Generalized Imputation Estimators for Factor Models

Nicholas Brown[*] and Kyle Butts[†]

September 9, 2022

This paper considers a general identification strategy of average treatment effect parameters in panel data settings under a linear factor model allowing for staggered treatment adoption, treatment effect heterogeneity, and a fixed number of time periods. Our model nests the classic two-way fixed effect model while allowing for selection into treatment based on common-contemporaneous shocks. We provide a unifying result for the identification of imputation-based estimators under a factor model. We propose a particular estimator, establish asymptotic properties, and provide statistical tests for the sufficiency of the two-way fixed effect model.

JEL Classification Number: C13, C21, C23, C26

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

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

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

1 — Introduction

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 heterogeneity that is constant over time, a set of time 'fixed effects' that capture macroeconomic trends which affect each unit equally, 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 outcomes, $y_{it}(\infty)$. We are interested in averages of individual treatment effects

$$\mathbb{E}\left[y_{it} - y_{it}(\infty) \mid \Omega\right] \tag{2}$$

where Ω is the specified set of post-treatment observations to average over. 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 units select into treatment based on contemporaneous shocks, the treated units no longer follow the same outcome trajectory as untreated units resulting in treatment effects being confounded by contemporaneous shocks.

This paper considers a more general 'parallel trends' type assumption that allows units to enter treatment based on their differential exposure to a set of unobservable but commonly experienced macroeconomic shocks. To accommodate this form of selection, we model untreated potential outcomes as

$$y_{it}(\infty) = \mu_i + \lambda_t + \boldsymbol{f}_t'\boldsymbol{\gamma}_i + \varepsilon_{it}, \tag{3}$$

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 we assume that $\mathbb{E}[\varepsilon_{it}] = 0$ for all (i, t).[1] One possible motivation for this model is that the factor is a macroeconomic shock and the factor-loading γ_i denotes a unit's exposure to the shock. Another possibility lets the γ_i represent time-invariant characteristics with a marginal effect on the outcome that changes over time.[2]

Current panel-data 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). 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.

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. 2021, Gardner 2021, Wooldridge 2021). However, these approaches only allow for level fixed effects and preclude interactions like in equation (3). We generalize these techniques by proposing an estimator that imputes the untreated potential outcomes under the more general (3).

To do so, we first remove the additive fixed effects with a double-demeaning transformation. Our treatment effect identification result then only requires root-N consistent estimates of $\boldsymbol{f}_t$.[3] We compute a matrix that projects the pre-treatment outcomes onto the estimated post-treatment factors, imputing the untreated poten-

1. For simplicity, covariates ignored at first and are added to (3) in Section 3. Note that this model for outcomes coincides with the standard TWFE model when $p = 0$ and with a 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 γ_i are unobserved worker characteristics and $\boldsymbol{f}_t$ are their time-varying prices or returns. See Bai (2009) for a collection of economic examples that justify the inclusion of a factor structure.

3. As discussed below in Section 2.2, some estimators of $\boldsymbol{f}_t$ can only identify a normalized version of $\boldsymbol{f}_t$. This is fine as our imputation procedure works with any normalization of the factors.

tial outcome. Averaging over the difference between the post-treatment observed outcome and the imputed potential outcomes gives a $\sqrt{N}$-consistent estimator of the average treatment effects.

There are two benefits of our imputation approach. First, it 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. Second, root-N consistent estimates of $\boldsymbol{f}_t$ are possible in short panels using a variety of approaches, such as instrumental variables (Callaway and Karami 2022, Ahn et al. 2013), common-correlated effects (Westerlund 2019), or projected principal components (Fan et al. 2016).[4] Since our identification result only relies on estimates of $\boldsymbol{f}_t$, we open up a broad set of tools from the factor-model literature on short-T estimation.

Below, we derive asymptotic properties of an imputation estimator using the method of Ahn et al. (2013) to estimate the factors. 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.[5] One advantage of this estimator is that we can form statistical tests for the sufficiency of the TWFE model, equation (1), for consistently estimating the ATTs. This is practically useful since difference-in-differences is simple to implement, so providing practical ways to test whether our more general model is necessary is valuable.

Our work contributes to an emerging literature on adjusting for parallel-trends violations in short panels. Freyaldenhoven et al. (2019) propose a similar instrumental variable type estimator in the presence of time-varying confounds. Their results rely importantly on homogeneous treatment effects and their simulations show that

4. Our imputation procedure also works in large panels because the identification results are independent of T.

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

heterogeneous treatment effects bias their estimates severly. Callaway and Karami (2022) allows for heterogeneous effects in short panels. They prove identification using a similar strategy of quasi-long-differencing and instrumental variables. They require time-invariant instruments whose effects on the outcome are constant over time. Their instruments would be valid in our estimator, but we allow for more general instruments including time-varying covariates.

