol{\Delta}}_{ig}' - \widehat{\boldsymbol{\tau}}_g\widehat{\boldsymbol{\tau}}_g' \xrightarrow{p} \mathbb{E}(\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g) \mid G_i = g) = \boldsymbol{\Delta}_g$$

The inverse exists with probability approaching one by Assumption 5.

□

B — The Quasi-Long-Differencing Estimator

We discuss identification and inference of the imputation estimator using the QLD estimator for the factors. We derive the results here because QLD is used in both the simulations and application.

B.1. Identification

We adapt the identifying assumptions from Ahn et al. (2013) to our setup, guaranteeing Assumption 4 holds. Part (i) of the assumption holds assuming that $\boldsymbol{F}$ is full rank, because there always exists a matrix that applies Gaussian row-reduction to a full rank matrix. For parts (ii) and (iii), we need

the following matrix to be full rank:

$$\boldsymbol{I}_{T-p} \otimes \mathbb{E}(\boldsymbol{w}_i \boldsymbol{\gamma}_i' \mid G_i = \infty) \tag{B1}$$

It implies that the instruments $\boldsymbol{w}_i$ are "strong" in the sense that they correlate with the factor loadings $\boldsymbol{\gamma}_i$. Unfortunately, this restriction is not easily testable like in the case of two-stage least squares because the variable being instrumented for is unobserved. We leave the question of testing for instrument strength in quasi-differencing for a future project. Part (iv) is an additional assumption that is routinely made in practice.

B.2. Asymptotic Variance

We now derive the analytical formulas for the asymptotic variance when quasi-differencing is used to estimate the factor space. Analytical standard errors can be obtained by replacing the population parameters with their estimators and expectations with the relevant sample average, e.g. expectations of the never-treated group are estimated using the average of the never-treated subsample. Conversely, one can average over the entire sample but multiply each observation by $D_{i\infty}$ and divide by N_∞/N. To get the gradient of the set of moment conditions that identify the factor space, we rewrite the moment function as

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{y}_i \otimes \boldsymbol{w}_i &= \text{vec}(\boldsymbol{w}_i \boldsymbol{y}_i' \boldsymbol{H}(\boldsymbol{\theta})') \\ &= (\boldsymbol{I}_{(T-p)} \otimes \boldsymbol{w}_i \boldsymbol{y}_i') \boldsymbol{K}_{(T-p)T} \text{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\end{aligned}$$

where $\boldsymbol{K}_{(T-p)T}$ is the $(T-p)T \times (T-p)T$ commutation matrix and we use the well-known relationship between vectorization and the Kronecker product[28]. Because $\text{vec}(\boldsymbol{H}(\boldsymbol{\theta})) = [\text{vec}(\boldsymbol{I}_{T-p})', \boldsymbol{\theta}']'$, the gradient of the moment function is

$$\left(\boldsymbol{I}_{(T-p)} \otimes \boldsymbol{w}_i \boldsymbol{y}_i'\right) \boldsymbol{K}_{T(T-p)} [\boldsymbol{0}'_{(T-p)^2 \times (T-p)p}, \boldsymbol{I}_{(T-p)p}]' \tag{B2}$$

The expected gradient is obtained by taking expectations conditional on being in the never-treated group.

28. See Exercise 10.18 of Abadir and Magnus (2005).

We now consider the gradient of the moment functions that determine the treatment effects with respect to the factor estimator for a given group treated at time g. The relevant part of the moment function for the purpose of finding the gradient is

$$\boldsymbol{F}_{t\geq g}(\boldsymbol{\theta})' \left(\boldsymbol{F}_{t<g}(\boldsymbol{\theta})'\boldsymbol{F}_{t<g}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{F}_{t<g}(\boldsymbol{\theta})'\boldsymbol{y}_{i,t<g} \tag{B3}$$

