SIMPLE DIFFERENCE-IN-DIFFERENCES ESTIMATION IN FIXED-T PANELS[*]

Nicholas Brown
Queen's University

Kyle Butts
University of Colorado Boulder

Joakim Westerlund[†]
Lund University
and
Deakin University

May 24, 2023

Abstract

The present paper proposes a new treatment effects estimator that is valid when the number of time periods is small, and the parallel trends condition holds conditional on covariates and unobserved heterogeneity in the form of interactive fixed effects. The estimator also allow the control variables to be affected by treatment and it enables estimation of the resulting indirect effect on the outcome variable. The asymptotic properties of the estimator are established and their accuracy in small samples is investigated using Monte Carlo simulations. The empirical usefulness of the estimator is illustrated using as an example the effect of increased trade competition on firm markups in China.

JEL Classification: C31, C33, C38.

Keywords: Difference-in-differences; interactive fixed effects; common correlated effects; fixed-T.

1 Introduction

A key assumption in treatment effects studies is that there cannot be any unobserved systematic differences between treated and untreated cross-sectional units in absence of treatment.

[*] 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.

This is the so-called “parallel trend” assumption, which has long been acknowledged to be controversial in practice. Yet there have been surprisingly few formal attempts to resolve the issue, despite the huge empirical literature that has emerged. The standard approach in the panel data context is to assume that any non-parallel trending can be captured using fixed effects. But then this assumption is known to be restrictive. Interactive effects can be used to allow for more general types of non-parallel trending. Here the time effects, or “common factors”, represent common trends and the individual effects, or “ factor loadings”, measure the extent to which the impact of these trends is equal, or parallel, across units.

Chan and Kwok (2022) allow for non-parallel trending in the form of interactive effects that are dealt with using a version of the principal components-based approach of Bai (2009). However, this method requires that the number of time periods, T, is large, and in treatment effects studies T is often small (see Bertrand et al., 2004, for a survey). The approach also requires solving a non-convex optimization problem, which means that it is not only computationally costly but it can also be difficult to get to converge, and even if it does converge it may not be to the global optimum (see Moon and Weidner, 2019). Callaway and Karami (2020) and Brown and Butts (2022) provide treatment effects estimators that are valid even if T is small. However, these estimators are based on generalized method of moments (GMM), which is computationally burdensome and rely on the availability of certain external instruments. Both estimators require that the number of unobserved factors is know, which is of course never the case in practice.

In this paper, we propose a new treatment effects estimator that is not only valid when T is fixed and the number of factors is unknown but that is also extremely simple to implement. Moreover, unlike most existing estimators, the new estimator is applicable even if the covariates are affected by the treatment status, which is likely to be the case in practice (see Caetano et al., 2022, for a discussion). It is therefore very attractive from an empirical point of view. This attractiveness is achieved by our novel use of the common correlated effects (CCE) approach of Pesaran (2006), which has a closed form, does not require T to be large and is valid provided only that the number of factors is not larger than the number of observables.

The object of interest is the average treatment effect on the treated (ATT), which is the average difference between the actual and counterfactual post-treatment outcomes of treated cross-section units. This average could be computed had it not been for the fact that the counterfactual outcome is unobserved. We therefore have to estimate, or "impute", it and this is where the CCE approach come in. The proposed CCE-based difference-in-differences (DD) estimator, dubbed "C^2ED^2" and pronounced "Cetoo-E-Detoo", is computed in four steps.[1] We begin by estimating the common factors using cross-sectional averages of the outcome variable and covariates from the never-treated sample, as prescribed by CCE. We then estimate the slope coefficients of the controls along with the heterogeneous factor loadings conditional on the first-step factor estimates. In the third step, we use the first- and second-step estimates to estimate untreated covariates in post-treatment periods. In the fourth and final step, we use the first- and second-step estimates together with the third-step estimated covariates to estimate counterfactual outcomes. The estimated ATT is the average difference between the observed treated and estimated counterfactual outcomes.

The new estimator is shown to be consistent and asymptotically normal under very general condition provided only that the number of cross-sectional units, N, is large enough, a results that is verified in finite samples by means of a small-scale Monte Carlo simulation study. As an empirical illustration, we consider as an example the effect of increased trade competition on the dispersion of markups in China.

The rest of the paper is structured as follows. Section 2 presents the model and defines the ATT, the estimation of which is the concern of Section 3. Sections 4, 5 and 6 contain the asymptotic, Monte Carlo and empirical studies, respectively. Section 7 concludes. All proofs are relegated to the online appendix.

2 The model

We are interested in estimating the ATT 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

[1] The name and its pronunciation are inspired by the Star Wars robot character R2-D2.

the possibility that the N units can be divided into groups within which treatment timing is the same. We follow Callaway and Sant'Anna (2021) in defining a treatment group by the time period in which they enter treatment. There are G such groups indexed by $g \in \mathcal{G} \subset \{2, ..., T\}$, which for notational convenience is also the period at which the units of group g enter treatment. Hence, if $\mathcal{G} = \{4, 8\}$, then there are $|\mathcal{G}| = 2$ groups, the first (second) of which enter treatment in time period $g = t = 4$ ($g = t = 8$). Treated units never leave their groups but remain exposed for all periods after entering treatment; that is, treatment is of the "absorbing state". A unit that is never treated is a member of group $g = \infty$. Treatment timing is randomly assigned conditional on the unobserved interactive effects. Let us therefore denote by $g_i \in \mathcal{G}^+ = \mathcal{G} \cup \{\infty\}$ a random variable stating the group membership of cross-sectional unit i, and by $\mathcal{I}_g = \{i : g_i = g \in \mathcal{G}^+\} \subset \{1, ..., N\}$ the set of cross-sectional units that are members of group g. The set of non-treated units is therefore denoted $\mathcal{I}_\infty$, and it is convenient to let $\mathcal{I}_\infty^c = \{1, ..., N\} \backslash \mathcal{I}_\infty$ denote the set of treated units. The number of cross-sectional unit within group g is given by $|\mathcal{I}_g|$. The start of the first treatment is henceforth denoted $g_{\min} = \min\{g_1, ..., g_N\}$.

Following the previous literature, we denote by $y_{i,t}(g)$ the "potential" outcome of cross-sectional unit i in period t when member of group $g \in \mathcal{G}^+$. Of course, we do not observe $y_{i,t}(g)$ simultaneously for all g; instead we observe $y_{i,t} = y_{i,t}(g_i)$, the realized outcome for unit i at time t. We may also observe covariates, whose outcome may again depend on treatment status. In our empirical application, the outcome variable is industry-level markup dispersion, treatment is China's ascension into WTO, and a key control variable is the dispersion in marginal-cost. Our analysis allows treatment to affect the dispersion of both prices and marginal-cost and quantify the effect of markup-dispersion on the outcome.

Let us therefore introduce the $m \times 1$ vector $\mathbf{x}_{i,t}(g)$, whose realized value is given by $\mathbf{x}_{i,t} = \mathbf{x}_{i,t}(g_i)$. The model for $y_{i,t}(\infty)$ that we will be considering is given by

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

where $\boldsymbol{\beta}_i$ is a $m \times 1$ vector of heterogeneous 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.[2] The interactive effects are given here by $\boldsymbol{\alpha}_i'\mathbf{f}_t$. The purpose of these is to capture non-parallel trending behaviour, that is, unobserved differences in trends between treated and untreated units in absence of treatment. 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.[3] Accurate estimation of $\boldsymbol{\alpha}_i$ is therefore not needed.

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. The C^{2}ED^2 approach enables inference on $\boldsymbol{\beta}$ but the main object of interest is as already pointed out the ATT.

We want to entertain the possibility that $\mathbf{x}_{i,t}(\infty)$ load on $\mathbf{f}_t$, because otherwise the factors can be ignored without cost.[4] Also, many variables are affected by common shocks, and it is not difficult to find empirical evidence in support of this (see, for example, Westerlund et al., 2019). Let us therefore assume that

$$\mathbf{x}_{i,t}(\infty) = \boldsymbol{\lambda}_i'\mathbf{f}_t + \mathbf{v}_{i,t}, \tag{2}$$

where $\boldsymbol{\lambda}_i$ is a $r \times m$ matrix of factor loadings and $\mathbf{v}_{i,t}$ is a $m \times 1$ vector of idiosyncratic errors.

