DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 23-05

Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels

Kyle Butts
University of Colorado Boulder

October 16, 2023

Department of Economics

University of Colorado Boulder
Boulder, Colorado 80309

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

Nicholas Brown† and Kyle Butts‡

October 16, 2023

[Most recent version]

We study inference on dynamic average treatment effect parameters for staggered interventions when parallel trends are only valid conditional on an unobserved interactive fixed effects model. The interactive fixed effects model allows for units to select into treatment based on differential-exposure to macroeconomic shocks. We propose a general imputation-style estimator that is consistent in settings with few pre-treatment time periods and under arbitrary treatment effect heterogeneity. Our identification strategy allows for a wide set of factor-model estimators including principal components, common correlated effects, quasi-differencing, and more. 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

*We would like to thank Stephane Bonhomme, Brantly Callaway, Brian Cadena, Peter Hull, and Jeffrey Wooldridge, as well as seminar participants from the University of Georgia, Florida State University, Queen's University, the 2022 Midwest Econometrics Group, the CU Boulder Econometrics Brownbag, the 2023 CEA Annual Meeting, and the 2023 IAAE Annual Conference for their insightful questions and comments.

†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 for estimating the dynamic effects of a binary treatment in linear panel data models. In many empirical settings, treatment is assigned non-randomly based on trends in economic variables, and the parallel trends assumption required by difference-in-differences is not plausible. 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 opening 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. In many settings, though, it is reasonable to assume 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 be declining, consumer trends for walkable neighborhoods cause certain neighborhoods to become increasingly demanded, and national industry growth rates impact counties differentially.

A recent growing literature models these kind of parallel trends deviations using an interactive fixed effects where there are common national shocks, but the exposure to the shock vary across locations. While interactive fixed effects relax the parallel trends assumptions relied on by difference-in-differences, the current literature requires long panels to estimate effects. Requiring long panels for estimation is often impractical because of (i) lack of data for many years of outcomes, (ii) strong assumptions like serially uncorrelated outcomes, or (iii) the presence of structural breaks, e.g. recessions or structural changes to the macroeconomy, rendering 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, $y_{it}(\infty)$, as an interactive fixed effect model

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

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, as we did in the above examples, the factors $\boldsymbol{F}_t$ as macroeconomic shocks 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 in outcome.

Our main treatment effect identification result only requires 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

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 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.

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, consistent estimation of $\boldsymbol{F}_t$ is possible through a variety of approaches, such as quasi-differencing (Ahn et al., 2013; Callaway and Karami, 2023), common correlated effects (Pesaran, 2006; Westerlund et al., 2019), or principal components (Bai, 2009; Fan et al., 2016; Westerlund, 2020; Chan and Kwok, 2022). 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. It is also consistent when the number of pre-treatment time periods is small.[3] One advantage of this estimator is that we can form statistical tests for the consistency of the two-way fixed effects (TWFE) estimator. These tests are practically useful since difference-in-differences is simple to implement.

Relation to Literature

Recent work has proposed 'imputation' estimators for treatment effects using non-treated and pre-treatment observations to 'impute' the untreated potential outcomes for the post-treatment observations (e.g. Borusyak et al., Forthcoming; Gardner, 2021; Wooldridge, 2021). However, these approaches only allow for level fixed effects and preclude interactions like in equation (1). Borusyak et al. (Forthcoming) 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.

3. Deriving the asymptotic distribution of treatment effects using large-T factor estimators is left for future work.

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 (Xu, 2017; Gobillon and Magnac, 2016), 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 (Imbens et al., 2021).[4] 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, but can still accommodate large-T and unit-heterogeneous estimation strategies.

Our work 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

4. 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.

our application and Section 6 leaves with some concluding remarks.

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 different units; 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, which we denote $y_{it}(g)$. 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 $t \geq g$. It is trivial to estimate other averages as well in our framework, 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 has 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. The Online Appendix discusses the inclusion of covariates and the subsequent relaxation of assumption 2. 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.

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 knives 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.

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.

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 group-time average treatment effects. Gardner (2021), Wooldridge (2021), and Borusyak et al. (Forthcoming) 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) and Xu (2017), or a generalized version of a factor model as in 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. 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 $G_i = g$. 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 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 $\text{ATT}(g, t)$.

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 $\boldsymbol{X}_1$. Gardner (2021) and Borusyak et al. (Forthcoming) implicitly uses the imputation matrix for an additive error model where $\boldsymbol{X}_1$ is the matrix of unit and time fixed effects and $\boldsymbol{X}_0$ is $\boldsymbol{X}_1$ with rows of zero whenever $d_{it} = 1$. When applying this matrix of factors to our outcomes, the post-treatment factors are multiplied by the factor loadings from the pre-treatment observations. In particular, we impute $y_{it}(\infty)$ by $\boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g}$ for $G_i = g$, similar to the bridge function identification scheme in Imbens et al. (2021). However, because we only need a conditional mean assumption, we can allow arbitrary correlation between the idiosyncratic errors.