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 ATTs when the outcomes are generated by a linear factor structure. Section 3 provides the main asymptotic normality result. We also study covariates with group-specific partial effects. Section 4 gives several specification tests for the underlying model. We demonstrate when TWFE is sufficient for consistency. We also derive a structural break test for the underlying common factor assumption. 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 — Theory

We assume a panel dataset with $i = 1, \ldots, N$ 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 $\{g_1, \ldots, g_G\} \equiv \mathcal{G} \subset \{2, \ldots, T\}$.[6] Following Callaway and Sant'Anna (2021), we denote $G_i = \infty$ for units that never receive treatment in the sample. Potential outcomes are a function of group-timing which we denote $y_{it}(g)$. For treatment indicators, we define the vector of treatment status $\boldsymbol{d}_i = (d_{i1}, ..., d_{iT})$ 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.

6. We do not allow units to start treatment in the first observed period because they have no untreated observations to use for imputation.

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 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 $\boldsymbol{X}$. Finally, we suppose there exist observed instruments $\boldsymbol{w}_i$ that will identify the factor space. We elaborate on these instruments in Section 2.2.

Assumption 1 (Sampling). The data $\{(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{d}_i, \boldsymbol{\gamma}_i, \mu_i, \boldsymbol{u}_i)\}$ 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). $\mathbb{E}\left[u_{it} \mid \boldsymbol{\gamma}_i, \mu_i, G_i\right] = 0$ for $t = 1, ..., T$.

It may seem that Assumption 4 is a stronger assumption than the standard parallel trends assumption. However, this assumption is more general 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 macroeconomic trends.

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 a given group g in a given period t would have expectation:

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

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. 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

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

These quantities represent the average effect of treatment for units that start treatment in period g when they are in period t. If there are too few units in a given group, then averaging the group-time ATTs is necessary for proper inference. For example, averaging over all post-treatment periods estimates an overall ATT (when using weights proportional to the number of units in (g,t)) and averaging over (i,t) where $t - G_i = \ell$ estimates event-study estimands ATT^{ℓ}.[8]

The key econometric challenge lies in that we do not observe $y_{it}(\infty)$ whenever

7. This assumption is seen in the imputation estimator proposed by Borusyak et al. (2021) for example. The following derivation is also shown in Callaway and Karami (2022), but we are repeating it here for expositional purposes.

8. Our theory also extends for other averages of τ_{gt} in the post period. See Callaway and Sant'Anna (2021) for more details.

$d_{it} = 1$. Our goal is to consistently estimate $\mathbb{E}[y_{it}(\infty) \mid G_i = g]$ under equation (3) to consistently estimate group-time average treatment effects. Borusyak et al. (2021) and Gardner (2021) implicitly rely on this insight in studying the two-way error model.

Intuitively, if we had a large number of pre-treatment periods (large T_0), we could separately estimate $\boldsymbol{\gamma}_i$, $\boldsymbol{f}_t$, and the fixed effects using untreated observations ($d_{it} = 0$) to produce an estimate for $\hat{y}_{it}(\infty) = \hat{\mu}_i + \hat{\lambda}_t + \hat{\boldsymbol{f}}_t' \hat{\boldsymbol{\gamma}}_i$. This strategy is studied by Gobillon and Magnac (2016) and Xu (2017). However, this technique requires the number of pre-periods T_0 to grow to infinity which is often undesirable in applied settings which typically have only a few pre-treatment periods[9].

Instead, we pursue identification noting that

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

Therefore, we only need to estimate the *average* of the fixed effects and factor-loadings among a treatment group. Instead of the typical requirement that we need a large number of pre-treatment periods, we require a large number of treated units.

2.1. ATT(g,t) Identification

Identification of average treatment effects will proceed in three steps. The first step is to manually remove the additive fixed effects. 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 $\hat{\tilde{y}}_{it}(\infty)$ to estimate treatment effects.

9. Ahn et al. (2001) show that least squares estimation of a single factor model is only consistent when the errors are white noise.

We first define the following averages:

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

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

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

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 only 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 of aggregating treatment effects (Goodman-Bacon 2021, Borusyak et al. 2021). Second, it preserves a common factor structure for all units and time periods[10]. This result is summarized in the following lemma:

Lemma 2.1. $\mathbb{E}\left[\tilde{y}_{it} \mid G_i = g\right] = \mathbb{E}\left[d_{it}\tau_{it} + (\boldsymbol{f}_t - \overline{\boldsymbol{f}}_{t<T_0})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_\infty) \mid G_i = g\right]$ 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 is important because it tells us that the factors and loadings retain a constant structure across the panel (both being only shifted by a constant). Any imputation method that wants to include

10. The TWFE imputation estimator of Gardner (2021) and Borusyak et al. (2021) would not share this property because they estimate μ_i and λ_t based on the full sample $d_{it} = 0$. Then the factor structure is different for the post-treated treatment group than the pre-treatment groups because the residualized factor model is different.

factor models should provide a similar result to Lemma 2.1. Otherwise, they may not be able to use untreated observations to remove the effects in the post-treatment periods. Since we are not interested in inference on the factors themselves, this form will suffice for the imputation process. Explicitly, the transformed outcomes take the form

$$\tilde{y}_{it} = d_{it}\tau_{it} + (\boldsymbol{f}_t - \overline{\boldsymbol{f}}_{\text{pre}})'(\boldsymbol{\gamma}_i - \overline{\boldsymbol{\gamma}}_\infty) + \tilde{u}_{it}.$$

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

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

where we define $\tilde{\boldsymbol{F}} = (\tilde{\boldsymbol{f}}_1, ..., \tilde{\boldsymbol{f}}_T)'$.

