Brown, Nicholas; Butts, Kyle

Working Paper

A unified framework for dynamic treatment effect estimation in interactive fixed effect models

Queen's Economics Department Working Paper, No. 1495

Provided in Cooperation with:
Queen's University, Department of Economics (QED)

Suggested Citation: Brown, Nicholas; Butts, Kyle (2022) : A unified framework for dynamic treatment effect estimation in interactive fixed effect models, Queen's Economics Department Working Paper, No. 1495, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/281099

Queen's Economics Department Working Paper No. 1495

A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models

Nicholas Brown
Queen's University

Kyle Butts
University of Colorado Boulder, Economics Department

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

11-2022

A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models*

Nicholas Brown† and Kyle Butts‡

November 30, 2022

We present a unifying identification strategy of dynamic average treatment effect parameters for staggered interventions when parallel trends are valid only after controlling for interactive fixed effects. This setting nests the usual parallel trends assumption, but allows treated units to have heterogeneous exposure to unobservable macroeconomic trends. We show that any estimator that is consistent for the unobservable trends up to a non-singular rotation can be used to consistently estimate heterogeneous dynamic treatment effects. This result can apply to data sets with either many or few pre-treatment time periods. We also demonstrate the robustness of two-way fixed effects imputation to certain parallel trends violations and provide a test for its consistency. A quasi-long-differencing estimator is proposed and implemented to estimate the effect of Walmart openings on local economic conditions.

JEL Classification Number: C13, C21, C23, C26

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

* We would like to thank Brantly Callaway, Peter Hull, Brian Cadena, and Jeffrey Wooldridge, as well as seminar participants from Queen's University, the 2022 Midwest Econometrics Group, and the CU Boulder Econometrics Brownbag for their insightful questions and comments.

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

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

1 — Introduction

Estimation of the effects of a treatment in panel settings often relies on a two-way fixed effect (TWFE) structure. The untreated potential outcomes for a unit i at time t are determined by a unit 'fixed effect' that captures the individual's heterogeneous characteristics, a set of time 'fixed effects' that capture macroeconomic trends, and a mean-zero error term u_{it}. This model is written as

$$y_{it}(\infty) = \mu_i + \lambda_t + u_{it}. \tag{1}$$

Individual treatment effects are defined as the contrast between the observed post-treatment outcomes, y_{it}, and untreated potential outcomes, $y_{it}(\infty)$. We are interested in averages of heterogeneous individual treatment effects that can vary arbitrarily over time. To estimate average treatment effects, researchers often invoke a 'parallel-trends' type restriction that the unobservable confounder, u_{it}, is unrelated to selection into treatment. When this assumption fails, the treated units no longer follow the same outcome trajectory as untreated units, resulting in treatment effects being confounded by contemporaneous shocks. For example, we consider the effects of Walmart opening a store on local employment in our empirical application. If Walmart chooses which counties to open stores in based on local economic trends, as found in Neumark et al. (2008), this error term assumption is implausible.

We consider a more general 'parallel trends' type assumption that allows units to enter treatment based on an interactive fixed effects structure:

$$y_{it}(\infty) = \mu_i + \lambda_t + \boldsymbol{f}_t'\boldsymbol{\gamma}_i + \varepsilon_{it}, \tag{2}$$

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 $\mathbb{E}[\varepsilon_{it}] = 0$ for all (i, t).[1] One possible motivation for this model

1. Note that this model for outcomes coincides with the standard TWFE model when $p = 0$ and with a

is that the factors $\boldsymbol{f}_t$ are macroeconomic shocks with factor loadings $\boldsymbol{\gamma}_i$ denoting a unit's exposure to the shocks. Another possibility lets the $\boldsymbol{\gamma}_i$ represent time-invariant characteristics with a marginal effect on the outcome $\boldsymbol{f}_t$ that changes over time.[2]

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

Recent work has proposed 'imputation' based estimators for treatment effects that use non-treated and pre-treatment observations to 'impute' the untreated potential outcomes for the post-treatment observations (e.g. Borusyak et al. 2022, Gardner 2021, Wooldridge 2021). However, these approaches only allow for level fixed effects and preclude interactions like in equation (2). We generalize these techniques by proposing an estimator that imputes the untreated potential outcomes under the more general (2).

To do so, we first remove the additive fixed effects with a double-demeaning transformation so that we explicitly nest the two-way error model. Our main treatment effect identification result then only requires consistent estimates of $\boldsymbol{f}_t$.[4] Using the fac-

TWFE model with unit-specific linear time trends coincides when $p = 1$ and $\boldsymbol{f}_t = t$.

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.

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

4. It is generally only possible to estimate a normalized version of $\boldsymbol{f}_t$. This is fine as our imputation procedure works with any normalization of the factors.

tors, we compute a matrix that projects the pre-treatment outcomes onto the estimated post-treatment factors, imputing the untreated potential outcome. Averaging over the difference between the post-treatment observed outcome and the imputed untreated potential outcomes gives a consistent estimator of average treatment effects.

There are major two benefits of our general identification argument. First, consistent estimation of $\boldsymbol{f}_t$ is possible through a variety of approaches, such as quasi-differencing (Ahn et al. 2013, Callaway and Karami 2022), common correlated effects (Pesaran 2006, Westerlund et al. 2019), or principal components (Bai 2009, Fan et al. 2016, Chan and Kwok 2022). These techniques allow the user to tailor their factor estimator to the specific data and problem under consideration, including how many pre-treatment time periods are available. Our identification result provides a recipe for using any consistent estimator of the factors to estimate treatment effects, opening up the large factor-model literature for causal inference methods. Second, our imputation method allows researchers to graph the estimated untreated potential outcomes and the observed outcomes for treated units, similar to a synthetic control plot. These plots provide a visual check for the parallel-trends type assumption that our estimator requires, making empirical analysis more transparent.

We derive asymptotic properties of an imputation estimator using the quasi-long-differencing (QLD) method of Ahn et al. (2013). The resulting estimator takes the form of a generalized method of moments (GMM) estimator, which allows estimation and inference to be handled easily with common statistical software. It is also consistent when the number of pre-treatment time periods is small.[5] One advantage of this estimator is that we can form statistical tests for the consistency of the TWFE estimator. These tests are practically useful since difference-in-differences is simple to implement and interpret.

Our work contributes to an emerging literature on adjusting for parallel-trends violations in short panels. Freyaldenhoven et al. (2019) propose a similar instrumental

5. Deriving the asymptotic distribution of treatment effects using other factor estimators is left for future work.

variable type estimator in the presence of time-varying confounds. 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. Callaway and Karami (2022) also allow for heterogeneous effects in short panels. They prove identification using a similar strategy to QLD and instrumental variables. They require time-invariant instruments whose effects on the outcome are constant over time. Their instruments would be valid in our QLD estimator, but we also allow for time-varying covariates as instruments. They also do not provide a general identification scheme like ours and so their results do not readily extend to other estimators.

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 when the outcomes are generated by a linear factor structure. Section 3 provides the main asymptotic theory for a particular QLD estimator. We also discuss practical concerns for practitioners. Section 4 gives several specification tests for the underlying model. We include a small Monte Carlo experiment in Section 5 to examine the finite-sample performance of our estimator. Finally, Section 6 contains our application and Section 7 leaves with some concluding remarks.

2 — Model and Identification

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

last period before the earliest treatment adoption.

We also introduce some matrix notation. For a vector $\boldsymbol{x}$ of length T, we use the subscript $\boldsymbol{x}_{t<g}$ to denote the first $g-1$ elements and $\boldsymbol{x}_{t\geq g}$ to refer to the last $T-g+1$ elements. This holds similarly for the rows of a matrix $\boldsymbol{X}$. We now state our main identifying assumptions.