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 Online 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 is important because it is well understood that $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are not separately identified (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.

Theorem 1 shows we can apply these conclusions to any interactive fixed effects estimator that achieves consistency by asymptotically spanning the factor space. Examples include the principal components estimator of Bai (2009)[8], the common correlated effects estimator of Pesaran (2006)[9], the differencing techniques of Ahn et al. (2001, 2013), Callaway and Karami (2023) and Brown (2023), or the internally generated instruments of Juodis and Sarafidis (2022) or Cui et al. (2021). As long as the factors are consistently estimated using the control sample, dynamic ATTs are identified as in Theorem 1, regardless of the normalization used for estimation. In fact, we do not even require the factors to be full rank, though this assumption is typically made in practice.

To present a general framework for the estimation of the factors, we formally present the identifying assumptions needed for factor space estimators:

Assumption 4. There exists a $q \times 1$ 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.

8. The PC estimator in Bai (2009) requires $T \to \infty$ for asymptotic normality. Westerlund (2020) provides conditions under which the PC estimator is fixed-T consistent. Regardless, the identification strategy still holds.

9. The CCE factor proxies do not converge to a full rank matrix. However, Westerlund et al. (2019) show that the residual maker matrix is consistent for the space orthogonal to the factors due to an implicit normalization. See Brown et al. (2023) for a proof of asymptotic normality in our setting under typical CCE assumptions.

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. The parameters $\boldsymbol{\theta}$ themselves are often the result of an underlying normalization like in Bai (2009), Ahn et al. (2013), Juodis and Sarafidis (2022), 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).

Assumption 4 is written with fixed-T estimation and inference in mind. As mentioned before, accommodating general 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. Further research should explicitly derive the large-T properties of our estimator using principal components in the first stage.

Remark 1. 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) (ALS). ALS propose a QLD transformation given by

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

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

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

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

$$\mathbb{E}(g_{i\infty}(\boldsymbol{\upsilon}) \mid D_{i\infty} = 1) = \mathbb{E}(\boldsymbol{w}_i \otimes \boldsymbol{H}(\boldsymbol{\upsilon})\boldsymbol{y}_i \mid D_{i\infty} = 1) = \boldsymbol{0} \tag{12}$$

where $\boldsymbol{w}_i$ is a vector of instruments that are exogenous with respect to the idiosyncratic error in Assumption 2 but correlated with $\boldsymbol{\gamma}_i$. 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. ■

2.2. 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 models; it may also provide efficiency by reducing the burden on first-stage factor estimators. 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. (Forthcoming), Gardner (2021), and Wooldridge (2021) that estimate the additive effects using the untreated sample will not maintain a common factor structure. 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 never-treated unit respectively satisfy

$$\sum_{t=1}^{g-1}(y_{it} - \widehat{\lambda}_t - \widehat{\mu}_i) = 0 \tag{13}$$

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

The control sample will remove more time averages than in every treated sample, meaning the factors are demeaned using different subsamples. As such, the transformed factors are not equal across groups and so we cannot then 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:

$$\overline{y}_{\infty,t} = \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i\infty} y_{it} \tag{15}$$

$$\overline{y}_{i,t\leq T_0} = \frac{1}{T_0} \sum_{t=1}^{T_0} y_{it} \tag{16}$$

$$\overline{y}_{\infty,t<T_0} = \frac{1}{N_\infty T_0} \sum_{i=1}^{N} \sum_{t=1}^{T_0} D_{i\infty} y_{it} \tag{17}$$

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., Forthcoming). Second, it preserves a common factor structure for all units and time periods[10]. The TWFE imputation estimator of Gardner (2021), Wooldridge (2021), and Borusyak et al. (Forthcoming) 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.

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

10. 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).

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

$$\tilde{y}_{it} = d_{it}\tau_{it} + (\boldsymbol{F}_t - \overline{\boldsymbol{F}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_{\infty}) + \tilde{u}_{it}. \tag{18}$$

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

$$\tilde{y}_{it} = d_{it}\tau_{it} + \tilde{\boldsymbol{F}}_t'\tilde{\boldsymbol{\gamma}}_i + \tilde{u}_{it}. \tag{19}$$

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 the 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 Online Appendix. We also provide simple tests for mean independence of the factor loadings in Remark 5, 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 standard 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.[11] Implementation requires replacing $\boldsymbol{P}(D_{ig}=1)$ 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 assumption 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

$$\min_{\boldsymbol{\theta},\boldsymbol{\tau}}\left(\sum_{i=1}^{N}\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right)'\widehat{\boldsymbol{\Delta}}^{-1}\left(\sum_{i=1}^{N}\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right) \tag{20}$$

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

11. We implicitly assume $\mathbb{P}(D_{ig_h}=1)$ is strictly between 0 and 1 for every $g_h\in\mathcal{G}\cup\{\infty\}$.