We are now ready to introduce the ATT. The treatment effect for unit i at time t when treated in time g is given by

$$\Delta_{i,g,t} = y_{i,t}(g) - y_{i,t}(\infty), \tag{3}$$

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 hope for is accurate estimation of

[2] The presence of $\boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty)$ in (1) is an allowance and not a requirement. If there are no regressors, we define $\boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) = 0$. 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.

[3] In fact, inference on $\boldsymbol{\alpha}_i$ and $\mathbf{f}_t$ is not even possible, as they are not separately identifiable.

[4] If $\mathbf{x}_{i,t}(\infty)$ does not load on $\mathbf{f}_t$, $\boldsymbol{\beta}_i$ can be estimated by ordinary least squares (OLS) as in Wooldridge (2005).

the ATT, which is the average $\Delta_{i,g,t}$ for group g;

$$\mathbb{E}(\Delta_{i,g,t}|g_i = g) = \Delta_{g,t} \tag{4}$$

for $t \geq g \in \mathcal{G}$. 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 $y_{i,t}(g) = y_{i,t}(\infty)$ for all not-yet-treated observations $t < g \in \mathcal{G}$.[5]

Most studies assume that the covariates are unaffected by the treatment and in this case the model for $y_{i,t}(g)$ can be obtained by simply inserting (1) into (3) (see, for example, Chan and Kwok, 2022). In the present paper, however, there is no such assumption. In order to be able to separate the part of the ATT that is due to the covariates from the part that is not, we define $\boldsymbol{\tau}_{i,g,t} = \mathbf{x}_{i,t}(g) - \mathbf{x}_{i,t}(\infty)$ and $\eta_{i,g,t} = \Delta_{i,g,t} - \boldsymbol{\tau}'_{i,g,t}\boldsymbol{\beta}_i$. In the terminology of the mediation literature (see, for example, Huber, 2014), $\eta_{i,g,t}$ is the "direct" effect of treatment and $\boldsymbol{\tau}'_{i,g,t}\boldsymbol{\beta}_i$ is the mediated effect of treatment through the covariates, henceforth referred to as the "indirect" effect. Hence, provided that $\boldsymbol{\tau}_{i,g,t}$ and $\boldsymbol{\beta}_i$ are independent, defining $\eta_{g,t} = \mathbb{E}(\eta_{i,g,t}|g_i = g)$ and $\boldsymbol{\tau}_{g,t} = \mathbb{E}(\boldsymbol{\tau}_{i,g,t}|g_i = g)$, the total ATT can be decomposed as follows:

$$\Delta_{g,t} = \eta_{g,t} + \boldsymbol{\tau}'_{g,t}\boldsymbol{\beta}, \tag{5}$$

where $\eta_{g,t}$ and $\boldsymbol{\tau}'_{g,t}\boldsymbol{\beta}$ are the direct and indirect ATTs, respectively.

3 The C^{2}ED^2 estimator

3.1 The total ATT

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 (4) by estimates. Note first that since $y_{i,t}(g)$ is observed for treated units in post-treatment periods,

[5] If treated units anticipate treatment up to s periods before g, shift treatment timing to $g - s$.

we have $y_{i,t} = y_{i,t}(g)$ for treated units post-treatment. Let us therefore turn to $y_{i,t}(\infty)$. We need to estimate this counterfactual for all treated units in post-treatment periods. CCE takes cross-sectional averages of the outcome and covariates as estimators of (the space spanned by) the factors. We tailor this procedure to the present treatment effect scenario where treatment status can affect both outcomes and covariates in unspecified ways. We use never-treated observations to estimate the factors. Then, for the treated units, we estimate the never-treated potential covariates, which are in turn used to estimate the never-treated potential outcomes. This method is detailed in the following four-step procedure to the estimation of $y_{i,t}(\infty)$.

Counterfactual estimation procedure:

1. Compute

   $$\widehat{\mathbf{f}}_t = \frac{1}{|\mathcal{I}_\infty|} \sum_{i \in \mathcal{I}_\infty} \mathbf{z}_{i,t} \tag{6}$$

   for all t, where $\mathbf{z}_{i,t} = [y_{i,t}, \mathbf{x}'_{i,t}]'$ is a $(m+1) \times 1$ vector containing all the observables. The above is the regular CCE estimator of $\mathbf{f}_t$ computed using the never-treated units only. The fact that $\widehat{\mathbf{f}}_t$ is computed based on the never-treated units only 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 t. In step 2 the pre-treatment estimates are used to estimate $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha}_i\}_{i=1}^N$, while in steps 3 and 4 the post-treatment estimates are used to impute $y_{i,t}(\infty)$ and $\mathbf{x}_{i,t}(\infty)$ in treatment periods.

2. Estimate the following regression by ordinary least squares (OLS) for all i and $t < g_{\min}$, where $g_{\min}$ again marks the start of the first treatment:

   $$y_{i,t} = \boldsymbol{\beta}'\mathbf{x}_{i,t} + \mathbf{a}_i'\widehat{\mathbf{f}}_t + u_{i,t}. \tag{7}$$

   Also, $\mathbf{a}_i$ 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 pre-treatment sample but where $\widehat{\mathbf{f}}_t$ comes from the subsample of untreated units.[6] Define the $(g_{\min}-1)\times 1$ vector $\mathbf{y}_i=[y_{i,1},...,y_{i,g_{\min}-1}]'$, and the $(g_{\min}-1)\times m$ matrices $\mathbf{x}_i=[\mathbf{x}_{i,1},...,\mathbf{x}_{i,g_{\min}-1}]'$ and $\widehat{\mathbf{f}}=[\widehat{\mathbf{f}}_1,...,\widehat{\mathbf{f}}_{g_{\min}-1}]'$. Let $\mathbf{M_A}=\mathbf{I}_{g_{\min}-1}-\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ for any $(g_{\min}-1)$-rowed matrix $\mathbf{A}$. In this notation, the CCE estimators of $\boldsymbol{\beta}$ and $\mathbf{a}_i$ in (7) are given by

$$\widehat{\boldsymbol{\beta}}=\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{x}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{M}_{\widehat{\mathbf{f}}}\mathbf{y}_i, \tag{8}$$

$$\widehat{\mathbf{a}}_i=(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{y}_i-\mathbf{x}_i\widehat{\boldsymbol{\beta}}), \tag{9}$$

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

3. Compute

$$\widehat{\mathbf{x}}_{i,t}(\infty)=\widehat{\boldsymbol{\lambda}}_i'\widehat{\mathbf{f}}_t \tag{10}$$

for all treated observations $i\in\mathcal{I}_\infty^c$ and $t\geq g_i$. Here, $\{\widehat{\mathbf{f}}_t\}_{t\geq g_{\min}}$ is from step 1 and

$$\widehat{\boldsymbol{\lambda}}_i=(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'\mathbf{x}_i, \tag{11}$$

where $\widehat{\mathbf{f}}$ and $\mathbf{x}_i$ are the same as in step 2. Note that $\widehat{\boldsymbol{\lambda}}_i$ is the OLS estimator of $\boldsymbol{\lambda}_i$ in the following regression, which is estimated for each $i\in\mathcal{I}_\infty^c$ individually and $t<g_{\min}$:

$$\mathbf{x}_{i,t}=\boldsymbol{\lambda}_i'\widehat{\mathbf{f}}_t+\mathbf{w}_{i,t}, \tag{12}$$

where $\mathbf{w}_{i,t}=\boldsymbol{\lambda}_i'(\mathbf{f}_t-\widehat{\mathbf{f}}_t)+\mathbf{v}_{i,t}$.

4. The sought counterfactual estimator is given by

$$\widehat{y}_{i,t}(\infty)=\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{x}}_{i,t}(\infty)+\widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t \tag{13}$$

which is again available for all treated observations. Here $\widehat{\boldsymbol{\beta}}$ and $\{\widehat{\mathbf{a}}_i\}_{i\in\mathcal{I}_\infty^c}$ are from step 2, $\{\widehat{\mathbf{f}}_t\}_{t\geq g_{\min}}$ is from step 1, and $\{\widehat{\mathbf{x}}_{i,t}(\infty)\}_{i\in\mathcal{I}_\infty^c,t\geq g_i}$ comes from step 3.

[6] 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 $\{\widehat{\mathbf{f}}_t\}_{t\geq g_{\min}}$ can be taken directly from step 1.

A few remarks are in order. First, 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.

Second, one can allow $\boldsymbol{\beta}$ to vary systematically across groups without affecting the asymptotic validity of the estimated ATT. 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 $\{\widehat{\boldsymbol{\beta}}_g\}_{g \in \mathcal{G}}$, which should then be inserted instead of $\widehat{\boldsymbol{\beta}}$ in step 3.