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 $g \in \mathcal{G}$.

This result is an immediate consequence of Assumptions 1 – 4 as $\mathbb{E}[\gamma_j \mid G_i = g] - \mathbb{E}[\gamma_i]$ 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-loading of each treatment group does not differ[11]. Asymptotic normality of our imputation procedure under a two-way error model is studied in Appendix Section A.

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* $\boldsymbol{P}(\boldsymbol{X}_1, \boldsymbol{X}_0) \equiv \boldsymbol{X}_1(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$. This matrix takes a similar form to a projection matrix but "imputes" on a different matrix $\boldsymbol{X}_1$ than the matrix used for estimating the regression coefficient $(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$. Gardner (2021) implicitly uses

11. This result echos Theorem 1 of Westerlund (2019)

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 $\tilde{\boldsymbol{F}}$ 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 $\boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g}$ with $g = G_i$.

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

$$\text{ATT}(g,t) = \mathbb{E}\left[\tilde{y}_{it} - \boldsymbol{P}(\tilde{\boldsymbol{f}}_t', \tilde{\boldsymbol{F}}_{t<g})\tilde{\boldsymbol{y}}_{i,t<g} \mid G_i = g\right] \tag{4}$$

for $t \geq g$

Theorem 2.1 shows that we can identify τ_{gt} if we know the factor imputation matrix. As we discussed previously, estimation of the factors and factor loadings are generally infeasible when T is small. However, our approach requires estimation of only the factors. We can estimate these consistently in fixed-T settings using the QLD approach of Ahn et al. (2013).

2.2. Factor 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 subsection 4.

It is well-known in the factor literature that neither the factors $\boldsymbol{f}_t$ nor the loadings $\boldsymbol{\gamma}_i$ are separately identifiable because both are unobserved. We, therefore, need to impose a normalization on the factors. The particular normalization does not affect our resulting imputation, so we follow Ahn et al. (2013) and use the normalization:

$$\tilde{F}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Theta} \\ -\boldsymbol{I}_p \end{pmatrix} \tag{5}$$

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

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{I}_{(T-p)} \\ \boldsymbol{\Theta}' \end{pmatrix}$$

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

Like Callaway and Karami (2022), we require instruments $\boldsymbol{w}_i$ 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 $\boldsymbol{\theta}$. We introduce three additional identifying assumptions:

Assumption 5 (Factor identification). The following rank assumptions for the untreated units, where $\boldsymbol{w}_i$ is a $L \times 1$ vector of instruments:

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

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

(iii) $\mathbb{E}\left[\boldsymbol{u}_i \mid \boldsymbol{w}_i, G_i = \infty\right] = \mathbf{0}$.

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 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.

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

$$\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\theta})'\tilde{\boldsymbol{y}}_i \otimes \boldsymbol{w}_i \mid G_i = \infty\right] = \mathbf{0} \tag{6}$$

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[12]. We treat p as known for the remainder of the paper.

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,

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

for $g \in \mathcal{G}$.

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

Combined with Theorem 2.1, Lemma 2.3 implies that the $\tau_{g,t}$s are identified under Assumptions 1-5. The following Section uses the moment conditions constructed in this section to consistently estimate the functionals of the treatment effects.

3 — Estimation and Inference

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

3.1. Asymptotic Normality

Equations (4) and (6) provide us the necessary moment conditions to estimate the ATTs. We collect them here in their unconditional form:

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

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

$$\vdots$$

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

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.[13] We need one final assumption to implement the asymptotically efficient GMM estimator:

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 $g \in \mathcal{G}$.

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

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 $\boldsymbol{\Delta} = \mathbb{E}\left[\boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau})\boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau})'\right]$ which is positive definite with probability approaching one by Assumptions 5 and 6. Then our GMM estimators of $(\boldsymbol{\theta}', \boldsymbol{\tau}')'$ solve

$$\min_{\boldsymbol{\theta}, \boldsymbol{\tau}} \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau}) \right)' \widehat{\boldsymbol{\Delta}}^{-1} \left(\sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta}, \boldsymbol{\tau}) \right) \tag{7}$$

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. Let $(\widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\tau}}')'$ solve equation (7). 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 the matrices $\boldsymbol{D}_g$ and $\boldsymbol{\Delta}_g$ are defined in the Appendix. 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. Estimation of $\boldsymbol{\tau}_g$ is not dependent on the first stage estimation of $\boldsymbol{\theta}$ when $\boldsymbol{D}_g = \mathbf{0}$. This typically occurs when the transformed factor loadings for group g center about zero. However, simpler fixed effects imputation suffices if this equality holds; see Corollary 2.1.

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:

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

This test can easily be carried 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 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.

