Difference-in-Differences via Common Correlated Effects*

Nicholas Brown
Queen's University

Kyle Butts
University of Colorado Boulder

Joakim Westerlund†
Lund University
and
Deakin University

January 21, 2023

Abstract

We study the effect of treatment on an outcome when parallel trends hold conditional on an interactive fixed effects structure. In contrast to the majority of the literature, we propose identification using time-varying covariates. We assume the untreated outcomes and covariates follow a common correlated effects (CCE) model, where the covariates are linear in the same common time effects. We then demonstrate consistent estimation of the treatment effect coefficients by imputing the untreated potential outcomes in post-treatment time periods. Our method accounts for treatment affecting the distribution of the control variables and is valid when the number of pre-treatment time periods is small. We also decompose the overall treatment effect into estimable direct and mediated components.

JEL Classification: C31, C33, C38.

Keywords: Difference-in-differences, interactive fixed effects, fixed-T, imputation.

*Westerlund would like to thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

†Corresponding author: Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

1 Introduction

Econometric analysis of treatment effects often relies on the so-called "parallel trends" assumption. This assumption generally states that the change in the counterfactual untreated outcomes is equal to a time-varying effect across units, regardless of treatment status. Recent work in the evaluation of treatment effects has sought to relax this condition by allowing units in the population to select into treatment based on an interactive fixed effects model. Letting $y_{i,t}(\infty)$ be the untreated potential outcome for unit i at time t, the new parallel trends assumptions take the form

$$y_{i,t}(\infty) = \sum_{j=1}^{r} f_{t,j}\gamma_{j,i} + \epsilon_{i,t}, \tag{1}$$

where $\epsilon_{i,t}$ has mean zero and $\sum_{j=1}^{r} f_{t,j}\gamma_{j,i}$ is unobserved.

The interactive effects structure in equation (1) is often called a "factor model", where the common time-effects $f_{t,r}$ are called factors and the heterogeneous unit-effects $\gamma_{r,i}$ are factor loadings. This model relaxes "parallel trends" by allowing the impact of time-effects to vary across individuals in a way that can be correlated with treatment. The usual two-way fixed effects (TWFE) estimator, which estimates fixed effects at the unit and time levels, is generally insufficient for estimating treatment effects when the untreated potential outcomes take the form of a factor model. As such, a new literature has focused on specifically controlling for such error structures while estimating treatment effect functions.

Chan and Kwok (2022) propose a class of principal components difference-in-differences (PCDID) estimators. They estimate the factors using least squares on the never-treated sample as in Bai (2009) then use these estimates as proxies in the treated sample. However, their asymptotic theory requires the number of time periods to grow to infinity. Such asymptotic sequences are problematic. First, they put restrictions on the time dynamics of the treatment effects. Second, such a setting is unrealistic in many applications where the number of pre- and post-treatment time periods is small. Callaway and Karami (2020) and Brown and Butts (2022) provide consistent estimators of treatment effect functions when the number of time periods is small. The former requires a specific set of time-invariant instruments. The latter relaxes this

requirement, but suggests an overidentified GMM problem which is computationally burdensome.

We provide an imputation-based estimator that is both consistent for fixed time periods and simple to implement. We utilize the popular common correlated effects (CCE) scheme first studied in Pesaran (2006). He proposed a regression-based estimator for eliminating unobserved factors in linear panel data models. Assuming a rich set of covariates that are linear in the same factors, he takes the cross-sectional averages of the covariates and the outcome, then treats them as fixed effects in a pooled regression. This pooled CCE (CCEP) estimator has been shown to be asymptotically normal when the number of time periods is fixed (Westerlund et al., 2019; Brown et al., 2022). The CCE model also allows a flexible method for parallel trends conditional on both interactive effects and covariates.

We propose the CCE difference-in-differences (CCEDID) estimator. First, we construct proxies of the common factors using cross-sectional averages of the outcome and independent variables from the never-treated sample. Second, we estimate the partial effects of the covariates along with the heterogeneous factor loadings. While inconsistent for the loadings because the number of time periods is fixed, our estimator proxies for the average value of the loadings conditional on treatment status, which Brown and Butts (2022) show is sufficient. Finally, we use the estimates of the covariate coefficients, factors, and factor loadings to estimate the counterfactual untreated potential outcomes in the post-treatment periods, then average over their difference from the observed treated outcomes. We prove that this estimator of the ATTs is consistent and asymptotically normal.

Our estimator allows the covariates to change with treatment status using the reduced form CCE model for their untreated status. We also study how treatment status impacts the final treatment effect estimators through the status of the covariates. By assuming a common factor model for the untreated covariates, we are able to leave the post-treatment observables completely arbitrary. We can also decompose the effect of the treatment on the outcome in terms of its direct level effect and its mediated effect through the covariates. While Chan and Kwok (2022) leave the relationship between the covariates and factors unrestricted, they assume treatment does not affect the distribution of the covariates. We provide a detailed explanation for

why this assumption can cause misleading estimates. We also provide a robust CCE-based estimator of the same effect that Chan and Kwok (2022) estimate under weaker assumptions.

The rest of the paper is divided into the following sections: Section 2 presents the model and estimable quantities of interest. Section 3 defines the CCEDID estimator and gives conditions for asymptotic normality in short panels. Section 4 demonstrates how to separately and consistently estimate the direct and mediated treatment effects. Section 5 provides a brief Monte Carlo experiment. Section 6 presents an empirical application of our estimator. Section 7 gives some concluding remarks.

2 Treatment model

We are interested in estimating the effect of a particular treatment on some outcome variable $y_{i,t}$, observable for $i = 1, ..., N$ cross-sectional units and $t = 1, ..., T$ time periods. We allow for the possibility that the N units can be divided into groups within which treatment timing is the same. We follow Callaway and Sant'Anna (2020) in defining a treatment group by the time period in which they enter treatment. There are G groups defined by $\{g_1, ..., g_G\} \subset \{2, ..., T\}$ where we assume without loss of generality that $g_1 < g_2 < ... < g_G$. A unit that is never treated is a member of group ∞.

Let G_i be a random variable with support $\{g_1, ..., g_G, \infty\}$ denoting group membership and define $D_{i,g}$ as the dummy variable that is 1 if i is in group g. Then we define $N_g = \sum_{i=1}^{N} D_{i,g}$ as the number of units in group g. The dummy variable $d_{i,t}$ is one if $t \geq G_i$ and zero otherwise. It is convenient to also define $T_0 = g_1 - 1$ as the last period before the first treatment takes place.

Following the previous literature, we denote by $y_{i,t}(g)$ as the "potential" outcome of unit $i \in \mathcal{I}_g$ in period t subject to treatment at period g. The term $y_{i,t}(\infty)$ is the potential outcome when the unit is never subject to treatment. In our paper, $y_{i,t}(\infty)$ is given by

$$y_{i,t}(\infty) = \boldsymbol{\beta}_i' \mathbf{x}_{i,t}(\infty) + \boldsymbol{\alpha}_i' \mathbf{f}_t + \varepsilon_{i,t}, \tag{2}$$

where $\mathbf{x}_{i,t}(\infty)$ is a $m \times 1$ vector of observable regressors associated with the untreated potential outcomes, $\boldsymbol{\beta}_i$ is a $m \times 1$ vector of slope coefficients, $\mathbf{f}_t$ is a $r \times 1$ vector of unobservable common factors, $\boldsymbol{\alpha}_i$ is a $r \times 1$ vector of factor loadings, and $\varepsilon_{i,t}$ is an idiosyncratic error term.

The interactive effects are given here by $\boldsymbol{\alpha}_i'\mathbf{f}_t$. The purpose of these is to capture unobserved differences between treated and untreated units in absence of treatment, often referred to as "non-parallel trending". In this terminology, the factors represent common trends and the loadings measure the extent to which the effect of these trends are equal (or "parallel") across units. We are not interested in inference on these effects.[1] Accurate estimation of $\boldsymbol{\alpha}_i$ is therefore not needed.

Most applications of DID estimators use time-constant covariates to motivate a more plausible *conditional*-parallel trends assumption. When covariates change over time, the validity of DID depends on whether or not treatment status effects the distribution of the covariates (Caetano et al., 2022). For example, if we are estimating the effect of a place-based policy on employment, parallel trends might only be plausible when conditioning on poverty rates since treatment is targeted for areas with increasing poverty rates. However, the goal of place-based policies is to indirectly improve poverty rates so that $x_{i,t}(g) \neq x_{i,t}(\infty)$. Then controlling linearly in the post-periods for the observed poverty rate $x_{i,t}$ 'absorbs' some of the treatment effect.[2] This indirect mechanism of treatment creates a dilemma where controlling for $x_{i,t}$ induces 'post-treatment bias' and not controlling for $x_{i,t}$ introduces 'omitted variables bias' (Aklin and Bayer, 2017). We discuss this problem in more details in subsection 4.2. Following Caetano et al. (2022), we propose to solve this dilemma by imputing and controlling for untreated potential value for the covariates, $x_{i,t}(\infty)$. Our paper is the first to consider this solution in a factor model setting.

Because CCE estimation generally requires covariates that change over i and t, we need to impose structure on their distribution with respect to treatment. We assume there are m time- and individual-varying covariates that admit a pure factor structure when not subject to treatment. These covariates are modeled

$$\mathbf{x}_{i,t}(\infty) = \boldsymbol{\lambda}_i'\mathbf{f}_t + \mathbf{v}_{i,t}, \tag{3}$$

where $\boldsymbol{\lambda}_i$ is a $r \times m$ matrix of factor loadings and $\boldsymbol{v}_{i,t}$ is a $m \times 1$ vector of idiosyncratic errors. The condition that $\mathbf{x}_{i,t}$ should load on the same factors as $y_{i,t}$ is actually quite natural given that

[1] One reason for this is that $\boldsymbol{\alpha}_i$ and $\mathbf{f}_t$ are not separately identifiable.
[2] This is what Angrist and Pischke (2009) call a 'bad control'.

the main reason for considering interactive effects is their likely correlation with the regressors, and the detrimental effect of this on estimation and inference. Hence, if $\mathbf{x}_{i,t}$ does not load on $\mathbf{f}_t$ the parameters $\boldsymbol{\beta}$ can be estimated via OLS as in Wooldridge (2005).

We assume the untreated potential covariates are generated according to equation (3). Under this setting, we allow treatment to affect the covariate value by proposing the treated potential covariates to be

$$x_{i,t}(g) = \boldsymbol{\tau}_{i,g,t} d_{i,t} + x_{i,t}(\infty) = \boldsymbol{\tau}_{i,g,t} d_{i,t} + \boldsymbol{\lambda}_i' \mathbf{f}_t + \mathbf{v}_{i,t}, \tag{4}$$

where $\boldsymbol{\tau}_{i,g,t} d_{i,t}$ can correlate arbitrarily with the outcome treatment effects. Our imputation procedure for observations with $d_{i,t} = 1$ needs to account for the fact that observed $x_{i,t}$ does not necessarily equal $x_{i,t}(\infty)$. In contrast to equation (3), the treated covariates actually place no restrictions on the distribution of $\mathbf{x}_{i,t}$ because we do not restrict the distribution of $\boldsymbol{\tau}_{i,g,t}$.

CCE estimation takes cross-sectional averages of the outcome and covariates to proxy for the space spanned by the factors. We tailor this intuition to the treatment effect case where treatment status can affect the distribution of both the outcomes and covariates in unspecified ways. Thus, to prevent 'post-treatment bias', we use only the never-treated outcomes to proxy for the factors. Then for the treated group, we impute the never-treated potential covariates which in turn are used to impute the never-treated potential outcomes. This method is detailed in the following section.

Unlike $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$ is often of some interest. However, since in the present paper T is fixed, we cannot estimate each individual slope accurately. The best that we can hope for is accurate estimation of $\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{\beta}_i)$. In fact, in many applications in economics (and elsewhere) we are not particularly interested in the marginal effect for a particular unit anyway and so we focus instead on the average marginal effect. Partial effects are random over individuals but assumed unaffected by treatment status. We could impose parsimony on the model by assuming group-specific slopes $\boldsymbol{\beta}_g$ for each treated group g. These effects could be estimated for each treatment group then aggregated to get an overall partial effect estimate.