Third, as Caetano et al. (2022) point out, the validity of estimates of the ATT depends on whether or not the covariates are affected by treatment status. For example, if we are estimating the effect of a certain policy aimed at reducing unemployment, we might want to control for the rate of poverty. But then such policies might indirectly reduce poverty, which means that the poverty rate covariate will absorb some of the treatment effect. This is what Angrist and Pischke (2009) call a "bad control". It creates a dilemma where including the covariate induces "post-treatment bias" and excluding it induces "omitted variables bias" (see Aklin and Bayer, 2017). In this paper we follow Caetano et al. (2022), and solve this dilemma by imputing and controlling for untreated potential covariates, $\mathbf{x}_{i,t}(\infty)$. In fact, we go a step further and allow for inference in this indirect effect.

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

$$\widehat{\Delta}_{i,g,t} = y_{i,t} - \widehat{y}_{i,t}(\infty) \tag{14}$$

for $i \in \mathcal{I}_g \subset \mathcal{I}_\infty^c$. The estimated ATT for group g at time t is obtained by averaging over the relevant treated group;

$$\widehat{\Delta}_{g,t} = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \widehat{\Delta}_{i,g,t}. \tag{15}$$

This is the C^{2}ED^2 estimator of $\Delta_{g,t}$.

It is important to note that the C^{2}ED^2 estimator does not involve any estimation of the number of factors, r. This is in stark contrast to existing principal components-based approaches such those of Chan and Kwok (2022), and Xu (2017), and GMM approaches such as those of Callaway and Karami (2020), and Brown and Butts (2022), where 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.

Asymptotic standard errors of estimates of the ATT are generally difficult to compute. 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 variance estimator considered by Pesaran (2006). The appropriate estimator to use in our case is

$$\widehat{\sigma}^2(\widehat{\Delta}_{g,t}) = \frac{1}{|\mathcal{I}_g| - 1} \sum_{i \in \mathcal{I}_g} (\widehat{\Delta}_{i,g,t} - \widehat{\Delta}_{g,t})^2. \tag{16}$$

In addition to being simple to compute, non-parametric standard errors are robust and they tend to perform well in small samples (see, for example, Chudik et al., 2011, Pesaran, 2006, and Westerlund and Kaddoura, 2022).

3.2 The direct and indirect ATTs

We demonstrate in Section 2 how the total ATT $\Delta_{i,g}$ can be decomposed into the direct ATT, $\eta_{i,g}$, and the indirect ATT, $\boldsymbol{\tau}_{i,g}'\boldsymbol{\beta}$. We now demonstrate how to estimate these constituent parts.

The estimator of $\boldsymbol{\tau}_{g,t}$ is completely analogous to that of $\Delta_{g,t}$, and is given by

$$\widehat{\boldsymbol{\tau}}_{g,t} = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} \widehat{\boldsymbol{\tau}}_{i,g,t}, \tag{17}$$

where $\widehat{\boldsymbol{\tau}}_{i,g,t} = \mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)$. In the empirical literature, significant estimates of $\widehat{\boldsymbol{\tau}}_{g,t}$ is sometimes taken as evidence of indirect treatment effects. However, even if the covariates are affected by treatment, this does not necessarily imply that the outcome is affected, as the effect of changing the covariates on the outcome is determined by their partial effects, here represented by $\boldsymbol{\beta}_i$. The proposed C^{2}ED^2 approach recognizes this possibility. Our estimate of the indirect ATT is given by the product $\widehat{\boldsymbol{\tau}}_{g,t}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is from step 2 of the counterfactual estimation procedure. Given $\widehat{\boldsymbol{\tau}}_{g,t}'\widehat{\boldsymbol{\beta}}$, the estimated direct ATT is given by

$$\widehat{\eta}_{g,t} = \widehat{\Delta}_{g,t} - \widehat{\boldsymbol{\tau}}_{g,t}'\widehat{\boldsymbol{\beta}}. \tag{18}$$

The variances of $\widehat{\boldsymbol{\tau}}_{g,t}$ and $\widehat{\eta}_{g,t}$ can be estimated non-parametrically in the following obvious way:

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\tau}}_{g,t}) = \frac{1}{|\mathcal{I}_g| - 1} \sum_{i \in \mathcal{I}_g} (\widehat{\boldsymbol{\tau}}_{i,g,t} - \widehat{\boldsymbol{\tau}}_{g,t})(\widehat{\boldsymbol{\tau}}_{i,g,t} - \widehat{\boldsymbol{\tau}}_{g,t})', \tag{19}$$

$$\widehat{\sigma}^2(\widehat{\eta}_{g,t}) = \frac{1}{|\mathcal{I}_g| - 1} \sum_{i \in \mathcal{I}_g} (\widehat{\eta}_{i,g,t} - \widehat{\eta}_{g,t})^2. \tag{20}$$

Note that $\widehat{\sigma}^2(\widehat{\eta}_{g,t})$ is a direct estimator of the variance of the estimated direct ATT. The corresponding estimator of the variance of the estimated indirect ATT is given by $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\tau}}_{g,t})\widehat{\boldsymbol{\beta}}$.

The above estimator of $\eta_{g,t}$ is of the plug-in type; it takes the definition of $\eta_{g,t}$ and plugs in estimates in places of true quantities. An alternative estimation approach is to take $\widehat{\Delta}_{g,t}$ but to replace $\widehat{\mathbf{x}}_{i,t}(\infty)$ with $\widehat{\mathbf{x}}_{i,t}$ when computing $\widehat{y}_{i,t}(\infty)$ in step 4 of the counterfactual estimation procedure. The fact that changing the way that the covariates enter in step 4 alters the object being estimated is important not only for the present paper but also when considering the works of others. As mentioned earlier, Chan and Kwok (2022) proposes a principal components-based estimator of the ATT that assumes that the covariates are unaffected by treatment and they use the observed covariates in their estimations. Logic based on our findings suggests that if the unaffected covariates assumption is false, Chan and Kwok's estimator will only capture the direct ATT. In the empirical illustration of Section 6, we elaborate on this point.

4 Asymptotic results

In this section, we study the asymptotic properties of the estimated total ATT and its direct and indirect parts. The conditions that we will be working under are given in Assumptions 1–9. Here and throughout, tr $\mathbf{A}$, rank $\mathbf{A}$ and $\|\mathbf{A}\| = \sqrt{\text{tr}\,(\mathbf{A}'\mathbf{A})}$ denote the trace, the rank, and the Frobenius (Euclidean) norm of the generic matrix $\mathbf{A}$, respectively. The symbols $\to_d$ and $\to_p$ signify convergence in distribution and probability, respectively.

Assumption 1. $g_{\min} > m + 2$.

Assumption 2. $\text{plim}_{N\to\infty}|\mathcal{I}_g|/N \in (0,1)$ for all $g \in \mathcal{G}^+$.

Assumptions 1 and 2 are sample size conditions. They ensure that g_{min} is large enough to ensure that the step-2 regression model in (7) 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 $|\mathcal{I}_g|$ is a random quantity.

Assumption 3. $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\nu}_i$, $\Delta_{i,g,t} = \Delta_{g,t} + \upsilon_{i,t}$, and $\boldsymbol{\tau}_{i,g,t} = \boldsymbol{\tau}_{g,t} + \boldsymbol{\zeta}_{i,t}$ where $\boldsymbol{\nu}_i$, $\upsilon_{i,t}$, 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 parameter condition that is largely the same as in Chan and Kwok (2022), and Gobillon and Magnac (2016). None of parameters are required to be heterogeneous, as the covariance matrices of $\boldsymbol{\nu}_i$, $\upsilon_{i,t}$ and $\boldsymbol{\zeta}_{i,t}$ need not be positive definite.