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

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

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

for $t = 1, .., T$. We allow for heterogeneous and dynamic treatment effects of any form, i.e. $y_{it}(g) = \tau_{igt} + y_{it}(\infty)$. ■

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

Assumption 4 (Selection into treatment). $\mathbb{E}[u_{it} \mid \boldsymbol{\gamma}_i, \mu_i, G_i] = 0$ for $t = 1, ..., T$. ■

Assumption 4 is more general than the standard difference-in-differences parallel trend assumption since we include the factor structure in our potential outcome model. In particular, it assumes that the error term is uncorrelated with treatment status *after* controlling for the factor loadings. Treatment can still be correlated with contemporaneous shocks so long as the shocks are 'common' across the sample. For example, our identification strategy is valid if workers select into a job training program based on their exposure to productivity shocks. We also allow arbitrary serial correlation among the idiosyncratic errors.[6] We also assume the time effects λ_t and $\boldsymbol{f}_t$ are constant. The alternative is to assume they are random and independent of the variables in Assumption 1, then derive the limiting theory as in Westerlund et al. (2019).

6. This condition may need to be strengthened if we have many time periods.

The two-way error model cannot accommodate differential exposure. Consider the standard TWFE parallel trends assumption that $\mathbb{E}[u_{it} - u_{it-1} \mid G_i] = 0$.[7] In the more general factor model, this assumption would imply our 'error term' for group g in period t would have expectation:

$$\begin{aligned}\mathbb{E}[y_{it} - y_{it-1} \mid G_i = g] &= \lambda_t + (\boldsymbol{f}_t - \boldsymbol{f}_{t-1})'\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] + \mathbb{E}[u_{it} - u_{it-1} \mid G_i = g] \\ &= \lambda_t + (\boldsymbol{f}_t - \boldsymbol{f}_{t-1})'\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]\end{aligned}$$

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 these two cases hold for all g and t, the TWFE model is correctly specified; we prove this result in the next section. In contrast, our Assumption 4 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.

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

$$\text{ATT}(g,t) = \tau_{gt} \equiv \mathbb{E}[y_{it}(g) - y_{it}(\infty) \mid G_i = g] \tag{3}$$

These quantities represent the average effect of treatment for units that start treatment in period g when they are in period t. It is trivial to estimate other averages as well in our framework, including averaging over all post-treatment observations to estimate an overall ATT, and averaging over (i,t) where $t - G_i = \ell$ to estimate event-study ATT$^{\ell}$'s. We discuss these and other extensions from Callaway and Sant'Anna (2021) in Section 3.

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 (2) to

7. The following derivation is also shown in Callaway and Karami (2022), but we are repeating it here for expositional purposes.

consistently estimate group-time average treatment effects. Gardner (2021), Wooldridge (2021), and Borusyak et al. (2022) implicitly rely on this insight in studying the two-way error model.

Prior attempts at estimating average treatment effects focus on finding conditions that allow for estimation of $\boldsymbol{\gamma}_i$ and $\boldsymbol{f}_t$ jointly, such as in Gobillon and Magnac (2016) and Xu (2017). These techniques require the number of pre-treatment periods to grow to infinity and often place restrictions on both the dynamics of the treatment effects' distribution and the serial dependence among the idiosyncratic errors. Instead, we pursue identification noting that

$$\mathbb{E}[y_{it}(\infty) \mid G_i = g] = \mathbb{E}[\mu_i \mid G_i = g] + \lambda_t + \boldsymbol{f}_t' \mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] \tag{4}$$

Therefore, we only need to estimate the *average* of the unit effects and factor loadings among a treatment group. Our methodology will always allow for fixed post-treatment time periods but can 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

Identification of ATTs will proceed in three steps. The first step is to remove the additive fixed effects with a double-demeaning transformation. The second step is to impute $\tilde{y}_{it}(\infty)$ in the post-treatment periods for each group, where $\tilde{y}_{it}$ denotes the outcome after the fixed effects are removed. The final step is averaging the contrast between $\tilde{y}_{it}$ and $\hat{\tilde{y}}_{it}(\infty)$.

We first define the following averages:

$$\overline{y}_{\infty,t} = \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i\infty} y_{it} \tag{5}$$

$$\overline{y}_{i,t\leq T_0} = \frac{1}{T_0} \sum_{t=1}^{T_0} y_{it} \tag{6}$$

$$\overline{y}_{\infty,t<T_0} = \frac{1}{N_\infty T_0} \sum_{i=1}^{N} \sum_{t=1}^{T_0} D_{i\infty} y_{it} \tag{7}$$

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. These quantities leverage a subsample of observations with $d_{it} = 0$ and are not contaminated by the treatment.

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. 2022). Second, it preserves a common factor structure for all units and time periods. The TWFE imputation estimator of Gardner (2021), Wooldridge (2021), and Borusyak et al. (2022) would not share this property because they estimate μ_i and λ_t based on the full sample $d_{it} = 0$.

This result is summarized in the following lemma:

Lemma 2.1. $\mathbb{E}[\tilde{y}_{it} \mid G_i = g] = \mathbb{E}\left[d_{it}\tau_{it} + (\boldsymbol{f}_t - \overline{\boldsymbol{f}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_\infty) \mid G_i = g\right]$ 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. ■

All proofs are contained in the Appendix. Lemma 2.1 demonstrates how to explicitly nest the TWFE model while allowing for a general common factor structure. Any imputation method that wants to include factor models while explicitly nesting the two-

way error structure should provide a similar result to Lemma 2.1. Otherwise, they may not be able to use pre-treatment observations to impute outcomes in the post-treatment periods. 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

$$\tilde{y}_{it} = d_{it}\tau_{it} + (\boldsymbol{f}_t - \overline{\boldsymbol{f}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_{\infty}) + \tilde{u}_{it}. \tag{8}$$

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

$$\tilde{y}_{it} = d_{it}\tau_{it} + \tilde{\boldsymbol{f}}_t'\tilde{\boldsymbol{\gamma}}_i + \tilde{u}_{it}. \tag{9}$$

Lemma 2.1 has the added benefit of showing us when the ATTs are identified by our TWFE transformation alone.

Corollary 2.1. Under Assumptions 1-4, ATT(g,t) is identified by the fixed effects imputation transformation if $\mathbb{E}[\gamma_i \mid G_i = g] = \mathbb{E}[\gamma_i]$ for all $g \in \mathcal{G} \cup \{\infty\}$. ■

This result is an immediate consequence of Assumptions 1 – 4 as $\mathbb{E}\left[\boldsymbol{\gamma}_j \mid G_i = g\right] - \mathbb{E}[\boldsymbol{\gamma}_i]$ for $j \neq i$ under random sampling. Corollary 2.1 tells us that TWFE imputation is sufficient to estimate the ATTs, even when the factor structure exists, so long as the average factor loadings do not differ systemically with treatment status. Asymptotic normality of our imputation procedure under a two-way error model is studied in the Online Appendix. We also provide simple tests for mean independence of the factor loadings in Section 4, i.e. we can test when the TWFE model suffices.

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

$$\boldsymbol{P}(\boldsymbol{X}_1, \boldsymbol{X}_0) \equiv \boldsymbol{X}_1(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0' \tag{10}$$

This matrix takes a similar form to a projection matrix but “imputes" the fitted values from regressing on $\boldsymbol{X}_0$ onto a different matrix $\boldsymbol{X}_1$. Gardner (2021) implicitly uses the

imputation matrix where X_1 is the matrix of unit and time fixed effects and X_0 is X_1 with rows of zero whenever $d_{it} = 1$.

Now that the transformed untreated outcomes display a pure-factor structure, we impute untreated potential outcomes for group g using $\boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\geq g}, \tilde{\boldsymbol{F}}_{t<g})$ where $\tilde{\boldsymbol{F}}_{t<g}$ is the first $g-1$ rows of $\tilde{\boldsymbol{F}} = (\tilde{\boldsymbol{f}}_1, ..., \tilde{\boldsymbol{f}}_T)'$ and $\tilde{\boldsymbol{F}}_{t\geq g}$ is the last $T-g+1$. When applying this matrix to outcomes, the post-treatment factors are multiplied by the factor loadings from the pre-treatment observations. In particular, we impute $\tilde{y}_{it}(\infty)$ by $\boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g}$ for $G_i = g$, similar to the bridge function identification scheme in Imbens et al. (2021). However, we allow arbitrary correlation between the idiosyncratic errors.