3.2. Inference of Aggregate Treatment Effects

As in Callaway and Sant'Anna (2021), we can form aggregates of our group-time average treatment effects. For example, event-study type coefficients would average over the τ_{gt} where $t - g = e$ for some relative event-time e with weights proportional to group membership. Consider a general aggregate estimand δ which we define as a weighted average of $ATT(g,t)$:

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

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

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

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

This form comes from the fact that the weight matrix is positive definite with probability approaching one.[14]. The first term on the right-hand side is the influence function and hence inference on aggregate quantities follows directly. This result allows for uniform confidence bands on event-study estimates as recommended by Freyaldenhoven et al. (Forthcoming).

3.3. Plotting Estimates

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

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

where the first term on the right-hand side imputes $\hat{\tilde{y}}_{it}(\infty)$ 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, you can aggregate over a group and plot the average of y_{it} and the average

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

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.

3.4. Including Covariates

We now discuss the inclusion of covariates in the untreated potential outcomes:

$$y_{it}(\infty) = \boldsymbol{x}_{it}\boldsymbol{\beta} + \mu_i + \lambda_t + \boldsymbol{f}_t'\boldsymbol{\gamma}_i + u_{it} \tag{8}$$

where $\boldsymbol{x}_{it}$ is a $K \times 1$ vector of covariates. All covariates must vary over i and t if we hope to identify their coefficients. We can jointly estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ using the moments

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

Given a consistent estimator for $\boldsymbol{\beta}$, identification of $\boldsymbol{\tau}$ follows just as it did without covariates. For each group $g \in \mathcal{G}$, define $\boldsymbol{X}_{i,t<g}$ and $\boldsymbol{X}_{i,t\geq g}$ as the pre- and post-treatment covariates. The treatment effects are identified by imputing the full error $\boldsymbol{F}\boldsymbol{\gamma}_i + \boldsymbol{u}_i$. Imputation on treatment group g then follows

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

As the work above demonstrates, all procedures in this paper can be easily modified to incorporate covariates by simply imputing the residual $\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$.

We may believe the slopes are specific to the treatment timing. If we further add $\mathbb{E}\left[\boldsymbol{\tau}_i \mid \boldsymbol{G}_i = g, \boldsymbol{X}_i\right] = \mathbb{E}\left[\boldsymbol{\tau}_i \mid \boldsymbol{G}_i = g\right]$ along with $\mathbb{E}\left[\boldsymbol{u}_i \mid \boldsymbol{w}_i, G_i = g\right] = \boldsymbol{0}$, which is

common in practice, we have the following valid moments:

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

where there are $(T-p)K(T-g+1)$ moment conditions for $(T-g+1)+K$ parameters. It is then easy to test the hypothesis $\boldsymbol{\beta}_\infty = \boldsymbol{\beta}_{g_G} = ... = \boldsymbol{\beta}_{g_1}$ by Theorem 3.1. Allowing $\boldsymbol{\beta}_\infty \neq \boldsymbol{\beta}_g$ implies that the slopes can change after exposure to treatment. As G is fixed asymptotically, we place no restrictions on the slopes $\boldsymbol{\beta}_g$ and thus allow the change $\boldsymbol{\beta}_g - \boldsymbol{\beta}_\infty$ to be artbitrary.

4 — Specification Testing

This section provides tests for different aspects of the model's functional form. These tests can help the researcher determine if a factor model is appropriate for their application.

4.1. Sufficiency of TWFE

A novel insight of our paper concerns the ability to test for a factor structure, once the additive effects have been removed.[15] 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 QLD procedure is unnecessary and may lead to a less efficient estimate of τ_{it}. It is also computationally more difficult to implement

15. 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. While it may seem like the full imputation estimator is less efficient under the null, a direct efficiency comparison requires substantive assumptions on the error $\boldsymbol{u}_i$. The tests presented in this section require no assumptions on $\boldsymbol{u}_i$ beyond those needed for identification in Section 2. For example, removing the factors from the error could yield a spherical covariance, making the factor imputation estimator efficient. But if the factor structure centers around zero, OLS is still consistent.

than a standard fixed effects regression. 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.

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 $g \in \mathcal{G}$. Our imputation approach allows us to identify these terms up to a rotation. To see how, let $\boldsymbol{A}^*$ be the rotation that imposes the Ahn et al. (2013) normalization. Then

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

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

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

common factors. We note that

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

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

Collect the moments

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

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

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

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

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

4.2. Testing Equality of Factors

An important assumption underlying our approach is that the factors, which affect both the pre- and post-treatment outcomes, are equal between the treated and un-

treated groups. This assumption may not hold if, for example, the control and treatment groups are geographically or sociologically separated. We, therefore, derive tests for equivalence of the factors.

We can only compare the pre-treatment factors because those are the ones the treated groups can identify. Testing each group sequentially may give misleading results, especially when there are few units per group. Therefore, we combine all treated groups and only compare the first T_0 factor observations before any group is treated. We define $D_i = \sum_{g \in \mathcal{G}} D_{ig}$ which is one if the unit is ever treated.