There are two leading cases to compute: $g-1 \geq T-p$ and $g-1 < T-p$. In the first case, the parameters $\boldsymbol{\theta}$ are entirely contained in the pre-treatment factor matrix. Then

$$\boldsymbol{F}_{t<g} = \begin{pmatrix} \boldsymbol{\Theta} \\ \boldsymbol{E} \end{pmatrix} \tag{B4}$$

where $\boldsymbol{E}$ is the first $(g-1)-(T-p)$ rows of $-\boldsymbol{I}_p$. Then the post-treatment factor matrix is just the lower $T-g+1$ rows of $-\boldsymbol{I}_p$ so we do not need to worry about differentiating it. In this setting,

$$\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} = -\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{B5}$$

We use the notation in Chapter 13 of Abadir and Magnus (2005) to obtain the differential:

$$-\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}(d\boldsymbol{\Theta})' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{B6}$$

$$\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\left((d\boldsymbol{\Theta})'\boldsymbol{\Theta}\right)\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{B7}$$

$$\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\left(\boldsymbol{\Theta}'(d\boldsymbol{\Theta})\right)\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{B8}$$

which can then be rewritten as

$$-\left(\boldsymbol{y}_{i,t<g}\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\begin{pmatrix}\boldsymbol{K}_{(T-p)p}(d\boldsymbol{\theta})' & \boldsymbol{K}_{((g-1)-(T-p)p}\text{vec}(\boldsymbol{E})'\end{pmatrix}' \tag{B9}$$

$$\left(\left(\boldsymbol{\Theta}\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\boldsymbol{K}_{(T-p)p}d\boldsymbol{\theta} \tag{B10}$$

$$\left(\left(\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\boldsymbol{\Theta}'\right)d\boldsymbol{\theta} \tag{B11}$$

The full gradient is then

$$-\left(\boldsymbol{y}_{i,t<g}\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\begin{pmatrix}\boldsymbol{K}'_{(T-p)p} & \boldsymbol{0}'_{((g-1)-(T-p)p\times(T-p)p}\end{pmatrix}' \tag{B12}$$

$$\left(\left(\boldsymbol{\Theta}\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\boldsymbol{K}_{(T-p)p} \tag{B13}$$

$$\left(\left(\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\boldsymbol{\Theta}'\right) \tag{B14}$$

when $g-1\geq T-p$.

The second case, when $g-1<T-p$, now has parameters in the post-treatment matrix $\boldsymbol{F}_{t\geq g}$. We redefine the parameters as $\boldsymbol{\Theta}=[\boldsymbol{\Theta}'_1,\boldsymbol{\Theta}'_2]'$ where $\boldsymbol{\Theta}_1$ is $(g-1)\times p$ and $\boldsymbol{\Theta}_2$ is $(T-p-g+1)\times p$. Now we write $\boldsymbol{F}_{t<g}=\boldsymbol{\Theta}_1$ and

$$\boldsymbol{F}_{t\geq g}=\begin{pmatrix}\boldsymbol{\Theta}_2\\ -\boldsymbol{I}_p\end{pmatrix} \tag{B15}$$

Because $\boldsymbol{\theta}\neq(\text{vec}(\boldsymbol{\Theta}_1)',\text{vec}(\boldsymbol{\Theta}_2)')'$, we define the matrices $\boldsymbol{E}_1=[\boldsymbol{I}_{g-1},\boldsymbol{0}_{(g-1)\times(T-p-g+1)}$ and $\boldsymbol{E}_2=[\boldsymbol{0}_{(T-p-g+1)\times(g-1)},\boldsymbol{I}_{(T-p-g+1)}]$ such that

$$\boldsymbol{\Theta}_1=\boldsymbol{E}_1\boldsymbol{\Theta} \tag{B16}$$

$$\boldsymbol{\Theta}_2=\boldsymbol{E}_2\boldsymbol{\Theta} \tag{B17}$$

Now we can rewrite the relevant portion of the moment function for the gradient as

$$\begin{pmatrix}\boldsymbol{E}_2\boldsymbol{\Theta}\\ -\boldsymbol{I}_p\end{pmatrix}\left(\boldsymbol{\Theta}'\boldsymbol{E}'_1\boldsymbol{E}_1\boldsymbol{\Theta}\right)^{-1}\boldsymbol{\Theta}'\boldsymbol{E}'_1\boldsymbol{y}_{i,t<g} \tag{B18}$$

