Dynamic Treatment Effect Estimation with Interactive Fixed Effects and Short Panels

Nicholas L. Brown[*] and Kyle Butts[†]

February 6, 2025

We study the estimation and inference of dynamic average treatment effect parameters when parallel trends hold conditional on interactive fixed effects and where units enter into treatment at different time periods. Our proposed generalized method of moments estimator consists of two parts: first, we estimate the unobserved time effects by applying the fixed-T consistent quasi-long-differencing estimator of Ahn et al. (2013) to the never-treated group. Second, we estimate the interactive fixed effects for treated groups post-treatment to recover their unobserved counterfactual outcomes. We subtract this quantity from the observed outcomes and average over group membership to achieve our estimator of the Average Treatment Effect on the Treated. 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. We also provide statistical software to implement our estimator in `Julia` and `R`.

JEL Classification Number: C13, C21, C23, C26

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

[*] Florida State University, Economics Department (nlb24c@fsu.edu)

[†] University of Arkansas, Economics Department (kbutts@walton.uark.edu)

1 — Introduction

Difference-in-differences estimators are one of the most popular causal inference tools. While computationally simple and easy to interpret, they rely on the strong parallel-trends assumptions. In many empirical settings, treatment is assigned non-randomly based on trends in economic variables, rendering this method unreliable. For example, in urban economics, place-based policies target locations with worsening labor markets (Neumark and Simpson, 2015), new apartments are built in appreciating neighborhoods (Asquith et al., 2021; Pennington, 2021), and firms open new stores in growing economies (Basker, 2005; Neumark et al., 2008). Estimation of treatment effects in this setting is confounded by the pre-existing economic trends. It is common 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 decline, consumer tastes for walkable neighborhoods caused certain existing neighborhoods to become increasingly demanded, and national macroeconomic changes benefited certain counties.

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

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

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

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

[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

factor loadings $\boldsymbol{\gamma}_i$ denoting a unit's exposure to the shocks. Another interpretation 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 (TWFE) event-study model suggests that Walmart opened stores in counties that had higher retail employment growth prior to the opening (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 the estimated positive impacts are due to pre-existing trends rather than the effect of Walmart per se. However, there seems to be a discrete jump when the Walmart opened. The goal then is to remove these pre-existing trends to isolate the treatment effect. It is plausible to assume that during their period of mass expansion, Walmart selected appealing locations based on their local demographic background and national economic trends, while ignoring transitory local economic shocks. Our framework allows this type of selection mechanism and effectively 'controls' for these pre-existing trends.

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

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.

Our general identification argument has two major benefits. First, fixed-T consistent estimation of $\boldsymbol{F}_t$ is possible through a variety of approaches, most popularly quasi-differencing (Ahn et al., 2001, 2013; Callaway and Karami, 2023) and cross-sectional averages (Westerlund et al., 2019; Juodis and Sarafidis, 2022a,b; Brown et al., 2023). Our identification result provides a recipe for using any consistent estimator of the factors to estimate treatment effects, as long as there are sufficiently many pre-treatment observations to identify the common factors. 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 that uses factors estimated from the never-treated group to recover the factor loadings for the treated groups. The resulting estimator takes the form of a generalized method of moments (GMM) estimator, which allows estimation and inference via common statistical software.[3] We implement the estimator using the quasi-long-differencing (QLD) transformation of Ahn et al. (2013) because it is consistent when the number of time periods is fixed.

The QLD estimator requires a set of 'instruments', or proxy variables, that correlate with the factor loadings $\boldsymbol{\gamma}_i$ to identify the factor space. In our application, we estimate the impact of Walmart opening in a county on local retail employment. To estimate the factor space, we use baseline county-level demographic variables, such as employment and educational characteristics, that are likely correlated with Walmart's opening decisions, itself driven by the common factor structure. Hatch and Clinton (2000) provide evidence that such characteristics drive long-term trends in both retail spending and employment nationally, and are therefore good proxies for the factor loadings, $\boldsymbol{\gamma}_i$. Unlike a traditional IV approach, these variables can be either internal or external, and must explicitly correlate with the unobserved effects.

[3] We provide easy to use open-source software for our proposed estimation strategy in Julia https://github.com/kylebutts/QLD.jl and an R package that accesses this routine. While there have been computational concerns around the QLD estimator, our software is highly optimized. In our empirical application with tens of thousands of observations, our procedure produces point estimates and analytic standard errors in less then 0.2 seconds.

Relation to Literature

Current estimators that allow for selection based on a factor model either require (i) the number of time periods available is large, e.g. synthetic control (Abadie, 2021), factor model imputation via principal components (Gobillon and Magnac, 2016; Xu, 2017; Bai and Ng, 2021), and the matrix completion method (Athey et al., 2021; Fernández-Val et al., 2021); or (ii) that an individual's error term u_{it} is uncorrelated over time (Feng, 2021; 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.

A recent set of papers has proposed 'imputation' in the TWFE setting (Borusyak et al., 2024; Gardner, 2021; Wooldridge, 2021).[5] While these estimators are consistent in fixed-T settings, these approaches only allow for level fixed effects and preclude interactions like in equation (1). Borusyak et al. (2024) allow a structure similar to equation (1) but requires the the factors $\boldsymbol{F}_t$ be observed. We generalize these techniques by proposing an estimator that imputes the untreated potential outcomes under the more general (1) with unobserved interactive effects.

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

[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.

[5] The imputation procedure has been proposed in various settings in causal inference (Imbens and Rubin, 2015).

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 when using the QLD estimator in the first stage. We include a small Monte Carlo experiment in Section 4 to examine the finite-sample performance of our estimator. Finally, Section 5 contains our application and Section 6 leaves with some concluding remarks.

2 — Model and Identification

We assume a balanced panel data set with units $i = 1, \ldots, N$ and periods $t = 1, \ldots, T$. Treatment turns on in different periods for units in different groups; we denote these groups by the period they start treatment. For each unit, we define G_i to be unit i's group with possible values $\{g_1, \ldots, g_G\} \equiv \mathcal{G} \subseteq \{2, \ldots, T\}$.[6] 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 can be a function of group-timing, so we denote $y_{it}(g)$ as the treated potential outcome for unit i at time t if they were treated at time g. 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:

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

These quantities represent the average effect of treatment at time t for units that start treatment in period g for $t \geq g$. Once these quantities are obtained, it is trivial to estimate other aggregations,

[6] The number of possible treatment cohorts is fixed asymptotically because it is strictly less than the number of time periods.

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$^{\ell}$'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 (Random sampling of outcomes). The random vectors $\{(\boldsymbol{d}_i, \boldsymbol{\gamma}_i, \boldsymbol{u}_i)\}$ are independent and identically distributed over i and have finite moments up to the fourth order. ■

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

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

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

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

Assumption 2 imposes a factor model for the untreated potential outcomes. We discuss the inclusion of covariates and the subsequent relaxation of assumption 2 in the Appendix. We allow for heterogeneous and dynamic treatment effects of any form, i.e. $y_{it}(g) = \tau_{igt} + y_{it}(\infty)$. We also allow arbitrary serial correlation among the idiosyncratic errors. We assume the common factors $\boldsymbol{F}_t$ are nonrandom parameters and the number of factors p is both known and fixed in the asymptotic analysis. In practice, the number of factors can be consistently estimated; see section 2.2.

Assumption 2 is more general than the standard parallel trends 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. For example, in our application, employment is measured at the county level, but Walmart ultimately determines treatment status (when and where a new Walmart opens). One potential common factor is skill-biased technological change: Walmart might anticipate continued development in computing technologies leading to automation of certain manual tasks. A county's exposure to these changes will be a (possibly) nonlinear function of workforce education levels, trade exposure, and unobserved variables like entrepreneurial talent. If Walmart believes a county is negatively exposed to future positive technological shocks, it may opt to not open a store, even if the county's contemporaneous economic outlook is positive. This form of selection is allowed by Assumption 2. On the other hand, Assumption 2 rules out Walmart opening stores based on location-specific shocks.

The two-way error model cannot generally accommodate differential exposure.[7] 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 would not hold. If either of the two cases hold for all g and t, the two-way error model is correctly specified.[8] However, these are knife-edge cases which are not the focus of the paper.

Our Assumption 2 allows for the factor loadings to be correlated with treatment timing and opens up treatment effect estimation for a much broader set of empirical questions. The key econometric challenge lies in that we do not observe $y_{it}(\infty)$ whenever $d_{it} = 1$. Our goal is to consistently estimate $\mathbb{E}[y_{it}(\infty) \mid G_i = g]$ under equation (1) to consistently estimate group-time average treatment effects. Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024) implicitly rely on this insight in studying the two-way error model.

Prior work on estimation of average treatment effects in a factor model setting focus on finding conditions that allow for estimation of $\boldsymbol{\gamma}_i$ and $\boldsymbol{F}_t$ jointly as in Gobillon and Magnac (2016), Xu (2017), and Bai and Ng (2021), or a generalized version of a factor model as in Feng (2021) and Arkhangelsky et al. (2021). These techniques require a large number of pre-treatment periods 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

[7] The following derivation is also shown in Callaway and Karami (2023), but we are repeating it here for exposition.

[8] We explicitly prove this result later.