We consider two estimators of the pre-sample factors, one using the untreated observations and one using the pre-treated observations. The rank condition on $\boldsymbol{F}$ in Assumption 3(i) means we can hope to identify the pre-treatment factors with the pre-treatment treated observations. We apply the Ahn et al. (2013) normalization to the pre-treatment factors, and define $\boldsymbol{H}^*(\boldsymbol{\theta})' = \left[\boldsymbol{I}_{(T_0-p)} \boldsymbol{\Theta}^*\right]$ where $\boldsymbol{\Theta}^*$ is $(T_0 - p) \times p$ matrix of free parameters.

Given the appropriate identifying assumptions on the treated units, the two sets of moments are then

$$\mathbb{E}\left[\boldsymbol{g}_i^0(\boldsymbol{\theta}_0)\right] = \mathbb{E}\left[\frac{(1-D_i)}{\mathbb{P}(D_i=0)} \boldsymbol{H}^*(\boldsymbol{\theta}_0)' \boldsymbol{y}_{i,t<T_0} \otimes \boldsymbol{w}_i\right] = \mathbf{0}_{T_0 \times 1}$$
$$\mathbb{E}\left[\boldsymbol{g}_i^1(\boldsymbol{\theta}_1)\right] = \mathbb{E}\left[\frac{D_i}{\mathbb{P}(D_i=1)} \boldsymbol{H}^*(\boldsymbol{\theta}_1)' \boldsymbol{y}_{i,t<T_0} \otimes \boldsymbol{w}_i\right] = \mathbf{0}_{T_0 \times 1}$$

which are the unconditional versions of the moments based on both respective subsamples, and $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are the vectorizations of the $(T_0 - p) \times p$ unrestricted parameters associated with the ALS normalization applied to $\boldsymbol{F}_{\text{pre}}$. We write the empirical analogs as $\boldsymbol{g}^1(\boldsymbol{\theta}_j) = \frac{1}{N_1} \sum_{i=1}^N D_1 \boldsymbol{g}_i^1(\boldsymbol{\theta}_1)$ and $\boldsymbol{g}^0(\boldsymbol{\theta}_0) = \frac{1}{N_0} \sum_{i=1}^N (1 - D_i) \boldsymbol{g}_i^0(\boldsymbol{\theta}_0)$ where N_0 and N_1 are the number of never-treated and treated individuals, respectively.

First, we must test whether the number of factors affecting both groups is the same. This can be achieved simply by estimating p separately using both subsamples. ALS provide tests for estimating p using their GMM estimator. Given that p is

the same for both sets of moment conditions, we are interested in testing the null hypothesis $H_0 : \boldsymbol{\theta}_0 - \boldsymbol{\theta}_1 = \mathbf{0}$. This condition suffices for testing the equality of the pretreatment factors for the untreated and pre-treated groups, which we denote $\boldsymbol{F}_{0,\text{pre}}$ and $\boldsymbol{F}_{1,\text{pre}}$ respectively. This fact holds because

$$\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1 \iff \boldsymbol{F}(\boldsymbol{\theta}_0) = \boldsymbol{F}(\boldsymbol{\theta}_1) \iff \boldsymbol{F}(\boldsymbol{\theta}_0)\boldsymbol{A} = \boldsymbol{F}(\boldsymbol{\theta}_1)\boldsymbol{A} \iff \boldsymbol{F}_0 = \boldsymbol{F}_1$$

where the second equivalence holds because the rotation matrix $\boldsymbol{A}$ is nonsingular, just as in Section 4.1.

We define the variance matrices as

$$\boldsymbol{S}_0(\boldsymbol{\theta}_0) = Var(\boldsymbol{g}_i^0)$$
$$\boldsymbol{S}_1(\boldsymbol{\theta}_1) = Var(\boldsymbol{g}_i^1)$$

with consistent estimators $\widehat{\boldsymbol{S}}_0$ and $\widehat{\boldsymbol{S}}_1$. Let

$$\boldsymbol{J}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{N_0}{N} g^0(\boldsymbol{\theta}_0)' \widehat{\boldsymbol{S}}_0^{-1} g^0(\boldsymbol{\theta}_0) + \frac{N_1}{N} g^1(\boldsymbol{\theta}_1)' \widehat{\boldsymbol{S}}_1^{-1} \boldsymbol{g}^1(\boldsymbol{\theta}_1) \tag{9}$$

Finally, define $\widehat{\boldsymbol{\theta}}$ as the estimator of $\boldsymbol{\theta}$ which uses both sets of moment conditions, and let $\widehat{\boldsymbol{\theta}}_0$, $\widehat{\boldsymbol{\theta}}_1$ be the estimators using the respective subsamples and their respective moment conditions.

Theorem 4.3. Suppose Assumption 3 holds conditional on D_i. Then under Assumptions 1-5 and the null hypothesis,

$$N * \left(J(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}) - J(\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1) \right) \xrightarrow{d} \chi^2_{((T_0 - p)p)} \tag{10}$$

as $N \to \infty$.

This result is a direct application of Theorem 5.8 from Hall (2004). He requires the partial sums $\sqrt{N}\boldsymbol{g}^0(\boldsymbol{\theta}_0)$ and $\sqrt{N}\boldsymbol{g}^1(\boldsymbol{\theta}_1)$ be uncorrelated, which holds under ran-

dom sampling. Further, we can replace N_0/N and N_1/N in equation (9) with their asymptotic counterparts $\mathbb{P}(D_i = 0)$ and $\mathbb{P}(D_i = 1)$ because they are multiplied by $O_p(N^{-1})$ terms. $\mathbb{P}(D_i = 0)$ takes the place of π in Hall (2004).