Before we continue onto Assumption 4, is it useful to first lay out some additional notation. Step 1 of the counterfactual estimation procedure 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 (1) and (2) we arrive at the following static factor model for $\mathbf{z}_{i,t}$:

$$\mathbf{z}_{i,t} = \boldsymbol{\Lambda}_i'\mathbf{f}_t + \mathbf{e}_{i,t}, \tag{21}$$

where $\boldsymbol{\Lambda}_i = [\boldsymbol{\alpha}_i + \boldsymbol{\lambda}_i\boldsymbol{\beta}_i, \boldsymbol{\lambda}_i]$ 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 $(m+1) \times 1$. This expression for $\mathbf{z}_{i,t}$ implies that $\widehat{\mathbf{f}}_t$ can be written in the following way:

$$\widehat{\mathbf{f}}_t = \frac{1}{|\mathcal{I}_\infty|}\sum_{i\in\mathcal{I}_\infty}\mathbf{z}_{i,t} = \frac{1}{|\mathcal{I}_\infty|}\sum_{i\in\mathcal{I}_\infty}\boldsymbol{\Lambda}_i'\mathbf{f}_t + \frac{1}{|\mathcal{I}_\infty|}\sum_{i\in\mathcal{I}_\infty}\mathbf{e}_{i,t}. \tag{22}$$

Assumptions 4–6 below ensure that the average $\mathbf{e}_{i,t}$ tends to zero as N increases and that the average $\boldsymbol{\Lambda}_i$ 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, $\varepsilon_{i,t}$, $\mathbf{v}_{i,t}$, ν_i, $\zeta_{i,t}$, and $\upsilon_{i,t}$ are mutually independent.

Assumption 6. $\text{rank}(|\mathcal{I}_\infty|^{-1} \sum_{i \in \mathcal{I}_\infty} \mathbf{\Lambda}_i) = r \leq m + 1$ almost surely.

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

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 $y_{i,t}$. If there are such regressors present, then these should be treated as observed factors, which can be appended to $\widehat{\mathbf{f}}_t$ in step 1 of the counterfactual estimation procedure, as we illustrate in Section 6. This is an advantage in the sense that while $\boldsymbol{\beta}_i$ and $\Delta_{i,g,t}$ are subject to the random parameter condition in Assumption 3, $\boldsymbol{\alpha}_i$ is not. Hence, unlike the coefficients of the observed covariates, 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. If there are common known deterministic terms, such as an intercept or a linear time trend, these can be inserted into $\widehat{\mathbf{f}}$ along with the cross-sectional averages. 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) $\dfrac{\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\sigma(\widehat{\Delta}_{g,t})} \to_d N(0,1)$,

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

where the definition of $\sigma^2(\widehat{\Delta}_{g,t})$ is provided in the appendix.

The proof of Theorem 1 is contained in the appendix, where we show that $\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})$ is asymptotically mixed normal, and that this implies that $\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})/\sigma(\widehat{\Delta}_{g,t})$ 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. The reason is 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 $\mathbf{f}_t$. The asymptotic distribution of $\widehat{\Delta}_{i,g,t} - \Delta_{i,g,t}$ is therefore 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 $\sigma^2(\widehat{\Delta}_{g,t})$ depends on whether $m + 1 = r$ or $m + 1 > r$, this dependence is successfully mimicked in large samples by $\widehat{\sigma}^2(\widehat{\Delta}_{g,t})$. We can therefore show that

$$\frac{\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\widehat{\sigma}^2(\widehat{\Delta}_{g,t})} = \frac{\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\sigma^2(\widehat{\Delta}_{g,t})} + o_p(1) \to_d N(0,1) \tag{23}$$

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).

The asymptotic distributions of the direct and indirect ATTs are a direct consequence of Theorem 1 and the consistency of $\widehat{\boldsymbol{\beta}}$, and are summarized in the following corollary.

Corollary 1. *Suppose that the conditions of Theorem 1 are met. Then, as $N \to \infty$,*

(a) $\dfrac{\sqrt{|\mathcal{I}_g|}(\widehat{\boldsymbol{\tau}}'_{g,t}\widehat{\boldsymbol{\beta}} - \boldsymbol{\tau}'_{g,t}\boldsymbol{\beta})}{\sqrt{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\tau}}_{g,t})\widehat{\boldsymbol{\beta}}}} \to_d N(0,1),$

(b) $\dfrac{\sqrt{|\mathcal{I}_g|}(\widehat{\eta}_{g,t} - \eta_{g,t})}{\widehat{\sigma}(\widehat{\eta}_{g,t})} \to_d N(0,1).$

5 Monte Carlo simulations

In this section, we present the results of a small-scale Monte Carlo study. The processes used to generate the potential treated outcome and covariates, $y_{i,t}(\infty)$ and $\mathbf{x}_{i,t}(\infty)$, respectively, are given by restricted versions of (1) and (2) that set $r = m = 2$ and $\mathbf{f}_t = [1,t]'$. Equation (2) is generated with $\boldsymbol{\lambda}_i = \mathbf{I}_2 + \mathbf{Z}_i$, where the elements of $\mathbf{Z}_i$ are drawn independently from $N(0,1)$, as are the elements of $\mathbf{v}_{i,t}$. Equation (1) is generated with $\boldsymbol{\beta}_i = \boldsymbol{\beta} = [1,1]'$ for all i and $\boldsymbol{\alpha}_i \sim \text{diag}(\boldsymbol{\lambda}_i) + \boldsymbol{\theta} d_i + N([0,0]',\mathbf{I}_2)$, where $d_i = 1(i \in \mathcal{I}^c_\infty)$ is a dummy that is one if cross-section unit i is treated and zero otherwise, and $\text{diag}(\boldsymbol{\lambda}_i)$ vectorizes the main diagonal of $\boldsymbol{\lambda}_i$. The term $\boldsymbol{\theta} d_i$ controls whether the parallel trend condition is met. If $\boldsymbol{\theta} = [0,0]'$, then $\mathbb{E}(\boldsymbol{\alpha}_i) = [1,1]'$ for all i and so trends are on average parallel, whereas if $\boldsymbol{\theta} = [0,1]'$, then $\mathbb{E}(\boldsymbol{\alpha}_i) = [1, 1+d_i]'$, and so the treated and untreated cross-sectional units are on different trend paths. The presence of $\text{diag}(\boldsymbol{\lambda}_i)$ makes $\boldsymbol{\alpha}_i$ correlated with $\boldsymbol{\lambda}_i$, which in turn means that $\mathbf{x}_{i,t}(g)$ is endogenous. The regression errors are allowed to be serially correlated through $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + u_{i,t}$, where $\varepsilon_{i,0} = 0$, $\rho = 0.75$ and $u_{i,t} \sim N(0,1)$.

The potential treated outcome and covariates are generated as $y_{i,t}(g) = \Delta_g + y_{i,t}(\infty)$ and $\mathbf{x}_{i,t}(g) = \boldsymbol{\tau}_g + \mathbf{x}_{i,t}(\infty)$, respectively, which means that in this data generating process the direct treatment effect is given by $\eta_g = \Delta_g - \boldsymbol{\tau}'_g\boldsymbol{\beta}$. We assume that there is just one treated group and randomly assign half of the cross-sectional units to this group. Consistent with the empirical illustration of Section 6, we set $N = 164$ and $T = 9$. Treatment starts in period seven, and so $g = g_{\min} = 7$. As for $\boldsymbol{\tau}_g$ and Δ_g, we consider two cases. In the first, $\Delta_g = 1$ and $\boldsymbol{\tau}_g = [0,0]'$, and therefore the direct ATT is given by $\eta_g = \Delta_g = 1$, whereas in the second, $\Delta_g = 2$ and $\boldsymbol{\tau}_g = [0,1]'$, which means that while $\eta_g = 1$ is the same as before, now there is also an indirect ATT equal to $\boldsymbol{\tau}'_g\boldsymbol{\beta} = 1$.

The C^{2}ED^2 procedure is implemented exactly as described in Section 3. We focus on the total ATT. The results for the direct and indirect ATTs were very similar and are available upon request. The C^{2}ED^2 results are compared to those obtained by using two-way fixed effects OLS with one treatment dummy for each of the three treatment periods, which represents the workhorse of the empirical treatment effects literature. We consider two specifications; one that accounts for the covariates and one that ignores them. For each estimator, we report the average bias and the mean squared error (MSE). The number of replications is set to 1,000.

INSERT TABLES 1 AND 2 ABOUT HERE

Tables 1 and 2 report the results for the cases when the parallel trend condition holds and when it fails, respectively. We begin by considering Table 1. Since in this case trends are parallel and the covariates are unaffected by the treatment, even the OLS estimator that omits the covariates is expected to be unbiased, which is just what we see in the table. The ranking of the three estimators in terms of MSE is also as expected with the C^{2}ED^2 estimator that accounts for both factors and covariates outperforming the competition. The covariate-augmented OLS estimator is biased when the indirect ATT is nonzero. This is due in part to the correlation between $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_i$, which causes an omitted variables bias when the covariates are included but the factors are not appropriately accounted for, in part to the fact that controlling for the covariates absorbs the indirect ATT, as pointed out in Section 3. According to Table 2, if trends are not parallel, OLS breaks down regardless of whether it is covariate-augmented or not, which is again just as expected because fixed effects OLS is inconsistent in this case.