Remark 1. The presence of $\boldsymbol{\beta}_i' \mathbf{x}_{i,t}$ in (2) is an allowance and not a requirement. If there are no regressors, we define $\boldsymbol{\beta}_i' \mathbf{x}_{i,t} = 0$. It is important to note, though, that if there are no regressors,

the number of factors can be at most one unless there are outside factor proxies ($r \leq 1$), as will be made clear in Section 3. ■

We now introduce this paper's object of interest. The treatment effect for unit i at time t treated in time g is given simply by

$$\Delta_{i,g,t} = y_{i,t}(g) - y_{i,t}(\infty). \tag{5}$$

Because we do not observe $y_{i,t}(g)$ and $y_{i,t}(\infty)$ simultaneously, $\Delta_{i,g,t}$ must be treated as unknown and estimated from the data. This brings us back to the discussion in the previous paragraph about $\boldsymbol{\beta}_i$; because T is fixed, the best that we can do is hope for is accurate estimation of the ATT. The particular ATT that we are interested in is the average $\Delta_{i,g,t}$ for group g;

$$\mathbb{E}(\Delta_{i,g,t}|G_i = g) = \Delta_{g,t} \tag{6}$$

for $t \geq g$ and $g \in \{g_1, ..., g_G\}$. Note that while there cannot be any systematic variation across units within groups, we do allow $\Delta_{g,t}$ to vary freely over time and across groups, which means that the effect of the treatment need not take place abruptly at time g but can be gradual in nature. The effect cannot take place prior to treatment, though, which is the so-called "no anticipation" condition. Formally, we require that $y_{i,t} = y_{i,t}(0)$ for all non-treated ($g = \infty$) and not-yet-treated ($G_i = g$ and $t < g$) observations[3].

It is reasonable to allow for the covariates to also enter the treated outcomes:

$$y_{i,t}(g) = \eta_{i,g,t} d_{i,t} + \mathbf{x}_{i,t}(g)'\boldsymbol{\beta}_i + \mathbf{f}_t'\boldsymbol{\gamma}_i + \epsilon_{i,t}, \tag{7}$$

where $\boldsymbol{\gamma}_i$ is unaffected by treatment because it is time-invariant and $\eta_{i,g,t}$ is a unit-time specific intercept that appears with treatment status and defines the error to be $\epsilon_{i,t}$. Under equations (2), (4), (5), and (7), we can decompose the overall treatment effect $\Delta_{i,g,t}$ as

$$\begin{aligned}
\Delta_{i,g,t} &= y_{i,t}(g) - y_{i,t}(\infty) \\
&= \eta_{i,g,t} + (\mathbf{x}_{i,t}(g) - \mathbf{x}_{i,t}(\infty))'\boldsymbol{\beta}_i + \epsilon_{i,t} \\
&= \eta_{i,g,t} + \boldsymbol{\tau}_{i,g,t}'\boldsymbol{\beta}_i + \epsilon_{i,t}
\end{aligned}$$

[3]This condition can be relaxed by redefining the treatment time to $p < g$, so long as there are enough pre-treatment time periods to run the CCE regression.

where we define

$$\boldsymbol{\tau}_{g,t} = \mathbb{E}(\boldsymbol{\tau}_{i,g,t}|G_i = g) \tag{8}$$

$$\eta_{g,t} = \mathbb{E}(\eta_{i,g,t}|G_i = g) \tag{9}$$

as the group-specific dynamic ATTs for the covariates and direct effect, respectively. These quantities are useful to define because we later demonstrate how to estimate them individually. Our proposed average of the difference between the treated and untreated potential outcome ($\Delta_{i,t}$) estimates the joint effect while allowing covariates to change with treatment, which we accomplish by imputing the untreated potential covariates in post-treatment periods.

We follow the mediation analysis literature by labeling the treatment-specific intercept $\eta_{i,g,t}$ as the direct effect of treatment and $\boldsymbol{\tau}'_{i,g,t}\boldsymbol{\beta}_i$ as the mediated effect of treatment through the covariates (sometimes called the indirect effect).[4] These two effects warrant some discussion. The definition of $\boldsymbol{\tau}_{i,g,t}$ in equation (4) allows each individual covariate to have its own effect on the outcome through $\beta_{i,k}$. Thus, mediated effects allow researchers to tie the effect of treatment back to changes in specific covariates. This decomposition is particularly useful for explaining the mechanism through which treatment affects the outcome (a significant estimated mediation effect for the k-th covariate).

Applied researchers will typically apply DID to time-varying covariates that they suspect to be the underlying mechanism of treatment. Significant estimates of treatment effects on $x_{k,i,t}$ (τ in our notation) are used as evidence that treatment effects partially operate through the specified channels. However, even a change in the covariates that is systemically related to treatment might not translate to the outcome if the partial effect is zero. We allow a more precise decomposition of overall effects into component parts, providing the means for rigorous testing of different channels. The decomposition also allows for various 'countervailing' effects to be considered. On the other hand, the direct effect, $\eta_{g,t}$, captures group-specific changes to the average level of the outcomes post-treatment. We think of this as either the treatment directly effecting outcomes or there being unobserved pathways that are uncorrelated with the (observed) intermediate causal mechanisms and unrelated to the heterogeneous factor load-

[4] See, for example, MacKinnon et al. (2007) and Huber (2014).

ings.

The hypothetical outcomes, $y_{i,t}(\infty)$ and $y_{i,t}(g)$, should not be confused with the actual outcome, $y_{i,t}$, the model of which can be inferred from (2) and (5). Note that for units eventually treated at time g,

$$y_{i,t} = y_{i,t}(\infty)(1 - d_{i,t}) + y_{i,t}(g)d_{i,t} = \Delta_{i,g,t}d_{i,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t}. \tag{10}$$

This model holds in general for all $i \leq N$ and $t \leq T$. Another point of interest concerns the covariates. Only the untreated potential covariates show up in the equation for a general outcome. This fact demonstrates that we need to be sure our model for $\mathbf{x}_{i,t}(\infty)$ is sufficient in the following imputation method.

3 The CCEDID estimator and its asymptotic properties

3.1 The estimator

The estimation of the ATT is carried out using a version of what Borusyak et al. (2021) refer to as the "imputation" approach, or what Xu (2017) refer to as the "generalized synthetic control" method, which is based on replacing all unknowns in the definition of $\Delta_{g,t}$ in (6) by estimates. Note first that since $y_{i,t}(g)$ is observed for treated units in post-treatment periods ($G_i = g$ and $t \geq g$), we have $y_{i,t} = y_{i,t}(g)$. Let us therefore turn to the $y_{i,t}(\infty)$. We need to estimate this counterfactual for all treated units in post-treatment periods. This is done in four steps:

Four-step estimation of counterfactual:

1.Compute

$$\widehat{\mathbf{f}}_t = \frac{1}{N_\infty}\sum_{i=1}^{N}\mathbf{z}_{i,t}D_{i,\infty} \tag{11}$$

for all $t \leq T$, where $z_{i,t} = [y_{i,t}, \mathbf{x}_{i,t}']'$ is a $(m+1) \times 1$ vector containing all the observables. The above is the regular CCE estimator of (the space spanned by) $\mathbf{f}_t$ computed using the never-treated units only. This is crucial since in the present paper both $y_{i,t}$ and $\mathbf{x}_{i,t}$ may depend on the treatment, and this in turn may well render CCE inconsistent. Equally important is the fact that $\widehat{\mathbf{f}}_t$ is computed for all time periods. The pre-treatment estimates

are used to estimate $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}_{i=1}^N$, while the post-treatment estimates are used to impute $\mathbf{x}_{i,t}(\infty)$ and $y_{i,t}(\infty)$ respectively for the treatment period.

2.Estimate

$$y_{i,t} = \boldsymbol{\beta}'\mathbf{x}_{i,t} + \mathbf{a}_i'\widehat{\mathbf{f}}_t + u_{i,t} \tag{12}$$

by OLS for all $i \leq N$ and $t \leq T_0$. Here, $\mathbf{a}_i$ is a $(m+1) \times 1$ vector of factor loadings and $u_{i,t} = \boldsymbol{\alpha}_i'\mathbf{f}_t - \mathbf{a}_i'\widehat{\mathbf{f}}_t + (\boldsymbol{\beta}_i - \boldsymbol{\beta})'\mathbf{x}_{i,t} + \varepsilon_{i,t}$ is a composite error term. The above OLS regression with $\widehat{\mathbf{f}}_t$ in place of $\mathbf{f}_t$ is regular CCE based on the full pretreatment sample but where $\widehat{\mathbf{f}}_t$ comes from the subsample of untreated units.[5] Define the $T_0 \times 1$ vector $\mathbf{y}_i = [y_{i,1}, ..., y_{i,T_0}]'$, and the $T_0 \times m$ matrices $\mathbf{x}_i = [\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,T_0}]'$ and $\widehat{\mathbf{f}} = [\widehat{\mathbf{f}}_1, ..., \widehat{\mathbf{f}}_{T_0}]'$. Let $\mathbf{M}_\mathbf{A} = \mathbf{I}_{T_0} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ for any T_0-rowed matrix $\mathbf{A}$. In this notation, the CCE estimators of $\boldsymbol{\beta}$ and $\mathbf{a}_i$ in (12) are given by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^N \mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{x}_i\right)^{-1} \sum_{i=1}^N \mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{y}_i, \tag{13}$$

$$\widehat{\mathbf{a}}_i = (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}), \tag{14}$$

where the latter estimator is computed for all $i \leq N$. This gives $\widehat{\boldsymbol{\beta}}$ and $\{\widehat{\mathbf{a}}_i\}_{i=1}^N$. The fact that $\widehat{\mathbf{a}}_i$ is computed for treated units is again important, because in step 3, $y_{i,t}(\infty)$ will be estimated for all treated units.

3.Estimate $\mathbf{x}_{i,t}(\infty)$ as

$$\widehat{\mathbf{x}}_{i,t}(\infty) = \widehat{\boldsymbol{\lambda}}_i'\widehat{\mathbf{f}}_t \tag{15}$$

for all i in group g and $t \geq g$. Here $\{\widehat{\mathbf{f}}_t\}_{t \geq T_g}$ is from step 1. For $\widehat{\boldsymbol{\lambda}}_i$, in the notation used in step 2 above,

$$\widehat{\boldsymbol{\lambda}}_i = (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'\mathbf{x}_i \tag{16}$$

is the OLS estimator of $\boldsymbol{\lambda}_i$ in the following model for all $i \leq N$ and $t \leq T_0$:

$$\mathbf{x}_{i,t} = \boldsymbol{\lambda}_i'\widehat{\mathbf{f}}_t + \mathbf{w}_{i,t}. \tag{17}$$

[5]Note that unlike when using the principal components method, in CCE there is no need to recompute $\widehat{\mathbf{f}}_t$ if the time period changes, and hence $\{\widehat{\mathbf{f}}_t\}_{t \geq T_g}$ can be taken directly from step 1.

4.The sought counterfactual estimator is given by

$$\widehat{y}_{i,t}(\infty) = \widehat{\boldsymbol{\beta}}'\widehat{\mathbf{x}}_{i,t}(\infty) + \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t \tag{18}$$

which is again available for all $i \in \mathcal{I}_G^c$ and $t \geq T_g$ with $g < G$. Here $\widehat{\boldsymbol{\beta}}$ and $\{\widehat{\mathbf{a}}_i\}_{i\in\mathcal{I}_G^c}$ are from step 2, $\{\widehat{\mathbf{f}}_t\}_{t\geq T_g}$ is from step 1, and $\{\widehat{\mathbf{x}}_{i,t}\}_{t\geq T_g}$ comes from step 3. ■

Remark 2. Because we implement the estimators $\widehat{\mathbf{f}}$ and $\widehat{\boldsymbol{\beta}}$ using only the pre-treatment observations, and the cross-sectional averages of $\mathbf{x}_i$ are used as factor proxies, our estimator of the average of $\boldsymbol{\beta}_i$ is invariant to the inclusion of common variables (e.g. covariates that only change over i). This fact of CCE estimation was first pointed out in Brown et al. (2022). Secular time effects can thus be captured in our definition of $\eta_{g,t}$ for their overall effect on the outcome. ■

Remark 3. One can allow $\boldsymbol{\beta}$ to vary systematically across groups without affecting the asymptotic results reported in Section 3.2. The only change needed is that the step-2 estimation of this coefficient has to be carried out group-wise, as opposed to just once for all N units. This gives $\{\widehat{\boldsymbol{\beta}}_g\}_{g=1}^G$, which should then be inserted instead of $\widehat{\boldsymbol{\beta}}$ in step 3. Because these parameters are all consistently estimated for fixed-T, this setting effectively allows more slope heterogeneity than the pooled CCE estimator. ■

Remark 4. While $\widehat{\boldsymbol{\beta}}$ is consistent, $\widehat{\mathbf{a}}_i$ is not and in fact remains random even asymptotically because T is fixed. Moreover, the asymptotic distribution is not centered at $\boldsymbol{\alpha}_i$ but at a certain rotation of $\mathbf{a}_i$. Interestingly, as we show in Section 3.2, these problems do not interfere with the consistency and asymptotic normality of the estimated ATT. This general identification scheme is formally presented in Brown and Butts (2022). ■

With $y_{i,t}(g)$ known and $y_{i,t}(\infty)$ imputed, the estimated treatment effect is given by

$$\widehat{\Delta}_{i,g,t} = y_{i,t} - \widehat{y}_{i,t}(\infty). \tag{19}$$

The estimated ATT for group g at time t is obtained by averaging over the relevant treated group;

$$\widehat{\Delta}_{g,t} = \frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\widehat{\Delta}_{i,g,t}, \tag{20}$$

which is again available for all $t \geq T_g$ and $g < G$. Equation (20) is the CCEDID estimator of $\Delta_{g,t}$.