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 treated at time g. Define $\boldsymbol{y}_{i,t<g}$ and $\boldsymbol{y}_{i,t\geq g}$ as respectively the first $(g-1)$ and last $(T-g+1)$ outcomes for unit i, or the 'pre-treatment' and 'post-treatment' outcomes. Define $\boldsymbol{F}$ to be the matrix of factors with rows given by $\boldsymbol{F}_t'$. We similarly define $\boldsymbol{F}_{t<g}$ and $\boldsymbol{F}_{t\geq g}$ as the first and last rows of matrix $\boldsymbol{F}$. Equation (3) implies

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

If the factors were observed, we could consistently estimate the mean values of the p-vector of average factor loadings for treated group $G_i = g$ via ordinary least squares; if $\text{Rank}(\boldsymbol{F}_{t<g}) = p$, the coefficient from the population regression of $\mathbb{E}[\boldsymbol{y}_{i,t<g}(\infty) \mid G_i = g]$ 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 $\mathbb{E}[y_{i,t}(\infty) \mid G_i = g]$ for $t \geq g$ by multiplying $\boldsymbol{F}_t'$ by the OLS estimate from the prior infeasible regression. We then obtain $\mathbb{E}[y_{it}(\infty) \mid G_i = g]$ for the post-treatment outcomes, which we can subtract from y_{it} and average over the respective sample to obtain an estimate of ATT(g, t).

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

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

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

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

$g \in \mathcal{G}$,

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

for $t \geq g$.

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

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

■

All proofs are contained in the Appendix. Theorem 1 shows that we can identify the ATTs if we know the factor matrix, provided there are enough pre-treatment observations. Because we require $\text{Rank}(\boldsymbol{F}_{t\leq T_0}) = p$, we need at least as many observations before any unit is treated as there are unobserved effects. The second part of the theorem states that any rotation of the true factor matrix can be used in the imputation matrix. This result is important because it is well understood that $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are not separately identified (Bai, 2009; Ahn et al., 2013; Xu, 2017). All of the estimators discussed so far can at best approximate the column space of the factors because both $\boldsymbol{F}_t$ and $\boldsymbol{\gamma}_i$ are unobserved. The second part of the theorem shows that our identification scheme allows for this class of estimators. To see how, note that $\boldsymbol{F} \in \text{col}(\boldsymbol{F}^*)$ implies the existence of a $m \times p$ matrix $\boldsymbol{A}$ such that $\boldsymbol{F}^*\boldsymbol{A} = \boldsymbol{F}$. Thus

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

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

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

Borusyak et al. (2024) who study the additive error model. In fact, Gardner (2021) and Borusyak et al. (2024) implicitly use the imputation matrix but with known factors.

Theorem 1 shows we can apply these conclusions to any estimator that achieves fixed-T consistency by asymptotically spanning the factor space. One such estimator that we focus on in this paper is the QLD estimator of Ahn et al. (2013).

2.2. *Quasi-Long-Differencing*

A leading example of a set of moment equations for factor-space estimation comes from Ahn et al. (2013). They normalize the factors as

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Theta} \\ -\boldsymbol{I}_p \end{pmatrix} \tag{9}$$

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

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

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

$$\mathbb{E}[\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{y}_i \otimes \boldsymbol{w}_i \mid G_i = \infty] = \boldsymbol{0} \tag{11}$$

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 elaborate on these conditions below). Essentially, we require the instruments to be strictly exogenous with the defactored errors, but correlate strongly with the factor loadings $\boldsymbol{\gamma}_i$. We also require there be at least as many instruments as factors. We discuss the practical selection of instruments $\boldsymbol{w}_i$ in Section 5. It is important to stress that variables in $\boldsymbol{w}$ are instruments for the factors and not for treatment, i.e. we do not view $\boldsymbol{w}$ as shifting units into treatment.

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, Ahn et al. (2013)

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

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. We can allow $\boldsymbol{w}_i$ to come from either external variables or covariates that have a linear effect on the outcome. We demonstrate the inclusion of covariates in the Appendix.

Because the QLD matrix is a function of p, we need a consistent estimator of the number of common factors. Ahn et al. (2013) propose a number of consistent estimators of p. First, they consider the usual Hansen-Sargan over-identifying test restriction. Let

$$J(\boldsymbol{\theta}) = \left(N_\infty^{-1}\sum_{i=1}^{N} D_{i\infty}\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\right)' \left(N_\infty^{-1}\sum_{i=1}^{N} D_{i\infty}\boldsymbol{g}_{i\infty}(\tilde{\boldsymbol{\theta}})\boldsymbol{g}_{i\infty}(\tilde{\boldsymbol{\theta}})\right)^{-1} \left(N_\infty^{-1}\sum_{i=1}^{N} D_{i\infty}\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})\right) \tag{12}$$

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

We make the following assumption on the instruments and factor matrix to guarantee the identification of the rotated factors $\boldsymbol{F}(\boldsymbol{\theta})$:

Assumption 4 (QLD identification).

(i) $\text{Rank}(\boldsymbol{F}) = \mathbb{E}[\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i' \mid G_i = \infty] = p \leq T_0$.

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

(ii) There exists a $k \times 1$ vector of observed variables $\boldsymbol{w}_i$ such that $\mathbb{E}[\boldsymbol{u}_i \mid G_i = \infty, \boldsymbol{\gamma}_i, \boldsymbol{w}_i] = 0$ and $\mathbb{E}\left[\boldsymbol{I}_{(T-p)} \otimes \boldsymbol{w}_i \boldsymbol{\gamma}_i' \mid G_i = \infty\right]$ has full rank.

Assumption 4 applies the 'Basic Assumptions' of Ahn et al. (2013) to our setting. Part (i) defines the number of factors, as also seen in Bai (2009).[11] We require the rank of the factor matrix to be strictly less than T because we need a baseline period to perform the QLD transformation. If $p = T$, then the QLD transformation is undefined.

Part (ii) requires instruments that are strongly correlated with the unobserved factor loadings. We do not put restrictions on these instruments; they may be time-varying or time constant. We do require that the instruments satisfy the two standard instrument requirements: relevancy and exogeneity. 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 characteristics that the researcher thinks might be driving differential trends. In our application, we use county-level demographic characteristics like share of college-educated residents as instruments. During this time, retail employment growth was strongly correlated with college-educated share, so baseline share is a natural 'proxy' for the time-invariant exposures γ_i (see Hatch and Clinton (2000) for discussions). The exogeneity restriction requires that the instrument values are uncorrelated with location-specific idiosyncratic shocks. For example, this condition would fail if individuals within a county decided to get a college education on the basis of short term economic fluctuations, which we find unlikely. We reiterate that $\boldsymbol{w}_i$ are instruments for the factors and not for treatment.

2.3. Two-Way Error Model

We now demonstrate how to explicitly control for additive effects before estimating the more general interactive effects. While the additive structure is a special case of the factor model, we consider the special case because manually eliminating the additive effects saves degrees of freedom to estimate the factor model and provides efficiency by reducing the burden on the QLD estimator.[12]

[11] Our assumptions implicitly require the factors to be 'strong' as in Bai (2009) and Ahn et al. (2013) in that all factors that affect at least one treated group affect the entire never-treated group, because we will use the never-treated group to predict the factors.

[12] Pesaran (2006) also allows for so-called "known factors" (like an additive intercept).

We might consider first using the TWFE imputation procedures of Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024) to obtain residuals that are free of the additive effects, then apply the QLD method to estimate the remaining interactive effects. However, such a procedure will not maintain a common factor structure between the untreated and treated groups, meaning their proposed estimators cannot be used as a first step before estimating the interactive effects. 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 we cannot use the control sample to estimate the factors for the treated samples.

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

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

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

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

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., 2024). Second, it preserves a common factor structure for all units and time periods[13]. The TWFE imputation estimator of Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024) would not share this property because they estimate μ_i and λ_t based on the full sample $d_{it} = 0$, while we use a specific subsample.

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 remove additive effects while preserving a common factor structure for the QLD estimation. 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:

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

Estimation can then proceed on $\tilde{y}_{it}$.

We can also see that if there is no time variation in the factors or if the factor loadings are mean independent of treatment status, the interactive effects structure will be zero in expectation, meaning TWFE will be consistent. The TWFE imputation methods mentioned above will all be simpler to estimate and likely more efficient. See the Appendix for tests of consistency of the TWFE estimator.

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

3 — Estimation and Inference

This section considers estimation of the group-time ATTs. Our moment conditions lead to a simple GMM estimator for which inference is standard and can be computed via routine packages in common statistical software. Further, we can use the moment conditions to test the fundamental features of the model.