We can now take the gradient with respect to the full set of parameters $\boldsymbol{\theta}$:

$$\begin{pmatrix} \boldsymbol{E}_2 d\boldsymbol{\Theta} \\ \mathbf{0}_{p\times p} \end{pmatrix} \left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} \tag{B19}$$

$$-\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} d\boldsymbol{\Theta}'\boldsymbol{E}'_1\boldsymbol{F}_{t<g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} \tag{B20}$$

$$-\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{E}_1 d\boldsymbol{\Theta}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} \tag{B21}$$

$$+\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} d\boldsymbol{\Theta}'\boldsymbol{E}'_1\boldsymbol{y}_{i,t<g} \tag{B22}$$

where we inserted $\boldsymbol{F}_{t<g}$ and $\boldsymbol{F}_{t\geq g}$ for $\boldsymbol{E}_1\boldsymbol{\Theta}$ and $\boldsymbol{E}_2\boldsymbol{\Theta}$ respectively to preserve space, noting that these matrices are actually functions of the parameters $\boldsymbol{\theta}$ and not the true, unobserved factors. We rewrite line (B19) so we can write the differential in terms of $\boldsymbol{\theta}$:

$$\begin{pmatrix} \boldsymbol{E}_2 d\boldsymbol{\Theta} \\ \mathbf{0}_{p\times p} \end{pmatrix} \left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} = \begin{pmatrix} \boldsymbol{E}_2 \\ \mathbf{0}_{p\times p} \end{pmatrix} d\boldsymbol{\Theta}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g} \tag{B23}$$

$$= \left(\left(\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right) \otimes \begin{pmatrix} \boldsymbol{E}_2 \\ \mathbf{0}_{p\times p} \end{pmatrix}\right) d\boldsymbol{\theta} \tag{B24}$$

We put this expression with the others to get the final gradient:

$$= \left(\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right) \otimes \begin{pmatrix} \boldsymbol{E}_2 \\ \mathbf{0}_{p\times p} \end{pmatrix} \tag{B25}$$

$$-\left(\boldsymbol{E}'_1\boldsymbol{F}_{t<g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right)' \otimes \left(\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1}\right) \boldsymbol{K}_{(T-p)p} \tag{B26}$$

$$-\left(\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{y}_{i,t<g}\right)' \otimes \left(\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}'_{t<g}\boldsymbol{E}_1\right) \tag{B27}$$

$$+\left(\boldsymbol{y}'_{i,t<g}\boldsymbol{E}_1\right) \otimes \left(\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}'_{t<g}\boldsymbol{F}_{t<g}\right)^{-1}\right) \boldsymbol{K}_{(T-p)p} \tag{B28}$$

C — Inference of Aggregate Treatment Effects

As in Callaway and Sant'Anna (2021), we can form aggregates of our group-time average treatment effects. For example, event-study type coefficients would average over the τ_{gt} where $t - g = e$ for

some relative event-time e with weights proportional to group membership. Consider a general aggregate estimand δ which we define as a weighted average of $ATT(g,t)$:

$$\delta = \sum_{g \in \mathcal{G}} \sum_{t > T_0} w(g,t) \tau_{gt} \tag{C1}$$

where the weights $w(g,t)$ are non-negative and sum to one. Table 1 of Callaway and Sant'Anna (2021) and the surrounding discussion describes various treatment effect aggregates and discuss explicit forms for the weights.

Our plug-in estimate for δ is given by $\hat{\delta} = \sum_{g \in \mathcal{G}} \sum_{t > T_0} \hat{w}(g,t) \hat{\tau}_{gt}$. Inference on this term follows directly from Corollary 2 in Callaway and Sant'Anna (2021) if we have the influence function for our τ_{gt} estimates. Rewriting our moment equations in an asymptotically linear form, we have:

$$\sqrt{N}\left((\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')' - (\boldsymbol{\theta}', \boldsymbol{\tau}')' \right) = - \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{D}' \boldsymbol{\Delta}^{-1} \boldsymbol{D})^{-1} \boldsymbol{D}' \boldsymbol{\Delta}^{-1} \boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau}) \right) + o_p(1). \tag{C2}$$

This form comes from the fact that the weight matrix is positive definite with probability approaching one[29]. The first term on the right-hand side is the influence function and hence inference on aggregate quantities follows directly. This result allows for use of the multiplier bootstrap to estimate standard errors in a computationally efficient manner.