Asymptotic standard errors of estimates of the ATT are generally difficult to implement. Many studies therefore resort to bootstrap inference (see, for example, Callaway and Karami (2020) and Xu (2017)), which can be computationally unattractive. We instead employ a version of the non-parametric standard error estimator considered by Chudik et al. (2011) and Pesaran (2006). The appropriate estimator to use in our case is

$$\widehat{\omega}_{g,t}^2 = \frac{1}{N_g - 1} \sum_{i=1}^{N} D_{i,g} \left(\widehat{\Delta}_{i,g,t} - \frac{1}{N_g} \sum_{j=1}^{N} D_{j,g} \widehat{\Delta}_{j,t} \right)^2. \tag{21}$$

In addition to being simple to compute, non-parametric standard errors tend to perform well in small samples (see, for example, Chudik et al. (2011), Pesaran (2006), and Westerlund and Kaddoura (2022)). It also means that the asymptotic variance can be consistently estimated whether or not the number of factor proxies outnumbers the true number of factors ($m + 1 = r$ or $m + 1 > r$).

Remark 5. It is important to note that the proposed CCEDID estimator does not involve any estimation of the number of factors, r. This is in stark contrast to existing principal components-based approaches such as Chan and Kwok (2022) and Xu (2017), as well as quasi-differencing approaches of Callaway and Karami (2020) and Brown and Butts (2022), where the existing asymptotic theory is based on treating r as known. This means that in empirical work, r has to be replaced by an estimator, and accurate estimation of this object is known to be a difficult; see, for example, Moon and Weidner (2015) and Breitung and Hansen (2021). The fact that the proposed estimator does not require estimation of r is therefore a great advantage in practice. ■

3.2 Asymptotic results

In this section, we study the asymptotic properties of $\widehat{\Delta}_{g,t}$ and $\widehat{\omega}_{g,t}^2$. The conditions that we will be working under are given in Assumptions 1–9. Here and throughout, tr $\mathbf{A}$, rank $\mathbf{A}$ and $\|\mathbf{A}\| = \sqrt{\text{tr}\left(\mathbf{A}'\mathbf{A}\right)}$ denote the trace, the rank, and the Frobenius (Euclidean) norm of the generic matrix

A, respectively. The symbols $\to_d$ and $\to_p$ signify convergence in distribution and probability, respectively.

Assumption 1. $T_0 > m + 1$. ■

Assumption 2. $N_g/N \to_p \mathbb{P}(D_{i,g} = 1) \in (0,1)$ for $g = g_1, ..., g_G, \infty$. ■

Assumptions 1 and 2 are sample size conditions. They ensure that T_0 is large enough to ensure that the step-2 regression model in (12) is feasible and also that each group is non-negligible as N increases, which is necessary for accurate estimation of the group-specific ATTs. We write Assumption 2 in terms of convergence in probability because $N_g = \sum_{i=1}^{N} D_{i,g}$ is a random quantity.

Assumption 3. $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\nu}_i$, $\Delta_{i,g,t} = \Delta_{g,t} + v_{i,t}$, and $\boldsymbol{\tau}_{i,g,t} = \boldsymbol{\tau}_{g,t} + \boldsymbol{\zeta}_{i,t}$ where $\boldsymbol{\nu}_i$, $v_{i,t}$, and $\boldsymbol{\zeta}_{i,t}$ are independently distributed across i and t with zero mean, and finite fourth-order cumulants. ■

Assumption 3 is a random coefficient condition that is largely the same as in Chan and Kwok (2022) and Gobillon and Magnac (2016). The slopes are not required to be heterogeneous, though, as the covariance matrices of $\boldsymbol{\nu}_i$ and $v_{i,t}$ need not be positive definite.

Before we continue onto Assumption 4, is it useful to first lay out some additional notation. Step 1 uses the cross-sectional averages of the observables in $\mathbf{z}_{i,t}$ for the untreated units to estimate the factors. This means that both $y_{i,t}$ and $\mathbf{x}_{i,t}$ have to be informative of those factors. By combining (10) and (3) we arrive at the following static factor model for $\mathbf{z}_{i,t}$:

$$\mathbf{z}_{i,t} = \boldsymbol{\mu}_{i,t} d_{i,t} + \boldsymbol{\Lambda}_i' \mathbf{f}_t + \mathbf{e}_{i,t}, \tag{22}$$

where $\boldsymbol{\mu}_{i,t} = [\Delta_{i,g,t} + \boldsymbol{\beta}_i' \boldsymbol{\tau}_{i,g,t}, \boldsymbol{\tau}_{i,g,t}']'$ is $(m+1) \times 1$, $\boldsymbol{\Lambda}_i = [\boldsymbol{\alpha}_i + \boldsymbol{\lambda}_i \boldsymbol{\beta}_i, \boldsymbol{\lambda}_i]$ is $r \times (m+1)$ and $\mathbf{e}_{i,t} = [\varepsilon_{i,t} + \boldsymbol{\beta}_i' \mathbf{v}_{i,t}, \mathbf{v}_{i,t}']'$ is $(m+1) \times 1$. Since $d_{i,t} = 0$ for untreated units, we have

$$\widehat{\mathbf{f}}_t = \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i,\infty} \mathbf{z}_{i,t} = \left(\frac{1}{N_\infty} \sum_{i=1}^{N} D_{i,\infty} \boldsymbol{\Lambda}_i \right)' \mathbf{f}_t + \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i,\infty} \mathbf{e}_{i,t}. \tag{23}$$

Assumptions 4–6 below ensure that the average $\mathbf{e}_{i,t}$ tends to zero as N increases and that the average of $\boldsymbol{\Lambda}_i$ has full row rank, which in turn ensure that $\widehat{\mathbf{f}}_t$ is consistent for the space spanned by $\mathbf{f}_t$.

Assumption 4. $\varepsilon_{i,t}$ and $\mathbf{v}_{i,t}$ are independently distributed across i with zero mean, and finite fourth-order cumulants. ■

Assumption 5. $\mathbf{f}_t$, G_i, $\varepsilon_{i,t}$, $\mathbf{v}_{i,t}$, $\boldsymbol{\nu}_i$, $\boldsymbol{\zeta}_{i,t}$, and $v_{i,t}$ are mutually independent. ■

Assumption 6. $\text{rank}(N_\infty^{-1}\sum_{i=1}^{N} D_{i,\infty}\boldsymbol{\Lambda}_i) = r \leq m+1$. ■

Assumption 7. The $r \times r$ matrix $\sum_{t=1}^{T} \mathbf{f}_t\mathbf{f}_t'$ is positive definite for all T. ■

Assumption 8. $N^{-1}\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{x}_i \to_p \boldsymbol{\Sigma}$ as $N \to \infty$, where the $m \times m$ matrix $\boldsymbol{\Sigma}$ is positive definite. ■

Note that Assumption 5 places no restrictions on the correlation between the treatment effects and the model's errors. This fact leaves the treated potential outcomes and covariates unrestricted with respect to their deviation from their untreated states. We also do not restrict the pattern of factor loadings among different treatment statuses. Brown and Butts (2022) show that TWFE is consistent for the ATTs when heterogeneity is mean independent of treatment assignment.

Assumptions 7 and 8 are standard non-collinearity conditions. Assumption 7 generalizes the usual "within assumption" in the individual fixed effects only model, which rules out time-invariant regressors. Assumption 7 rules out more general "low-rank" regressors, as it is almost always done in models with interactive effects (see Moon and Weidner (2015) for a discussion). The exclusion restriction is not very restrictive, though, as it does not rule out low rank regressors in the model for $y_{i,t}$ in (10). If there are such regressors present in (10), then these should be treated as observed factors, which can be appended to $\widehat{\mathbf{f}}_t$ in step 1.This is an advantage in the sense that while $\boldsymbol{\beta}_i$ and $\Delta_{i,g,t}$ are subject to the random coefficient condition in Assumption 3, $\boldsymbol{\alpha}_i$ is not. Hence, unlike the coefficients of the observed regressors, the coefficients of low rank regressors are not restricted in any way. The disadvantage of this observed factor treatment of low rank regressors is that we cannot estimate their coefficients.

An important point about Assumptions 1–8 is that the time series properties of $\mathbf{f}_t$, $\varepsilon_{i,t}$, $\mathbf{v}_{i,t}$ and $\Delta_{i,g,t}$ are essentially unrestricted. Chan and Kwok (2022) allow for non-stationary factors and regressors (in a large-T setting) but the regression errors have to be stationary, which is

tantamount to requiring that the observables are cointegrated with the factors. Assumptions 1–8 are more general in this regard.

One implication of this generality is that as long as $m+1 \geq r$ there is no need to model the deterministic component of the data, as deterministic regressors can be treated as additional (unknown) factors to be estimated from the data. However, it is common in practice to include typical known factors like an intercept or a time trend. This can easily be accomplished by inserting them into $\widehat{\mathbf{f}}$ along with the cross-sectional averages of the data. As with the dynamics, the type of heteroskedasticity that can be permitted is not restricted in any way.

We are now ready to state Theorem 1, which contains our two main results.

Theorem 1. *Under Assumptions 1–8, as* $N \to \infty$,

(a) $\dfrac{\sqrt{N_g}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\omega_{g,t}} \to_d N(0,1)$,

(b) $\widehat{\omega}^2_{g,t} \to_p \omega^2_{g,t}$,

where the definition of $\omega^2_{g,t}$ *is provided in the appendix.*

The proof of Theorem 1 is contained in the appendix, where we show that $\sqrt{N_g}(\widehat{\Delta}_{g,t} - \Delta_{g,t})$ is asymptotically mixed normal, and that this implies that $\sqrt{N_g}(\widehat{\Delta}_{g,t} - \Delta_{g,t})/\omega_{g,t}$ is asymptotically standard normal. This result is unintuitive given the inconsistency of $\widehat{\mathbf{a}}_i$ in step 2 of the counterfactual estimation procedure, as mentioned earlier in Remark 3. However, Brown and Butts (2022) show that the asymptotic distribution of $\widehat{\mathbf{a}}_i$ is centered at a rotated version of $\mathbf{a}_i$, and that the effect of this rotation is absorbed in the estimation of $\mathbf{f}$. The asymptotic distribution of $\widehat{\Delta}_{i,g,t} - \Delta_{i,g,t}$ is correctly centered at zero despite the inconsistency, and it is independent across i. Asymptotic normality is therefore possible after averaging over the relevant subsample.

Another point about Theorem 1 is that it holds even if r is unknown, provided only that $m+1 \geq r$, so that the number of factors is not under-specified. As we show in the proof, while $\omega^2_{g,t}$ depends on whether $m+1 = r$ or $m+1 > r$, this dependence is successfully mimicked in large samples by $\widehat{\omega}^2_{g,t}$. We can therefore show that

$$\frac{\sqrt{N_g}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\widehat{\omega}_{g,t}} = \frac{\sqrt{N_g}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\omega_{g,t}} + o_p(1) \to_d N(0,1) \tag{24}$$

as $N \to \infty$. Asymptotically valid inference is therefore possible for any r satisfying $m + 1 \geq r$. This robustness is particularly important given the well-known bias problem of post-selection estimators (Leeb and Pötscher, 2005).

4 Estimating direct and mediated effects

We now discuss estimation and inference of the decomposition of the treatment effects studied in Section 3. We also demonstrate how using observed covariates, while potentially ruling out estimation of the overall treatment effect, can produce a more robust estimator of the direct treatment effect.

4.1 Decomposing treatment effects

This section considers estimation and inference for the constituent parts of the overall treatment effect. We demonstrate in Section 2 how the overall treatment effect $\Delta_{i,g,t}$ can be decomposed into the direct effect $\eta_{i,g,t}$ and the mediated effect $\boldsymbol{\tau}'_{i,g,t}\boldsymbol{\beta}_i$. We now demonstrate how to consistently estimate these constituent parts of the treatment effect.