6 Empirical illustration

One of the channels through which competition may affect gains from trade is via changes in markups, which measures the ability of firms to charge prices above their marginal costs. As is well-known, first-best efficiency is obtained when markups are the same across goods. Of course, in practice markups are never the same and this raises the possibility of so-called "pro-competitive" effects of trade, which is the idea that trade liberalization through increased competition drive down both the level and dispersion of markups, leading to increased efficiency. Moreover, welfare improves when consumers benefit from lower markups of the goods they consume and when producers gain from higher markups in foreign markets.

Since its accession into the World Trade Organization (WTO) in the end of 2001, China's role in the world economy has grown enormously. As a result, the pro-competitive effects of China's WTO accession have attracted considerable attention, so much so that there is by now a separate strand of literature devoted to them. The bulk of the evidence seem to suggest that both the level and dispersion of markups have gone down following the WTO entry, and that this development has had important welfare effects (see, for example, Hsu et al., 2020).

The purpose of the current application is to contribute to the above mentioned literature. This is done in two ways. First, we account for general forms of unobserved heterogeneity. The standard approach in the literature is to exploit differences in tariffs across industries. The basic idea is to split the sample of industries into a treatment and a control group, where the former is assumed to be relatively more exposed to the WTO accession. Given that pre-WTO tariffs varied greatly across industries, the argument goes on to say that industries that had previously been protected with relatively high tariffs experienced greater tariff reduction. They should therefore be relatively more exposed. The effect of the WTO accession is then estimated via a standard DD-style OLS regression in which markup is regressed onto a dummy variable that takes on the value one for treated industries in post-WTO periods, control variables, and industry and time fixed effects.

While popular, the standard approach to WTO evaluation has (at least) two drawbacks. One drawback is that it requires that in absence of treatment the difference between the treatment and control groups is constant over time. Trends therefore have to be parallel, which is known to be restrictive. A very commonly cited reason is that certain industries have more lobbying power for protection. Tariffs may be granted in response to domestic special interest groups, the pressure of which may vary over time (see, for example, Fan et al., 2018, Deng et al., 2018, and Xiang et al., 2017). Differences in lobbying power may therefore cause the treatment and control groups to differ systematically over time even if China had not joined the WTO in 2001.[7] Such differences are problematic as they render the fixed effects OLS estimator inconsistent, as the Monte Carlo results of Section 5 illustrates. The main problem is that many sources of possible non-parallel trending are unknown and lack good proxies. For this reason, in lack of better alternatives, it is common to control for industry-specific linear time trends (see, for example, Liu and Qiu, 2016, and Mao and Xu, 2019). Deterministic trends can account for some non-parallel trending but not all. Moreover, results tend to be highly sensitive to the inclusion of such trends, which reinforces the sentiment in the literature that non-parallel trending is an important issue.[8]

Another drawback of the standard approach is that it is not designed to deal with the case when both the outcome and the covariates are affected by treatment. This is important because the literature has identified many channels through which the WTO accession may affect markups (see Mao and Xu, 2019, Fan et al., 2018, Deng et al., 2018, Liu and Ma, 2021, and Brandt et al., 2017, to mention a few). Two common examples are the price- and cost-change channels. Markup is defined as the ratio of price to marginal cost. This means that markup changes can emanate from price changes, cost changes, or both. It is therefore common to include one of these variables as a covariate and also to estimate the effect of the WTO accession on them (see, for example, Mao and Xu, 2019, Fan et al., 2018, and Lu and Yu, 2015). But then we know from Section 3 that treatment-affected covariates require special treatment or else the estimated ATT will be misleading. Specifically, the inclusion of such covariates will absorb the indirect ATT. Some researchers seem to be aware of this. The following quotation, taken from Fan et al. (2018, page 116), is quite suggestive: "If the marginal-cost channel indeed plays a role, then once the marginal costs are included as an explanatory variable, we would witness attenuation of the impact of input tariffs on markups." However, it is not until recently that researchers in econometrics have considered the possibility of treatment-affected covariates, and there is still much to do (see Caetano et al., 2022). Empirical researchers there-

[7] Similarly, policymakers may lower tariffs selectively only in industries that are able to compete with relatively less expensive imports, for example, in industries experiencing a productivity boom (see Brandt et al., 2017).

[8] Some studies include common controls that are thought to be highly correlated with various kinds of protectionism, such as wage rates, employment, exports, and imports (see, for example, Hsu et al., 2020). Again the results tend to be very sensitive.

fore have little or no option but to either ignore the problem or to exclude all potentially bad controls from their specifications.

The present paper is not the first to point to these shortcomings, but it is the first to consider an econometric approach that is designed to deal with both in a rigorous way. The C^2ED^2 approach allows for interactive effects in which there may be unobserved differences between cross-sectional units that change over time as a result of common shocks. The parallel trend condition is therefore not required, which is a substantial advantage when compared to the standard fixed effects-based approach. Another advantage of the approach that we exploit in this section is that it not only allows for covariates that may be affected by treatment but that it makes it possible to assess the relative importance of the direct and indirect treatment channels. It should therefore be well suited for the problem at hand.

The data set that we use is taken from Lu and Yu (2015) (see also Deng et al., 2018, who use the same data), and comprise 164 industries (three-digit Chinese industrial classification) observed over the 1998–2005 period. The smallness of T here, which is a feature of most data sets in the literature, means that it is important to use techniques that work even if T is not large. The Monte Carlo results reported in Section 5 suggest that the proposed C^2ED^2 approach should work well. Following Lu and Yu (2015), the outcome variable is markup dispersion, as measured by the markup Theil index (in logs). Industries are assigned to the treatment and control groups based on whether they faced tariffs above or below the sample median in 2001.

Our preference to focus on the Lu and Yu (2015) study is motivated in part by their analysis of the price- and cost-change channels (see their Section E). As a proxy for marginal costs, the authors use productivity (TFP). The ATT is estimated via an OLS regression that in addition to fixed effects, controls and the treatment variable includes the TFP Theil index as a covariate to account for cost dispersion effects. The authors argue that this should allow them to partially isolate the price-change channel. In order to assess the ATT of the WTO accession on costs, the authors run a second OLS regression with the TFP Theil index as dependent variable and the treatment variable as a covariate. The estimated ATTs are significant, which is taken as evidence to suggest that both channels are operational. The purpose of this illustration is to assess the accuracy of this last conclusion.

The above discussion suggests that in terms of the notation of Section 2, in this section $y_{i,t}$ is the markup Theil index, and $\mathbf{x}_{i,t}$ is the TFP Theil index. The estimated factors in $\widehat{\mathbf{f}}_t$ are made up of the cross-sectional averages of these variables. A constant is included as an observed factor (as explained in Section 4), which is tantamount to allowing for industry fixed effects. We therefore allow for one known and two unknown factors.

INSERT FIGURE 1 ABOUT HERE

The estimated direct and indirect ATTs are reported in Figure 1. The estimates are reported for each year and averaged over all the post- and pre-treatment periods, as is customary in the literature. Both types are reported together with 95% confidence intervals. The first thing to note is that the both the direct and indirect ATTs are estimated to be negative, suggesting that markup dispersion decreased more after 2001 in industries with relatively high tariffs in 2001. Given that industries with higher tariffs in 2001 experienced greater tariff reduction after 2002, these results imply that the WTO accession reduced markup dispersion. We also note that the pre-treatment estimates are all very close to zero, which means that in this period there were no differences in the markup Theil index that depended on group membership.

While insignificant in 2002 and 2003, the year-specific total ATTs reported in Figure 1 (a) are significant in 2004 and 2005. The point estimate in 2003 is notably noisy. A possible reason for this is that the industry classification system changed in 2003, as noted by, for example, Chen et al. (2019), and Lu and Yu (2015). The estimated average ATT during the whole post-treatment period is about -0.1 and significant, which consistent with the results of Chen et al. (2019).