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 $\boldsymbol{D}_{g_h}(\boldsymbol{D}_{\infty}'\boldsymbol{\Delta}_{\infty}^{-1}\boldsymbol{D}_{\infty})^{-1}\boldsymbol{D}_{g_k}'$. ■

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. Because we have proved asymptotic normality, one can also use the usual nonparametric bootstrap. We derive an asymptotically linear representation of the ATT estimates in the Appendix that also allow for the multiplier bootstrap as in Callaway and Karami (2023).

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), strictly so 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

$$\widehat{\boldsymbol{\Delta}}_g = \frac{1}{N_g - 1} \sum_{i=1}^{N} D_{ig} \left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{g_G} \right) \left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{g_G} \right)' \tag{21}$$

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 2 (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:

$$H_0 : \tau_{g,q_g} = ... = \tau_{g,g-1} = 0 \tag{22}$$

■

Remark 3 (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.[12] 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. (Forthcoming), these can be imposed on the moment conditions as well.

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

We can also derive pre-treatment "placebo" effects by estimating a coefficient on the pre-treatment time periods. The imputation matrix that carries out this estimation is the usual projection matrix $\boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g})$. Under the no anticipation assumption,

$$\mathbb{E}((\boldsymbol{I}_g - \boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g}))\, \boldsymbol{y}_{i,t\leq g} \mid G_i = g) = \mathbf{0} \tag{23}$$

so that the properly standardized vector of pre-treatment residuals is asymptotically normal. ■

Remark 4 (Plotting Estimates)**.** The proposed estimator can be used to produce estimates for $y_{it}(\infty)$ in all periods for the treated observations:

$$\hat{y}_{it}(\infty) = P(\boldsymbol{F}_t, \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} + \overline{y}_{\infty,t} + \overline{y}_{i,t<T_0} - \overline{y}_{\infty,t<T_0} \tag{24}$$

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[13]. 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 5 (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 $\tilde{y}_{it}$.

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

Theorem 4. If Assumption 1 and 2 hold with $\boldsymbol{F}_t'\boldsymbol{\gamma}_i = \mathbf{0}$ almost surely, then $p = 0$. ■

If the null hypothesis is true, the more computationally burdensome QLD procedure is unnecessary for estimating the ATTs.[14] 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 Online Appendix for an additional test of the equality of the factor loadings' conditional means. ■

4 — Simulations

We present a brief simulation study to compare our estimator to different TWFE specifications. We specifically study the quasi-differencing factor estimation approach of Ahn et al. (2013) in the first stage because it is used in our empirical example. See Brown et al. (2023) for simulation evidence for common correlated effects as the first stage estimator. We consider the setting where $T = 8$ and treatment turns on starting in period 6 implying $T_0 = 5$. We draw $N = 200$ observations, which is a relatively small number for a nonlinear estimation problem.

We generate untreated potential outcomes following equation (1). We consider the setting with one factor that we generate as a time-trend $f_t = t$.[15] We generate the time fixed effects as $\zeta_t = 0.75 * \zeta_{t-1} + \nu_t$ where $\nu_t \sim N(0, 1)$. We generate the unit fixed effects as iid with $\mu_i \sim N(0, 4)$ and the factor loadings to be correlated with the unit fixed-effects by drawing from $\gamma_i \sim N(\mu_i, 1)$. The error term is generated as an $AR(1)$ process with correlation coefficient 0.75 and is uncorrelated with treatment status. We generate individual-level treatment effect heterogeneity by defining individual treatment effects $\tau_{i\ell}$ to be τ_ℓ times the unit fixed effect but then re-scale the individual effects to have mean equal to $\tau_6 = 1$, $\tau_7 = 2$, and $\tau_8 = 3$ and for the variance of $\tau_{i\ell}$ to be one. For example, $\tau_{i6} = (\mu_i + 2)/2$.