We can estimate $\boldsymbol{\tau}_{g,t}$ for post-treatment time periods by averaging the difference between the observed covariates and their imputed untreated counterfactual. We define this estimator as

$$\widehat{\boldsymbol{\tau}}_{g,t} = \frac{1}{N_g} \sum_{i=1}^{N} D_{i,g}(\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)), \tag{25}$$

which is simply the overall treatment effect estimator applied to the covariates. The estimated effect on $\boldsymbol{x}$ is what researchers typically use as evidence of a causal mechanism (albeit typically under a more restrictive TWFE model). This analysis is incomplete, as the effect of changing the covariate *on the outcome* is determined by the partial effect of the covariates on the outcome variable, which in our model is given by $\boldsymbol{\beta}_i$. Our estimate of the mediated effect is the product of $\widehat{\boldsymbol{\tau}}_{g,t}\widehat{\boldsymbol{\beta}}$ with each component of the vector being the estimated mediated effect of the k-th covariate. It is important to see that a consistent estimator requires a pure factor structure in the untreated potential covariates. This fact precludes the use of nonlinear functions of covariates, like squares and interactions, as well as covariates that are likely nonlinear in the common

factors, like discrete or bounded variables. We discuss how to weaken this requirement in Section 4.3.

The direct effect of treatment is easy to obtain once we have a consistent estimator of the mediated effect. To get a consistent estimator of the direct effect, we define

$$\widehat{\eta}_{g,t} = \widehat{\Delta}_{g,t} - \widehat{\boldsymbol{\tau}}'_{g,t}\widehat{\boldsymbol{\beta}} = \frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}(y_{i,t} - \mathbf{x}'_{i,t}\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{f}}'_t\widehat{\mathbf{a}}_i). \tag{26}$$

That is, our estimated direct effect is our estimated overall effect minus the mediated effects that operate through the included covariates.

We can use non-parametric standard errors for inference on the decomposed treatment effects just like in the case of the overall treatment effect. Let $\widehat{\boldsymbol{\tau}}_{i,g,t} = \mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)$ so that $\widehat{\boldsymbol{\tau}}_{g,t}$ is the group average of the $\widehat{\boldsymbol{\tau}}_{i,g,t}$ terms. We define the $m \times m$ covariance estimator of the treatment effect on $\mathbf{x}_{i,t}$ in equation (25) as

$$\widehat{\boldsymbol{\Omega}}_{g,t} = \frac{1}{N_g - 1}\sum_{i=1}^{N} D_{i,g}\left(\widehat{\boldsymbol{\tau}}_{i,g,t} - \widehat{\boldsymbol{\tau}}_{g,t}\right)\left(\widehat{\boldsymbol{\tau}}_{i,g,t} - \widehat{\boldsymbol{\tau}}_{g,t}\right)' \tag{27}$$

By setting $\widehat{\eta}_{i,g,t} = y_{i,t} - \mathbf{x}_{i,t}\widehat{\beta} - \widehat{\mathbf{f}}'_t\widehat{\boldsymbol{\gamma}}_i$, we similarly define the estimator of the asymptotic variance for the direct effect as

$$\widehat{\omega}^2_{\eta_{g,t}} = \frac{1}{N_g - 1}\sum_{i=1}^{N} D_{i,g}\left(\widehat{\eta}_{i,g,t} - \widehat{\eta}_{g,t}\right)^2 \tag{28}$$

which is identical to the non-parametric variance estimator in equation (21) but replacing the imputed covariates with their observed counterparts.

Using the definitions above, we can now present asymptotic convergence results for the direct and indirect treatment effects.

Theorem 2. *Under Assumptions 1–8 , as* $N \to \infty$,

(a) $\widehat{\boldsymbol{\Omega}}_{g,t}^{-1/2}\sqrt{N_g}\left(\widehat{\boldsymbol{\tau}}_{g,t} - \boldsymbol{\tau}_{g,t}\right) \to_d N(0, \mathbf{I}_K)$,

(b) $\dfrac{\sqrt{N_g}\left(\widehat{\eta}_{g,t} - \eta_{g,t}\right)}{\widehat{\omega}_{\eta_{g,t}}} \to_d N(0,1)$.

The proof of Theorem 2 follows directly from the work in the proof of Theorem 1. To estimate the mediated effect on treatment, we only have to pre-multiply the indirect treatment

effect estimator by $\widehat{\boldsymbol{\beta}}'$ and adjust the standard errors accordingly. Because $\widehat{\boldsymbol{\beta}}$ is time constant, the estimator is equivalent to averaging over $\left(\mathbf{x}_{i,t}' - \widehat{\mathbf{x}}_{i,t}(\infty)'\right)\widehat{\boldsymbol{\beta}}$. The relevant asymptotic variance estimator is $\sqrt{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\Omega}}_{g,t}\widehat{\boldsymbol{\beta}}}$.

4.2 Consequences of using observed covariates

This section discusses the problems with identification when a researcher uses the observed covariates $\mathbf{x}_{i,t}$ in the imputation stage for $\widehat{y}_{i,t}$ instead of the correctly imputed untreated potential covariates. We define this alternative estimator of group g's ATT at time t as $\tilde{\Delta}_{g,t}$ for $t \geq g$. We showed in Section 4.1 that using observed covariates only allows us to estimate the direct effect.

To see this, note that

$$\tilde{\Delta}_{g,t} = \frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\left(y_{i,t} - \mathbf{x}_{i,t}'\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{f}}_t'\widehat{\boldsymbol{\gamma}}_i\right) = \frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\left(\Delta_{i,g,t} + \mathbf{x}_{i,t}'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \left(\mathbf{f}_t'\boldsymbol{\gamma}_i - \widehat{\mathbf{f}}_t'\widehat{\boldsymbol{\gamma}}_i\right)\right) \quad (29)$$

We demonstrate in the proof of Theorem 1 that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$ and $\frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\left(\widehat{\mathbf{f}}_t'\widehat{\boldsymbol{\gamma}}_i - \mathbf{f}_t'\boldsymbol{\gamma}_i\right) = o_p(1)$. Because $\frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\mathbf{x}_{i,t} = O_p(1)$, $\tilde{\Delta}_{g,t}$ is only consistent for the direct effect of treatment on the outcomes. This fact demonstrates that the interpretation of the imputation estimator that uses observed covariates is as an estimator of the direct effect; such is the case of the PC-DID estimator of Chan and Kwok (2022). Researchers who implement interactive fixed effects imputation estimators with time-varying covariates must therefore be careful in interpreting their results.

We can now study the conditions under which $\tilde{\Delta}_{g,t}$ is consistent for the overall effect. Intuitively, the only time $\tilde{\Delta}_{g,t}$ is consistent for the overall ATT (and not just the direct effect) is when the treatment does not affect the outcome via the covariates. To demonstrate this fact, we write $\tilde{\Delta}_{g,t}$ in terms of our overall estimator from Theorem 1:

$$\tilde{\Delta}_{g,t} = \widehat{\Delta}_{g,t} - \frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}(\widehat{\mathbf{x}}_{i,t}(\infty) - \mathbf{x}_{i,t})'\widehat{\boldsymbol{\beta}} \quad (30)$$

Note that $\tilde{\Delta}_{g,t}$ is numerically identical to our direct effect estimate, $\widehat{\eta}_{g,t}$, in the previous section.

Thus $\tilde{\Delta}_{g,t}$ is consistent for the overall ATT $\Delta_{g,t}$ when $\frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}(\widehat{\mathbf{x}}(\infty) - \mathbf{x}_{i,t})'\widehat{\boldsymbol{\beta}} = o_p(1)$.

We demonstrate in equation (A.58) of the Appendix that

$$\frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}(\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)) = \frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\boldsymbol{\tau}_{i,g,t} + o_p(1) \tag{31}$$

Equation (30) then becomes

$$\tilde{\Delta}_{g,t} = \widehat{\Delta}_{g,t} - \left(\frac{1}{N_g}\sum_{i=1}^{N} D_{i,g}\boldsymbol{\tau}_{i,g,t}\right)' \widehat{\boldsymbol{\beta}} + o_p(1) \tag{32}$$

Because we know the limits of both terms in the above equation, we can demonstrate when replacing the imputed potential covariates with observed covariates leads to a consistent estimator of $\Delta_{g,t}$.

Theorem 3. *Under Assumptions 1–8, $\tilde{\Delta}_{g,t} \to_p \Delta_{g,t}$ as $N \to \infty$ if and only if $\boldsymbol{\tau}_{g,t} = \mathbf{0}$ or $\boldsymbol{\beta} = \mathbf{0}$.*

The conditions for consistency of $\tilde{\Delta}_{g,t}$ are unsurprising. They essentially require the mediated effect on treatment to be zero. The former case, where covariates are unaffected by treatment ($\boldsymbol{\tau}_{g,t} = \mathbf{0}$), is similar to what Chan and Kwok (2022) assume for the PCDID estimators. The alternative case occurs when the covariates are uninformative for $y_{i,t}$. This setting implies the $\mathbf{x}_{i,t}$ are irrelevant to the mean of $y_{i,t}$. However, the covariates can still be used as factor proxies in this setting.

4.3 Robust direct effect estimation

The results in this paper have so far depended on covariates admitting a common factor structure. This assumption may be too strong for some microeconometric applications. We now discuss consistency under a more general model due to Brown et al. (2022). We show that the direct effect is still estimable under the more general model, but that we can no longer identify the mediated effect without further restrictions.

Brown et al. (2022) consider the model in equation (22) but without treatment effects. When $\overline{\boldsymbol{\Lambda}}$ is full rank (like in Assumption 6), they show that

$$\mathbf{F} = \overline{\mathbf{Z}}\overline{\boldsymbol{\Lambda}}'\left(\overline{\boldsymbol{\Lambda}}\overline{\boldsymbol{\Lambda}}'\right)^{-1} + o_p(1) \tag{33}$$

where $\overline{\mathbf{Z}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i$ and $\mathbf{Z}_i$ is the $T \times (m+1)$ matrix of stacked $\mathbf{z}_{i,t}$. They assume random sampling in the cross section so that all population equations are written in moment conditions.

They then consider the general model

$$\mathbf{F} = \mathbf{\Psi B} \tag{34}$$

where $\mathbf{F}$ is the full $T \times r$ matrix of stacked factors, $\mathbf{\Psi}$ is a $T \times q$ matrix of observed or estimable factor proxies and $\mathbf{B}$ is an arbitrary $q \times r$ matrix. In the case of the classic CCE model, $\mathbf{\Psi} = \mathbb{E}(\mathbf{Z}_i)$ and $\mathbf{B} = \mathbb{E}(\mathbf{\Lambda})$. However, they place no restrictions on the rank of $\mathbf{B}$ and so the model in equation (34) is strictly more general than CCE. This model allows the DGP of the covariates to be essentially unrestricted, allowing for polynomial functions and interactions, as well as count and limited variables.

Stacking the outcomes over time, and assuming $\mathbf{\Psi} = \mathbb{E}(\mathbf{Z}_i)$, equation (34) implies

$$y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\beta}_i + \mathbb{E}(\mathbf{z}_{i,t})\boldsymbol{\rho}_i + \epsilon_{i,t} \tag{35}$$

where $\boldsymbol{\rho}_i = \mathbf{B}\boldsymbol{\gamma}_i$. We can adopt this model to our current setting by assuming it holds only for the untreated potential outcomes. We write this model as

$$y_{i,t}(\infty) = \mathbf{x}_{i,t}(\infty)'\boldsymbol{\beta}_i + \mathbb{E}(\mathbf{z}_{i,t}|G_i = \infty)\boldsymbol{\rho}_i + \epsilon_{i,t} \tag{36}$$

which assumes nothing about the distribution of the treated potential outcomes and covariates. The treated potential outcome for a member of group g at time $t \geq g$ is

$$\begin{aligned} y_{i,t}(g) &= \eta_{i,g,t} + y_{i,t}(\infty) \\ &= \eta_{i,g,t} + \mathbf{x}_{i,t}(g)'\boldsymbol{\beta}_i + \mathbb{E}(\mathbf{z}_{i,t}|G_i = \infty)\boldsymbol{\rho}_i + \epsilon_{i,t} \end{aligned}$$

We can no longer identify aggregates of the overall treatment effect $\eta_{i,g,t} + \boldsymbol{\tau}_{i,g,t}'\boldsymbol{\beta}_i$ because we no longer have a model for the untreated potential covariates, thus leaving the potential channel for the mediated effect unspecified. However, we can still estimate the direct effect of treatment because $\widehat{\eta}_{t,g}$ corresponds to the treatment effect estimator using observed covariates as in Section 4.1. Brown et al. (2022) show that the model in equation (34) implies consistency of the pooled CCE estimator under a similar random slope assumption as our current Assumptions 3 and 5. Instead of being consistent for the factors, $\frac{1}{N_\infty}\sum_{i=1}^{N} D_{i,\infty}\mathbf{z}_{i,t}$ is not consistent for $\mathbf{f}_t$, but for $\mathbb{E}(\mathbf{z}_{i,t}|G_i = \infty)$, which is all that we require according to equation (35). Further, $\widehat{\mathbf{a}}_i$ is

consistent for a rotation of the transformed loadings $\boldsymbol{\rho}_i$ and not $\boldsymbol{\gamma}_i$, which again is all we require by equation (35).