Theorem 2.1. Suppose $\tilde{\boldsymbol{F}}$ is known and $\text{Rank}(\tilde{\boldsymbol{F}}_{t\leq T_0}) = p$. Under Assumptions 1-4 for $g \in \mathscr{G}$,

$$\text{ATT}(g,t) = \mathbb{E}\left[\tilde{y}_{it} - \boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g} \mid G_i = g\right] \tag{11}$$

for $t \geq g$. ■

Theorem 2.1 shows that we can identify τ_{gt} if we know the factor matrix. However, all of the estimators discussed earlier only estimate the factors up to an unknown rotation because both $\boldsymbol{f}_t$ and $\boldsymbol{\gamma}_i$ are unobserved. Fortunately for us, our results are invariant to this rotation:

Theorem 2.2. Let $\boldsymbol{A}$ be a nonsingular $p \times p$ matrix. Then for any $g > T_0$,

$$\boldsymbol{P}(\tilde{\boldsymbol{f}}_t'\boldsymbol{A}, \tilde{\boldsymbol{F}}_{t<g}\boldsymbol{A}) = \boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g}) \tag{12}$$

■

Theorems 2.1 and 2.2 are in fact very general results. We can apply these conclusions to any interactive fixed effects estimator that achieves consistency by estimating a rotation of the factor space. Examples include the common correlated effects estimator of Pesaran (2006), the principal components estimator of Bai (2009), or the QLD transformation of Ahn et al. (2001, 2013). As long as the factors are consistently estimated using

the control sample, dynamic ATTs are identified as in Theorem 2.1, regardless of the normalization used for estimation. We explore the QLD method in the following section because it provides overidentifying conditions that we use for practical tests of the model's assumptions.

Chan and Kwok (2022) propose a principal components difference-in-differences estimator for unit-specific treatment effects, ruling out dynamic effects because of the large time series assumed in the asymptotic arguments. Our results show that estimation of the factors can be carried out using untreated observations and then applied to any post-treatment period for a time-specific ATT. This result relaxes their time-homogeneity assumption while also allowing for empirical examples where there are many pre-treatment observations.

2.2. Quasi-Long-Differencing Identification

This section considers identification of the factors in a fixed-T environment using the approach of Ahn et al. (2013). We reiterate that it is not the only method to identify the factors and any estimator that is consistent for the factors would work in Theorem 2.1. Each estimator has different identifying assumptions which may be more or less plausible in different contexts. For example, if one wanted to utilize a common correlated effects approach, they would require identifying assumptions like those in Westerlund et al. (2019).

The advantage of our proposed estimator is two-fold. First, the estimator takes the form of a generalized method of moments estimator which makes asymptotic inference a result of simple theory. Second, this estimator will allow us to form an easy-to-implement statistical test for the sufficiency of the two-way fixed effect model in Section 4.

We proved in Theorem 2.2 that the given normalization of the factors does not affect our resulting imputation, so we follow Ahn et al. (2013) and impose the following p^2

normalizations:

$$\tilde{F}(\theta) = \begin{pmatrix} \Theta \\ -I_p \end{pmatrix} \tag{13}$$

where Θ is a $(T-p) \times p$ matrix of unrestricted parameters and $\theta = \text{vec}(\Theta)$. Given this normalization, the **quasi-long-differencing (QLD)** matrix is

$$H(\theta) = \begin{pmatrix} I_{(T-p)} \\ \Theta' \end{pmatrix} \tag{14}$$

For any given θ, $H(\theta)'\tilde{F}(\theta) = 0$.

Like Callaway and Karami (2022), we require instruments w_i to identify the factor model. Section 3 describes how to include covariates in the selection assumption. Naturally, these covariates would also serve as instruments to identify θ. We introduce three additional identifying assumptions:

Assumption 5 (Factor identification). Let w_i be a $L \times 1$ vector of instruments that is randomly sampled along with the other data and has finite fourth moments. Then

(i) $\text{Rank}(Var(\tilde{\gamma}_i \mid G_i = \infty)) = \text{Rank}(\tilde{F}_{t<T_0}) = p < T_0$.

(ii) The matrix $\mathbb{E}\left[I_{(T-p)} \otimes w_i \tilde{\gamma}_i' \mid G_i = \infty\right]$ has full column rank.[8]

(iii) $\mathbb{E}[u_i \mid w_i, G_i = \infty] = 0$.

■

Assumption 5 is our adaptation of BA.3 from Ahn et al. (2013) and gives identification of the normalized factors. Assumption 5(ii) and (iii) inform what instruments are allowed. Part (iii) implies they are exogenous with respect to the idiosyncratic error. We can weaken the strict exogeneity assumption to allow for instruments that are only valid

8. '$\otimes$' denotes a Kronecker product, where each element of $I_{(T-p)}$ is multiplied by $w_i \tilde{\gamma}_i$.

in certain time periods so that (iii) is not as restrictive as it seems. Second, we require the instruments to correlate with the demeaned factor loadings. We can allow covariates that vary over time and individual, or just across individuals, giving us a broad selection of potential instruments that nests those in Callaway and Karami (2022).

Given Assumption 5, we now show that the never-treated individuals can be used to identify the parameters in $\boldsymbol{\theta}$.

Lemma 2.2. Under Assumptions 1-5 and given p is known and $p+1 < T$, $\boldsymbol{\theta}$ is identified by

$$\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\theta})'\tilde{\boldsymbol{y}}_i \otimes \boldsymbol{w}_i \mid G_i = \infty\right] = \mathbf{0} \tag{15}$$

■

The proof is an immediate consequence of Lemma 2.1 and Section 2 of Ahn et al. (2013). A key identifying assumption is that p is known to the researcher. Ahn et al. (2013) provide consistent tests of p under Assumptions 1-5. Further, simulation evidence suggests that overestimating the number of factors does not lead to bias in the parameters of interest[9]. We treat p as known for the remainder of the theory. Our empirical application includes transparent evidence for our selection of p.

Lemma 2.2 tells us that $\boldsymbol{\theta}$ can be identified, but says nothing about the actual $\tilde{\boldsymbol{F}}$. However, as $\boldsymbol{\theta}$ is generated by a rotation of $\tilde{\boldsymbol{F}}$, we can use $\boldsymbol{\theta}$ to identify the column space of $\tilde{\boldsymbol{F}}$.

Lemma 2.3. Under Assumption 5,

$$\boldsymbol{P}(\tilde{\boldsymbol{F}}(\boldsymbol{\theta})_{t\geq g}, \tilde{\boldsymbol{F}}(\boldsymbol{\theta})_{t<g}) = \boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\geq g}, \tilde{\boldsymbol{F}}_{t<g}) \tag{16}$$

for $g \in \mathscr{G}$. ■

Lemma 2.3 is a direct consequence of Theorem 2.2. Combined with Theorem 2.1, Lemma 2.3 implies that the $\tau_{g,t}$'s are identified under Assumptions 1-5. The original

9. See Ahn et al. (2013), Breitung and Hansen (2021), and Brown (2022)

QLD estimator of Ahn et al. (2013) and the pooled QLD estimator of Brown (2022) may not have the same asymptotic variances for different normalizations of the factors. However, our estimator of treatment effects does according to this lemma.