We generate a covariate $w_i = \gamma_i + \xi_i$ where ξ_i is white-noise measurement error. w_i will be used as a covariate in some TWFE specifications and as our instrument for our factor-model estimation. In the baseline simulation, we consider the case where $\xi_i \sim N(0, 1)$, which creates a

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

15. In this particular case, if the researcher knew that f_t took this form, then including unit-specific time-trends would fix this problem. However, we emphasize that f_t is generally not observable. We include this simple form of f_t so that the expected bias of TWFE is easy to compute: $t * (\mathbb{E}(\gamma_i \mid D_i = 1) - \mathbb{E}(\gamma_i \mid D_i = 0))$.

signal-to-noise ratio for the instrument of $1/2$. In a set of simulations, we vary the level of noise to see how the instrument strength affects estimates. These results will allow us to compare our methods to those that use noisy measurements of unobserved heterogeneity.

We consider three data-generating processes. First, we consider the true two-way error model where there are no interactive effects. In this case, the two-way fixed effects estimator should be unbiased. Second, we generate outcomes with the factor model described above. Treatment is then assigned completely randomly with probability of treatment at 50% for all units. This implies that the factor loadings are uncorrelated with treatment status, which Corollary 1 shows is sufficient for the TWFE imputation procedure to be consistent. Third, we generate treatment with probability increasing in the factor loading such that parallel trends fail (since treated units are more exposed to the time-trend in f_t). In particular, we form the term

$$\pi_i = 0.5 + \frac{\gamma_i}{\max_i \gamma_i - \min_i \gamma_i} \tag{25}$$

We normalize this term by the mean of π_i so that the unconditional probability of treatment stays at 50%.

We estimate event-study treatment effects using four estimators. First, we estimate the classical two-way error model using ordinary least squares (OLS), i.e. the TWFE estimator. Second, we estimate the two-way error model using the imputation estimator proposed by Borusyak et al. (Forthcoming) and Gardner (2021).[16] Third, we augment the two-way error model by including a noisy measure of the factor loadings. This is sometimes done by applied researchers in an attempt to control for confounders. That is, they model outcomes as

$$y_{it} = \mu_i + \lambda_t + w_i \beta_t + u_{it} \tag{26}$$

where w_i is a time-invariant covariate and β_t allows for trends to vary based on w_i. In the case where $w_i = \gamma_i$, 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

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

the QLD transformation of Ahn et al. (2013) to estimate the factors.[17] The covariate w_i is our instrument in the first stage to estimate the QLD parameters. See Remark 1.

Results are presented in table 1. Each panel presents results from each of the three data-generating processes described above. For each estimate, we present the average bias for the estimate as well as the mean-squared error. For Panel A where the outcomes are generated under the two-way fixed effect model (i.e. without a factor structure), all estimators are unbiased for the treatment effects, but the more robust factor imputation pays an efficiency cost with larger mean-squared error. However, this flips in Panel B where outcomes are generated under a factor model but with parallel trends holding for the two-way error model. In this case, all estimators are still unbiased but the factor imputation estimator is the most efficient because it absorbs the factor-structure that is present in the error term for the two-way error model.

Turning to where parallel trends does not hold in Panel C, we see that only our factor-imputation estimator is unbiased. This result emphasizes that our estimator is robust for parallel trend violations coming from differential exposure to macroeconomic factors. The magnitude of bias present in the two-way error models is growing from τ_6 to τ_8 due to the factor being a linear time-trend, implying parallel trend deviations grow worse over time.

It is worth noting that while including $w_i\beta_t$ in the model does remove some bias, 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 where the covariate w_i has different amount of noise added in. In particular, we choose different values of $\text{Var}(\xi_i)$ to have different signal-to-noise measures. The signal-to-noise definition is

$$\text{signal to noise ratio} = \frac{\text{Var}(\gamma_i)}{\text{Var}(\gamma_i) + \text{Var}(\xi_i)} \tag{27}$$

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,

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

Table 1 — Monte Carlo Simulation

Panel A: Two-way error model.

	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	0.00	0.01	-0.00	0.02	0.00	0.02
TWFE Imputation	0.01	0.01	0.00	0.02	0.01	0.02
TWFE Imputation with Covariates	0.01	0.01	0.00	0.02	0.01	0.02
Factor Imputation	-0.00	0.04	-0.01	0.11	-0.01	0.24

Panel B: Factor Model. Parallel Trends Hold

	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	0.00	0.11	0.00	0.43	0.01	0.95
TWFE Imputation	0.00	0.94	0.00	1.67	0.01	2.60
TWFE Imputation with Covariates	0.00	0.17	0.00	0.29	0.01	0.44
Factor Imputation	-0.00	0.02	-0.00	0.03	0.00	0.05

Panel C: Factor Model. Parallel Trends Do Not Hold

	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	-1.63	2.77	-3.27	11.05	-4.90	24.84
TWFE Imputation	-4.90	24.81	-6.53	44.12	-8.16	68.93
TWFE Imputation with Covariates	-0.92	1.06	-1.22	1.88	-1.53	2.93
Factor Imputation	0.01	0.03	0.01	0.05	0.02	0.09

Notes. This table presents a set of simulations with 10000 simulations. Each panel contains one of three data-generating processes described in the text. Each row in a panel consists of one of the four treatment effect estimators as described in the text. The columns report average bias and mean-squared error for the three post-treatment treatment effects.