The only departure from our original procedure under this model comes when we are interested in inference on $\boldsymbol{\beta}$. Brown et al. (2022) show that without the CCE assumption guaranteeing independence between $(\mathbf{I}_{T_0} - \mathbf{f}_t(\mathbf{f}_t'\mathbf{f}_t)^{-1}\mathbf{f}_t')\mathbf{X}_i$ and $\boldsymbol{\gamma}_i$, the cluster-robust standard errors from Westerlund et al. (2019) are inconsistent. They derive analytic standard errors that correct for this first-stage estimation uncertainty. We could also use a nonparametric bootstrap while re-estimating $\widehat{\mathbf{f}}$ with every bootstrap sample.

5 Monte Carlo simulations

We now present our main simulation results. Our simulations set $N = 200$ and $T = 8$, a relatively small number of cross-sectional units. Treatment turns on at time period 6 for treated individuals and hence $T_0 = 5$. The data is generated as follows. First, there are two factors, $m = 2$, with $f_1 = [1, \ldots, 1]'$ being a vector of ones allowing for unit fixed effects and $f_2 = [1, \ldots, T]'$ allowing for unit-specific linear time-trend.

There are two covariates, $K = 2$, both generated according to (4) with factor loadings λ_i being a 2×2 matrix with

$$\lambda_i \sim \begin{bmatrix} N(1,1) & N(0,1) \\ N(0,1) & N(1,1) \end{bmatrix}$$

The outcome's factor-loadings are generated $\mathbf{a}_i \sim N(\text{diag}(\lambda_i), I_2)$. Outcomes are generated as follows:

$$y_{i,t} = \Delta * d_{i,t} + \boldsymbol{\beta}_0'(\mathbf{x}_{i,t}(\infty) + \boldsymbol{\tau} d_{i,t}) + \mathbf{a}_i'\mathbf{f}_t + \varepsilon_{i,t}$$

with $\beta_0 = [1, 1]$. The error term is an $AR(1)$ process, i.e. $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + u_{it}$ with $u_{it} \sim N(0,1)$. We set $\rho = 0.75$.

In the first set of simulations, we assign treatment randomly with unconditional probability of treatment of 50%. This implies parallel trends holds since the average of factor-loadings are the same in the treated and control group. This simulation aims to show the importance of not controlling for observed $\mathbf{x}_{i,t}$ as described in the previous section.

In the second set of simulations, treatment is assigned with probability increasing in the

second factor-loading, $a_{i,2}$, such that parallel trends fail (since treated units are more exposed to the time-trend in f_2). In particular, we form the term

$$\pi_i = 0.5 + \frac{a_{i2}}{\max_i a_{i2} - min_i a_{i2}}$$

Then, we normalize this by the mean of π_i, $\pi_i / (\frac{1}{N}\sum_i \pi_i)$, so that the unconditional probability of treatment stays at 50%.

We run three data-generating processes. First, we generate the data with a direct effect of $\Delta = 0$ and an effect on $\mathbf{x}$ of $\boldsymbol{\tau} = [0, 0]$ such that the true effect of treatment is zero. Second, we allow for a direct effect of $\Delta = 1$ but don't allow treatment to effect the distribution of covariates ($\boldsymbol{\tau} = [0, 0]$). Last, we set $\Delta = 1$ and $\tau = [0, 1]$ to have a mediated effect of 1 ($\boldsymbol{\beta}'\boldsymbol{\tau}$) and a total effect of 2.

We estimate three specifications. First, we estimate a two-way fixed effect event-study model

$$y_{it} = \mu_i + \eta_t + \sum_{k=6}^{8} \tau_k d_{it}^k + \text{error}_{it}. \tag{37}$$

where d_{it}^k is an indicator equaling one if unit i is treated and $t = k$. Second, we do what is commonly done in empirical applications and run the same event-study model while controlling for the time-varying covariates

$$y_{it} = \mu_i + \eta_t + \boldsymbol{X}_{it}\boldsymbol{\beta}_t + \sum_{k=6}^{8} \tau_k d_{it}^k + \text{error}_{it}. \tag{38}$$

Last, we run our proposed CCEDID procedure. For each estimator, we report the average bias of our estimates $\mathbb{E}\left[\widehat{\tau}_k - \tau_k\right]$ as well as the mean-squared error, $MSE(\widehat{\tau}_k) \equiv \mathbb{E}\left[(\tau_k - \widehat{\tau}_k)^2\right]$ for $k = 6, 7, 8$.

Table 1 contain results for when parallel trends holds. First note that all three estimates produce unbiased estimates for treatment effects when treatment does not affect the distribution of $\mathbf{x}$ (Panels A and B). However, note that since $CCEDID$ is absorbing the full factor-model, the mean-square error can be significantly smaller.[6] In panel C, we allow treatment to effect the distribution of x_2. As discussed above, controlling for observed $\boldsymbol{x}_{it}$ is problematic since

[6] After controlling for time-dummies, $(a_{i2} - \bar{a}_2) f_{t,2}$ enters the error term where $\bar{a}_2$ is the cross-sectional average of a_{i2}. The variance of this term grows with t and hence the mean-square error is largest in period 8.

Table 1: Monte Carlo: Parallel Trends Hold

Panel A: No Effect. $\eta = [0, 0, 0]$ and $\tau = [0, 0, 0]$.

	$\mathbb{E}(\widehat{\Delta}_6 - \Delta_6)$	$\text{MSE}(\widehat{\Delta}_6)$	$\mathbb{E}(\widehat{\Delta}_7 - \Delta_7)$	$\text{MSE}(\widehat{\Delta}_7)$	$\mathbb{E}(\widehat{\Delta}_8 - \Delta_8)$	$\text{MSE}(\widehat{\Delta}_8)$
TWFE	0.04	0.90	0.08	3.55	0.11	7.88
TWFE with Covariates	-4.05	16.92	-4.14	18.69	-4.19	20.66
CCEDID	0.04	0.52	0.09	1.08	0.06	1.60

Panel B: Direct Effect. $\eta = [1/3, 2/3, 1]$ and $\tau = [0, 0, 0]$.

	$\mathbb{E}(\widehat{\Delta}_6 - \Delta_6)$	$\text{MSE}(\widehat{\Delta}_6)$	$\mathbb{E}(\widehat{\Delta}_7 - \Delta_7)$	$\text{MSE}(\widehat{\Delta}_7)$	$\mathbb{E}(\widehat{\Delta}_8 - \Delta_8)$	$\text{MSE}(\widehat{\Delta}_8)$
TWFE	0.04	0.95	0.08	3.70	0.12	8.29
TWFE with Covariates	0.05	0.49	0.08	1.57	0.11	3.35
CCEDID	-0.03	0.49	-0.05	0.93	-0.05	1.54

Panel C: Direct and Mediated Effect. $\eta = [1/3, 2/3, 1]$ and $\tau = [1/3, 2/3, 1]$.

	$\mathbb{E}(\widehat{\Delta}_6 - \Delta_6)$	$\text{MSE}(\widehat{\Delta}_6)$	$\mathbb{E}(\widehat{\Delta}_7 - \Delta_7)$	$\text{MSE}(\widehat{\Delta}_7)$	$\mathbb{E}(\widehat{\Delta}_8 - \Delta_8)$	$\text{MSE}(\widehat{\Delta}_8)$
TWFE	0.04	0.90	0.08	3.55	0.11	7.88
TWFE with Covariates	-4.05	16.92	-4.14	18.69	-4.19	20.66
CCEDID	0.04	0.52	0.09	1.08	0.06	1.60

Notes.

Table 2: Monte Carlo: Parallel Trends Do Not Hold

Panel A: No Effect. $\eta = [0, 0, 0]$ and $\tau = [0, 0, 0]$.

	$\mathbb{E}(\widehat{\Delta}_6 - \Delta_6)$	$\text{MSE}(\widehat{\Delta}_6)$	$\mathbb{E}(\widehat{\Delta}_7 - \Delta_7)$	$\text{MSE}(\widehat{\Delta}_7)$	$\mathbb{E}(\widehat{\Delta}_8 - \Delta_8)$	$\text{MSE}(\widehat{\Delta}_8)$
TWFE	-0.66	1.23	-1.32	4.89	-1.99	10.99
TWFE with Covariates	-0.32	0.48	-0.62	1.74	-0.88	3.72
CCEDID	-0.00	0.44	-0.05	0.87	-0.06	1.47

Panel B: Direct Effect. $\eta = [1/3, 2/3, 1]$ and $\tau = [0, 0, 0]$.

	$\mathbb{E}(\widehat{\Delta}_6 - \Delta_6)$	$\text{MSE}(\widehat{\Delta}_6)$	$\mathbb{E}(\widehat{\Delta}_7 - \Delta_7)$	$\text{MSE}(\widehat{\Delta}_7)$	$\mathbb{E}(\widehat{\Delta}_8 - \Delta_8)$	$\text{MSE}(\widehat{\Delta}_8)$
TWFE	-0.65	1.31	-1.31	5.25	-1.95	11.70
TWFE with Covariates	-0.24	0.47	-0.50	1.66	-0.79	3.54
CCEDID	-0.05	0.47	-0.08	1.00	-0.08	1.54

Panel C: Direct and Mediated Effect. $\eta = [1/3, 2/3, 1]$ and $\tau = [1/3, 2/3, 1]$.

	$\mathbb{E}(\widehat{\Delta}_6 - \Delta_6)$	$\text{MSE}(\widehat{\Delta}_6)$	$\mathbb{E}(\widehat{\Delta}_7 - \Delta_7)$	$\text{MSE}(\widehat{\Delta}_7)$	$\mathbb{E}(\widehat{\Delta}_8 - \Delta_8)$	$\text{MSE}(\widehat{\Delta}_8)$
TWFE	-0.61	1.26	-1.22	4.89	-1.83	11.08
TWFE with Covariates	-4.42	20.05	-4.87	25.26	-5.25	30.77
CCEDID	0.01	0.47	0.02	0.82	0.01	1.45

Notes.

it absorbs the mediated effect of that covariate. Hence, TWFE with covariates produces very biased estimates for the treatment effect, even in this extreme case where parallel trends hold unconditionally.

Table 2 contains the results of simulations where prallel trends do not hold. In this case, both TWFE estimators are biased across panels. Our proposed CCEDID procedure performs well in all these cases with near-zero bias and consistently the lowest mean-squared error.

6 Empirical illustration

TO BE ADDED.

7 Conclusion

We derive a consistent estimator of ATTs when untreated potential outcomes are generated by an interactive fixed effects error. Our identification strategy, based on the popular common correlated effects model, relies on time- and individual-varying covariates that admit a pure factor structure. We use the cross-sectional averages of the data to impute the untreated potential outcomes in post-treatment time periods. Our main consistency result allows for a fixed number of time periods, but can easily extend to when there are many pre-treatment observations.

While most treatment effect analyses omit time-varying covariates due to their possible correlation with treatment status, we explicitly allow treatment to affect the covariates' distribution in an arbitrary way. This model allows us to decompose the effect of treatment via a direct effect on the level of the outcomes and a mediated effect through the covariates and their slopes. Such a decomposition allows researchers to perform inference on the possible mechanisms of an intervention through different relevant channels. We also demonstrate how estimators based on time-varying controls that do not allow indirect effects, such as the principal components estimator of Chan and Kwok (2022), are only consistent for the direct effect of treatment unless either the covariates are independent of treatment status or have zero effect on the outcome. This effect is consistently estimated by our CCE estimator under a weaker model proposed by Brown et al. (2022).

References

K. M. Abadir and J. R. Magnus. *Matrix algebra*, volume 1. Cambridge University Press, 2005.

M. Aklin and P. Bayer. How can we estimate the effectiveness of institutions? solving the post-treatment versus omitted variable bias dilemma. Technical report, Working paper, 2017.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

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

K. Borusyak, X. Jaravel, and J. Spiess. Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*, 2021.

J. Breitung and P. Hansen. Alternative estimation approaches for the factor augmented panel data model with small t. *Empirical Economics*, 60(1):327–351, 2021.

N. Brown and K. Butts. A unified framework for dynamic treatment effect estimation in interactive fixed effect models. *Working Paper*, 2022.

N. L. Brown, P. Schmidt, and J. M. Wooldridge. Simple alternatives to the common correlated effects model. *arXiv preprint arXiv:2112.01486*, 2022.