D — Inference in Two-Way Fixed Effect Model

We derive the asymptotic distribution of our imputation estimator based off of the two-way error model in equation (1). First, we note that this estimator can be written in terms of the imputation matrix from Section 2. In particular, let $\mathbf{1}_t$ be a $T \times 1$ vector of ones up the t'th spot, with all zeros after. Define $\overline{\boldsymbol{y}}_\infty = (\overline{y}_{\infty,1}, ..., \overline{y}_{\infty,T})'$ be the full vector of never-treated cross-sectional averages. Then our imputation transformation can be written as

$$\tilde{\boldsymbol{y}}_i = \left[\boldsymbol{I}_T - \boldsymbol{P}(\mathbf{1}_T, \mathbf{1}_{T_0}) \right] (\boldsymbol{y}_i - \overline{\boldsymbol{y}}_\infty) \tag{D1}$$

29. This is a well-known expansion for analyzing the asymptotic properties of GMM estimators. See Chapter 14 of Wooldridge (2010) for example.

where the t^{th} component of the above T-vector is

$$d_{it}\tau_{it} + \tilde{u}_{it}, \tag{D2}$$

with $\tilde{u}_{it}$ is defined as the same transformation as $\tilde{y}_{it}$.

The imputation step of our estimator is a just-identified system of equations. As such, we do not need to worry about weighting in implementation and inference comes from standard theory of M-estimators. In fact, we have the following closed-form solution for the estimator of a group-time average treatment effect:

$$\widehat{\tau}_{gt} = \frac{1}{N_g} \sum_i D_{ig} \tilde{y}_{it}, \tag{D3}$$

where $N_g = \sum_i D_{ig}$ is the number of units in group g.

The following theorem characterizes estimation under the two-way error model:

Theorem D1. Assume untreated potential outcomes take the form of the two-way error model given in equation (1). Suppose Assumptions 1 and 3 hold, as well as Assumption 2 with $\boldsymbol{\gamma}_i = 0$. Then for all (g, t) with $g > t$, $\widehat{\boldsymbol{\tau}}_{gt}$ is conditionally unbiased for $\mathbb{E}(\tau_{it} \mid D_{ig} = 1)$, has the linear form

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{D4}$$

and

$$\sqrt{N_1}(\widehat{\tau}_{gt} - \tau_{gt}) \xrightarrow{d} N(0, V_1 + V_0) \tag{D5}$$

as $N \to \infty$, where V_1 and V_0 are given below and $\tau_{gt} = \mathbb{E}(y_{it}(g) - y_{it}(\infty) \mid D_{ig} = 1)$ is the group-time average treatment effect (on the treated). ■

Theorem (D1) demonstrates the simplicity of our imputation procedure under the two-way error model. While the general factor structure requires more care, estimation and inference will yield a similar result.

Proof of Theorem D1

The transformed post-treatment observations are

$$\tilde{y}_{it} = \tau_{it} + u_{it} - \overline{u}_{\infty,t} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0} \tag{D6}$$

To show unbiasedness, take expectation conditional on $D_{ig} = 1$. This expected value is

$$\mathbb{E}(\tau_{it} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0} \mid D_{ig} = 1) = \mathbb{E}(\tau_{it} \mid D_{ig} = 1) \tag{D7}$$

by Assumption 2 and 3.

For consistency, note that averaging over the sample with $D_{ig} = 1$, subtracting τ_{gt}, and multiplying $\sqrt{N_g}$ gives

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}) + \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(-\overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{D8}$$

which is two normalized sums of uncorrelated iid sequences that have mean zero (by iterated expectations) and finite fourth moments.

Rewriting the second term in terms of the original averages $\frac{1}{N_\infty} \sum_{i=1}^{N} -u_{i,t} + \overline{u}_{i,t<T_0}$ gives:

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}) + \sqrt{\frac{N_g}{N_\infty}} \left(\frac{1}{\sqrt{N_\infty}} \sum_{i=1}^{N} D_{i\infty}(-u_{i,t} + \overline{u}_{i,t<T_0}) \right) \tag{D9}$$

Since these terms are mean zero and uncorrelated, we find the variance of each term separately.