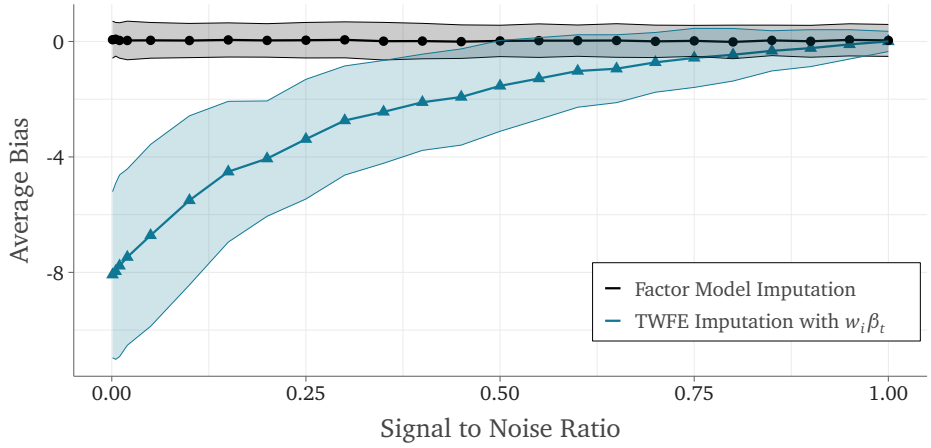


Figure 1 — Bias of TWFE Imputation with Covariates

Notes. This figure plots the average and empirical 95% confidence intervals for treatment effect estimates in the final period, $\hat{\tau}_8$. 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.

i.e. $w_i = \gamma_i$, the bias is completely removed. Regardless, the factor model imputation estimator is unbiased in all cases. This experiment echos the results of Kejriwal et al. (2021). However, we note that our results are still generous to 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 in the two-way error model. Indeed, we observe significant differences in both employment trends for treated 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. For this reason, we revisit this question with an alternative strategy to answer this question.

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.[18] We use a geocoded dataset 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. [19] 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.[20] Walmart is a more vertically integrated business, so we expect Walmart to compete in the retail and the wholesale sectors (Basker, 2005).

First, we estimate the two-way fixed effect imputation estimator proposed by Borusyak et al. (Forthcoming) 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),

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

19. 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.

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

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 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 remark 1. 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.[21] We use baseline shares to prevent our instruments from picking up on contemporaneous economic shocks that could be correlated with Walmart opening, i.e. to avoid violations of the exclusion restriction. 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.

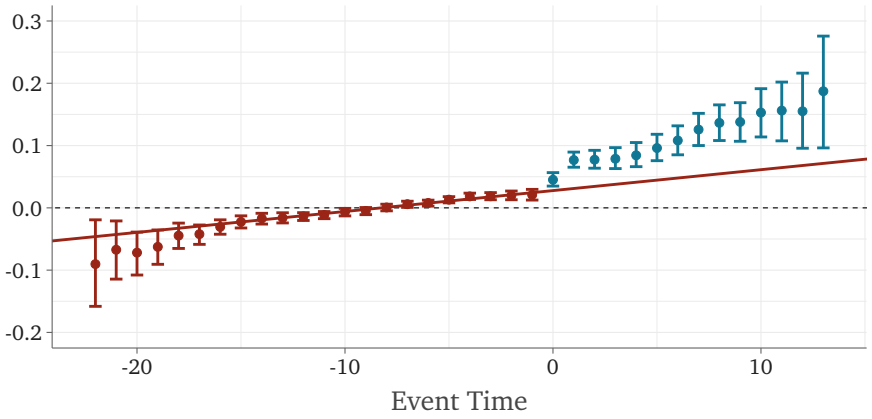
The results of our estimator are presented in panel (b) of figure 2 and figure 3.[22] For retail