C. Caetano, B. Callaway, S. Payne, and H. S. Rodrigues. Difference in differences with time-varying covariates. *arXiv:2202.02903 [econ]*, Feb 2022. URL http://arxiv.org/abs/2202.02903. arXiv: 2202.02903.

B. Callaway and S. Karami. Treatment effects in interactive fixed effects models. *arXiv:2006.15780 [econ]*, Jun 2020. URL http://arxiv.org/abs/2006.15780. arXiv: 2006.15780.

B. Callaway and P. H. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, page S0304407620303948, Dec 2020. ISSN 03044076. doi: 10.1016/j.jeconom.2020.12.001.

M. K. Chan and S. S. Kwok. The pcdid approach: difference-in-differences when trends are potentially unparallel and stochastic. *Journal of Business & Economic Statistics*, 40(3):1216–1233, 2022.

A. Chudik, M. H. Pesaran, and E. Tosetti. Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal*, 14(1):C45–C90, 2011.

L. Gobillon and T. Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551, Jul 2016. ISSN 0034-6535, 1530-9142. doi: 10.1162/REST_a_00537.

M. Huber. Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6):920–943, 2014.

A. Juodis, H. Karabiyik, and J. Westerlund. On the robustness of the pooled cce estimator. *Journal of Econometrics*, 220(2):325–348, 2021.

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz. Mediation analysis. *Annual review of psychology*, 58:593, 2007.

H. R. Moon and M. Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579, 2015.

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

J. Westerlund and Y. Kaddoura. Cce in heterogenous fixed-t panels. *The Econometrics Journal*, 2022.

J. Westerlund, Y. Petrova, and M. Norkute. Cce in fixed-t panels. *Journal of Applied Econometrics*, 34(5):746–761, 2019.

J. M. Wooldridge. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics*, 87(2):385–390, 2005.

Y. Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, Jan 2017. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2016.2.

Appendix

Proof of Theorem 1.

We start with part (a). We begin by considering the step-1 estimator of $\mathbf{f}_t$. In so doing, it is useful to denote by $\bar{\mathbf{a}}_t = N_\infty^{-1} \sum_{i=1}^{N} D_{i,\infty} \mathbf{a}_{i,t}$ the cross-sectional average of any vector $\mathbf{a}_{i,t}$ for the group of untreated units. In this notation, $\widehat{\mathbf{f}}_t = \bar{\mathbf{z}}_t$. By inserting (3) into (11), and noting that $B_t = 0$ in the pretreatment sample when $t \leq T_0$, we obtain

$$\widehat{\mathbf{f}}_t = \bar{\mathbf{z}}_t = \overline{\mathbf{\Lambda}}' \mathbf{f}_t + \bar{\mathbf{e}}_t, \tag{A.39}$$

If $m+1 = r$, $\overline{\mathbf{\Lambda}}$ is full rank and invertible, which means that (A.39) can be rewritten as

$$\overline{\mathbf{\Lambda}}^{-1\prime} \widehat{\mathbf{f}}_t = \mathbf{f}_t + \overline{\mathbf{\Lambda}}^{-1\prime} \bar{\mathbf{e}}_t. \tag{A.40}$$

Because $\|\bar{\mathbf{e}}_t\| = O_p(N^{-1/2})$ under Assumption 4, we have

$$\overline{\mathbf{\Lambda}}^{-1\prime} \widehat{\mathbf{f}}_t = \mathbf{f}_t + O_p(N^{-1/2}) \tag{A.41}$$

and hence $\overline{\mathbf{\Lambda}}^{-1\prime} \widehat{\mathbf{f}}_t$ is consistent for $\mathbf{f}_t$. In practice, we never observe $\overline{\mathbf{\Lambda}}$. However, since $\boldsymbol{\alpha}_i' \mathbf{f}_t = \boldsymbol{\alpha}_i' \overline{\mathbf{\Lambda}}^{-1\prime} \widehat{\mathbf{f}}_t + O_p(N^{-1/2})$, it is enough if we know $\widehat{\mathbf{f}}_t$, because $\overline{\mathbf{\Lambda}}^{-1}$ is subsumed in the estimation of the coefficient of $\widehat{\mathbf{f}}_t$, which is $\mathbf{a}_i$ in our notation.

The above analysis is not possible when $m+1 > r$ since $\overline{\mathbf{\Lambda}}$ is no longer invertible. However, we still need something similar to (A.40), because it determines the object that is being estimated. The way we approach this issue is the same as in Juodis et al. (2021), and others. In particular, we begin by partitioning $\mathbf{\Lambda}_i$ as $\overline{\mathbf{\Lambda}} = [\overline{\mathbf{\Lambda}}_r, \overline{\mathbf{\Lambda}}_{-r}]$, where $\overline{\mathbf{\Lambda}}_{-r}$ is $r \times (m+1-r)$ and $\overline{\mathbf{\Lambda}}_r$ is $r \times r$ and full rank. Note that this partition is without loss of generality under Assumption 6. We then introduce the following $(m+1) \times (m+1)$ rotation matrix, which is chosen such that $\overline{\mathbf{\Lambda}}\overline{\mathbf{H}} = [\mathbf{I}_r, \mathbf{0}_{r \times (m+1-r)}]$ and that is going to play the same role as $\overline{\mathbf{\Lambda}}^{-1}$ under $m+1 = r$:

$$\overline{\mathbf{H}} = \begin{bmatrix} \overline{\mathbf{\Lambda}}_r^{-1} & -\overline{\mathbf{\Lambda}}_r^{-1} \overline{\mathbf{\Lambda}}_{-r} \\ \mathbf{0}_{(m+1-r) \times r} & \mathbf{I}_{m+1-r} \end{bmatrix} = [\overline{\mathbf{H}}_r, \overline{\mathbf{H}}_{-r}], \tag{A.42}$$

where $\overline{\mathbf{H}}_r = [\overline{\mathbf{\Lambda}}_r^{-1\prime}, \mathbf{0}_{r \times (m+1-r)}]'$ is $(m+1) \times r$, while $\overline{\mathbf{H}}_{-r} = [-\overline{\mathbf{\Lambda}}_{-r}' \overline{\mathbf{\Lambda}}_r^{-1\prime}, \mathbf{I}_{m+1-r}]'$ is $(m+1) \times (m+1-r)$. If $m+1 = r$, we define $\overline{\mathbf{H}} = \overline{\mathbf{H}}_r = \overline{\mathbf{\Lambda}}_r^{-1} = \overline{\mathbf{\Lambda}}^{-1}$. We further introduce the

$(m+1)\times(m+1)$ matrix $\mathbf{D}_N = \text{diag}(\mathbf{I}_r, \sqrt{N}\mathbf{I}_{m+1-r})$ with $\mathbf{D}_N = \mathbf{I}_{m+1}$ if $m+1=r$. By pre-multiplying $\widehat{\mathbf{f}}_t$ by $\mathbf{D}_N\overline{\mathbf{H}}'$, we obtain

$$\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}_t = \widehat{\mathbf{f}}_t^0 = \mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}'\mathbf{f}_t + \mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{e}}_t = \mathbf{f}_t^0 + \overline{\mathbf{e}}_t^0, \tag{A.43}$$

where $\mathbf{f}_t^0 = [\mathbf{f}_t', \mathbf{0}'_{(m+1-r)\times 1}]'$ and $\overline{\mathbf{e}}_t^0 = [\overline{\mathbf{e}}_t'\overline{\mathbf{H}}_r, \sqrt{N}\overline{\mathbf{e}}_t'\overline{\mathbf{H}}_{-r}]' = [\overline{\mathbf{e}}_{r,t}^{0\prime}, \overline{\mathbf{e}}_{-r,t}^{0\prime}]'$ are both $(m+1)\times 1$ with $\overline{\mathbf{e}}_{r,t}^0$ and $\overline{\mathbf{e}}_{-r,t}^0$ being $r\times 1$ and $(m+1-r)\times 1$, respectively. Hence, since $\|\overline{\mathbf{e}}_{r,t}^0\| = O_p(N^{-1/2})$ and $\|\overline{\mathbf{e}}_{-r,t}^0\| = O_p(1)$, when $m+1>r$ we are no longer estimating $\mathbf{f}_t$ but rather $\mathbf{f}_t^+ = [\mathbf{f}_t', \overline{\mathbf{e}}_{-r,t}^{0\prime}]'$;

$$\widehat{\mathbf{f}}_t^0 = \mathbf{f}_t^0 + \overline{\mathbf{e}}_t^0 = \begin{bmatrix} \mathbf{f}_t \\ \mathbf{0}_{(m+1-r)\times 1} \end{bmatrix} + \begin{bmatrix} \overline{\mathbf{e}}_{r,t}^0 \\ \overline{\mathbf{e}}_{-r,t}^0 \end{bmatrix} = \mathbf{f}_t^+ + O_p(N^{-1/2}), \tag{A.44}$$

The fact that $\mathbf{f}_t$ is included in $\mathbf{f}_t^+$ suggests that asymptotically CCEDID should be able to account for the unknown factors even if $m+1>r$. By ensuring the existence of $\overline{\mathbf{H}}$, Assumption 6 makes this possible. However, we also note that because of the presence of $\overline{\mathbf{e}}_{-r,t}^0$, the asymptotic distribution theory will in general depend on whether $m+1=r$ or $m+1>r$.

It is useful to be able to use the above notation not only when $m+1>r$ but also when $m+1=r$. We therefore define $\widehat{\mathbf{f}}_t^0 = \overline{\mathbf{\Lambda}}^{-1\prime}\widehat{\mathbf{f}}_t$, $\mathbf{f}_t^0 = \mathbf{f}_t$ and $\overline{\mathbf{e}}_t^0 = \overline{\mathbf{\Lambda}}^{-1\prime}\overline{\mathbf{e}}_t$ if $m+1=r$, so that we are back in (A.40).

Let us now consider $\widehat{\Delta}_{i,g,t}$, which, unlike $\widehat{\mathbf{f}}_t$, is computed based on treated units in post-treatment periods ($i\in\mathcal{I}_g$, $g<G$ and $t\geq T_g$). From Assumption 3, and the definitions of $y_{i,t}$ and $\widehat{y}_{i,t}(0)$,

$$\begin{aligned}
\widehat{\Delta}_{i,g,t} &= y_{i,t} - \widehat{y}_{i,t}(\infty) \\
&= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} - [\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{x}}_{i,t}(\infty) + \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t] \\
&= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} - (\widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,t} + \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t) + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] \\
&= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - (\widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'\mathbf{f}_t) + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] + \varepsilon_{i,t} \\
&= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - (\widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'\mathbf{f}_t) + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] + \varepsilon_{i,t},
\end{aligned} \tag{A.45}$$

where the last equality makes use of the fact that $D_iB_t = 1$ for all $i\in\mathcal{I}_g$, $g<G$ and $t\geq T_g$. Consider $\widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'\mathbf{f}_t$. While the $(m+1)\times r$ matrix $\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}'$ is not necessarily square under Assumption 6, it has full column rank. This means that we can compute its Moore–Penrose inverse, which is given by $(\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}')^+ = (\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}')' = [\mathbf{I}_r, \mathbf{0}_{r\times(m+1-r)}]$, such that

$(\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}')^+\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}' = \mathbf{I}_r$. Hence, $\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}'\mathbf{f}_t = [\mathbf{f}_t', \mathbf{0}'_{(m+1-r)\times 1}]' = \mathbf{f}_t^0$ and we also have $\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}_t = \widehat{\mathbf{f}}_t^0$. Making use of this, and letting $\widehat{\mathbf{a}}_i^0 = (\mathbf{D}_N\overline{\mathbf{H}}')^{-1\prime}\widehat{\mathbf{a}}_i = (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\widehat{\mathbf{a}}_i$ and $\boldsymbol{\alpha}_i^0 = (\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}')^{+\prime}\boldsymbol{\alpha}_i = \mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}'\boldsymbol{\alpha}_i = [\boldsymbol{\alpha}_i', \mathbf{0}_{1\times(m+1-r)}]'$,

$$\begin{aligned}\widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'\mathbf{f}_t &= \widehat{\mathbf{a}}_i'(\mathbf{D}_N\overline{\mathbf{H}}')^{-1}\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'(\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}')^+\mathbf{D}_N\overline{\mathbf{H}}'\overline{\mathbf{\Lambda}}'\mathbf{f}_t \\ &= \widehat{\mathbf{a}}_i^{0\prime}\widehat{\mathbf{f}}_t^0 - \boldsymbol{\alpha}_i^{0\prime}\mathbf{f}_t^0 \\ &= \boldsymbol{\alpha}_i^{0\prime}(\widehat{\mathbf{f}}_t^0 - \mathbf{f}_t^0) + (\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0)'\widehat{\mathbf{f}}_t^0, \end{aligned} \tag{A.46}$$

from which it follows that

$$\widehat{\Delta}_{i,g,t} = \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i^{0\prime}(\widehat{\mathbf{f}}_t^0 - \mathbf{f}_t^0) - (\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0)'\widehat{\mathbf{f}}_t^0 + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(0)] + \varepsilon_{i,t} \tag{A.47}$$