3.1. Asymptotic Normality

Equations (7) and (11) provide us with the necessary moment conditions to estimate the ATTs. We collect them here in their unconditional form:

$$\begin{aligned}\mathbb{E}[\boldsymbol{h}_{i\infty}(\boldsymbol{\theta})] &= \mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty}=1)}\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{y}_i \otimes \boldsymbol{w}_i\right] = \mathbf{0}\\ \mathbb{E}[\boldsymbol{h}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})] &= \mathbb{E}\left[\frac{D_{ig_G}}{\mathbb{P}(D_{ig_G}=1)}\left(\boldsymbol{y}_{i,t\geq g_G} - \boldsymbol{P}(\boldsymbol{F}_{t\geq g_G}(\boldsymbol{\theta}),\boldsymbol{F}_{t<g_G}(\boldsymbol{\theta}))\boldsymbol{y}_{i,t<g_G} - \boldsymbol{\tau}_{g_G}\right)\right] = \mathbf{0}\\ &\vdots\\ \mathbb{E}[\boldsymbol{h}_{i1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})] &= \mathbb{E}\left[\frac{D_{ig_1}}{\mathbb{P}(D_{ig_1}=1)}\left(\boldsymbol{y}_{i,t\geq g_1} - \boldsymbol{P}(\boldsymbol{F}_{t\geq g_1}(\boldsymbol{\theta}),\boldsymbol{F}_{t<g_1}(\boldsymbol{\theta}))\boldsymbol{y}_{i,t<g_1} - \boldsymbol{\tau}_{g_1}\right)\right] = \mathbf{0}\end{aligned}$$

where $\boldsymbol{\tau}_g = (\tau_{gg}, ..., \tau_{gT})'$ is the vector of post-treatment treatment effects. We stack these over g as $\boldsymbol{\tau} = (\boldsymbol{\tau}'_{g_1}, ..., \boldsymbol{\tau}'_{g_G})'$. The first set of moment conditions identify the factor space by Assumption 4 and the remaining moments identify the τ_{gt} via our imputation method. Implementation requires replacing $\mathbb{P}(D_{ig}=1)$ with its sample counterpart N_g/N.

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

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

The moment functions $\boldsymbol{h}_i(\boldsymbol{\theta},\boldsymbol{\tau}) = (\boldsymbol{h}_{i\infty}(\boldsymbol{\theta})', \boldsymbol{h}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})', ..., \boldsymbol{h}_{ig_1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})')'$ are collected into a vector. We define $\boldsymbol{\Delta} = \mathbb{E}[\boldsymbol{h}_i(\boldsymbol{\theta},\boldsymbol{\tau})\boldsymbol{h}_i(\boldsymbol{\theta},\boldsymbol{\tau})']$ which is positive definite by Assumptions 4 and 5. Then our GMM estimator $(\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')'$ solves

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

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

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

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

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

Remark 1 (Inference)**.** Valid inference is easy to obtain because we use a GMM framework. Numerical standard errors are computed and reported by most routine statistical packages and is reproduced by the `Julia` and `R` packages that we provide. The nonparametric panel bootstrap is also available, though we derive in the Appendix an asymptotically linear representation of the estimator so one can utilize the multiplier bootstrap for uniform inference as in Callaway and Karami (2023).

We derive the functional forms of $\boldsymbol{D}_g$ in the Appendix for the computation of analytic standard errors. Estimating $\boldsymbol{D}_g$ requires we replace expectations with sample averages and unknown parameters with their estimators. We can also estimate the variance of the moment functions $\boldsymbol{\Delta}_g$ using a nonparametric variance estimator similar to Pesaran (2006):

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

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 quantity is then consistent for $\boldsymbol{\Delta}_g$:

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

■

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.[14] Researchers can perform heterogeneity analyses by estimating the ATTs by gender, race, age, or other observed characteristics. All one needs to do to estimate such aggregate effects is to correctly specify the unconditional treatment effect moment conditions. If there are *a priori* restrictions on treatment effects as in Borusyak et al. (2024), these can be imposed on the moment conditions as well. ■

Remark 4 (Assessing Model Fit)**.** Our key identifying assumption is that after subtracting off (the estimated) factor model, there is no remaining confounders in the post-treatment periods that are correlated with treatment. As is common in the difference-in-differences literature, we can assess the plausibility of this assumption using pre-treatment “placebo" effects. This is done by extending the projection matrix into the pre-treatment periods $\boldsymbol{P}(\boldsymbol{F}_{t\leq g}, \boldsymbol{F}_{t\leq g})$, which gives the usual projection matrix for $\boldsymbol{F}_{t\leq g}$. Under the no anticipation assumption,

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

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

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

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

$$\hat{y}_{it}(\infty) = \boldsymbol{P}(\boldsymbol{F}_t', \boldsymbol{F}_{t<g})\boldsymbol{y}_{i,t<g} + \overline{y}_{\infty,t} + \overline{y}_{i,t<T_0} - \overline{y}_{\infty,t<T_0} \tag{24}$$

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[15]. 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 by group and plot the average of y_{it} and the average of $\hat{y}_{it}(\infty)$ separately for each $g \in \mathcal{G}$. 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 $\tilde{y}_{it}$. If Assumption 1 and 2 hold with $\boldsymbol{F}_t'\boldsymbol{\gamma}_i = \boldsymbol{0}$ almost surely for all (i, t), then $p = 0$. One should then proceed with a more efficient estimator

[15] Leave this part out if you do not remove the additive effects by hand.

that is consistent under the additive model (Gardner, 2021; Wooldridge, 2021; Borusyak et al., 2024). ■

4 — Simulations

We present a brief simulation study to compare our estimator to alternatives in the literature. Since the focus of this paper is to propose a fixed-T estimator, while the majority of estimators are large-T based, we will present simulations with $T_0 = 4$ and $T_0 = 12$. There is a single post-treatment period where the effect of treatment is $\tau_{T_0+1} = 1$. We draw $N = 300$ observations, which is a moderately small number for a nonlinear estimation problem. We consider four different DGPs detailed in Table 1, where cross-sectional variables are iid.

We generate two factor loadings, $\mu_i \sim N(1, 1)$ and $\gamma_i \sim N(1, 1)$. We generate the first factor $f_{1t} = t$ as a linear time-trend so that the difference in trends is easy to characterize. We generate the second factor as an AR(1) process: $f_{2t} = 0.75 f_{2,t-1} + v_t$ where v_t is serially independent and has variance 0.5 for all t. We draw the 13×2 matrix $\boldsymbol{F} = (f_1, f_2)$ once for all simulations. When $T_0 = 4$, we use the last 5 rows so that the post-period $f_{1,T}$ and $f_{2,T}$ are the same as when $T_0 = 12$.

DGP 1 uses a two-way error model, $\mu_i + f_{2t}$, which is consistent with the usual parallel trends assumption. In DGPs 2-4, we use the multiplicative model $f_{1t}\mu_i + f_{2t}\gamma_i$. In DGPs 1-2, we assign treatment randomly with probability 1/2. Parallel trends holds in both DGPs because $(\mu_i, \gamma_i)' \perp\!\!\!\perp D_i$. In DGPs 3 and 4, we generate treatment as $\mathbb{1}(\mu_i > 0)$ so that treated units are more exposed to the linear time-trend than untreated units, inducing non-parallel trends.[16]

In DGPs 1-3, the idiosyncratic error ϵ_{it} is drawn as $N(0, 1)$ iid over time. In DGP 4, we replace ϵ_{it} with an AR(1) process ζ_{it}, where $\zeta_{it} = 0.75\zeta_{i,t-1} + z_{it}$. The innovation z_{it} is iid over time with error variance $(1 - 0.75^2)$ so that the variance of ζ_{it} is also 1. DGP 4 aims to to highlight the problems that autocorrelated errors can cause for the existing large-T factor model estimators.

We consider nine alternative estimators. The first two are versions of the TWFE imputation estimators of Gardner (2021), Wooldridge (2021), and Borusyak et al. (2024), and should only perform well in the first two DGPs when parallel trends holds. The estimator 'TWFE' assumes an untreated potential outcome model of $y_{it}(0) = c_i + \theta_t + u_{it}$. $\mathbb{E}[c_i \mid D_i = 1]$ and θ_t are estimated

[16] To help make sense of our bias estimates, DGPs 3 and 4 have a difference in $y_{it}(0)$ between the treated and untreated groups in the post-period of $(\mathbb{E}[\mu_i \mid D_i = 1] - \mathbb{E}[\mu_i \mid D_i = 0])\, f_{1T} \approx (1.8 - 0.20) * 13 = 20.8$.

Table 1 — Simulation DGPs

	Error Term	Selection
DGP1	$\mu_i + f_{2t} + \epsilon_{it}$	Random
DGP2	$f_{1t}\mu_i + f_{2t}\gamma_i + \epsilon_{it}$	Random
DGP3	$f_{1t}\mu_i + f_{2t}\gamma_i + \epsilon_{it}$	$\mathbb{1}(\mu_i > 0)$
DGP4	$f_{1t}\mu_i + f_{2t}\gamma_i + \zeta_{it}$	$\mathbb{1}(\mu_i > 0)$

Notes. This table summarizes the data-generating processes used in the simulations. In all simulations, the probability of treatment is 50%.

via OLS on the untreated sample.[17] Most modern treatment effect analyses also condition on observed covariates. We generate two “observed covariates" as

$$w_{i1} = \mu_i + \xi_{i1} \tag{25}$$

$$w_{i2} = \gamma_i + \xi_{i2} \tag{26}$$

where ξ_{i1} and ξ_{i2} are mutually independent $N(0,1)$ errors. If researchers believe that their observed covariates are correlated with the unobserved drivers of non-parallel trending, they often control for the baseline values while allowing for time-varying slopes. Let $\boldsymbol{w}_i = (w_{i1}, w_{i2})'$; then the estimator ‘TWFE with $\boldsymbol{w}_i'\boldsymbol{\beta}_t$’ is the TWFE imputation estimator of Borusyak et al. (2024) that estimates time-varying slopes on the covariates $\boldsymbol{w}_i$: $y_{it}(0) = c_i + \theta_t + \boldsymbol{w}_i\beta_t + u_{it}$ using the untreated sample. In a second set of simulations below, we will generate $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})'$ with varying amounts of noise to highlight issues with using noisy ‘proxies’ in a TWFE model.