The first term has asymptotic variance

$$V_1 = \mathbb{E}\left(\Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big) \Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big)' \mid D_{ig} = 1 \right) \tag{D10}$$

and the second term has asymptotic variance

$$V_0 = \frac{\mathbb{P}(D_{ig} = 1)}{\mathbb{P}(D_{i\infty} = 1)} \mathbb{E}\left(\left(\overline{u}_{i,t<T_0} - u_{i,t} \right) \left(\overline{u}_{i,t<T_0} - u_{i,t} \right)' \mid D_{i\infty} = 1 \right) \tag{D11}$$

The result follows from the independence of the two sums.

E — Including Covariates

We now discuss the inclusion of covariates in the untreated potential outcome mean model. Allowing for covariates further weakens our parallel trends assumption by allowing selection to hold on unobserved heterogeneity as well as observed characteristics. Identifying the effects of covariates requires some kind of time and unit variation because we manually remove the level fixed effects.

A common inclusion in the treatment effects literature is time-constant variables with time-varying slopes. Suppose $\boldsymbol{x}_i$ is $1 \times K$ vector of time-constant covariates. We could write the mean model of the untreated outcomes as

$$\mathbb{E}(y_{it}(\infty) \mid x_i, \mu_i, \boldsymbol{\gamma}_i, D_i) = \boldsymbol{x}_i \boldsymbol{\beta}_t + \mu_i + \lambda_t + \boldsymbol{F}_t' \boldsymbol{\gamma}_i \tag{E1}$$

which allows observable covariates to have trending partial effects; covariates with constant slopes are captured by the unit effect. After removing the additive fixed effects, $\boldsymbol{x}_i \boldsymbol{\beta}_t$ will take the same form as the residuals of factor structure. Estimating $\boldsymbol{\theta}$ can be done jointly with the time-varying coefficients by applying the QLD transformation to the vector of $\tilde{y}_{it} - \tilde{x}_i \tilde{\beta}_t$. We cannot identify the underlying partial effects because of the time-demeaning, but we can include them for the sake of strengthening the parallel trends assumption.

Time-constant covariates (or time-varying covariates fixed at their pre-treatment value) are often employed because there is little worry that they are affected by treatment. However, we could also include time- and individual-varying covariates of the form $\boldsymbol{x}_{it}$ that are allowed to have identifiable constant slopes if we assume their distribution is unaffected by treatment status. Let $\boldsymbol{x}_{it}$ be a $1 \times K$ vector of covariates that vary over i and t. We can jointly estimate a $K \times 1$ vector

of parameters $\boldsymbol{\beta}$ along with $\boldsymbol{\theta}$ using the moments

$$\mathbb{E}\Big(\boldsymbol{H}(\boldsymbol{\theta})'(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}) \otimes \boldsymbol{w}_i \mid G_i = \infty\Big) = \mathbf{0} \tag{E2}$$

where $\tilde{\boldsymbol{X}}_i$ is the $T \times K$ matrix of stacked covariates after our double-demeaning procedure.

We could also allow slopes to vary across groups and estimate them via the group-specific pooled regression $D_{ig}y_{it}$ on $D_{ig}\boldsymbol{x}_{it}$ with unit-specific slopes on $D_{ig}\tilde{\boldsymbol{F}}(\widehat{\boldsymbol{\theta}})_t$ for $t = 1, ..., g-1$. Then we include the covariates and their respective slopes into the moment conditions

$$\mathbb{E}\Big((\tilde{\boldsymbol{y}}_{i,t\geq g} - \tilde{\boldsymbol{X}}_{i,t\geq g}\boldsymbol{\beta}_g) - \boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\geq g}, \tilde{\boldsymbol{F}}_{t<g})(\tilde{\boldsymbol{y}}_{i,t<g} - \tilde{\boldsymbol{X}}_{i,t<g}\boldsymbol{\beta}_g) - \boldsymbol{\tau}_g \mid G_i = g\Big) = \mathbf{0} \tag{E3}$$

We note that the above expression requires treatment to not affect the evolution of the covariates, a strong assumption in practice. Chan and Kwok (2022) make a similar assumption for their principal components difference-in-differences estimator. We study this assumption in the context of the common correlated effects model in Brown et al. (2023).