5 — Simulations

XXX

6 — Application

XXX

7 — Conclusions

We consider identification and inference of functions of treatment effects in a linear panel data model when the number of time periods is small. 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 treatment effect coefficients.

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

While we study the quasi-long-differencing transformation, any method for estimating the factor space suffices to estimate the ATTs. Other approaches, like common correlated effects or principal components, can also be implemented. Such a

range of techniques allow for flexibility in estimation for applied researchers. Further work can demonstrate both theoretical and finite-sample properties of these various estimators of the factors, and how they affect to ATT estimation.

A — Inference in Two-Way Fixed Effect Model

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

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

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

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

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

The imputation equation in Section 2.3 is a just-identified system of equations. As such, we do not need to worry about weighting in implementation and inference comes from standard theory of M-estimators. In fact, we have the following closed-form solution for the estimator of a group-time average treatment effect:

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

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

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

Theorem A.1. Assume untreated potential outcomes take the form of the two-way error model given in (1). Suppose Assumptions 1, 3(iii), and 4 hold with $\boldsymbol{\gamma}_i = 0$. Then for all (g, t) with $g > t$, $\widehat{\boldsymbol{\tau}}_{gt}$ is conditionally unbiased for $\mathbb{E}\left[\tau_{it} \mid D_{ig} = 1\right]$, has the linear

form

$$\sqrt{N_g}\big(\widehat{\tau}_{gt} - \tau_{gt}\big) = \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{ig}(\tau_{it} - \tau_{gt} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0}) \tag{11}$$

and

$$\sqrt{N_1}(\widehat{\tau}_{gt} - \tau_{gt}) \xrightarrow{d} N(0, V_1 + V_0) \tag{12}$$

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

Theorem (A.1) demonstrates the simplicity of our imputation procedure under the two-way error model. While the general factor structure requires more care, estimation and inference will yield a similar result.

Proof of Theorem A.1

The transformed post-treatment observations are

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

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

$$\mathbb{E}\left[\tau_{it} + u_{it} - \overline{u}_{i,t<T_0} - \overline{u}_{\infty,t} + \overline{u}_{\infty,t<T_0} \mid D_{ig} = 1\right] = \mathbb{E}\left[\tau_{it} \mid D_{ig} = 1\right]$$

by Assumption 3 and 4.

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

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

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

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

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

Since these terms are mean zero and uncorrelated (by Assumption 1), we find the variance of each term separately.

The first term has asymptotic variance

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

and the second term has asymptotic variance

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

The result follows from the independence of the two sums.

B — Common Correlated Effects

As noted in the main text, our identification results hold regardless of the particular estimator of the factor space. We briefly consider identification of the treatment effect coefficients under a common correlated effects scheme.

We suppose there exists a $1 \times K$ vector of covariates $\boldsymbol{x}_{it}$. We stack $\boldsymbol{x}_{it}$ over t to get the $T \times K$ matrix $\boldsymbol{X}_i$. Through what follows, we assume $\boldsymbol{X}_i$ is randomly sampled and has finite fourth moments. For simplicity, we consider only one treatment period which starts after T_0. membership into the treatment group is denoted by the dummy variable D_i.

B.1. Brown, Schmidt, and Wooldridge CCE

We assume $\boldsymbol{X}_i$ satisfies the model of Brown et al. (2022):

Assumption 7 (BSW Model). Let $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_i]$.

(i) There exists a $K \times p$ matrix $\boldsymbol{\Lambda}$ such that

$$\boldsymbol{F} = \boldsymbol{\mu}\boldsymbol{\Lambda} \tag{13}$$

(ii) $\text{Rank}(\boldsymbol{\mu}_{t \leq T_0}) = K < T_0$.

Assumption 7 is motivated by the common correlated effects literature. The usual CCE model assumes that $\text{Rank}(\mathbb{E}[\boldsymbol{\Gamma}_i]) = p \leq K$. When this is true, we can always rewrite

$$\boldsymbol{\mu} = \boldsymbol{F}\mathbb{E}[\boldsymbol{\Gamma}_i] \Rightarrow \boldsymbol{F} = \boldsymbol{\mu}\mathbb{E}[\boldsymbol{\Gamma}_i]' \left(\mathbb{E}[\boldsymbol{\Gamma}_i]\mathbb{E}[\boldsymbol{\Gamma}_i]'\right)^{-1}$$

and redefine $\boldsymbol{\Lambda} = \mathbb{E}[\boldsymbol{\Gamma}_i]' \left(\mathbb{E}[\boldsymbol{\Gamma}_i]\mathbb{E}[\boldsymbol{\Gamma}_i]'\right)^{-1}$. However, the BSW model puts no restrictions on the rank of $\boldsymbol{\Lambda}$, so it is not true that one can recover the CCE model from the BSW, making BSW strictly weaker.