To compare our QLD imputation estimator to an already popular factor model based imputation estimator, we include two versions of the generalized synthetic control estimator of Xu (2017). He proposes an imputation estimator that estimates the unobserved factors and factor loadings via Bai (2009)’s least squares principal components algorithm. He estimates the factors and factor loadings using the untreated observations. However, this method generally requires $T_0 \to \infty$ for consistency and may not perform well in our setting. The estimator ‘Generalized Synth (p known)’ is the Xu (2017) estimator but with $p = 2$ treated as known. The estimator ‘Generalized Synth (p

[17] When we say ‘untreated sample’, we mean the collection of both groups’ pre-treatment observations along with the never-treated group’s post-treatment observations.

unknown)' is the same estimator but uses the mean square prediction error cross validation test of Xu (2017) to estimate the number of factors. Additionally, we include the augmented synthetic control estimator of Ben-Michael et al. (2021), which combines the classic synthetic control method with a form of 'bias correction' estimating a factor model for untreated outcomes, similar to Xu (2017).

Finally, we include three versions of the estimator proposed in Section 3. The first is an infeasible estimator that simply treats $\boldsymbol{F}$ as known. This estimator, 'Factor Imputation ($\boldsymbol{F}$ known)', does not require a first-stage estimator of the factors and takes the form

$$\widehat{\boldsymbol{\tau}} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \left(\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}(\boldsymbol{F}_{t>T_0}, \boldsymbol{F}_{t\leq T_0}) \boldsymbol{y}_{i,t\leq T_0} \right) \tag{27}$$

We then include two versions that use QLD to estimate the factors, meaning we use the GMM estimator in equation (20). The first, 'QLD (p known)', treats $p = 2$ as known. The second, 'QLD (p estimated)', uses the sequential testing method for p as described in Section 2.2 at the 10% significance level. Both use moment conditions in equation (11) to estimate the factor parameters with instruments w_{i1} and w_{i2}.

4.1. Main Results

Results are presented in Table 2 for $T_0 = 4$ and Table 3 for $T_0 = 12$. For each data-generating process, we report the bias, root mean squared error (RMSE), and empirical coverage of 95% confidence intervals for the treatment effect estimator (where $\tau_{T_0+1} = 1$ in each setting). The second panel of the table describes the data generating processes corresponding to Table 1.

The TWFE estimator performs well in terms of bias and RMSE for DGPs 1 and 2. Even though we generate a factor model for DGP 2, treatment is randomly assigned and so parallel trends holds. The factor model is subsumed in the error and inflates the variance, but does not cause substantial bias. For DGPs 3 and 4, TWFE performs unacceptably poorly. We can see that performance worsens moving from $T_0 = 4$ to $T_0 = 12$ even though the post-period difference-in-$Y(0)$ is the same. Increasing the number of time periods increases deviations from treated and untreated units due to the linear trend, in line with recent work by **?**. It is interesting to note that the bias of the TWFE estimator falls substantially, almost by half, when the proxy variables are used as controls.

Table 2 — Simulation with $T_0 = 4$

	DGP 1			DGP 2			DGP 3			DGP 4		
Estimator	**Bias**	**RMSE**	**Cov.**	**Bias**	**RMSE**	**Cov.**	**Bias**	**RMSE**	**Cov.**	**Bias**	**RMSE**	**Cov.**
TWFE	0.00	0.13	94.7%	0.01	0.32	94.2%	3.98	3.99	0.0%	3.98	3.99	0.0%
TWFE with $\boldsymbol{w}_i\beta_t$	0.00	0.13	94.6%	0.00	0.25	94.7%	2.93	2.94	0.0%	2.92	2.93	0.0%
Augmented Synthetic Control	0.00	0.16	82.8%	0.01	0.23	82.7%	1.01	1.52	77.6%	0.87	1.21	54.6%
Generalized Synth (p known)	-0.02	0.24	95.0%	0.06	0.23	92.8%	0.29	0.53	91.2%	0.58	0.62	27.4%
Generalized Synth (p estimated)	-0.11	0.19	86.9%	0.73	1.25	63.0%	0.39	0.50	71.9%	0.46	0.51	50.5%
Factor Imputation ($\boldsymbol{F}$ known)	0.00	0.09	94.9%	0.00	0.10	95.2%	0.00	0.10	94.9%	-0.00	0.08	94.0%
QLD (p known)	-0.01	0.17	95.5%	0.00	0.14	95.6%	0.00	0.55	95.2%	-0.01	0.42	94.7%
QLD (p estimated)	-0.03	0.16	92.6%	0.00	0.14	95.5%	0.00	0.55	95.2%	-0.01	0.42	94.7%
Model	TWFE			Factor			Factor			Factor		
Treatment	Random			Random			$\mathbb{1}(\gamma_i > 0)$			$\mathbb{1}(\gamma_i > 0)$		
Parallel Trends	✓			✓								
AR(1) Error Term										✓		

Notes. This table presents a set of simulations with 1000 iterations. Each row in a panel consists of a treatment effect estimator as described in the text. There are 4 different data-generating processes as described in the main text with key details listed in the second portion of the table. The columns present the bias and mean-squared error of the estimate of $\hat{\tau}$ relative to the true effect of $\tau_{T_0+1} = 1$. Additionally, confidence intervals are formed and 'Cov.' presents the percent of 95% confidence intervals containing the true treatment effect.

Table 3 — Simulation with $T_0 = 12$

	DGP 1			DGP 2			DGP 3			DGP 4		
Estimator	**Bias**	**RMSE**	**Cov.**	**Bias**	**RMSE**	**Cov.**	**Bias**	**RMSE**	**Cov.**	**Bias**	**RMSE**	**Cov.**
TWFE	-0.00	0.12	95.2%	-0.01	0.76	95.5%	10.38	10.39	0.0%	10.38	10.39	0.0%
TWFE with $\boldsymbol{w}_i\beta_t$	-0.00	0.12	95.1%	0.01	0.56	94.5%	7.61	7.63	0.0%	7.61	7.63	0.0%
Augmented Synthetic Control	0.00	0.13	92.9%	-0.00	0.19	92.2%	0.75	1.34	87.8%	0.82	1.11	66.1%
Generalized Synth (p known)	-0.03	0.15	93.5%	-0.01	0.13	95.5%	-0.00	0.30	94.9%	0.08	0.25	92.5%
Generalized Synth (p estimated)	-0.52	0.54	3.5%	-0.01	0.13	95.2%	0.29	0.42	82.5%	0.27	0.34	75.5%
Factor Imputation ($\boldsymbol{F}$ known)	-0.00	0.09	95.5%	-0.00	0.09	95.5%	0.00	0.10	95.2%	-0.00	0.08	94.7%
QLD (p known)	-0.06	0.15	92.4%	-0.01	0.13	95.5%	0.03	0.52	94.9%	0.00	0.41	95.4%
QLD (p estimated)	-0.08	0.17	86.6%	-0.01	0.13	95.5%	0.04	0.54	94.4%	0.01	0.42	95.3%
Model	TWFE			Factor			Factor			Factor		
Treatment	Random			Random			$\mathbb{1}(\gamma_i > 0)$			$\mathbb{1}(\gamma_i > 0)$		
Parallel Trends	✓			✓								
AR(1) Error Term										✓		

Notes. This table presents a set of simulations with 1000 iterations. Each row in a panel consists of a treatment effect estimator as described in the text. There are 4 different data-generating processes as described in the main text with key details listed in the second portion of the table. The columns present the bias and mean-squared error of the estimate of $\hat{\tau}$ relative to the true effect of $\tau_{T_0+1} = 1$. Additionally, confidence intervals are formed and 'Cov.' presents the percent of 95% confidence intervals containing the true treatment effect.

We investigate this phenomenon at the end of the section.

The augmented and generalized synthetic control perform much better than TWFE, but typically worse than QLD. In both tables, the bias, RMSE, and empirical coverage are within acceptable levels for DGPs 1 and 2. Again, this is unsurprising because we would expect least squares methods to perform well when parallel trends holds. They perform worse in DGPs 3 and 4, especially compared to our factor imputation estimators. These alternate factor estimators perform slightly better when $T_0 = 12$ compared to $T_0 = 4$ due to having a more precise estimate of the factor loadings. Still, the generalized synthetic control estimator performs worse than the QLD, except for the case where p is known and there is no autocorrelated errors (DGP 3, table 3).

The oracle estimator that uses the true $\boldsymbol{F}$ performs better than all other estimators in every metric for each simulation exercise. The QLD estimators also perform well, with acceptable biases and empirical coverage in each setting. When $T_0 = 4$, QLD with estimated p outperforms generalized synthetic control, even when it takes p as known. We can also see that estimating p does not impose much of a cost on the QLD first-stage estimator. This is primarily because in the vast amount of simulations, the estimator estimates the correct p. Last, our confidence intervals perform well in all cases, having approximately correct coverage even in the case of autocorrelated errors.

4.2. Using a Noisy Proxy

As mentioned earlier, practitioners recognize that treated and untreated units may be differentially exposed to the same common trends. Instead of estimating a factor model as we suggest, they will often include a linear model with time-constant covariates interacted with time-varying slopes, i.e. the 'TWFE with $\boldsymbol{w}_i'\boldsymbol{\beta}_t$' estimator in Tables 2 and 3. We now demonstrate why this estimator performed poorly for DGPs 3 and 4, and when it can be used as a suitable alternative.