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

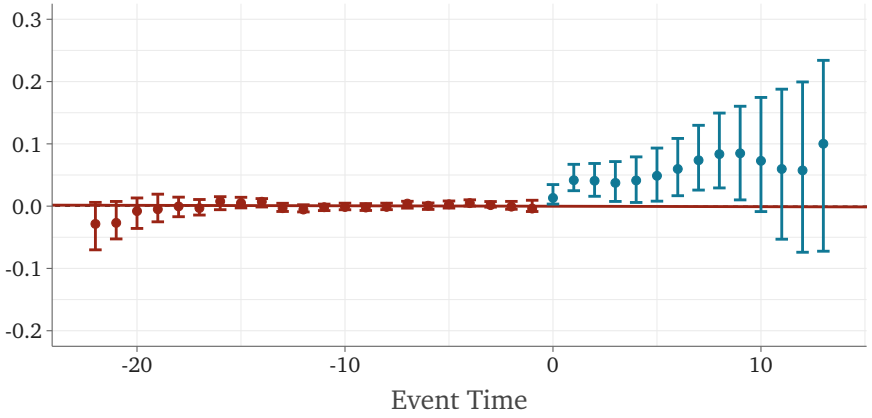
22. We carry out the test to determine the correct number of factors p following the discussion in Ahn et al. (2013). For retail, the p-value of the over-identification test were as follows: p = 0 with a p-value of 1.56e-5; p = 1 with a p-value of 0.001; p = 2 with a p-value of 0.133. Since $p = 2$ is 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



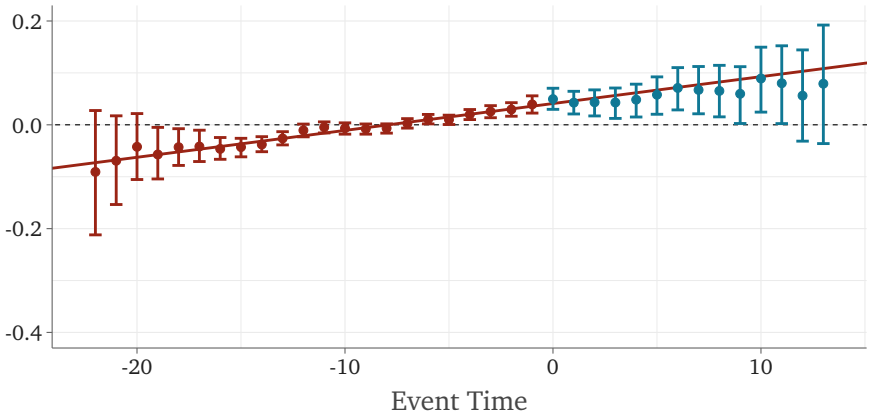
(b) Generalized Imputation Estimator



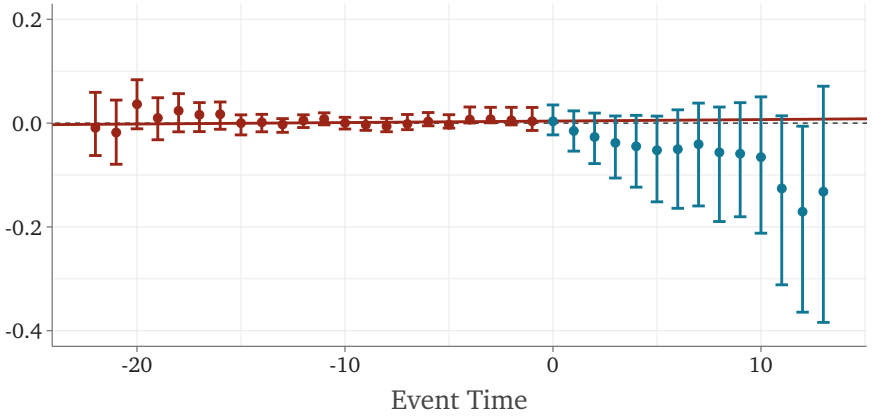
Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log retail employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (Forthcoming). Panel (b) estimates effects using the generalized imputation 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

(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 wholesale employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (Forthcoming). 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.

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 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 $\boldsymbol{F}$.[23] 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.

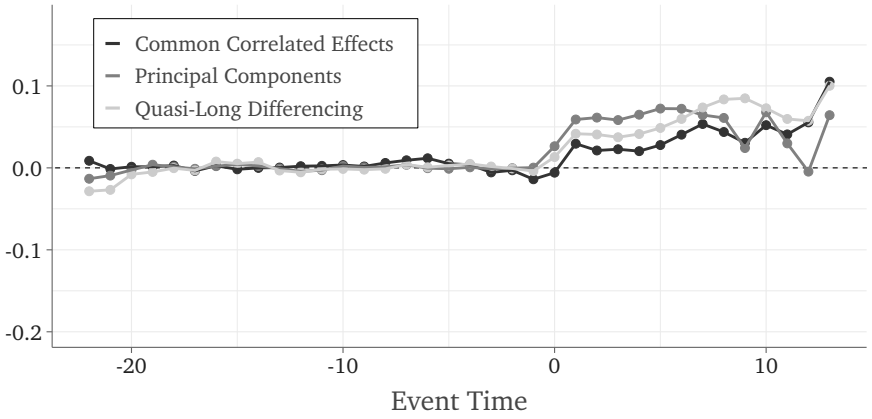
The results of each estimator are presented in figure 4. All three estimators are effective at removing underlying trends that the treated counties experienced. Moreover, the estimated effects are similar between estimators suggesting that all three are doing a good job at estimating the underlying factors. 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.