In their analysis, Brown et al. (2022) assume $\text{Rank}(\boldsymbol{\mu}) = K$, which is not implied by CCE. In fact, assuming the factors are full rank, the CCE model implies $\text{Rank}(\boldsymbol{\mu}) = p \leq K$ which means that cross-sectional averages of $\boldsymbol{X}_i$ may converge to a reduced rank limit. We do not address this problem here and continue to assume that $\boldsymbol{X}_i$ has full rank in expectation, which is reasonable in microeconometrics. An analysis of the classical case in a fixed-T setting can be found in Westerlund et al. (2019).

It is important to note that the Ahn et al. (2013) factor-identifying assumptions do not overlap with the Brown et al. (2022) assumptions. The QLD model places no restrictions on the relationship between covariates and factors. Also, the QLD parameters are not identified under the BSW assumptions as Brown et al. (2022) place no restrictions on $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Lambda}$ is not necessarily invertible. Thus, neither of the two models imply the other, so their study in this context leads to different estimators which can be chosen by the researcher. However, we should note that classical CCE allows for identification of the QLD parameters as in Brown (2022).

We begin with the pure factor potential outcome model $y_{it}(\infty) = \boldsymbol{f}_t\boldsymbol{\gamma}_i + u_{it}$. Under the BSW model, we can rewrite $y_{it}(\infty) = \boldsymbol{\mu}_t\boldsymbol{\rho}_i + u_{it}$ where $\boldsymbol{\rho}_i = \boldsymbol{\Lambda}\boldsymbol{\gamma}_i$. Then we can identify the ATTs with

$$\mathbb{E}\left[\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}(\boldsymbol{\mu}_{t>T_0}, \boldsymbol{\mu}_{t\leq T_0})\boldsymbol{y}_{i,t\leq T_0} \mid D_i = 1\right] = \boldsymbol{0}$$

The moment conditions hold as

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{\mu}_{t>T_0}, \boldsymbol{\mu}_{t\leq T_0})\boldsymbol{y}_{i,t\leq T_0} &= \boldsymbol{P}(\boldsymbol{\mu}_{t>T_0}, \boldsymbol{\mu}_{t\leq T_0})(\boldsymbol{\mu}_{t\leq T_0}\boldsymbol{\rho}_i + \boldsymbol{u}_{i,t\leq T_0}) \\ &= \boldsymbol{\mu}_{t>T_0}\boldsymbol{\rho}_i + \boldsymbol{P}(\boldsymbol{\mu}_{t>T_0}, \boldsymbol{\mu}_{t\leq T_0})\boldsymbol{u}_{i,t\leq T_0}\end{aligned}$$

which, in expectations conditional on $D_i = 1$, is equal to $\boldsymbol{y}_{i,t>T_0}(\infty)$.

The obvious estimator of $\boldsymbol{\tau}$ is

$$\frac{1}{N_1}\sum_{i=1}^{N} D_i\left(\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}(\overline{\boldsymbol{X}}_{t>T_0}, \overline{\boldsymbol{X}}_{t\leq T_0})\boldsymbol{y}_{i,t\leq T_0}\right)$$

which we denote $\widehat{\boldsymbol{\tau}}_{CCE}$ We first prove consistency of the above estimator.

Theorem B.1. Under Assumptions 1-4 and Assumption 6, $\widehat{\boldsymbol{\tau}}_{CCE} \xrightarrow{p} \boldsymbol{\tau}$.

Proof.

$$\begin{aligned}\widehat{\boldsymbol{\tau}}_{CCE} &= \frac{N}{N_1}\frac{1}{N}\left(\sum_{i=1}^{N} D_i\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}(\overline{\boldsymbol{X}}_{t>T_0}, \overline{\boldsymbol{X}}_{t\leq T_0}))\sum_{i=1}^{N} D_i\boldsymbol{y}_{i,t\leq T_0}\right) \\ &\xrightarrow{p} \mathbb{E}\left[\boldsymbol{y}_{i,t>T_0} \mid D_i = 1\right] - \boldsymbol{\mu}_{t>T_0}(\boldsymbol{\mu}'_{t\leq T_0}\boldsymbol{\mu}_{t\leq T_0})^{-1}\boldsymbol{\mu}_{t\leq T_0}\mathbb{E}\left[\boldsymbol{\mu}_{t\leq T_0}\boldsymbol{\rho}_i \mid D_i = 1\right] \\ &= \mathbb{E}\left[\boldsymbol{y}_{i,t>T_0} - \boldsymbol{\mu}_{t>T_0}\boldsymbol{\rho}_i \mid D_i = 1\right] \\ &= \mathbb{E}\left[\boldsymbol{y}_{i,t>T_0}(1) - \boldsymbol{F}\boldsymbol{\gamma}_i \mid D_i = 1\right] \\ &= \mathbb{E}\left[\boldsymbol{\tau}_i \mid D_i = 1\right] \\ &= \boldsymbol{\tau}\end{aligned}$$

as $\boldsymbol{F}\boldsymbol{\gamma}_i = \boldsymbol{\mu}_x\boldsymbol{\Lambda}\boldsymbol{\gamma}_i \equiv \boldsymbol{\mu}_x\boldsymbol{\rho}_i$ by BSW and $\mathbb{E}\left[\boldsymbol{y}_{i,\text{post}}(\infty) \mid D_i = 1\right] = \mathbb{E}\left[\boldsymbol{F}\boldsymbol{\gamma}_i \mid D_i = 1\right]$ by