3 — Estimation and Inference

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

3.1. Asymptotic Normality

Equations (11) and (15) provide us the necessary moment conditions to estimate the ATTs. We collect them here in their unconditional form:

$$\mathbb{E}[\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})] = \mathbb{E}\left[\frac{D_{i\infty}}{\mathbb{P}(D_{i\infty}=1)}\boldsymbol{H}(\boldsymbol{\theta})'\tilde{\boldsymbol{y}}_i \otimes \boldsymbol{w}_i\right] = \mathbf{0}$$

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

$$\vdots$$

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

where $\boldsymbol{\tau}_g = (\tau_{gg},...,\tau_{gT})'$ is the vector of post-treatment treatment effects. We stack these over g as $\boldsymbol{\tau} = (\boldsymbol{\tau}'_{g_1},...,\boldsymbol{\tau}'_{g_G})'$. The first set of moment conditions identify θ and the remaining moments identify the τ_{gt} via our imputation method.[10] Implementation requires replacing $P(D_{ig}=1)$ with its sample counterpart N_g/N. We need one final assumption to implement the asymptotically efficient GMM estimator:

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

Assumption 6. $\mathbb{E}\left[\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g)\boldsymbol{g}_{ig}(\boldsymbol{\theta},\boldsymbol{\tau}_g) \mid G_i = g\right]$ is positive definite for each $g \in \mathcal{G}$. ■

Assumption 6 makes sure the variance of the moments is not rank deficient after removing the factors. We collect the moment functions into the vector $\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau}) = (\boldsymbol{g}_{i\infty}(\boldsymbol{\theta})', \boldsymbol{g}_{ig_G}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_G})', ..., \boldsymbol{g}_{ig_1}(\boldsymbol{\theta},\boldsymbol{\tau}_{g_1})')'$. We define $\boldsymbol{\Delta} = \mathbb{E}[\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})']$ which is positive definite by Assumptions 5 and 6. Then our GMM estimators of $(\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')'$ solve

$$\min_{\boldsymbol{\theta},\boldsymbol{\tau}} \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right)' \widehat{\boldsymbol{\Delta}}^{-1} \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta},\boldsymbol{\tau})\right) \tag{17}$$

where $\widehat{\boldsymbol{\Delta}} \xrightarrow{p} \boldsymbol{\Delta}$ uses an initial consistent estimator of $(\boldsymbol{\theta}', \boldsymbol{\tau}')'$. We now present the main theoretical result.

Theorem 3.1. Under Assumptions 1-6, $\sqrt{N}\left((\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')' - (\boldsymbol{\theta}', \boldsymbol{\tau}')'\right)$ is jointly asymptotically normal and

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

where $\boldsymbol{D}_g$ is the gradient of group g's moment function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}_g$ is the variance of group g's moment function. Further, the asymptotic covariance between $\sqrt{N}(\widehat{\boldsymbol{\tau}}_{g_h} - \boldsymbol{\tau}_{g_k})$ and $\sqrt{N}(\widehat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k)$ is given by $\boldsymbol{D}_{g_h}(\boldsymbol{D}_\infty' \boldsymbol{\Delta}_\infty^{-1} \boldsymbol{D}_\infty)^{-1} \boldsymbol{D}_{g_k}'$. ■

The asymptotic distribution of $\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$ generally depends on the estimation of $\boldsymbol{\theta}$ in the first stage by the term $D_g(D_\infty' \Delta_\infty^{-1} D_\infty)^{-1} D_g'$. We can see directly from Theorem 3.1 that a smaller Avar$(\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}))$ leads to a smaller Avar$(\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g))$ (in the matrix sense), strictly so when $\boldsymbol{D}_g$ has full rank. This result also suggests that more efficient estimation of the factors is an important avenue of future work and demonstrates why our general identification result is so powerful: we can use different estimators of the

factors if we believe we can achieve substantial efficiency gains. Further, estimation of $\boldsymbol{\tau}_g$ is not dependent on the first stage estimation of $\boldsymbol{\theta}$ when $\boldsymbol{D}_g = \mathbf{0}$. A sufficient condition for this equality occurs when the transformed factor loadings for group g center about zero.

We also note that the estimator of $\boldsymbol{\tau}$ in Theorem 3.1 is asymptotically equivalent to the two-step estimator that estimates $\boldsymbol{\theta}$ on its own then treats $\widehat{\boldsymbol{\theta}}$ as given. This result follows from Theorem 2.2 of Prokhorov and Schmidt (2009) because the moments estimating θ and the moments estimating treatment effects are estimated using mutually exclusive subsamples. Direct computation of $\widehat{\boldsymbol{\theta}}$ frees the researcher to investigate different aggregates of the treatment parameters without jointly estimating $\boldsymbol{\theta}$, saving greatly on computational time because the ATT estimators have a closed-form solution. Further, if we choose the instruments so that the number of moments is equal to p, the two-step and joint estimators of $(\boldsymbol{\theta}, \boldsymbol{\tau})$ are numerically equivalent (Prokhorov and Schmidt 2009). However, inference on $\boldsymbol{\tau}$ requires accounting for the affect of estimating $\boldsymbol{\theta}$ in the first stage per Theorem 3.1. Inference can be accomplished via standard GMM statistical packages or the bootstrap. One can use the usual non-parametric bootstrap, re-estimating $\boldsymbol{\theta}$ in each sample, or the multiplier bootstrap using the influence function that we derive in the Online Appendix.

When the gradient $\boldsymbol{D}_g = \mathbf{0}$ for a given g, the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}_g)$ is just $\boldsymbol{\Delta}_g$. This quantity is simple to estimate via a nonparametric variance estimator. Let

$$\widehat{\boldsymbol{\Delta}}_g = \frac{1}{N_g - 1} \sum_{i=1}^{N} D_{ig} \left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{g_G}\right)\left(\widehat{\boldsymbol{\Delta}}_{ig} - \widehat{\boldsymbol{\tau}}_{g_G}\right)' \tag{18}$$

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

Theorem 3.2. Under Assumptions 1-6, $\widehat{\boldsymbol{\Delta}}_g^{-1} \xrightarrow{p} \boldsymbol{\Delta}_g^{-1}$.

It is important to note that while the estimator in equation (18) is always consistent for $\boldsymbol{\Delta}_g^{-1}$, this quantity is not generally equal to the asymptotic variance. One must take

care to include the contribution of estimating $\boldsymbol{\theta}$ into the standard errors.

Remark 3.1 (Limited Anticipation)**.** Assumption 3 implies treated individuals do not anticipate treatment and adjust their behavior prior to the intervention. Suppose treated individuals from group g anticipate the intervention in period $q_g < g$. We could simply redefine 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:

$$H_0 : \tau_{g,q_g} = ... = \tau_{g,g-1} = 0 \tag{19}$$

This test can be easily carried out using standard statistical packages once estimation is finished.

In fact, the above test is just one of many that can be carried out on the ATTs. As $\text{ATT}(g,t)$ is $\sqrt{N}$-consistently estimated by $\widehat{\tau}_{gt}$, and all standard errors come from known theory on GMM estimation, we can test any well-defined nonlinear function of the parameters using canned statistical packages. ■

Remark 3.2 (Other Aggregate Treatment Effects)**.** Our estimation method can handle other aggregations of $\tilde{y}_{it} - \hat{\tilde{y}}_{it}(\infty)$. For example, one could aggregate over all post-treatment (i,t) to estimate an overall ATT or over event time indicators to estimate event-study aggregates. Researchers can perform heterogeneity analyses by aggregating for units with different values of X_i like gender, race, or age to estimate a conditional ATT. Since our estimator is a GMM estimator, all one needs to do to estimate such aggregate effects is to correctly specify the weights that define the unconditional moments. For example, a researcher may be interested in the effect of treatment among all units one period after entering treatment. This event study coefficient is defined

$$\text{ATT}^{\ell} \equiv \mathbb{E}\left[y_{it}(G_i) - y_{it}(\infty) \mid t - G_i = \ell\right] \tag{20}$$

Simply define the group membership $D_{i,t-G_i=\ell}$ as one if the unit satisfies the condition $t - G_i = \ell$ for some t. Then to estimate the effect, we use the unconditional moments

$$\mathbb{E}\left[\frac{D_{i,t-G_i=\ell}}{\mathbb{P}(D_{i,t-G_i=\ell}=1)}\left(\tilde{y}_{it} - \boldsymbol{P}(\tilde{\boldsymbol{f}}_t, \tilde{\boldsymbol{F}}_{t<G_i})\tilde{\boldsymbol{y}}_{i,t<G_i}\right)\right] = 0 \tag{21}$$

where the probability is replaced with the observed proportion of the sample satisfying the condition. Stacking these treatment effect moments allows for estimating the vector of event-study coefficients. The only difference from before is that the variance matrix of the moment functions is not diagonal because of overlapping post-treatment samples. Inference via bootstrap of via the influence function follows through as well.