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 5. The standard errors on point estimates are far smaller, with estimates becoming strongly significant in wholesale Employment. 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.

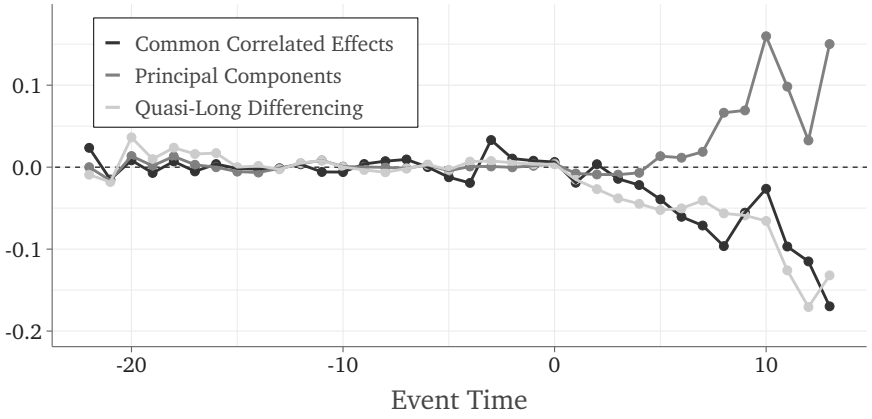
23. This imputation estimator is proposed by Xu (2017) in the context of large panels. The author uses an alternative identification strategy that fails to work in short-panels.

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

(a) log **Retail Employment**

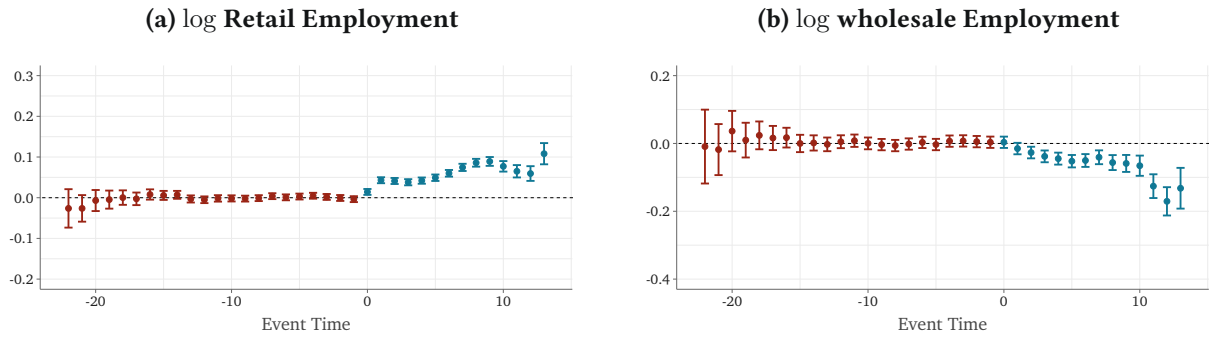


(b) log **wholesale Employment**



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 — 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.

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 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. Further work can 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.

References

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

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.

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.

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

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.

Cui, Guowei, Milda Norkutė, Vasilis Sarafidis, and Takashi Yamagata. 2021. "Two-stage instrumental variable estimation of linear panel data models with interactive effects." *The Econometrics Journal* 25 (2): 340–361. 10.1093/ectj/utab029.

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.

Fan, Jianqing, Yuan Liao, and Weichen Wang. 2016. "Projected principal component analysis in factor models." *Annals of statistics* 44 (1): 219. 10.1214/15-aos1364.

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."

Juodis, Artūras, and Vasilis Sarafidis. 2022. "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, and Evan Totty. 2021. "The Efficacy of Ability Proxies for Estimating the Returns to Schooling: A Factor Model-Based Evaluation." 10.2139/ssrn.3843260.

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.

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.

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. 10.1017/pan.2016.2.

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

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

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

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:

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

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

□

Proof of Lemma 2.1

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 3.1

Asymptotic normality is a consequence of well-known large sample GMM theory. See, for example, Hansen (1982).

We only need to derive the asymptotic variances. 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.