The intuition for imputation estimators in the microeconometric setting is that we may not be consistent for the unobserved factor loadings, but we can estimate them on average if we knew the factors. We generated $\boldsymbol{w}_i$ as being unbiased in the factor loadings. However, just having an unbiased estimator of the loadings does not mean we can consistently estimate the slopes $\boldsymbol{\beta}_t$ unless the estimator of the loadings is very precise. To demonstrate this, we plot the bias of the TWFE estimator with proxy $\boldsymbol{w}_i'\boldsymbol{\beta}_t$ along with the bias of the feasible QLD estimator (with

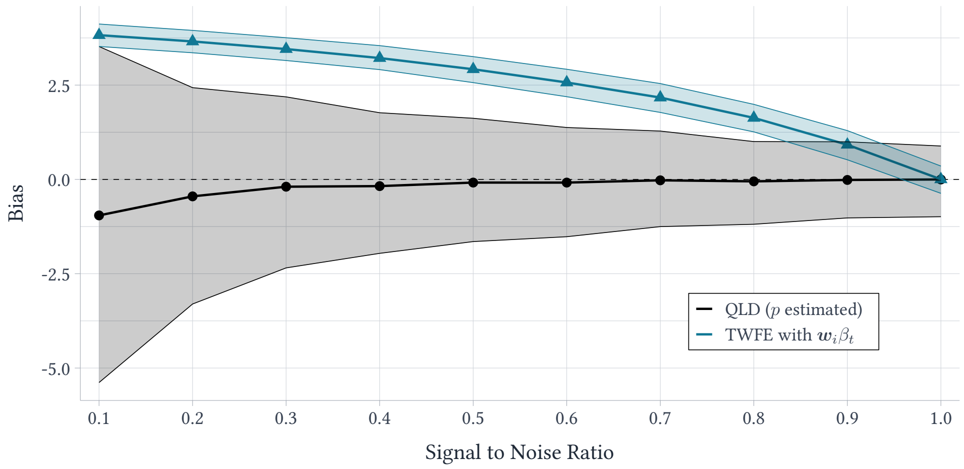


Figure 1 — Bias of TWFE Imputation with Covariates

Notes. This figure plots the bias of the 'TWFE with $\boldsymbol{w}_i'\boldsymbol{\beta}_t$' and 'QLD ($p$ estimated)' estimators when varying the signal to noise ratio of the observed covariates. The shaded regions represent values within the 2.5th and 97.5th percentile of all estimated values at the given specification. Along the x axis, we vary the signal to noise ratios of the observed covariates $(w_{i1}, w_{i2})'$ by changing the variance of the noise terms ξ_{i1} and ξ_{i2}. When the signal to noise ratio is 1, the covariates are equal to the factor loadings. A smaller noise ratio corresponds to a larger error variance. we run 1000 simulations for each value of the signal to noise ratio.

estimated p) at different levels of "signal to noise ratio", which we define as

$$\text{signal to noise ratio 1} = \frac{\text{Var}(\mu_i)}{\text{Var}(\mu_i) + \text{Var}(\xi_{i1})} \tag{28}$$

$$\text{signal to noise ratio 2} = \frac{\text{Var}(\gamma_i)}{\text{Var}(\gamma_i) + \text{Var}(\xi_{i2})} \tag{29}$$

When this ratio is one, $w_{i1} = \mu_i$ and $w_{i2} = \gamma_i$, so that the factor loadings are observed. At the smallest value we use, 0.1, the variance of ξ is 9 relative to the variance of $\mu = 1$. As we add more noise, the instrument becomes weaker and so we increase $N = 500$ for this set of simulations.

At one extreme, where the signal to noise ratio is approximately 0, i.e. ξ_{i1} and ξ_{i2} are white noise, the bias for the TWFE imputation estimator with linear model $\boldsymbol{w}_i'\boldsymbol{\beta}_t$ is the same as the TWFE imputation estimator that does not include covariates. At the other extreme, where the signal to noise ratio is approximately 1, i.e. $w_{i1} = \mu_i$ and $w_{i2} = \gamma_i$, the bias is completely removed. Except in settings with very large amounts of noise, the factor model imputation estimator remains unbiased because it only requires the covariates to be correlated with the factor loadings (Assumption 4). This experiment echos the results of Kejriwal et al. (2024). However, we note that our results are still generous to estimators that use such noisy measure because we generate the observables as unbiased estimators of the factor loadings.

5 — Application

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

We construct a dataset following the description in Basker (2005). In particular, we use the County Business Patterns dataset from 1964 and 1977-1999, subsetting to counties that (i) had more than 1500 employees overall in 1964 and (ii) had non-negative aggregate employment growth between 1964 and 1977.[18] We use a geocoded data set of Walmart openings from Arcidiacono et al. (2020) to construct our treatment variable. Our treatment dummy is equal to one if the county has any Walmart in that year and our group variable denotes the year of entrance for the *first* Walmart in the county. [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 vertically integrated business, so we expect Walmart to compete in both the retail and wholesale sectors (Basker, 2005).

[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 first implement the TWFE imputation estimator proposed by Borusyak et al. (2024) and estimate event-study effects on (log) retail and wholesale employment. In particular, we use the following model

$$\log(y_{it}) = \mu_i + \lambda_t + \sum_{\ell=-22}^{13} \tau^\ell d_{it}^\ell + u_{it} \tag{30}$$

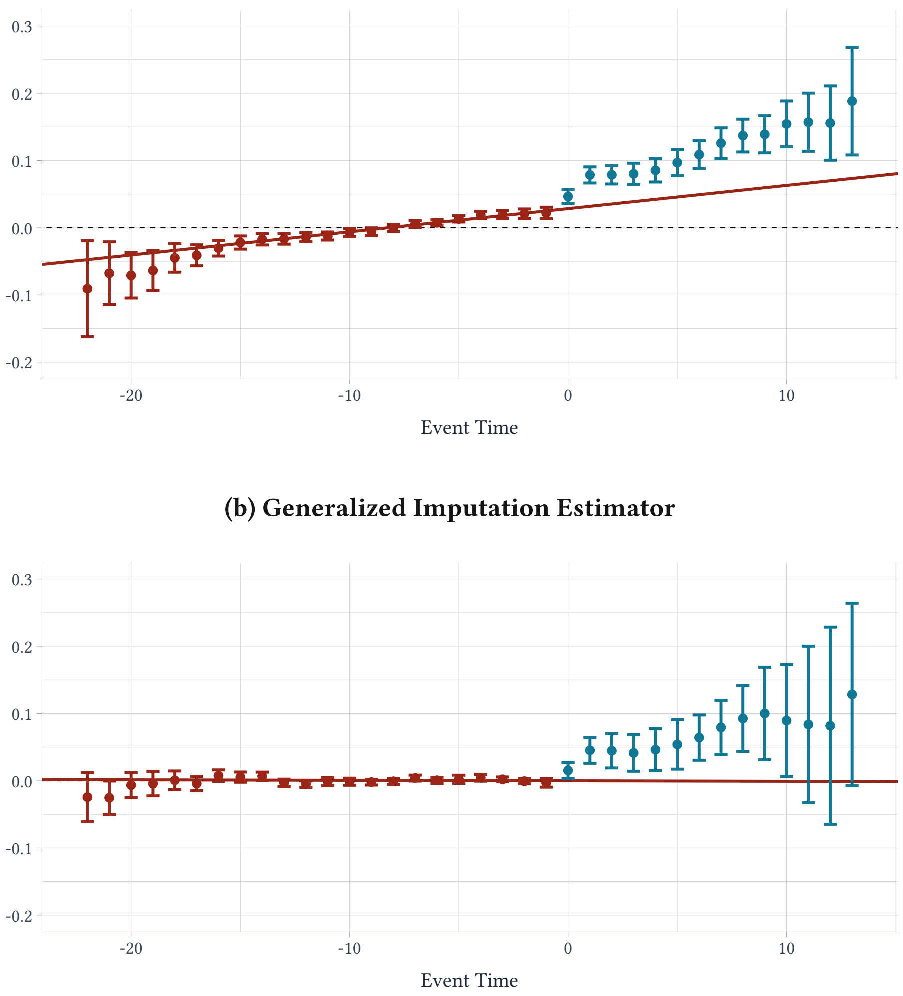
where i denotes county, t denotes year, y_{it} is either retail or wholesale employment, and $d_{it}^\ell = 1(t - g_i = \ell)$ are indicator variables denoting event-time. Results of the event-study estimates are presented in panel (a) of Figure 2 and Figure 3.

For both retail and wholesale employment, counties receiving Walmarts had faster employment growth relative to the control counties, emphasizing our concern over endogenous opening decisions. In the spirit of Freyaldenhoven et al. (Forthcoming) and Rambachan and Roth (2023), we draw the line of best fit for the 15 most-recent pre-treatment estimates ($\hat{\tau}^\ell$ for $-15 \leq \ell < 0$) and extend it into the post-treatment estimates. For both retail and wholesale employment, the 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 to estimate the factors as described in Section 2.2. to estimate the factor parameters $\boldsymbol{\theta}$, we need a set of instruments that satisfy the two standard instrument requirements as described by Assumption 4: 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. We specifically use the 1980 baseline values of the following variables as

Figure 2 — Effect of Walmart on County log Retail Employment



(a) TWFE Imputation Estimator

(b) Generalized Imputation Estimator

Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log retail employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Borusyak et al. (2024). Panel (b) estimates effects using the QLD 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.