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

Remark 3.3 (Pre-Treatment 'Placebo' Effects)**.** We can also derive pre-treatment "placebo" effects by estimating a coefficient on the pre-treatment time periods. The imputation matrix that carries out this estimation is $\boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\leq g}, \tilde{\boldsymbol{F}}_{t\leq g})$ which is just the projection matrix from regressing on the pre-treatment factors. Under the no anticipation assumption,

$$\mathbb{E}\left[\left(\boldsymbol{I}_g - \boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\leq g}, \tilde{\boldsymbol{F}}_{t\leq g})\right)\tilde{\boldsymbol{y}}_{i,t\leq g} \mid G_i = g\right] = \boldsymbol{0} \tag{22}$$

so that the properly standardized vector of pre-treatment residuals is asymptotically multivariate normal. As with ATT estimation, one must take care to control for estimation of the factors in constructing valid standard errors. ■

3.2. Plotting Estimates

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

$$\hat{y}_{it}(\infty) = P(\tilde{\boldsymbol{f}}_t, \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g} + \overline{y}_{\infty,t} + \overline{y}_{i,t<T_0} - \overline{y}_{\infty,t<T_0} \tag{23}$$

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

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

3.3. 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

$$\mathbb{E}[y_{it}(\infty) \mid x_i, \mu_i, \boldsymbol{\gamma}_i, D_i] = \boldsymbol{x}_i \boldsymbol{\beta}_t + \mu_i + \lambda_t + \boldsymbol{f}_t' \boldsymbol{\gamma}_i \tag{24}$$

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

$$\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\theta})'(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}) \otimes \boldsymbol{w}_i \mid G_i = \infty\right] = \mathbf{0} \tag{25}$$

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

$$\mathbb{E}\left[(\tilde{\boldsymbol{y}}_{i,t\geq g} - \tilde{\boldsymbol{X}}_{i,t\geq g}\boldsymbol{\beta}_g) - \boldsymbol{P}(\tilde{\boldsymbol{F}}_{t\geq g}, \tilde{\boldsymbol{F}}_{t<g})(\tilde{\boldsymbol{y}}_{i,t<g} - \tilde{\boldsymbol{X}}_{i,t<g}\boldsymbol{\beta}_g) - \boldsymbol{\tau}_g \mid G_i = g\right] = \mathbf{0} \tag{26}$$

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 leave evaluation of this assumption in factor-augmented linear models to future research.

4 — TWFE Specification Testing

A novel insight of our paper concerns the ability to test for a factor structure.[11] We consider the following hypotheses:

$$H_0 : y_{it}(\infty) = \mu_i + \lambda_t + u_{it}$$
$$H_A : y_{it}(\infty) = \mu_i + \lambda_t + \boldsymbol{f}_t' \boldsymbol{\gamma}_i + u_{it}$$

If the null hypothesis is true, the more computationally burdensome QLD procedure is unnecessary for estimating the ATTs.[12] Therefore, we think this test is of practical importance for researchers. We discuss in the previous section how Ahn et al. (2013) provide consistent estimation of p. Those tests have a new interpretation under this null hypothesis when testing for p on the residuals $\tilde{y}_{it}$.

Theorem 4.1. Under the null hypothesis $H_0 : y_{it}(\infty) = \mu_i + \mu_t + u_{it}$ and Assumptions 1 and 3, $p = 0$. ■

Failure to reject the null hypothesis implies that the two-way error model is sufficient for capturing all heterogeneity in the potential outcomes. Under the untreated model, one could use our imputation approach from Section 2, or an approach that uses all untreated outcomes to estimate μ_i and λ_t. One can even carry out this test without implementing a QLD procedure. The imputed residuals are mean zero under the null hypothesis so the usual overidentifying test is implemented by setting $\boldsymbol{H}(\boldsymbol{\theta})' = \boldsymbol{I}_T$.

Even if the two-way error model is unrepresentative of the factor structure, Corollary 2.1 shows that mean independence of the factor loadings with respect to treatment timing is sufficient for consistency of TWFE. Specifically, we need $\mathbb{E}[\boldsymbol{\gamma}_i] = \mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g]$ for all $g \in \mathscr{G}$. Our imputation approach allows us to identify these terms up to a rotation.

11. It is theoretically possible to compare the difference between our imputation estimator from Theorem 3.1 to the TWFE imputation estimator via a generalized Hausman test. We refer the reader to Wooldridge (2010) for example.

12. Even if TWFE is consistent, it is not necessarily more efficient than our procedure. See Section 5 for example.

To see how, let $\boldsymbol{A}^*$ be the rotation that imposes the Ahn et al. (2013) normalization. Then

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

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

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

$$\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] - \mathbb{E}[\boldsymbol{\gamma}_i] = \boldsymbol{0} \iff (\boldsymbol{A}^*)^{-1}(\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] - \mathbb{E}[\boldsymbol{\gamma}_i]) = \boldsymbol{0} \tag{27}$$

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

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

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 $g \in \mathscr{G}$. We include estimation of the factors for convenience, so that one does not

need to directly calculate the effect of first-stage estimation on the asymptotic variances of conditional means.

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

Theorem 4.2. If $\mathbb{E}[\boldsymbol{\gamma}_i \mid G_i = g] = \mathbb{E}[\boldsymbol{\gamma}_i]$ for all $g \in \mathcal{G}$, then

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*_{g_G} = ... = \boldsymbol{\gamma}^*_{g_1} \tag{28}$$

■

We also provide a test for the common factor assumption in the Online Appendix. Under the null, the factors that enter both the treated and never-treated group outcomes span the same column space. Our test takes the form of a structural breaks test and is valid when T is fixed.

5 — Simulations

We present a brief simulation study to compare our estimator to different TWFE specifications. We consider the setting where $T = 8$ and treatment turns on starting in period 6 implying $T_0 = 5$. We draw $N = 200$ observations, which is a relatively small number for a nonlinear estimation problem. We generate untreated potential outcomes following equation (2). We consider the setting with one factor that we generate as a time-trend $f_t = t$.[13] We generate the time fixed effects as $\theta_t = 0.75 * \theta_{t-1} + \nu_t$ where $\nu_t \sim N(0,1)$. We generate the unit fixed effects as iid with $\mu_i \sim N(0,4)$ and the factor loadings to be correlated with the unit fixed-effects by drawing from $\gamma_i \sim N(\mu_i, 1)$. The error term is generated as an $AR(1)$ process with correlation coefficient 0.75 and is uncorrelated with treatment status. We generate individual-level treatment effect heterogeneity by

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

defining individual treatment effects $\tau_{i\ell}$ to be τ_ℓ times the unit fixed effect but then re-scale the individual effects to have mean equal to $\tau_6 = 1$, $\tau_7 = 2$, and $\tau_8 = 3$ and for the variance of $\tau_{i\ell}$ to be one.

We generate a covariate $w_i = \gamma_i + \xi_i$ where ξ_i is white-noise measurement error. w_i will be used as a covariate in some TWFE specifications and as our instrument for our factor-model estimation. In the baseline simulation, we consider the case where $\xi_i \sim N(0, 1)$ which creates a signal-to-noise ratio of $1/2$. In a set of simulations, we vary the level of noise to see how the instrument strength affects estimates. These results will allow us to compare our methods to those that use noisy measurements of unobserved heterogeneity.