In order to assess to what extent the decrease in markup dispersion is due to decreases in TFP dispersion as predicted by the marginal cost channel we look at the estimated indirect ATTs. According to the results reported in Figure 1 (b), the estimated direct ATTs are negative and significant in the post-treatment period and insignificant in the pre-treatment period. Lu and Yu (2015) estimate the ATT on the TFP Theil index and find it to be significantly negative; however, their approach does not allow them to infer whether this negative response of the TFP Theil index has an effect on the markup Theil index. According to our results, the estimated indirect ATTs are sizable, accounting for almost half of the total ATTs. This is important in itself but also for what it means for the results reported by Lu and Yu (2015), which are based on including the TFP Theil index as a covariate. In particular, we know from before that this type of conditioning will absorb the indirect effect. In this case, since both ATTs are estimated to be negative and the magnitude of the indirect ATTs are about half of the direct ATTs, conditioning on the TFP Theil index will lead to an underestimation of the total ATTs by about 50%. This illustrates quite clearly the importance of being able to account for the fact that treatment may affect not only the outcome variable but also the covariates.

7 Conclusion

In this paper we propose a new ATT estimator dubbed "C^2ED^2" that is applicable even when the parallel trends condition fails because of the presence of unobserved heterogeneity in the form of interactive fixed effects. Our identification strategy, based on the popular CCE approach, relies on the presence of covariates that load on the same factors as the outcome variable. This allows us to use the cross-sectional averages of the observables to impute the untreated potential outcomes in post-treatment time periods. The covariates are allowed to depend on the treatment status, and if they do C^2ED^2 makes it possible separate the direct ATT that is unrelated to the covariates from the indirect ATT that works through those covariates. The estimator is shown to be consistent and asymptotically normal, thereby enabling standard inference, provided only that the number of cross-sectional units, N, is large, which is a great advantage in practice because in the literature many data sets involve only a few time periods.

References

AKLIN, M. AND P. BAYER (2017): "How can we estimate the effectiveness of institutions? Solving the post-treatment versus omitted variable bias dilemma," .

ANGRIST, J. D. AND J.-S. PISCHKE (2009): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

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

BERTRAND, M., E. DUFLO, AND S. MULLAINATHAN (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249–275.

BORUSYAK, K., X. JARAVEL, AND J. SPIESS (2021): "Revisiting event study designs: Robust and efficient estimation," ArXiv:2108.12419.

BRANDT, L., J. VAN BIESEBROECK, L. WANG, AND Y. ZHANG (2017): "WTO accession and performance of Chinese manufacturing firms," *American Economic Review*, 107, 2784–2820.

BREITUNG, J. AND P. HANSEN (2021): "Alternative estimation approaches for the factor augmented panel data model with small T," *Empirical Economics*, 60, 327–351.

BROWN, N. AND K. BUTTS (2022): "A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models," .

CAETANO, C., B. CALLAWAY, S. PAYNE, AND H. S. RODRIGUES (2022): "Difference in Differences with Time-Varying Covariates," ArXiv: 2202.02903.

CALLAWAY, B. AND S. KARAMI (2020): "Treatment Effects in Interactive Fixed Effects Models," ArXiv: 2006.15780.

CALLAWAY, B. AND P. H. SANT'ANNA (2021): "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, S0304407620303948.

CHAN, M. K. AND S. S. KWOK (2022): "The PCDID approach: difference-in-differences when trends are potentially unparallel and stochastic," *Journal of Business & Economic Statistics*, 40, 1216–1233.

CHEN, W., X. CHEN, C.-T. HSIEH, AND Z. SONG (2019): "A forensic examination of China's national accounts," *Brookings Papers on Economic Activity*, 77–141.

CHUDIK, A., M. H. PESARAN, AND E. TOSETTI (2011): "Weak and strong cross-section dependence and estimation of large panels," *The Econometrics Journal*, 14, C45–C90.

DENG, X., R. JING, AND Z. LIANG (2018): "Trade liberalisation and domestic brands: Evidence from China's accession to the WTO," *World Economy*, 43, 2237–2262.

FAN, H., X. GAO, Y. A. LI, AND T. A. LUONG (2018): "Trade liberalization and markups: Micro evidence from China," *Journal of Comparative Economics*, 46, 103–130.

GOBILLON, L. AND T. MAGNAC (2016): "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *Review of Economics and Statistics*, 98, 535–551.

HSU, W.-T., Y. LU, AND G. L. WU (2020): "Competition, markups, and gains from trade: A quantitative analysis of China between 1995 and 2004," *Journal of International Economics*, 122, 103266.

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

LEEB, H. AND B. M. PÖTSCHER (2005): "Model selection and inference: Facts and fiction," *Econometric Theory*, 21, 21–59.

LIU, Q. AND L. D. QIU (2016): "Intermediate input imports and innovations: Evidence from Chinese firms' patent filings," *Journal of International Economics*, 103, 166–6–183.

LIU, Z. AND H. MA (2021): "Input trade liberalisation and markup distribution: Evidence from China," *Economic Inquiry*, 59, 344–360.

LU, Y. AND L. YU (2015): "Trade liberalization and markup dispersion: evidence from China's WTO accession," *American Economic Journal: Applied Economics*, 7, 221–253.

MAO, Q. AND J. XU (2019): "Input trade liberalisation, institution and markup: Evidence from China's accession to the WTO," *World Economy*, 42, 3537–3568.

MOON, H. R. AND M. WEIDNER (2015): "Linear regression for panel with unknown number of factors as interactive fixed effects," *Econometrica*, 83, 1543–1579.

——— (2019): "Nuclear Norm Regularized Estimation of Panel Regression Models," .

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

WESTERLUND, J. AND Y. KADDOURA (2022): "CCE in heterogenous fixed-T panels," *Econometrics Journal*.

WESTERLUND, J., Y. PETROVA, AND M. NORKUTE (2019): "CCE in fixed-T panels," *Journal of Applied Econometrics*, 34, 746–761.

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

XIANG, X., F. CHEN, C.-Y. HO, AND W. YUE (2017): "Heterogeneous effects of trade liberalisation on firm-level markups: Evidence from China," *World Economy*, 40, 1667–1686.

XU, Y. (2017): "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 25, 57–76.

Table 1: Monte Carlo results when trends are parallel.

	BIAS($\widehat{\Delta}_7$)	MSE($\widehat{\Delta}_7$)	BIAS($\widehat{\Delta}_8$)	MSE($\widehat{\Delta}_8$)	BIAS($\widehat{\Delta}_9$)	MSE($\widehat{\Delta}_9$)
Direct effect only						
OLS	-0.00	1.06	-0.02	4.17	-0.03	9.32
OLS with covariates	-0.01	0.54	-0.01	1.85	-0.01	3.92
C^2ED^2	-0.01	0.58	-0.02	1.07	-0.03	1.72
Direct and indirect effects						
OLS	0.01	1.07	0.02	4.20	0.03	9.38
OLS with covariates	-4.19	18.21	-4.28	20.31	-4.35	23.07
C^2ED^2	-0.02	0.57	-0.03	1.07	-0.04	1.69