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] 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 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 treatment effect point estimates are smaller than those estimated by TWFE with an estimated effect on employment of around 6% on average. Evaluated at the median baseline retail employment of 1417 employees, this result 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.

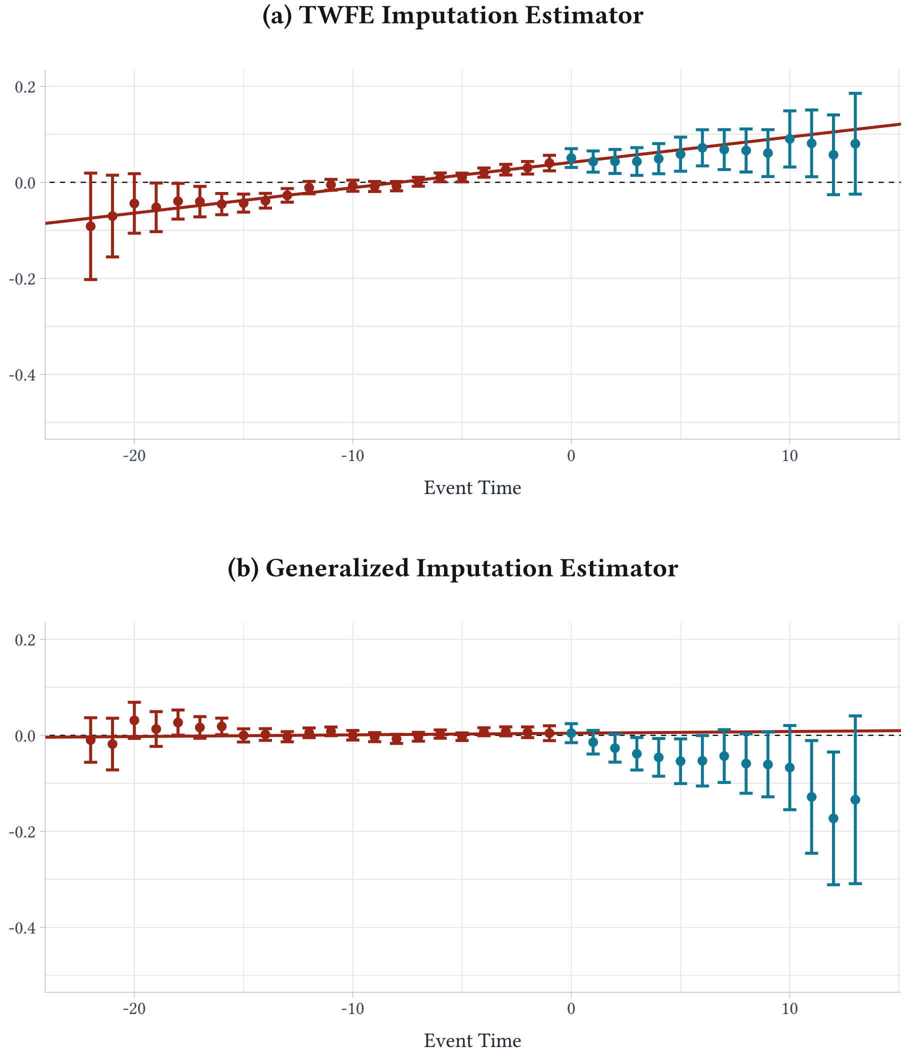
We see a similar story with wholesale employment, where our estimator removes most of the pre-trend violations. In this case, however, the estimated effects flip signs with an estimated effect of around -6%, although they are not statistically significant at the 5% level. Evaluated at the 1977 median wholesale employment of 410, this suggests a decrease of about 25 jobs, similar to the findings in Basker (2005).

Our estimator allows for any 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 QLD 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 assumed linear in the same common

[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 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. (2024). Panel (b) estimates effects using the generalized imputation estimator we propose in Section 3 with $p = 1$ and using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree. The red lines correspond to a linear estimate of pre-treatment point estimates for event time -15 to -1 and is extended into the post-treatment periods.

factors as the outcome variable:

$$X_{it} = \boldsymbol{\alpha}_i' \boldsymbol{F}_t + \nu_{it}. \tag{31}$$

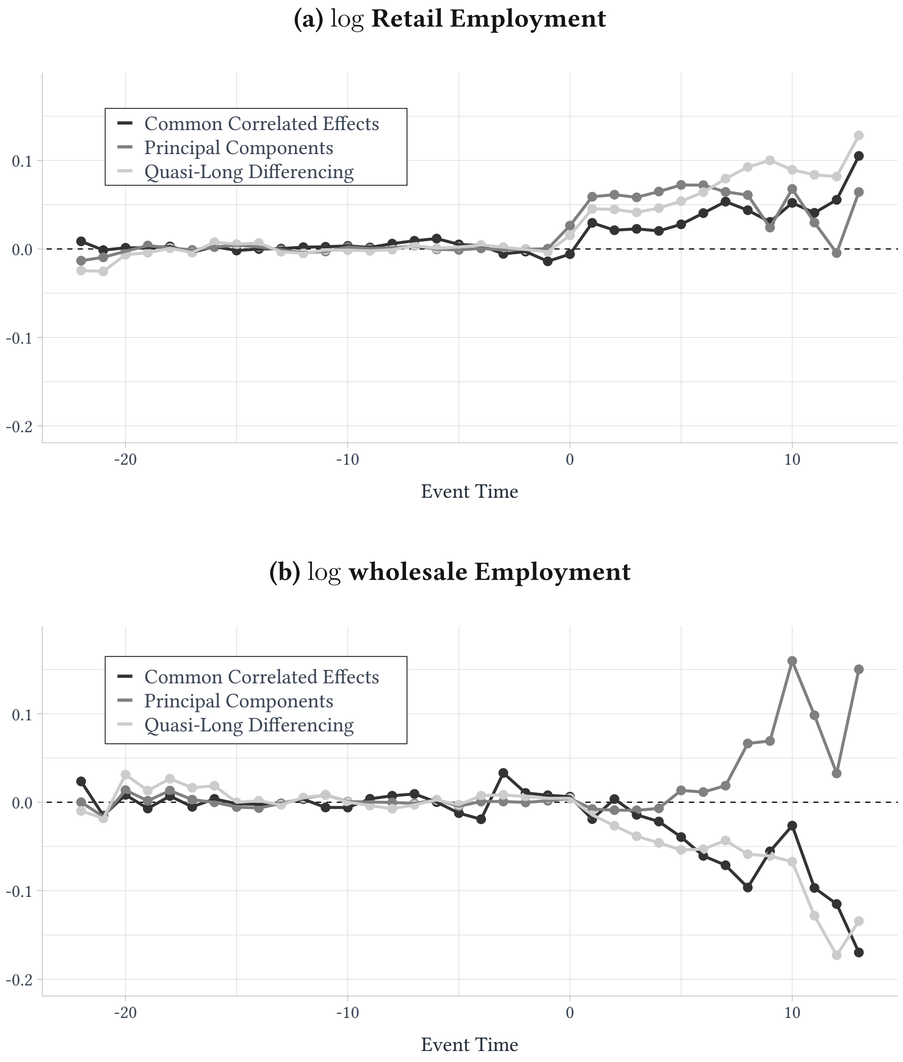
where $\boldsymbol{\alpha}_i$ is a $p \times 1$ vector of covariate-specific factor loadings and ν_{it} is a mean zero idiosyncratic shock. Under this assumption, the cross-sectional averages of $\boldsymbol{X}$ (averaged over the never-treated group) consistently span the column space of $\boldsymbol{F}$. 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. Lastly, we use the principal components estimator of Bai (2009) to impute the factors, as proposed by Xu (2017). This estimator does not require instruments/proxies or additional covariates. However, this advantage comes at the cost of requiring a large number of time periods, which may be infeasible to assume in our application.

The results of each estimator are presented in Figure 4. All three are effective at removing underlying trends that the treated counties experienced before treatment. 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 QLD 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. These results suggest we may not have a large enough time series to meaningfully estimate the factors via principal components.

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

As discussed in Section 4, a common approach in empirical work is to include a set of time-

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

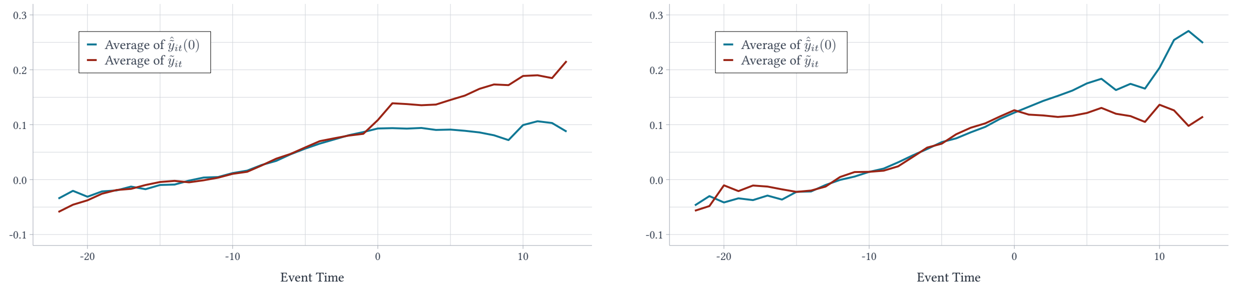


(a) log **Retail Employment**

(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 QLD estimator proposed in Ahn et al. (2013). Details of the estimation procedures appear in the text.