We consider three data-generating processes. First, we consider the true TWFE model where there is no factor model. In this case, the two-way fixed effects model should be unbiased. Second, we generate outcomes with the factor model described above. Treatment is then assigned completely randomly with probability of treatment at 50% for all units. This implies that the factor loadings are uncorrelated with treatment status, which corollary 2.1 shows is sufficient for the TWFE imputation procedure to be consistent. Third, we generate treatment with probability increasing in the factor loading such that parallel trends fail (since treated units are more exposed to the time-trend in f_t). In particular, we form the term

$$\pi_i = 0.5 + \frac{\gamma_i}{\max_i \gamma_i - \min_i \gamma_i} \tag{29}$$

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

We estimate event-study treatment effects using four estimators. First, we estimate the classical TWFE model using ordinary least squares (OLS). Second, we estimate the TWFE model using the imputation estimator proposed by Borusyak et al. (2022). Third, we augment the TWFE model by including a noisy measure of the factor loadings. This is sometimes done by applied researchers in an attempt to control for confounders. That

is, they model outcomes as

$$y_{it} = \mu_i + \lambda_t + w_i\beta_t + u_{it} \tag{30}$$

where w_i is a time-invariant covariate and β_t allows for trends to vary based on w_i. In the case where $w_i = \gamma_i$, i.e. the factors are observable, this model is correctly specified. However, when $\text{Var}(\xi_i) > 0$, i.e. the covariates are noisy measures for the underlying factor loadings, model (30) will only partially absorb the factor model.[14] Last, we use our proposed factor-model imputation estimator.

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

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

It is worth noting that while including $w_i\beta_t$ in the model does remove some bias, the estimates still perform worse than our imputation procedure due to w_i being a

14. Kejriwal et al. (2021) make a similar point about controlling for imperfect measures of latent ability when estimating the returns to schooling.

Table 1 — Monte Carlo Simulation

Panel A: TWFE Model.

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

Panel B: Factor Model. Parallel Trends Hold

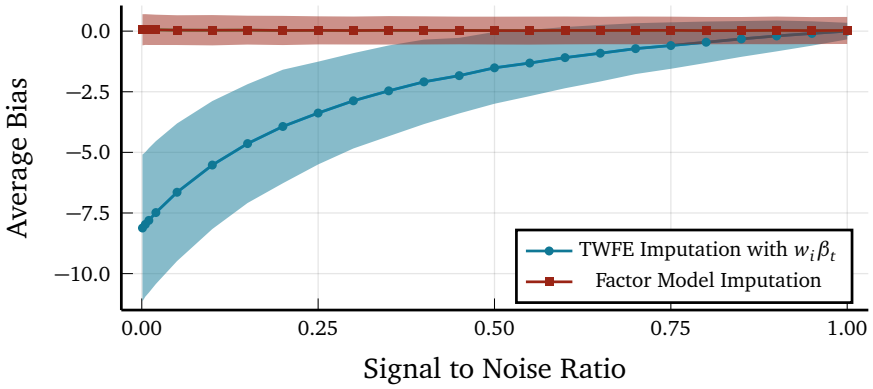
	Bias $(\hat{\tau}_6)$	MSE $(\hat{\tau}_6)$	Bias $(\hat{\tau}_7)$	MSE $(\hat{\tau}_7)$	Bias $(\hat{\tau}_8)$	MSE $(\hat{\tau}_8)$
TWFE	0.00	0.11	0.00	0.43	0.01	0.95
TWFE Imputation	0.00	0.94	0.00	1.67	0.01	2.60
TWFE Imputation with Covariates	0.00	0.17	0.00	0.29	0.01	0.44
Factor Imputation	-0.00	0.02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

Panel C: Factor Model. Parallel Trends Do Not Hold

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

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

Figure 1 — Bias of TWFE Imputation with Covariates



Notes. *This figure plots the average and empirical 95% confidence intervals for treatment effect estimates in the final period, $\hat{\tau}_8$. We estimate the TWFE imputation estimator that includes $w_i\beta_t$ linearly in the model and our the factor imputation we propose using w_i instead as an instrument. We vary the signal to noise ratios of w_i to make it a better or worse measure for the factor loading. For each signal to noise ratio, we run 5000 simulations.*

noisy measure. To highlight the problems with noisy proxies for factor loadings, Figure 1 presents a set of simulation results where the covariate w_i has different amount of noise added in. In particular, we choose different values of $\text{Var}(\xi_i)$ to have different signal-to-noise measures. The signal-to-noise definition is

$$\text{signal to noise ratio} = \frac{\text{Var}(\gamma_i)}{\text{Var}(\gamma_i) + \text{Var}(\xi_i)} = \frac{1}{1 + \text{Var}(\xi_i)/\text{Var}(\gamma_i)} \tag{31}$$

For each signal to noise ratio, we estimate the TWFE imputation estimator with covariates and the factor model imputation estimator. Figure 1 presents the results of estimates for τ_8. At one extreme, where the signal to noise ratio is approximately 0, i.e. w_i is white noise, the estimated bias for the TWFE imputation estimator is the same as the TWFE imputation estimator that does not include covariates. At the other extreme, where the signal to noise ratio is approximately 1, i.e. $w_i = \gamma_i$, the bias is completely removed. Regardless, the factor model imputation estimator is unbiased in all cases. This echos the results of Kejriwal et al. (2021). However, we note that our results are still generous to estimators that use such noisy measure because we generate w_i as an

unbiased estimator of γ_i. The instrument requirement in Assumption 5 does not require this mean restriction for identification of $\boldsymbol{\theta}$.

6 — Application

We now present an empirical application that revisits the literature on estimating county-level labor markets effects of Walmart store openings. 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 macroeconomic trends, then the non-treated counties would fail to be a valid counterfactual group in the TWFE model. Indeed, we observe significant differences in both retail and wholesale retail employment trends for treated counties in our data.[15]

Volpe and Boland (2022) point to conflicting results on retail employment with two leading papers finding effects of opposite signs. Employing different instrumental variable strategies, Basker (2005) finds positive effects on retail employment while Neumark et al. (2008) finds negative effects. For this reason, we revisit this question with an alternative strategy to reconcile results.

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.[16] We use a geocoded dataset of Walmart openings from Arcidiacono et al. (2020) to construct our treatment variable. Our treatment dummy is equal to one if the county has any Walmart in that year and our group variable denotes the year of entrance for the *first* Walmart in the county. [17] We drop any county that was treated with $g \leq T_0 = 1985$ so that we we have

15. Wholesale retail employment corresponds to NAICS 2-digit code 42 and retail employment corresponds with NAICS 2-digit codes 44 and 45.

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

17. For our sample 82.4% of our counties receive ≤ 1 Walmart and another 10.4% receive two Walmarts

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).

First, we estimate the two-way fixed effect imputation estimator proposed by Borusyak et al. (2022) and estimate event-study effects on (log) retail and wholesale retail employment. In particular, we use the following model

$$\log(y_{it}) = \mu_i + \lambda_t + \sum_{\ell=-22}^{13} \tau^\ell d_{it}^\ell + u_{it} \tag{32}$$

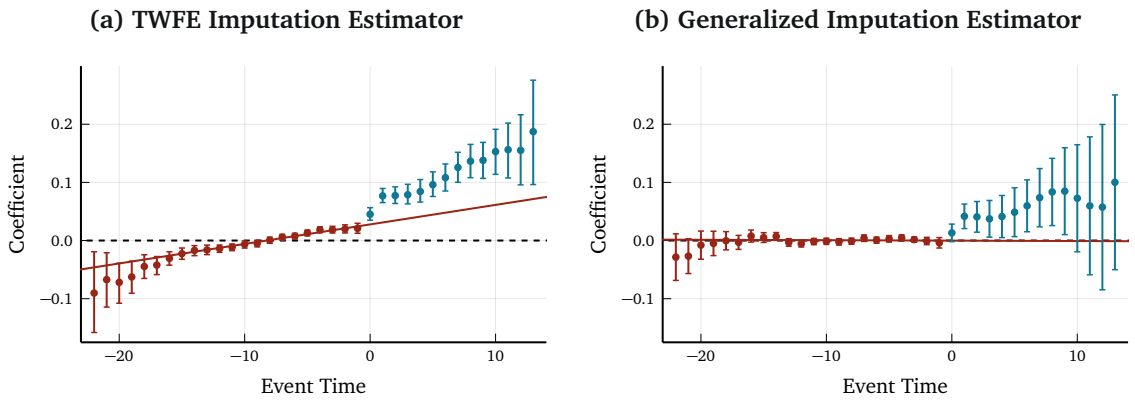
where i denotes county, t denotes year, y_{it} is either retail or wholesale retail employment, and $d_{it}^\ell = 1(t - g_i) = \ell$ are indicator variables denoting event-time. Results of the event-study estimates are presented in panel (a) of Figure 2 and Figure 3.