Assumption 4. □

Asymptotic normality is more complicated, because we may have to account for the effect of the factor proxies in the imputation matrix. For notational convenience, we write $\widehat{\boldsymbol{P}} = \boldsymbol{P}(\overline{\boldsymbol{X}}_{t>T_0}, \overline{\boldsymbol{X}}_{t\leq T_0})$.[16] First write everything in terms of the unconditional moments:

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

Let $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{\mu}_{t>T_0}, \boldsymbol{\mu}_{t\leq T_0})$. Add and subtract $\boldsymbol{P}\boldsymbol{y}_{i,t\leq T_0}$ within the sum to get

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\tau}}_{CCE} - \boldsymbol{\tau}) &= \frac{N}{N_1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\left((\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}\boldsymbol{y}_{i,t\leq T_0} - \boldsymbol{\tau}) + (\boldsymbol{P} - \widehat{\boldsymbol{P}})\boldsymbol{y}_{i,t\leq T_0}\right)\right) \\ &= \frac{N}{N_1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\left(\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}\boldsymbol{y}_{i,t\leq T_0} - \boldsymbol{\tau}\right)\right) + \sqrt{N}(\boldsymbol{P} - \widehat{\boldsymbol{P}})\frac{N}{N_1}\left(\frac{1}{N}\sum_{i=1}^{N} D_i\boldsymbol{y}_{i,t\leq T_0}\right)\end{aligned}$$

The first sum is comprised of mean zero, iid terms:

$$\begin{aligned}\mathbb{E}\left[D_i(\boldsymbol{y}_{i,t>T_0} - \boldsymbol{P}\boldsymbol{y}_{i,t\leq T_0} - \boldsymbol{\tau})\right] &= \mathbb{E}\left[\mu_{t>T_0}\boldsymbol{\rho}_i + \boldsymbol{u}_{i,t>T_0} + \boldsymbol{\tau}_i - \boldsymbol{\mu}_{t>T_0}\boldsymbol{\rho}_i - \boldsymbol{P}\boldsymbol{u}_{i,t\leq T_0} - \boldsymbol{\tau} \mid D_i = 1\right]\mathbb{P}(D_i = 1) \\ &= \mathbb{E}\left[\boldsymbol{\tau}_i - \boldsymbol{\tau} \mid D_i = 1\right]\mathbb{P}(D_i = 1) \\ &= \mathbf{0}\end{aligned}$$

as $\boldsymbol{\tau} \equiv \mathbb{E}\left[\boldsymbol{\tau}_i \mid D_i = 1\right] = \mathbb{E}\left[\boldsymbol{y}_{i,t>T_0}(1) - \boldsymbol{y}_{i,t>T_0}(\infty) \mid D_i = 1\right]$. Thus the first sum is asymptotically normal.

The second sum is easier to deal with theoretically. $\sqrt{N}(\text{vec}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}))$ is asymptotically normal by the CLT and Assumption 1. Then applying the continuous mapping theorem twice, $\sqrt{N}(\widehat{\boldsymbol{P}} - \boldsymbol{P})\frac{N}{N_1}\left(\frac{1}{N}\sum_{i=1}^{N} D_i\boldsymbol{y}_{i,t\leq T_0}\right)$ is asymptotically normal.

16. Authors in the CCE literature will often write $\widehat{\boldsymbol{F}} = \overline{\boldsymbol{X}}$ as shorthand, because the cross-sectional averages are proxies for the factors.

C — 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 $\overline{f}_{t<T_0}$ 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), $y_{it} = d_{it}y_{it}(1) + (1-d_{it})y_{it}(\infty) = d_{it}\tau + y_{it}(\infty)$. Then

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

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

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

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 Lemma 2.3

Let $\boldsymbol{A}^*$ be the $p \times p$ rotation that generates the Ahn et al. (2013) normalization for $\tilde{\boldsymbol{F}}$. Note that both inverses in the permutation matrix definition exist for every g because Rank$(\tilde{\boldsymbol{F}}(\boldsymbol{\theta}))$ = Rank$(\tilde{\boldsymbol{F}})$. Since

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

we have

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

□

Proof of Theorem 3.1

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

The only work we need to do is 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.

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

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:

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

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

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

The rest of the first column of matrices takes the form

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

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

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

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

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

□

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.

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.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2021. "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.

Brown, Nicholas L., Peter Schmidt, and Jeffrey M. Wooldridge. 2022. "Simple Alternatives to the Common Correlated Effects Model." 10.13140/RG.2.2.12655.76969/1.

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.

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.

Hall, Alastair R. 2004. *Generalized Method of Moments*. OUP Oxford.

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*.

Westerlund, Joakim. 2019. "On Estimation and Inference in Heterogeneous Panel Regressions with Interactive Effects." *Journal of Time Series Analysis* 40 (5): 852–857.

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.