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



(a) log **Retail Employment**

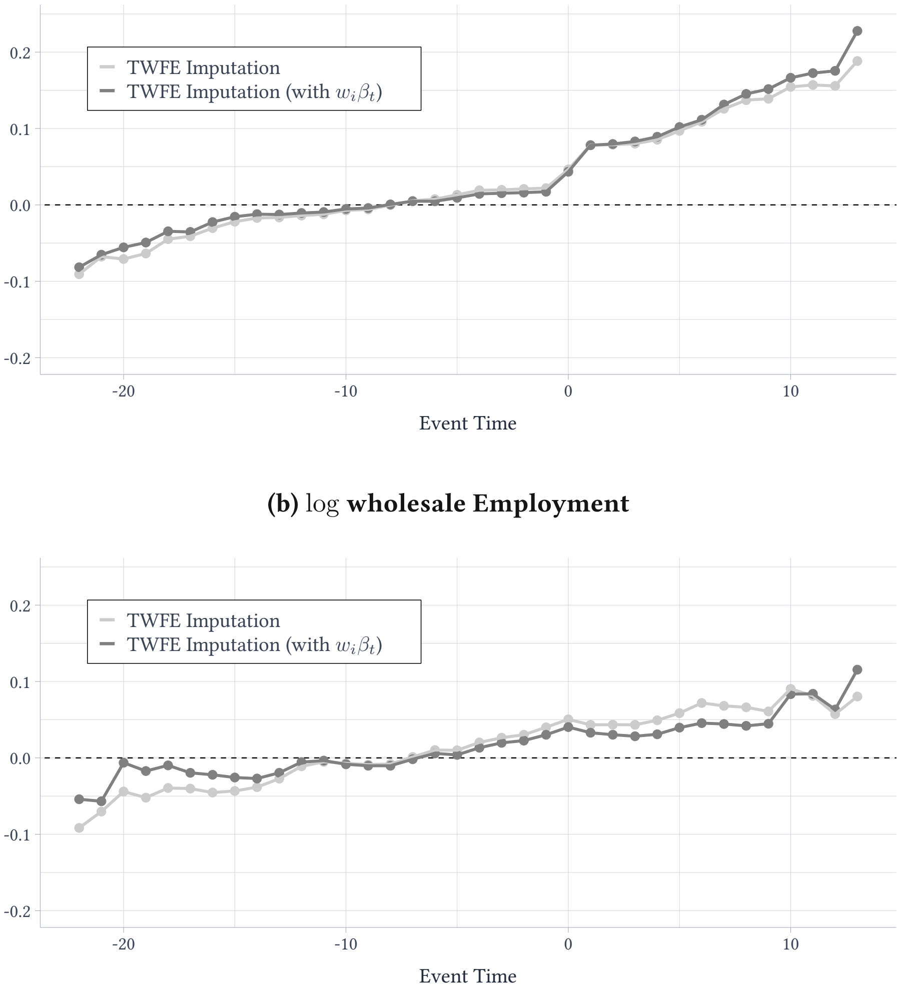
(b) log **wholesale Employment**

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

invariant covariates interacted with time-period-specific coefficients, $\boldsymbol{w}_i'\boldsymbol{\beta}_t$, to capture some forms of non-parallel trends (Abadie, 2005; Sant'Anna and Zhao, 2020). Mirroring our simulations, we rerun our TWFE model using our QLD instruments as the controls $\boldsymbol{w}_i$. Figure 6 presents the results. Including these variables in our TWFE model fails to absorb the non-parallel trends we think are present in the estimates. This result may imply that $\boldsymbol{w}_i$ is a correlated but noisy measures of the underlying factor loadings and causes attenuation bias in estimates of $\boldsymbol{\beta}_t$, which ultimately fails to absorb the non-parallel trends.

To highlight the importance of the uncertainty from estimation of the factors in the first stage, we recreate confidence intervals from our generalized imputation estimator with the QLD first stage while treating the factor estimates as the true unobserved factors. Results are given in Figure 7. The standard errors on point estimates are far smaller, with estimates becoming strongly significant in wholesale employment. We believe the robust standard errors are relatively large because there are few never-treated counties relative to the entire sample of treated counties, and hence estimation of $\boldsymbol{F}$ is imprecise. This result shows an important step for future research in finding more efficient factor estimators.

Figure 6 — Time-interacted covariates in TWFE model

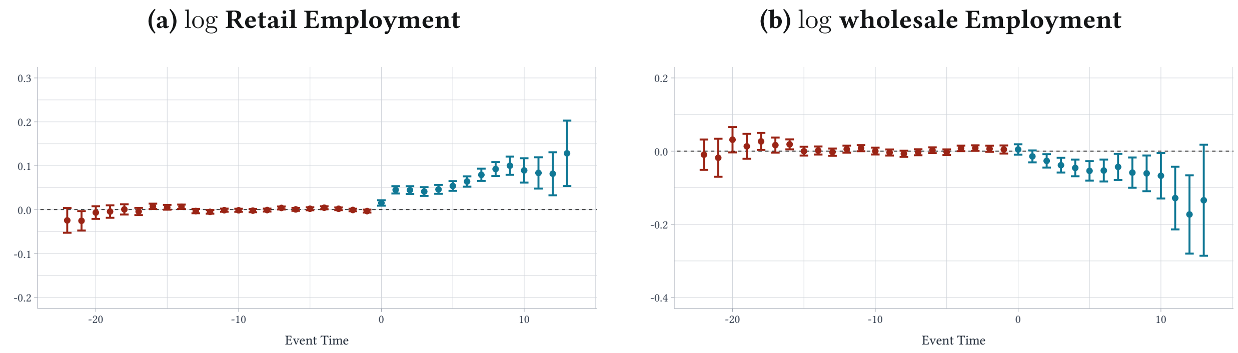


(a) log **Retail Employment**

(b) log **wholesale Employment**

Notes. This figure reproduces estimates from figures 2 and 3 and additionally plots estimates modifying the TWFE model to include a set of time-invariant covariates interacted with time-specific coefficients, $\boldsymbol{w}_i'\boldsymbol{\beta}_t$. The $\boldsymbol{\beta}_t$ parameters are estimated using the untreated sample (never-treated units and pre-treatment treated units).

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


(a) log **Retail Employment**

(b) log **wholesale Employment**

Notes. This figure recreates estimates from panel (b) of Figure 2 and Figure 3 with bootstrapped confidence intervals holding the first-stage estimate of θ fixed in repeated samples.

6 — Conclusions

We consider identification and inference of heterogeneous treatment effects in a linear panel data model. We 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 QLD, internally generated instruments, common correlated effects, or principal components, allowing for both large and small numbers of pre-treatment time periods. While we specifically consider the QLD estimator of Ahn et al. (2013), further work should demonstrate both theoretical and finite-sample properties of these various estimators of the factors and how they affect ATT estimation, especially for larger time series. The GMM imputation framework should also be examined in the context of unbalanced panels.

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 and demonstrate findings consistent with the IV estimation strategy of Basker (2005). Our results suggest that the factor imputation estimator 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. We leave this possibility for future study.

Acknowledgments

We would like to thank Stephane Bonhomme, Brantly Callaway, Brian Cadena, Peter Hull, Jeffrey Wooldridge, and Taylor Jaworski as well as seminar participants from the University of Georgia, the University of Arkansas, Louisiana State University, the University of Notre Dame, Carleton University, the University of Kentucky, Clemson University, 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. We also thank the Associate Editor for their careful review of our paper, as well as the thoughtful reports from two anonymous referees. All errors are our own.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Feng, Yingjie. 2021. “Causal Inference in Possibly Nonlinear Factor Models.” `https://arxiv.org/abs/2008.13651`.

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

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

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

Gardner, John. 2021. “Two-Stage Difference-in-Differences.”

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

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

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

Hatch, Julie, and Angela Clinton. 2000. “Job Growth in the 1990s: a Retrospect.” *Monthly Lab. Rev.* 123 3.

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

Appendix to

"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

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

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 1

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

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

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

The definition of τ_{it} is the difference between treated and untreated potential outcomes for unit i at time t, so for any (i,t), $y_{it} = d_{it}y_{it}(1) + (1-d_{it})y_{it}(\infty) = d_{it}\tau_{it} + y_{it}(\infty)$. Then

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

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

□

Proof of Theorem 2

We can appeal to standard large sample GMM theory as in Hansen (1982) due to the types of first-stage factor estimators we consider. We do not consider true “fixed effects" estimators where the number of parameters grows with the sample size. The IV and cross-sectional averages approaches are based on eliminating the factors (which are fixed in the asymptotic analysis) by reducing them to a smaller set of parameters. For example, while the CCE estimator can be implemented as a pooled regression where unit dummies are interacted with cross-sectional averages, the estimator itself takes a form similar to the within transformation in the linear fixed effects model. In fact, we prove asymptotic unbiasedness of dynamic ATT estimators using the CCE estimator in the first stage (Brown et al., 2023)[23].

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

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

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

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

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

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

We now derive the asymptotic variance of the full estimator under a general first-step estimator of the factors. Note that $\boldsymbol{h}_{i\infty}(\boldsymbol{\theta}) \otimes \boldsymbol{h}_{ig}(\boldsymbol{\theta}, \boldsymbol{\tau}_g) = \mathbf{0}$ (from the D_{ig} terms) and $\boldsymbol{h}_{ih}(\boldsymbol{\theta}, \boldsymbol{\tau}_h) \otimes \boldsymbol{h}_{ik}(\boldsymbol{\theta}, \boldsymbol{\tau}_k) = \mathbf{0}$ 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.