For both retail and wholesale retail employment, counties receiving Walmarts had faster employment trends relative to the control counties, emphasizing our concern over endogenous opening decisions. In the spirit of Freyaldenhoven et al. (Forthcoming) and Rambachan and Roth (2022), we draw the line of best fit for the 15 most-recent pre-treatment estimates (τ^ℓ for $-15 \leq \ell < 0$) and extend it into the post-treatment estimates. For both retail and wholesale retail 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). The goal of our generalized imputation estimator is to remove the pre-existing economic trends in order to isolate the true treatment effects.

We now implement the generalized imputation estimator proposed in Section 3. For our instruments $\boldsymbol{w}_i$, we use the following variables at their 1980 baseline values: 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,

in the sample, alleviating some concerns of making treatment binary.

Figure 2 — Effect of Walmart on County log Retail Employment



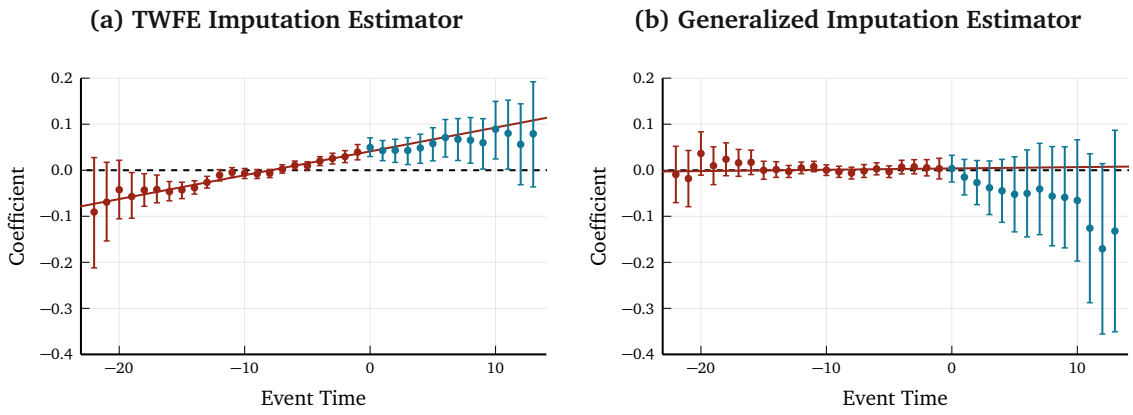
Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log retail employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Gardner (2021). Panel (b) estimates effects using the generalized imputation estimator we propose in Section 3 with $p = 2$ and using the following instruments: 1980 share of population employed in manufacturing, 1980 shares of population below and above poverty line; 1980 shares of population employed in private-sector and by the government, 1980 shares of population with high-school degree and college degree. The red lines correspond to a linear estimate of pre-treatment point estimates for event time -15 to -1 and is extended into the post-treatment periods.

and shares of population with high-school and college degrees. All of these values are obtained from 1980 Census Tables accessed from Manson (2020). Intuitively, we think these instruments capture the general macroeconomic trends that are driving differential economic growth in the 1980s and 1990s. We use baseline shares to prevent our instruments from picking up on contemporaneous economic shocks that could be correlated with Walmart opening. 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 3.1.

The results of our estimator are presented in panel (b) of Figure 2 and Figure 3.[18] 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

18. 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 retail 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 Retail Employment

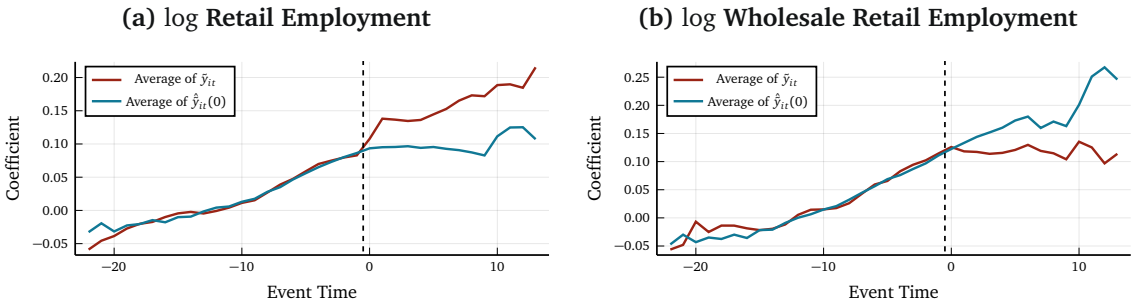


Notes. This figure plots point estimates and bootstrapped 95% confidence intervals for event-study treatment effects on log wholesale retail employment. Panel (a) estimates effects using the TWFE imputation estimator proposed in Gardner (2021). Panel (b) estimates effects using the generalized imputation 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.

point estimates are smaller than estimated by the TWFE model with an estimated effect on employment of around 6% on average in the post-treatment periods. Evaluated at the median baseline retail employment of 1417 employees, this would imply an increase in about 85 jobs which is in line with the estimates of Basker (2005) and Stapp (2014) who use alternative instrumental variables strategies. It is important to note that post-treatment estimates are noisier than the TWFE estimates largely due to estimating θ. This problem is at the worst for the furthest event-times largely due to very few counties being averaged over in the last few bins. We view this as a worthy trade-off since the point estimates are much less likely to be biased.

Turning to wholesale retail employment, we see a similar story with our estimator removing most of the pre-trend violations. In this case, however, the estimated effects flip signs with an estimated effect of around -6% although they are not statistically significant. Evaluated at the 1977 median wholesale retail employment of 410, this

Figure 4 — Synthetic Control Style Plot of the Effect of Walmart on County Employment



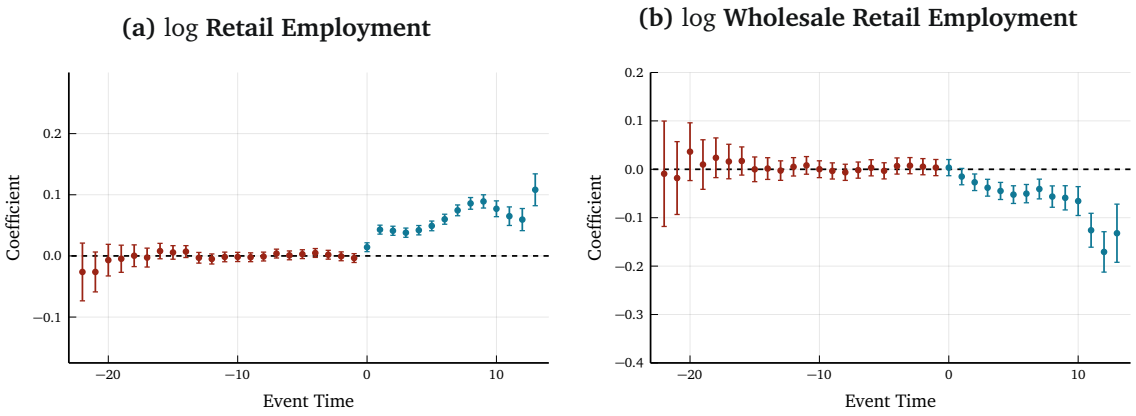
Notes. *This figure plots the observed* $\tilde{y}_{it}$ *and the imputed* $\hat{\tilde{y}}_{it}(0)$ *for treated units averaged over event time* $\ell = t - g_i$. *We impute within-transformed potential outcome using the generalized imputation estimator we propose in Section 3 using the following instruments: 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.*

suggests a decrease of about 25 jobs which is very similar to what Basker (2005) finds. Overall, we find effects very much in line with those reported in Basker (2005).

As we discuss in Section 3.2, 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 retail employment in Figure 4. In pre-treatment ($\ell < 0$), the imputed estimate, our 'synthetic control' follows closely with the observed $\tilde{y}_{it}$ giving us confidence in our ability to approximate the factor structure. In the post-periods, we see the observed counties and the imputed untreated version of the counties pulling apart. The gap between the two are our estimated treatment effects.