Amongst the terms appearing on the right-hand side of the above equation, the one involving $\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0$ requires most work. We therefore start with this. Note first that since $\widehat{\mathbf{a}}_i$ is estimated based on the pretreatment period only, $D_i = 0$. By using this and $\overline{\mathbf{\Lambda}}\overline{\mathbf{H}}_r = \mathbf{I}_r$, we get

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta}_i + \widehat{\mathbf{f}}\overline{\mathbf{H}}_r\boldsymbol{\alpha}_i - (\widehat{\mathbf{f}} - \mathbf{f}\overline{\mathbf{\Lambda}})\overline{\mathbf{H}}_r\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i = \mathbf{x}_i\boldsymbol{\beta}_i + \widehat{\mathbf{f}}\overline{\mathbf{H}}_r\boldsymbol{\alpha}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \tag{A.48}$$

where the T_0-rowed matrices $\mathbf{f}$ and $\boldsymbol{\varepsilon}_i$ are defined analogously to $\mathbf{y}_i$, $\mathbf{x}_i$ and $\widehat{\mathbf{f}}$. Note also that in the notation of the step-2 regression in (12), we have $\mathbf{a}_i = \overline{\mathbf{H}}_r\boldsymbol{\alpha}_i$. By inserting this and (A.48) into the expression given for $\widehat{\mathbf{a}}_i$ in step 2,

$$\begin{aligned}\widehat{\mathbf{a}}_i &= (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}) \\ &= (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{x}_i\boldsymbol{\beta}_i + \widehat{\mathbf{f}}\mathbf{a}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{a}_i + (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i], \end{aligned} \tag{A.49}$$

implying

$$\begin{aligned}\widehat{\mathbf{a}}_i^0 &= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\widehat{\mathbf{a}}_i \\ &= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i + (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i] \\ &= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i + (\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}'\widehat{\mathbf{f}}\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}'[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i] \\ &= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i + (\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}^{0\prime}[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i] \end{aligned} \tag{A.50}$$

where $\widehat{\mathbf{f}}^0 = [\widehat{\mathbf{f}}_1^0, ..., \widehat{\mathbf{f}}_{T_0}^0]' = \widehat{\mathbf{f}}\overline{\mathbf{H}}\mathbf{D}_N$ is $T_0 \times (m+1)$. Consider the first term on the right-hand side. A direct calculation using the rules for the inverse of a partitioned matrix (see, for example, Abadir and Magnus (2005), Exercise 5.16) reveals that

$$(\mathbf{D}_N\overline{\mathbf{H}})^{-1} = \begin{bmatrix} \overline{\mathbf{\Lambda}}_r & \overline{\mathbf{\Lambda}}_{-r} \\ \mathbf{0}_{(m+1-r)\times r} & N^{-1/2}\mathbf{I}_{m+1-r} \end{bmatrix}, \tag{A.51}$$

so that

$$(\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\overline{\mathbf{H}}_r = \begin{bmatrix} \overline{\mathbf{\Lambda}}_r & \overline{\mathbf{\Lambda}}_{-r} \\ \mathbf{0}_{(m+1-r)\times r} & N^{-1/2}\mathbf{I}_{m+1-r} \end{bmatrix} \begin{bmatrix} \overline{\mathbf{\Lambda}}_r^{-1} \\ \mathbf{0}_{(m+1-r)\times r} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0}_{(m+1-r)\times r} \end{bmatrix}. \tag{A.52}$$

This implies

$$(\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i = \begin{bmatrix} \boldsymbol{\alpha}_i \\ \mathbf{0}_{(m+1-r)\times 1} \end{bmatrix} = \boldsymbol{\alpha}_i^0, \tag{A.53}$$

leading to the following expression for $\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0$:

$$\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0 = (\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}^{0\prime}[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i]. \tag{A.54}$$

We similarly have

$$\widehat{\mathbf{x}}_{i,t}(\infty) = \widehat{\boldsymbol{\lambda}}_i'\widehat{\mathbf{f}}_t = \mathbf{x}_i'\widehat{\mathbf{f}}(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}_t = \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0, \tag{A.55}$$

from which it follows that

$$\widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] = \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0]. \tag{A.56}$$

By inserting the above expressions into the one given earlier for $\widehat{\Delta}_{i,g,t}$, we get

$$\begin{aligned}
\widehat{\Delta}_{i,g,t} &= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i^{0\prime}(\widehat{\mathbf{f}}_t^0 - \mathbf{f}_t^0) - (\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0)'\widehat{\mathbf{f}}_t^0 + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] + \varepsilon_{i,t} \\
&= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i'\overline{\mathbf{e}}_{r,t}^0 \\
&\quad - [-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i]'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0 + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0] + \varepsilon_{i,t} \\
&= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i'\overline{\mathbf{e}}_{r,t}^0 + \varepsilon_{i,t} - (\mathbf{x}_i\boldsymbol{\beta}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i)'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0.
\end{aligned} \tag{A.57}$$

Further use of $\widehat{\mathbf{f}} = \widehat{\mathbf{f}}^0\mathbf{D}_N^{-1}\overline{\mathbf{H}}^{-1}$ gives

$$\mathbf{x}_i = \mathbf{f}\boldsymbol{\lambda}_i + \mathbf{v}_i = \widehat{\mathbf{f}}\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i - (\widehat{\mathbf{f}} - \mathbf{f}\overline{\mathbf{\Lambda}})\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i + \mathbf{v}_i = \widehat{\mathbf{f}}^0\mathbf{D}_N^{-1}\overline{\mathbf{H}}^{-1}\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\lambda}_i + \mathbf{v}_i, \tag{A.58}$$

for $t \leq T_0$. If, on the other hand, $t > T_0$, then

$$\mathbf{x}_{i,t} = \boldsymbol{\tau}_{i,g,t} + \boldsymbol{\lambda}_i' \mathbf{f}_t + \mathbf{v}_{i,t} = \boldsymbol{\tau}_{i,g,t} + \boldsymbol{\lambda}_i' \overline{\mathbf{H}}_r' \overline{\mathbf{H}}^{-1\prime} \mathbf{D}_N^{-1} \widehat{\mathbf{f}}_t^0 - \boldsymbol{\lambda}_i' \overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t}. \tag{A.59}$$

These two last results imply

$$\begin{aligned}
&\mathbf{x}_{i,t} - \mathbf{x}_i' \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0 \\
&= \boldsymbol{\tau}_{i,g,t} + \boldsymbol{\lambda}_i' \overline{\mathbf{H}}_r' \overline{\mathbf{H}}^{-1\prime} \mathbf{D}_N^{-1} \widehat{\mathbf{f}}_t^0 - \boldsymbol{\lambda}_i' \overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t} - (\widehat{\mathbf{f}}^0 \mathbf{D}_N^{-1} \overline{\mathbf{H}}^{-1} \overline{\mathbf{H}}_r \boldsymbol{\lambda}_i - \overline{\mathbf{e}}_r^0 \boldsymbol{\lambda}_i + \mathbf{v}_i)' \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0 \\
&= \boldsymbol{\tau}_{i,g,t} - \boldsymbol{\lambda}_i' \overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t} - (-\overline{\mathbf{e}}_r^0 \boldsymbol{\lambda}_i + \mathbf{v}_i)' \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0,
\end{aligned} \tag{A.60}$$

and so we arrive at the following expression for $\widehat{\Delta}_{i,g,t}$:

$$\begin{aligned}
\widehat{\Delta}_{i,g,t} &= \eta_{i,g,t} + \boldsymbol{\beta}_i' (\boldsymbol{\tau}_{i,g,t} - \boldsymbol{\lambda}_i' \overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t}) - \boldsymbol{\alpha}_i' \overline{\mathbf{e}}_{r,t}^0 + \varepsilon_{i,t} \\
&\quad - [(-\overline{\mathbf{e}}_r^0 \boldsymbol{\lambda}_i + \mathbf{v}_i) \boldsymbol{\beta}_i - \overline{\mathbf{e}}_r^0 \boldsymbol{\alpha}_i + \varepsilon_i]' \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0 \\
&= \Delta_{i,g,t} - (\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)' \overline{\mathbf{e}}_{r,t}^0 + \boldsymbol{\beta}_i' \mathbf{v}_{i,t} + \varepsilon_{i,t} \\
&\quad - [-\overline{\mathbf{e}}_r^0 (\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i \boldsymbol{\beta}_i + \varepsilon_i]' \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0.
\end{aligned} \tag{A.61}$$

where $\Delta_{i,g,t} = \eta_{i,g,t} + \boldsymbol{\beta}_i' \boldsymbol{\tau}_{i,g,t}$ as defined at the end of Section 2.

The above expression for $\widehat{\Delta}_{i,g,t}$ is the cleanest possible without exploiting the fact that N is large. Hence, in what remains we are going to let $N \to \infty$. We begin by considering $\widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0$. Define $\mathbf{f}^+ = [\mathbf{f}_1^+, ... \mathbf{f}_{T_0}^+]' = [\mathbf{f}, \overline{\mathbf{e}}_{-r}^0]$, a $T_0 \times (m+1)$ matrix. We have already shown that $\widehat{\mathbf{f}}^0 = \mathbf{f}^+ + O_p(N^{-1/2})$. By using this and the results provided in the proof of Lemma A.1 in Westerlund et al. (2019), we have that $\|\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0 - \mathbf{f}^{+\prime} \mathbf{f}^+\| = O_p(N^{-1/2})$ and, more importantly,

$$\|(\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} - (\mathbf{f}^{+\prime} \mathbf{f}^+)^{-1}\| = O_p(N^{-1/2}), \tag{A.62}$$

where

$$\mathbf{f}^{+\prime} \mathbf{f}^+ = \begin{bmatrix} \mathbf{f}'\mathbf{f} & \mathbf{f}'\overline{\mathbf{e}}_{-r}^0 \\ \overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{f} & \overline{\mathbf{e}}_{-r}^{0\prime}\overline{\mathbf{e}}_{-r}^0 \end{bmatrix}, \tag{A.63}$$

$$(\mathbf{f}^{+\prime} \mathbf{f}^+)^{-1} = \begin{bmatrix} (\mathbf{f}'\mathbf{f})^{-1} + (\mathbf{f}'\mathbf{f})^{-1} \mathbf{f}'\overline{\mathbf{e}}_{-r}^0 (\overline{\mathbf{e}}_{-r}^{0\prime} \mathbf{M}_{\mathbf{f}} \overline{\mathbf{e}}_{-r}^0)^{-1} \overline{\mathbf{e}}_{-r}^{0\prime} \mathbf{f} (\mathbf{f}'\mathbf{f})^{-1} & -(\mathbf{f}'\mathbf{f})^{-1} \mathbf{f}'\overline{\mathbf{e}}_{-r}^0 (\overline{\mathbf{e}}_{-r}^{0\prime} \mathbf{M}_{\mathbf{f}} \overline{\mathbf{e}}_{-r}^0)^{-1} \\ -(\overline{\mathbf{e}}_{-r}^{0\prime} \mathbf{M}_{\mathbf{f}} \overline{\mathbf{e}}_{-r}^0)^{-1} \overline{\mathbf{e}}_{-r}^{0\prime} \mathbf{f} (\mathbf{f}'\mathbf{f})^{-1} & (\overline{\mathbf{e}}_{-r}^{0\prime} \mathbf{M}_{\mathbf{f}} \overline{\mathbf{e}}_{-r}^0)^{-1} \end{bmatrix}. \tag{A.64}$$