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

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:

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

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 of the GMM estimator is given by $(\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D})^{-1}$. $\boldsymbol{\Delta}$ is a block diagonal matrix so its inverse is trivial to compute. First, we have

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

The transpose of the gradient matrix is

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

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 $\boldsymbol{D} = \operatorname{diag}\{\boldsymbol{\Delta}_g^{-1}\}_{g\in\mathcal{G}}$. 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

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

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

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

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

where $\boldsymbol{K}_{(T-p)T}$ is the $(T-p)T \times (T-p)T$ commutation matrix and we use the well-known relationship between vectorization and the Kronecker product[24]. Because $\text{vec}(\boldsymbol{H}(\boldsymbol{\theta})) = [\text{vec}(\boldsymbol{I}_{T-p})', \boldsymbol{\theta}']'$, the gradient of the moment function is

$$\left(\boldsymbol{I}_{(T-p)} \otimes \boldsymbol{w}_i\boldsymbol{y}_i'\right) \boldsymbol{K}_{T(T-p)}[\mathbf{0}'_{(T-p)^2\times(T-p)p}, \boldsymbol{I}_{(T-p)p}]' \tag{A4}$$

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

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

$$\boldsymbol{F}_{t\geq g}(\boldsymbol{\theta})' \left(\boldsymbol{F}_{t<g}(\boldsymbol{\theta})'\boldsymbol{F}_{t<g}(\boldsymbol{\theta})\right)^{-1} \boldsymbol{F}_{t<g}(\boldsymbol{\theta})'\boldsymbol{y}_{i,t<g} \tag{A5}$$

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

$$\boldsymbol{F}_{t<g} = \begin{pmatrix} \boldsymbol{\Theta} \\ \boldsymbol{E} \end{pmatrix} \tag{A6}$$

where $\boldsymbol{E}$ is the first $(g-1)-(T-p)$ rows of $-\boldsymbol{I}_p$. Then the post-treatment factor matrix is

[24] See Exercise 10.18 of Abadir and Magnus (2005).

just the lower $T-g+1$ rows of $-\boldsymbol{I}_p$ so we do not need to worry about differentiating it. In this setting,

$$\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1}\boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g}=-\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A7}$$

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

$$-\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}(d\boldsymbol{\Theta})' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A8}$$

$$\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\left((d\boldsymbol{\Theta})'\boldsymbol{\Theta}\right)\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A9}$$

$$\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\left(\boldsymbol{\Theta}'(d\boldsymbol{\Theta})\right)\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g} \tag{A10}$$

which can then be rewritten as

$$-\left(\boldsymbol{y}_{i,t<g}\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\begin{pmatrix}\boldsymbol{K}_{(T-p)p}(d\boldsymbol{\theta})' & \boldsymbol{K}_{((g-1)-(T-p)p}\text{vec}(\boldsymbol{E})'\end{pmatrix}' \tag{A11}$$

$$\left(\left(\boldsymbol{\Theta}\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\boldsymbol{K}_{(T-p)p}d\boldsymbol{\theta} \tag{A12}$$

$$\left(\left(\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\boldsymbol{\Theta}'\right)d\boldsymbol{\theta} \tag{A13}$$

The full gradient is then

$$-\left(\boldsymbol{y}_{i,t<g}\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\begin{pmatrix}\boldsymbol{K}_{(T-p)p}' & \boldsymbol{0}_{((g-1)-(T-p)p\times(T-p)p}'\end{pmatrix}' \tag{A14}$$

$$\left(\left(\boldsymbol{\Theta}\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\right)\boldsymbol{K}_{(T-p)p} \tag{A15}$$

$$\left(\left(\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\begin{pmatrix}\boldsymbol{\Theta}' & \boldsymbol{E}'\end{pmatrix}\boldsymbol{y}_{i,t<g}\right)'\otimes\left(\boldsymbol{\Theta}'\boldsymbol{\Theta}+\boldsymbol{E}'\boldsymbol{E}\right)^{-1}\boldsymbol{\Theta}'\right) \tag{A16}$$

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

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

$$\boldsymbol{F}_{t\geq g}=\begin{pmatrix}\boldsymbol{\Theta}_2\\ -\boldsymbol{I}_p\end{pmatrix} \tag{A17}$$

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

$$\boldsymbol{\Theta}_1 = \boldsymbol{E}_1\boldsymbol{\Theta} \tag{A18}$$

$$\boldsymbol{\Theta}_2 = \boldsymbol{E}_2\boldsymbol{\Theta} \tag{A19}$$

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

$$\begin{pmatrix} \boldsymbol{E}_2\boldsymbol{\Theta} \\ \\ -\boldsymbol{I}_p \end{pmatrix} \left(\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{E}_1\boldsymbol{\Theta}\right)^{-1} \boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{y}_{i,t<g} \tag{A20}$$

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

$$\begin{pmatrix} \boldsymbol{E}_2 d\boldsymbol{\Theta} \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} \left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A21}$$

$$-\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} d\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{F}_{t<g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A22}$$

$$-\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{E}_1 d\boldsymbol{\Theta}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A23}$$

$$+\boldsymbol{F}_{t\geq g}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} d\boldsymbol{\Theta}'\boldsymbol{E}_1'\boldsymbol{y}_{i,t<g} \tag{A24}$$

where we inserted $\boldsymbol{F}_{t<g}$ and $\boldsymbol{F}_{t\geq g}$ for $\boldsymbol{E}_1\boldsymbol{\Theta}$ and $\boldsymbol{E}_2\boldsymbol{\Theta}$ respectively to preserve space, noting that these matrices are actually functions of the parameters $\boldsymbol{\theta}$ and not the true, unobserved factors. We rewrite line (A21) so we can write the differential in terms of $\boldsymbol{\theta}$:

$$\begin{pmatrix} \boldsymbol{E}_2 d\boldsymbol{\Theta} \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} \left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} = \begin{pmatrix} \boldsymbol{E}_2 \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} d\boldsymbol{\Theta}\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g} \tag{A25}$$

$$= \left(\left(\left(\boldsymbol{F}_{t<g}'\boldsymbol{F}_{t<g}\right)^{-1} \boldsymbol{F}_{t<g}'\boldsymbol{y}_{i,t<g}\right) \otimes \begin{pmatrix} \boldsymbol{E}_2 \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix}\right) d\boldsymbol{\theta} \tag{A26}$$

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

$$= \left(\left(\boldsymbol{F}'_{t<g} \boldsymbol{F}_{t<g} \right)^{-1} \boldsymbol{F}'_{t<g} \boldsymbol{y}_{i,t<g} \right) \otimes \begin{pmatrix} \boldsymbol{E}_2 \\ \\ \boldsymbol{0}_{p\times p} \end{pmatrix} \tag{A27}$$

$$- \left(\boldsymbol{E}'_1 \boldsymbol{F}_{t<g} \left(\boldsymbol{F}'_{t<g} \boldsymbol{F}_{t<g} \right)^{-1} \boldsymbol{F}'_{t<g} \boldsymbol{y}_{i,t<g} \right)' \otimes \left(\boldsymbol{F}_{t\geq g} \left(\boldsymbol{F}'_{t<g} \boldsymbol{F}_{t<g} \right)^{-1} \right) \boldsymbol{K}_{(T-p)p} \tag{A28}$$

$$- \left(\left(\boldsymbol{F}'_{t<g} \boldsymbol{F}_{t<g} \right)^{-1} \boldsymbol{F}'_{t<g} \boldsymbol{y}_{i,t<g} \right)' \otimes \left(\boldsymbol{F}_{t\geq g} \left(\boldsymbol{F}'_{t<g} \boldsymbol{F}_{t<g} \right)^{-1} \boldsymbol{F}'_{t<g} \boldsymbol{E}_1 \right) \tag{A29}$$

$$+ \left(\boldsymbol{y}'_{i,t<g} \boldsymbol{E}_1 \right) \otimes \left(\boldsymbol{F}_{t\geq g} \left(\boldsymbol{F}'_{t<g} \boldsymbol{F}_{t<g} \right)^{-1} \right) \boldsymbol{K}_{(T-p)p} \tag{A30}$$

□

Proof of Theorem 3

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

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

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

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

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 term that is $O_p(N^{-1/2})$. Assumption 1 and the weak law of large numbers imply

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

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:

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

This form comes from the fact that the weight matrix is positive definite with probability approaching one[25]. 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.

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

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

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

Theorem 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

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

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:

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

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

The first term has asymptotic variance

$$V_1 = \mathbb{E}\left[\Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big)\Big(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0}\Big)' \mid D_{ig} = 1\right] \tag{C9}$$

and the second term has asymptotic variance

$$V_0 = \frac{\mathbb{P}(D_{ig} = 1)}{\mathbb{P}(D_{i\infty} = 1)}\,\mathbb{E}\left[\Big(\overline{u}_{i,t<T_0} - u_{i,t}\Big)\Big(\overline{u}_{i,t<T_0} - u_{i,t}\Big)' \mid D_{i\infty} = 1\right] \tag{C10}$$

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

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

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

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

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

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

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; if the means are equal, then TWFE will be sufficient given the formulation in Lemma 1. 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

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

■