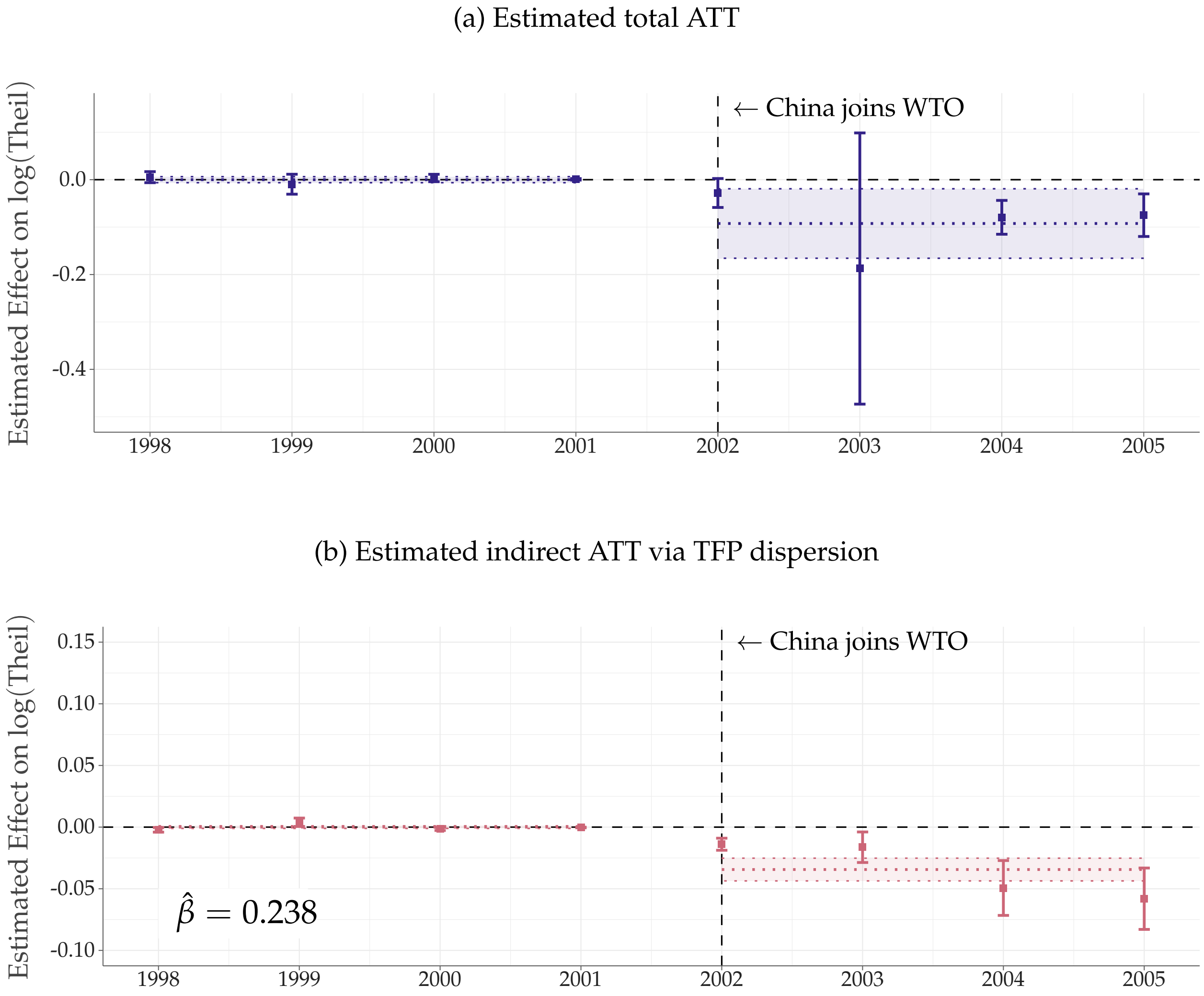
Notes: Data are generated for $N = 164$ cross-sections and $T = 9$ time periods to match the sample used in the empirical illustration. Treatment starts in period $g_{\min} = 7$. $\widehat{\Delta}_7$–$\widehat{\Delta}_9$ are the estimated total ATT for the post-treatment time periods. "BIAS($\widehat{\Delta}_t$)" and "MSE($\widehat{\Delta}_t$)" refer to the bias and MSE of the estimated ATT at post-treatment time period t, respectively. "OLS" and "OLS with covariates" refers to the two-way fixed effects OLS estimator without and with covariates, respectively. The results are reported for two data generating processes; one in which there is only a direct effect and one in which there is both direct and indirect effects.

Table 2: Monte Carlo results when trends are not parallel.

	BIAS($\widehat{\Delta}_7$)	MSE($\widehat{\Delta}_7$)	BIAS($\widehat{\Delta}_8$)	MSE($\widehat{\Delta}_8$)	BIAS($\widehat{\Delta}_9$)	MSE($\widehat{\Delta}_9$)
Direct effect only						
OLS	4.00	17.07	8.00	68.09	12.00	153.23
OLS with covariates	4.01	16.59	8.01	66.08	12.02	148.37
C^2ED^2	-0.03	1.17	-0.06	2.26	-0.06	3.55
Direct and indirect effects						
OLS	4.00	17.10	8.01	68.44	12.01	153.96
OLS with covariates	-0.19	0.68	3.71	15.76	7.63	62.27
C^2ED^2	-0.06	1.20	-0.06	2.36	-0.06	3.64

Notes: See Table 1 for an explanation.

Figure 1: Estimated ATTs of China's WTO accession in 2001 on the markup Theil index.



(a) Estimated total ATT

(b) Estimated indirect ATT via TFP dispersion

Notes: The figures present ATT estimates and 95% confidence intervals for the effect of China's WTO accession in 2001 on the dispersion of markups as measured by the markup Theil index. The treatment group comprise all industries that in 2001 had above-median tariff rates. Estimates are computed using the C^2ED^2 estimator with the TFP Theil index as a covariate. A constant is included as an observed factor. Figure (a) presents estimates of the total ATT and figure (b) presents the estimated indirect ATT operating through the TFP Theil index. $\hat{\beta}$ in figure (b) refers to the estimated slope on the TFP Theil index in the markup Theil index regression.

Online Appendix to "Simple Difference-in-Differences Estimation in Fixed-T Panels"

Nicholas Brown
Queen's University

Kyle Butts
University of Colorado Boulder

Joakim Westerlund*
Lund University
and
Deakin University

May 24, 2023

Abstract

This appendix provides (i) the proof of Theorem 1 reported in the main paper, and (ii) discussion of some of the assumptions.

1 Proof of Theorem 1

We start with part (a) of the theorem. We begin by considering the step-1 estimator of $\mathbf{f}_t$. In so doing, it is useful to denote by $\overline{\mathbf{a}}_t = (|\mathcal{I}_\infty|)^{-1}\sum_{i\in\mathcal{I}_\infty}\mathbf{a}_{i,t}$ the cross-sectional average of any vector $\mathbf{a}_{i,t}$ for the group of untreated units ($g=\infty$). In this notation, $\widehat{\mathbf{f}}_t = \overline{\mathbf{z}}_t$. Making use of this and the expression given for $\mathbf{z}_{i,t}$ in the main paper,

$$\widehat{\mathbf{f}}_t = \overline{\mathbf{z}}_t = \overline{\mathbf{\Lambda}}'\mathbf{f}_t + \overline{\mathbf{e}}_t \tag{A.1}$$

for the pretreatment sample $t \leq g_{\min}$. Here, $\overline{\mathbf{\Lambda}}$ and $\overline{\mathbf{e}}_t$ are the cross-sectional averages of $\mathbf{\Lambda}_i = [\boldsymbol{\alpha}_i + \boldsymbol{\lambda}_i\boldsymbol{\beta}_i, \boldsymbol{\lambda}_i]$ and $\mathbf{e}_{i,t} = [\varepsilon_{i,t} + \boldsymbol{\beta}_i'\mathbf{v}_{i,t}, \mathbf{v}_{i,t}']'$, respectively. If $m+1=r$, then the $r\times(m+1)$ matrix $\overline{\mathbf{\Lambda}}$ is square and invertible, which means that (A.1) can be rewritten as

$$\overline{\mathbf{\Lambda}}^{-1\prime}\widehat{\mathbf{f}}_t = \mathbf{f}_t + \overline{\mathbf{\Lambda}}^{-1\prime}\overline{\mathbf{e}}_t. \tag{A.2}$$

[*] 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.

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

$$\overline{\boldsymbol{\Lambda}}^{-1\prime}\widehat{\mathbf{f}}_t = \mathbf{f}_t + O_p(N^{-1/2}) \tag{A.3}$$

and hence $\overline{\boldsymbol{\Lambda}}^{-1\prime}\widehat{\mathbf{f}}_t$ is consistent for $\mathbf{f}_t$. In practice, we never observe $\overline{\boldsymbol{\Lambda}}$. However, since $\boldsymbol{\alpha}_i'\mathbf{f}_t = \boldsymbol{\alpha}_i'\overline{\boldsymbol{\Lambda}}^{-1\prime}\widehat{\mathbf{f}}_t + O_p(N^{-1/2})$, it is enough if we know $\widehat{\mathbf{f}}_t$, because $\overline{\boldsymbol{\Lambda}}^{-1}$ 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 $\overline{\boldsymbol{\Lambda}}$ is no longer invertible. However, we still need something similar to (A.2), because it determines the object that is being estimated. The way we approach this issue is the same as in Westerlund et al. (2019), and others. In particular, we begin by partitioning $\boldsymbol{\Lambda}_i$ as $\overline{\boldsymbol{\Lambda}} = [\overline{\boldsymbol{\Lambda}}_r, \overline{\boldsymbol{\Lambda}}_{-r}]$, where $\overline{\boldsymbol{\Lambda}}_{-r}$ is $r \times (m+1-r)$ and $\overline{\boldsymbol{\Lambda}}_r$ 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 $\overline{\boldsymbol{\Lambda}\mathbf{H}} = [\mathbf{I}_r, \mathbf{0}_{r\times(m+1-r)}]$ and that is going to play the same role as $\overline{\boldsymbol{\Lambda}}^{-1}$ under $m+1=r$:

$$\overline{\mathbf{H}} = \begin{bmatrix} \overline{\boldsymbol{\Lambda}}_r^{-1} & -\overline{\boldsymbol{\Lambda}}_r^{-1}\overline{\boldsymbol{\Lambda}}_{-r} \\ \mathbf{0}_{(m+1-r)\times r} & \mathbf{I}_{m+1-r} \end{bmatrix} = [\overline{\mathbf{H}}_r,\, \overline{\mathbf{H}}_{-r}], \tag{A.4}$$