The expression for $(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}$ is again obtained by using the rules for the inverse of a partitioned matrix. The fact that $\|(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^{0})^{-1} - (\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\| = O_p(N^{-1/2})$ together with $\widehat{\mathbf{f}}^{0} = \mathbf{f}^{+} + O_p(N^{-1/2})$ imply that

$$\begin{aligned}\widehat{\mathbf{f}}_t^{0\prime}(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^{0})^{-1}\widehat{\mathbf{f}}^{0\prime} &= \widehat{\mathbf{f}}_t^{0\prime}[(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^{0})^{-1} - (\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}]\widehat{\mathbf{f}}^{0\prime} + \widehat{\mathbf{f}}_t^{0\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\widehat{\mathbf{f}}^{0\prime} \\ &= \widehat{\mathbf{f}}_t^{0\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\widehat{\mathbf{f}}^{0\prime} + O_p(N^{-1/2}) \\ &= \mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}^{+\prime} + O_p(N^{-1/2}). \end{aligned} \tag{A.65}$$

where, defining $\mathbf{M}_{\mathbf{f}}$ analogously to $\mathbf{M}_{\widehat{\mathbf{f}}'}$

$$\begin{aligned}&\mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}^{+\prime} \\ &= [\mathbf{f}_t', \bar{\mathbf{e}}_{-r,t}^{0\prime}] \begin{bmatrix} (\mathbf{f}'\mathbf{f})^{-1} + (\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'\bar{\mathbf{e}}_{-r}^{0}(\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\bar{\mathbf{e}}_{-r}^{0})^{-1}\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{f}(\mathbf{f}'\mathbf{f})^{-1} & \\ -(\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\bar{\mathbf{e}}_{-r}^{0})^{-1}\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{f}(\mathbf{f}'\mathbf{f})^{-1} & \end{bmatrix} \\ &\quad -(\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'\bar{\mathbf{e}}_{-r}^{0}(\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\bar{\mathbf{e}}_{-r}^{0})^{-1} \\ &\quad (\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\bar{\mathbf{e}}_{-r}^{0})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{f}' \\ \bar{\mathbf{e}}_{-r}^{0\prime} \end{bmatrix} \\ &= \mathbf{f}_t'(\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'[\mathbf{I}_{T_0} - \bar{\mathbf{e}}_{-r}^{0}(\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\bar{\mathbf{e}}_{-r}^{0})^{-1}\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}] + \bar{\mathbf{e}}_{-r,t}^{0\prime}(\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\bar{\mathbf{e}}_{-r}^{0})^{-1}\bar{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}. \end{aligned} \tag{A.66}$$

The fact that $\|\widehat{\mathbf{f}}_t^{0\prime}(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^{0})^{-1}\widehat{\mathbf{f}}^{0\prime} - \mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}^{+\prime}\| = O_p(N^{-1/2})$ implies

$$\begin{aligned}\widehat{\Delta}_{i,g,t} &= \Delta_{i,g,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\bar{\mathbf{e}}_{r,t}^{0} + \boldsymbol{\beta}_i'\mathbf{v}_{i,t} + \varepsilon_{i,t} \\ &\quad - [-\bar{\mathbf{e}}_{r}^{0}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i]'\mathbf{f}^{+}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}_t^{+} + O_p(N^{-1/2}) \\ &= \Delta_{i,g,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\bar{\mathbf{e}}_{r,t}^{0*} + \boldsymbol{\beta}_i'\mathbf{v}_{i,t}^{*} + \varepsilon_{i,t}^{*} + O_p(N^{-1/2}), \end{aligned} \tag{A.67}$$

where

$$\mathbf{a}_{i,t}^{*} = \mathbf{a}_{i,t} - \mathbf{a}_i'\mathbf{f}^{+}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}_t^{+} = \mathbf{a}_{i,t} - \sum_{s=1}^{T_0}\mathbf{a}_{i,s}\mathbf{f}_s^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}_t^{+} \tag{A.68}$$

for any vector $\mathbf{a}_{i,t}$ with T_0-rowed stack $\mathbf{a}_i = [\mathbf{a}_{i,1}, ..., \mathbf{a}_{i,T_0}]'$. In words, $\mathbf{a}_{i,t}^{*}$ is the limiting "defactored" version of $\mathbf{a}_{i,t}$.

We now make use of the above expression for $\widehat{\Delta}_{i,g,t}$ to evaluate $\widehat{\Delta}_{g,t}$. In so doing, it is important to note that the order of the reminder incurred when replacing $\widehat{\mathbf{f}}_t^{0\prime}(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^{0})^{-1}\widehat{\mathbf{f}}^{0\prime}$ with $\mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}^{+\prime}$ is the same even after averaging over group g and multiplying by $\sqrt{N_g}$. In order to appreciate this, we make use of the fact that $\|\sqrt{N_g}\bar{\mathbf{e}}_{r}^{0}\| = O_p(1)$, and since $\mathbf{v}_i$ and $\boldsymbol{\beta}_i$ are independent with $\mathbf{v}_i$ mean zero and independent also across i, we also have $\|N_g^{-1/2}\sum_{i=1}^{N}D_{i,g}\mathbf{v}_i\boldsymbol{\beta}_i\| =$

$O_p(1)$. It follows that

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}[-\overline{\mathbf{e}}_r^0(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i] \right\| \\
&\leq \|\sqrt{N_g}\overline{\mathbf{e}}_r^0\| \left\| \frac{1}{N_g} \sum_{i=1}^{N} D_{i,g}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) \right\| + \left\| \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}\mathbf{v}_i\boldsymbol{\beta}_i \right\| + \left\| \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}\boldsymbol{\varepsilon}_i \right\| = O_p(1).
\end{aligned} \tag{A.69}$$

We can therefore show that

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}[-\overline{\mathbf{e}}_r^0(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i]'[\mathbf{f}^+(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\mathbf{f}_t^+ - \widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0] \right\| \\
&\leq \left\| \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}[-\overline{\mathbf{e}}_r^0(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i] \right\| \|\mathbf{f}^+(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\mathbf{f}_t^+ - \widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0\| \\
&= O_p(N^{-1/2}),
\end{aligned} \tag{A.70}$$

which means that the reminder incurred when replacing $\widehat{\mathbf{f}}_t^{0\prime}(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}^{0\prime}$ with $\mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\mathbf{f}^{+\prime}$ is $O_p(N^{-1/2})$ after averaging over group g and multiplying by $\sqrt{N_g}$.

For $\Delta_{i,g,t}$, we make use of the fact that $\Delta_{i,g,t} = \Delta_{g,t} + v_{i,t}$ and $\boldsymbol{\tau}_{i,g,t} = \boldsymbol{\tau}_{g,t} + \boldsymbol{\zeta}_{i,t}$ for $i \in \mathcal{I}_g$ by Assumption 3, giving

$$\begin{aligned}
\Delta_{i,g,t} &= \eta_{i,g,t} + \boldsymbol{\beta}_i'\boldsymbol{\tau}_{i,g,t} = \Delta_{g,t} + v_{i,t} + (\boldsymbol{\beta} + \boldsymbol{\nu}_i)'(\boldsymbol{\tau}_{g,t} + \boldsymbol{\zeta}_{i,t}) \\
&= \Delta_{g,t} + \boldsymbol{\beta}'\boldsymbol{\tau}_{g,t} + v_{i,t} + \boldsymbol{\beta}'\boldsymbol{\zeta}_{i,t} + \boldsymbol{\nu}_i'(\boldsymbol{\tau}_{g,t} + \boldsymbol{\zeta}_{i,t}) \\
&= \Delta_{g,t}^0 + v_{i,t}^0,
\end{aligned} \tag{A.71}$$

where $\Delta_{g,t}^0 = \Delta_{g,t} + \boldsymbol{\beta}'\boldsymbol{\nu}_{i,t}$ and $v_{i,t}^0 = v_{i,t} + \boldsymbol{\beta}'\boldsymbol{\zeta}_{i,t} + \boldsymbol{\nu}_i'\boldsymbol{\tau}_{i,g,t}$.

By putting everything together,

$$\begin{aligned}
\sqrt{N_g}(\widehat{\Delta}_{g,t} - \Delta_{g,t}^0) &= \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}(\widehat{\Delta}_{i,g,t} - \Delta_{g,t}^0) \\
&= \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}(\widehat{\Delta}_{i,g,t} - \Delta_{i,g,t}^0 + v_{i,t}^0) \\
&= \frac{1}{\sqrt{N_g}} \sum_{i=1}^{N} D_{i,g}[v_{i,t}^0 - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\overline{\mathbf{e}}_{r,t}^{0*} + \boldsymbol{\beta}_i'\mathbf{v}_{i,t}^* + \varepsilon_{i,t}^*] + O_p(N^{-1/2}).
\end{aligned} \tag{A.72}$$

Moreover, Assumption 2 gives us $N_g/N \to_p \mathbb{P}(D_{i,g} = 1)$. Hence, if we also define $\mathbf{a}_g =$

$\lim_{N\to\infty} N_g^{-1}\sum_{i=1}^N D_{i,g}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i)$, the above expression for $\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta^0_{g,t})$ becomes

$$\begin{aligned}
&\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta_{g,t})\\
&=\frac{1}{\sqrt{N_g}}\sum_{i=1}^N D_{i,g}(v^0_{i,t}+\boldsymbol{\beta}_i'\mathbf{v}^*_{i,t}+\varepsilon^*_{i,t})-\sqrt{\frac{N_g}{N}}\frac{1}{N_g}\sum_{i=1}^N D_{i,g}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i)'\sqrt{N}\overline{\mathbf{e}}^{0*}_{r,t}+O_p(N^{-1/2})\\
&=\frac{1}{\sqrt{N_g}}\sum_{i=1}^N D_{i,g}(v^0_{i,t}+\boldsymbol{\beta}_i'\mathbf{v}^*_{i,t}+\varepsilon^*_{i,t})-\sqrt{\mathbb{P}(D_{i,g}=1)}\mathbf{a}_g'\sqrt{N}\overline{\mathbf{e}}^{0*}_{r,t}+o_p(1). \qquad \text{(A.73)}
\end{aligned}$$

All the terms on the right-hand side of the above equation are mean zero and independent across i (conditionally on $\mathbf{f}$). They are therefore asymptotically normal by a central limit law for independent variables. However, they are not uncorrelated with each other, which complicates the calculation of the asymptotic variance. Let us therefore define $\omega^2_{g,t}=\text{var}(\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta_{g,t})|\mathcal{C})$, where $\mathcal{C}$ is the sigma-field generated by $\mathbf{f}$. The asymptotic distribution of $\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta_{g,t})$ as $N\to\infty$ can now be stated in the following way:

$$\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta_{g,t})\to_d MN(0,\omega^2_{g,t}), \qquad \text{(A.74)}$$

where $MN(\cdot,\cdot)$ signifies a mixed normal distribution that is normal conditionally on $\mathcal{C}$. This means that the conditional distribution of $\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta_{g,t})/\omega_{g,t}$ is also the unconditional distribution. Hence,

$$\frac{\sqrt{N_g}(\widehat{\Delta}_{g,t}-\Delta_{g,t})}{\omega_{g,t}}\to_d N(0,1), \qquad \text{(A.75)}$$

as required for part (a).

It remains to prove (b) and the consistency of $\widehat{\omega}^2_{g,t}$. From before,

$$\widehat{\Delta}_{i,g,t}=\Delta^0_{g,t}+v^0_{i,t}-(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0*}_{r,t}+\boldsymbol{\beta}_i'\mathbf{v}^*_{i,t}+\varepsilon^*_{i,t}+O_p(N^{-1/2}), \qquad \text{(A.76)}$$

$$\frac{1}{N_g}\sum_{i\in\mathcal{I}_g}\widehat{\Delta}_{i,g,t}=\Delta^0_{g,t}+\frac{1}{N_g}\sum_{i=1}^N D_{i,g}[v^0_{i,t}-(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0*}_{r,t}+\boldsymbol{\beta}_i'\mathbf{v}^*_{i,t}+\varepsilon^*_{i,t}]+O_p(N^{-1/2}). \qquad \text{(A.77)}$$

It follows that if we let $z_{i,t}=v^0_{i,t}-(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i+\boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0*}_{r,t}+\boldsymbol{\beta}_i'\mathbf{v}^*_{i,t}+\varepsilon^*_{i,t}$, then

$$\widehat{\Delta}_{i,g,t}-\frac{1}{N_g}\sum_{j=1}^N D_{j,g}\widehat{\Delta}_{j,g,t}=z_{i,t}-\frac{1}{N_g}\sum_{j=1}^N D_{j,g}z_{j,t}+O_p(N^{-1/2}). \qquad \text{(A.78)}$$

Hence, since $d_{i,t}$ is again independent across i, by a law of large numbers for independent variables,

$$\begin{aligned}\widehat{\omega}_{g,t}^2 &= \frac{1}{N_g - 1}\sum_{i=1}^{N} D_{i,g}\left(\widehat{\Delta}_{i,g,t} - \frac{1}{N_g}\sum_{j=1}^{N} D_{j,g}\widehat{\Delta}_{j,g,t}\right)^2 \\ &= \frac{1}{N_g - 1}\sum_{i=1}^{N} D_{i,g}\left(z_{i,t} - \frac{1}{N_g}\sum_{j=1}^{N} D_{j,g}z_{j,t}\right)^2 + O_p(N^{-1/2}) \to_p \omega_{g,t}^2 \end{aligned} \tag{A.79}$$

as $N \to \infty$ (see Pesaran, 2006, page 985, for a similar argument). This establishes part (b) and hence the proof of the theorem is complete. ■