To highlight the importance of the uncertainty from estimation of Θ, in Figure 5 we recreate point estimates from our generalized imputation estimator using the non-parametric standard errors that are derived in Theorem 3.2. The standard errors on point estimates are far smaller with estimates becoming strongly significant in Wholesale Retail Employment. This result shows an important step for future research in finding more efficient estimates of the factors. For instance, we consider the common correlated effects estimator in a follow-up paper. The CCE model generally implies that

Figure 5 — Generalized Imputation Estimator for Effect of Walmart on County Employment with Naive Standard Errors



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

the nonparametric standard errors are valid when there is a common factor model for time-varying covariates.

7 — Conclusions

We consider identification and inference of functions of heterogeneous treatment effects in a linear panel data model. We show how to relax the usual parallel trends assumption by introducing a linear factor model in the error. Our main identification result shows that a consistent estimator of the unobserved factors is all that one needs to estimate the dynamic treatment effect coefficients. This result is general and can be implemented by a number of modern interactive fixed effects estimators, such as quasi-long-differencing, common correlated effects, or principal components, allowing for both large and small numbers of pre-treatment time periods. Further work can demonstrate both theoretical and finite-sample properties of these various estimators of the factors and how they affect to ATT estimation.

While a factor model nests the usual two-way fixed effects error structure, we explicitly model the level fixed effects in addition to the factors. This setting allows us

to provide useful tests for the sufficiency 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 unifying identification result for imputation estimators of ATTs.

We demonstrate how to implement the quasi-long-differencing transformation to estimate ATTs. This method provides a number of benefits in applications with a small number of time periods. Namely, it allows us to easily test fundamental features of the model, like no treatment anticipation and systemic differences in heterogeneity among treated groups. Inference is straightforward and can be applied to interesting aggregates as discussed in Callaway and Sant'Anna (2021).

A — Proofs

Proof of Lemma 2.1

We first derive the averages defined in Section 2.1 in terms of the potential outcome framework:

$$\overline{y}_{\infty,t} = \frac{1}{N_\infty}\sum_{i=1}^{N} D_{i\infty} y_{it} = \overline{\mu}_\infty + \lambda_t + \boldsymbol{f}_t \overline{\boldsymbol{\gamma}}_\infty + \overline{u}_{t,\infty}$$

$$\overline{y}_{i,t\leq T_0} = \frac{1}{T_0}\sum_{t=1}^{T_0} y_{it} = \mu_i + \overline{\lambda}_{t<T_0} + \overline{\boldsymbol{f}}_{t<T_0}\boldsymbol{\gamma}_i + \overline{u}_{i,t<T_0}$$

$$\overline{y}_{\infty,t<T_0} = \frac{1}{N_\infty T_0}\sum_{i=1}^{N}\sum_{t=1}^{T_0} D_{i\infty} y_{it} = \overline{\mu}_\infty + \overline{\lambda}_{t<T_0} + \overline{\boldsymbol{f}}_{t<T_0}\overline{\boldsymbol{\gamma}}_\infty + \overline{u}_{\infty,t<T_0}$$

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 $\mathbb{E}\left[u_{it} - \overline{u}_{i,t<T_0} \mid G_i = g\right] = 0$ by Assumption 4 and $\mathbb{E}\left[\overline{u}_{\infty,t<T_0} - \overline{u}_{t,\infty} \mid G_i = g\right] = \mathbb{E}\left[\overline{u}_{\infty,t<T_0} - \overline{u}_{t,\infty}\right] = 0$ by random sampling and iterated expectations.

□

Proof of Theorem 2.1

$$\mathbb{E}\left[\tilde{y}_{it} - \boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g} \mid G_i = g\right] = \mathbb{E}\left[\tilde{y}_{it}(1) \mid G_i = g\right] - \mathbb{E}\left[\boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g} \mid G_i = g\right]$$

We use the fact that

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

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 Lemma 2.1 and Assumption 2.

□

Proof of Theorem 2.2

Let $\boldsymbol{A}$ be a $p \times p$ rotation matrix and let $\boldsymbol{F}^* = \boldsymbol{F}\boldsymbol{A}$. Note that both inverses in the permutation matrix definition exist for every g because $\text{Rank}(\tilde{\boldsymbol{F}}^*) = \text{Rank}(\tilde{\boldsymbol{F}})$. Since

$$\boldsymbol{F}\boldsymbol{A}^* = \begin{pmatrix} \tilde{\boldsymbol{F}}_{t<g}\boldsymbol{A} \\ \tilde{\boldsymbol{F}}_{t\geq g}\boldsymbol{A} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{F}}^*_{t<g} \\ \tilde{\boldsymbol{F}}^*_{t\geq g} \end{pmatrix}$$

the result holds for any post-treatment row of the factors. Then we have

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

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

□

Proof of Theorem 3.1

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

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

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

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

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

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

Given we use the optimal weight matrix, the overall asymptotic variance is given by $(\boldsymbol{D}'\boldsymbol{\Delta}^{-1}\boldsymbol{D})^{-1}$. Δ is a block diagonal matrix so its inverse is trivial to compute. First, we

have

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

The transpose of the gradient matrix is

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

so that we get

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

We write this matrix as

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

where $\boldsymbol{A} = \sum_{g\in\mathcal{G}\cup\{\infty\}} \boldsymbol{D}'_g\boldsymbol{\Delta}_g^{-1}\boldsymbol{D}_g$ and $\boldsymbol{D} = \text{diag}\{\boldsymbol{\Delta}_g^{-1}\}_{g\in\mathcal{G}}$. We then apply Exercise 5.16 of Abadir and Magnus (2005) to get the final inverse. The top left corner of the inverse is

$\boldsymbol{F}^{-1}$ where

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

The rest of the first column of matrices takes the form

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

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

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

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

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

□

Proof of Theorem 3.2

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

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

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

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

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

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

The inverse exists with probability approaching one by Assumption 6.

□

References

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

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

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.

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.

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.

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.

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

Basker, Emek. 2005. "Job creation or destruction? Labor market effects of Wal-Mart expansion." *Review of Economics and Statistics* 87 (1): 174–183.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2022. "Revisiting Event Study Designs: Robust and Efficient Estimation."Technical report.

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. 2022. "Moment-based Estimation of Linear Panel Data Models with Factor-augmented Errors."Technical report.

Callaway, Brantly, and Sonia Karami. 2022. "Treatment effects in interactive fixed effects models with a small number of time periods." *Journal of Econometrics*.

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

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.

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.

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

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. Forthcoming. "Visualization, identification, and estimation in the linear panel event-study design.." *Advances in Economics and Econometrics: Twelfth World Congress*.

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. https://doi.org/10.1257/aer.20180609.

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

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.

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

Imbens, Guido, Nathan Kallus, and Xiaojie Mao. 2021. "Controlling for Unmeasured Confounding in Panel Data Using Minimal Bridge Functions: From Two-Way Fixed Effects to Factor Models." *arXiv preprint arXiv:2108.03849*.

Kejriwal, Mohitosh, Xiaoxiao Li, and Evan Totty. 2021. "The Efficacy of Ability Proxies for Estimating the Returns to Schooling: A Factor Model-Based Evaluation." *Available at SSRN 3843260*.

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

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.

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

Prokhorov, Artem, and Peter Schmidt. 2009. "GMM redundancy results for general missing data problems." *Journal of Econometrics* 151 (1): 47–55.

Rambachan, Ashesh, and Jonathan Roth. 2022. "A More Credible Approach to Parallel Trends."Technical report, Working Paper.

Stapp, Jacob. 2014. "The Walmart Effect: Labor Market Implications in Rural and Urban Counties." *SS-AAEA Journal of Agricultural Economics* 2014 (318-2016-9525): .

Volpe, Richard, and Michael A Boland. 2022. "The Economic Impacts of Walmart Supercenters." *Annual Review of Resource Economics* 14 43–62.

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

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

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators."Technical report.

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