F — Testing Mean Equality of Factor Loadings

We develop this test in the context of the QLD estimation of Ahn et al. (2013). Specifically, we need $\mathbb{E}[\boldsymbol{\gamma}_i] = \mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g)$ for all $g \in \mathcal{G}$. Our imputation approach allows us to identify these terms up to a rotation. To see how, let $\boldsymbol{A}^*$ be the rotation that imposes the Ahn et al. (2013) normalization. Then

$$\begin{aligned}
\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta})_{t<g})\,\mathbb{E}(\boldsymbol{y}_{i,t<g} \mid G_i = g) &= \left(\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}(\boldsymbol{\theta})_{t<g}\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}_{t<g}\,\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) \\
&= \left(\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}(\boldsymbol{\theta})_{t<g}\right)^{-1}\boldsymbol{F}(\boldsymbol{\theta})'_{t<g}\boldsymbol{F}(\boldsymbol{\theta})_{t<g}(\boldsymbol{A}^*)^{-1}\,\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) \\
&= (\boldsymbol{A}^*)^{-1}\,\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g)
\end{aligned}$$

where $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{F}\boldsymbol{A}^*$.

It is irrelevant that the means of the factor loadings are only known up to a nonsingular transformation, because $\boldsymbol{A}^*$ is the same for each $g \in \mathcal{G}$ by virtue of the common factors. We note

that

$$\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) - \mathbb{E}[\boldsymbol{\gamma}_i] = \mathbf{0} \iff (\boldsymbol{A}^*)^{-1}(\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) - \mathbb{E}[\boldsymbol{\gamma}_i]) = \mathbf{0} \tag{F1}$$

The results above show how we can identify $(\boldsymbol{A}^*)^{-1}\,\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g)$ by imputing the pre-treatment observations onto an identify matrix.

Collect the moments

$$\begin{gathered}
\mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty}=1)}\boldsymbol{H}(\boldsymbol{\theta})\tilde{\boldsymbol{y}}_i \otimes \boldsymbol{w}_i\right] = \mathbf{0} \\
\mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty}=1)}\left(\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta}))\boldsymbol{y}_i - \boldsymbol{\gamma}^*\right)\right] = \mathbf{0} \\
\mathbb{E}\left[\frac{D_{ig_G}}{\mathbb{P}(D_{ig_G}=1)}\left(\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta})_{t<g_G})\boldsymbol{y}_{i,t<g_G} - \boldsymbol{\gamma}^*_{g_G}\right)\right] = \mathbf{0} \\
\vdots \\
\mathbb{E}\left[\frac{D_{ig_1}}{\mathbb{P}(D_{ig_1}=1)}\left(\boldsymbol{P}(\boldsymbol{I}_p, \boldsymbol{F}(\boldsymbol{\theta})_{t<g_1})\boldsymbol{y}_{i,t<g_1} - \boldsymbol{\gamma}^*_{g_G}\right)\right] = \mathbf{0}
\end{gathered}$$

The parameters $(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}^*_{g_G}, ..., \boldsymbol{\gamma}^*_{g_1})$ represent the rotated means of the factor loadings. $\boldsymbol{\gamma}$ is the unconditional mean $(\boldsymbol{A}^*)^{-1}\,\mathbb{E}[\boldsymbol{\gamma}_i]$ and $\boldsymbol{\gamma}_g$ is the conditional mean $(\boldsymbol{A}^*)^{-1}\,\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g)$ for $g \in \mathcal{G}$. We include estimation of the factors for convenience, so that one does not need to directly calculate the effect of first-stage estimation on the asymptotic variances of conditional means.

Joint GMM estimation of the above parameters, including $\boldsymbol{\theta}$, then allows one to test combinations of the rotated means. Specifically, we have the following result:

Theorem F2. If $\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) = \mathbb{E}[\boldsymbol{\gamma}_i]$ for all $g \in \mathcal{G}$, then

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*_{g_G} = ... = \boldsymbol{\gamma}^*_{g_1} \tag{F2}$$

■