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

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:

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

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 $(\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D})^{-1}$. Δ 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

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

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 3.2

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 — 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 $ATT(g,t)$:

$$\delta = \sum_{g \in \mathcal{G}} \sum_{t > T_0} w(g,t) \tau_{gt} \tag{B1}$$

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:

$$\sqrt{N}\left((\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')' - (\boldsymbol{\theta}', \boldsymbol{\tau}')' \right) = - \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\boldsymbol{D}' \boldsymbol{\Delta}^{-1} \boldsymbol{D})^{-1} \boldsymbol{D}' \boldsymbol{\Delta}^{-1} \boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau}) \right) + o_p(1). \tag{B2}$$

This form comes from the fact that the weight matrix is positive definite with probability approaching one[24]. 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.

C — 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

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

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

$$\tilde{\boldsymbol{y}}_i = [\boldsymbol{I}_T - \boldsymbol{P}(\mathbf{1}_T, \mathbf{1}_{T_0})] (\boldsymbol{y}_i - \overline{\boldsymbol{y}}_\infty) \tag{C1}$$

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

$$d_{it}\tau_{it} + \tilde{u}_{it}, \tag{C2}$$

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:

$$\widehat{\tau}_{gt} = \frac{1}{N_g} \sum_i D_{ig} \tilde{y}_{it}, \tag{C3}$$

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

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

Theorem C1. 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

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{C4}$$

and

$$\sqrt{N_1}(\widehat{\tau}_{gt} - \tau_{gt}) \xrightarrow{d} N(0, V_1 + V_0) \tag{C5}$$

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 (C1) 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 C1

The transformed post-treatment observations are

$$\tilde{y}_{it} = \tau_{it} + u_{it} - \overline{u}_{\infty,t} - \overline{u}_{i,t<T_0} + \overline{u}_{\infty,t<T_0} \tag{C6}$$

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

$$\mathbb{E}(\tau_{it} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0} \mid D_{ig} = 1) = \mathbb{E}(\tau_{it} \mid D_{ig} = 1) \tag{C7}$$

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

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}) + \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(-\overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{C8}$$

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:

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}) + \sqrt{\frac{N_g}{N_\infty}} \left(\frac{1}{\sqrt{N_\infty}} \sum_{i=1}^{N} D_{i\infty}(-u_{i,t} + \overline{u}_{i,t<T_0}) \right) \tag{C9}$$

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

The first term has asymptotic variance

$$V_1 = \mathbb{E}\left(\Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big) \Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big)' \mid D_{ig} = 1 \right) \tag{C10}$$

and the second term has asymptotic variance

$$V_0 = \frac{\mathbb{P}(D_{ig} = 1)}{\mathbb{P}(D_{i\infty} = 1)} \mathbb{E}\left(\left(\overline{u}_{i,t<T_0} - u_{i,t} \right) \left(\overline{u}_{i,t<T_0} - u_{i,t} \right)' \mid D_{i\infty} = 1 \right) \tag{C11}$$

The result follows from the independence of the two sums.

D — 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

$$\mathbb{E}(y_{it}(\infty) \mid x_i, \mu_i, \boldsymbol{\gamma}_i, D_i) = \boldsymbol{x}_i \boldsymbol{\beta}_t + \mu_i + \lambda_t + \boldsymbol{F}_t' \boldsymbol{\gamma}_i \tag{D1}$$

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

$$\mathbb{E}\Big(\boldsymbol{H}(\boldsymbol{\theta})'(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}) \otimes \boldsymbol{w}_i \mid G_i = \infty\Big) = \mathbf{0} \tag{D2}$$

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

$$\mathbb{E}\Big((\tilde{\boldsymbol{y}}_{i,t\geq g} - \tilde{\boldsymbol{X}}_{i,t\geq g}\boldsymbol{\beta}_g) - \boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\geq g}, \tilde{\boldsymbol{F}}_{t<g})(\tilde{\boldsymbol{y}}_{i,t<g} - \tilde{\boldsymbol{X}}_{i,t<g}\boldsymbol{\beta}_g) - \boldsymbol{\tau}_g \mid G_i = g\Big) = \mathbf{0} \tag{D3}$$

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).

E — 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

$$\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) - \mathbb{E}(\boldsymbol{\gamma}_i) = \mathbf{0} \iff (\boldsymbol{A}^*)^{-1}(\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) - \mathbb{E}(\boldsymbol{\gamma}_i)) = \mathbf{0} \tag{E1}$$

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 E2. If $\mathbb{E}(\boldsymbol{\gamma}_i \mid G_i = g) = \mathbb{E}(\boldsymbol{\gamma}_i)$ for all $g \in \mathcal{G}$, then

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*_{g_G} = ... = \boldsymbol{\gamma}^*_{g_1} \tag{E2}$$

■