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

$$\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}_t = \widehat{\mathbf{f}}_t^0 = \mathbf{D}_N\overline{\mathbf{H}}'\overline{\boldsymbol{\Lambda}}'\mathbf{f}_t + \mathbf{D}_N\overline{\mathbf{H}}'\bar{\mathbf{e}}_t = \mathbf{f}_t^0 + \bar{\mathbf{e}}_t^0, \tag{A.5}$$

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

$$\widehat{\mathbf{f}}_t^0 = \mathbf{f}_t^0 + \bar{\mathbf{e}}_t^0 = \begin{bmatrix} \mathbf{f}_t \\ \mathbf{0}_{(m+1-r)\times 1} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{e}}_{r,t}^0 \\ \bar{\mathbf{e}}_{-r,t}^0 \end{bmatrix} = \mathbf{f}_t^+ + O_p(N^{-1/2}), \tag{A.6}$$

The fact that $\mathbf{f}_t$ is included in $\mathbf{f}_t^+$ suggests that asymptotically C^{2}ED^2 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 $\overline{\mathbf{e}}^0_{-r,t}$, 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 $\widehat{\mathbf{f}}^0_t = \overline{\boldsymbol{\Lambda}}^{-1\prime}\widehat{\mathbf{f}}_t$, $\mathbf{f}^0_t = \mathbf{f}_t$ and $\overline{\mathbf{e}}^0_t = \overline{\boldsymbol{\Lambda}}^{-1\prime}\overline{\mathbf{e}}_t$ if $m+1=r$, so that we are back in (A.2).

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 ($i \in \mathcal{I}_g \subset \mathcal{G}$ and $t \geq g_{\min}$). Note first that because we are considering treated units in post-treatment periods, $y_{i,t}(g) = y_{i,t}$ and $\mathbf{x}_{i,t}(g) = \mathbf{x}_{i,t}$. Further use of the definitions of $\Delta_{i,g,t}$, $y_{i,t}(\infty)$ and $\boldsymbol{\tau}_{i,g,t}$, leads to the following model for $y_{i,t}$:

$$\begin{aligned} y_{i,t} &= \Delta_{i,g,t} + y_{i,t}(\infty) = \Delta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} \\ &= \Delta_{i,g,t} + \boldsymbol{\beta}_i'(\mathbf{x}_{i,t} - \boldsymbol{\tau}_{i,g,t}) + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} = (\Delta_{i,g,t} - \boldsymbol{\beta}_i'\boldsymbol{\tau}_{i,g,t}) + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} \\ &= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t}. \end{aligned} \tag{A.7}$$

It follows that

$$\begin{aligned} \widehat{\Delta}_{i,g,t} &= y_{i,t} - \widehat{y}_{i,t}(\infty) \\ &= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} - [\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{x}}_{i,t}(\infty) + \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t] \\ &= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t} - (\widehat{\boldsymbol{\beta}}'\mathbf{x}_{i,t} + \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t) + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] \\ &= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - (\widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'\mathbf{f}_t) + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] + \varepsilon_{i,t}. \end{aligned} \tag{A.8}$$

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

$$\begin{aligned} \widehat{\mathbf{a}}_i'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'\mathbf{f}_t &= \widehat{\mathbf{a}}_i'(\mathbf{D}_N\overline{\mathbf{H}}')^{-1}\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}_t - \boldsymbol{\alpha}_i'(\mathbf{D}_N\overline{\mathbf{H}}'\overline{\boldsymbol{\Lambda}}')^+\mathbf{D}_N\overline{\mathbf{H}}'\overline{\boldsymbol{\Lambda}}'\mathbf{f}_t \\ &= \widehat{\mathbf{a}}^{0\prime}_i\widehat{\mathbf{f}}^0_t - \boldsymbol{\alpha}^{0\prime}_i\mathbf{f}^0_t \\ &= \boldsymbol{\alpha}^{0\prime}_i(\widehat{\mathbf{f}}^0_t - \mathbf{f}^0_t) + (\widehat{\mathbf{a}}^0_i - \boldsymbol{\alpha}^0_i)'\widehat{\mathbf{f}}^0_t, \end{aligned} \tag{A.9}$$

from which it follows that

$$\widehat{\Delta}_{i,g,t} = \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i^{0\prime}(\widehat{\mathbf{f}}_t^0 - \mathbf{f}_t^0) - (\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0)'\widehat{\mathbf{f}}_t^0 + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] + \varepsilon_{i,t}. \qquad \text{(A.10)}$$

Amongst the terms appearing on the right-hand side of this last 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 pre-treatment period only, we have $y_{i,t} = y_{i,t}(\infty) = \boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) + \boldsymbol{\alpha}_i'\mathbf{f}_t + \varepsilon_{i,t}$ or, in terms of the stacked vector notation introduced in step 2 of the counterfactual estimation procedure outlined in the main paper, $\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta}_i + \mathbf{f}\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{y}_i$, $\mathbf{x}_i$, $\mathbf{f}$ and $\boldsymbol{\varepsilon}_i$ are all $(g_{\min} - 1)$-rowed. By using this and $\overline{\boldsymbol{\Lambda}\mathbf{H}}_r = \mathbf{I}_r$, we get

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta}_i + \widehat{\mathbf{f}}\overline{\mathbf{H}}_r\boldsymbol{\alpha}_i - (\widehat{\mathbf{f}} - \mathbf{f}\overline{\boldsymbol{\Lambda}})\overline{\mathbf{H}}_r\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i = \mathbf{x}_i\boldsymbol{\beta}_i + \widehat{\mathbf{f}}\overline{\mathbf{H}}_r\boldsymbol{\alpha}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i. \qquad \text{(A.11)}$$

We also note that $\mathbf{a}_i$ in step 2 can be expressed in terms of $\overline{\mathbf{H}}_r$ and $\boldsymbol{\alpha}_i$ as $\mathbf{a}_i = \overline{\mathbf{H}}_r\boldsymbol{\alpha}_i$. By inserting this and (A.11) into the expression given for $\widehat{\mathbf{a}}_i$ in step 2,

$$\begin{aligned}
\widehat{\mathbf{a}}_i &= (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}) \\
&= (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'(\mathbf{x}_i\boldsymbol{\beta}_i + \widehat{\mathbf{f}}\mathbf{a}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}) \\
&= \mathbf{a}_i + (\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i], \qquad \text{(A.12)}
\end{aligned}$$

implying

$$\begin{aligned}
\widehat{\mathbf{a}}_i^0 &= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\widehat{\mathbf{a}}_i \\
&= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i + (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}'[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i] \\
&= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i + (\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}'\widehat{\mathbf{f}}\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{D}_N\overline{\mathbf{H}}'\widehat{\mathbf{f}}'[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i] \\
&= (\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i + (\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}^{0\prime}[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i] \qquad \text{(A.13)}
\end{aligned}$$

where $\widehat{\mathbf{f}}^0 = [\widehat{\mathbf{f}}_1^0, ..., \widehat{\mathbf{f}}_{g_{\min}-1}^0]' = \widehat{\mathbf{f}}\overline{\mathbf{H}}\mathbf{D}_N$ is $(g_{\min} - 1) \times (m + 1)$. 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

$$(\mathbf{D}_N\overline{\mathbf{H}})^{-1} = \begin{bmatrix} \overline{\boldsymbol{\Lambda}}_r & \overline{\boldsymbol{\Lambda}}_{-r} \\ \mathbf{0}_{(m+1-r)\times r} & N^{-1/2}\mathbf{I}_{m+1-r} \end{bmatrix}, \qquad \text{(A.14)}$$

so that

$$(\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\overline{\mathbf{H}}_r = \left[\begin{array}{cc} \overline{\mathbf{\Lambda}}_r & \overline{\mathbf{\Lambda}}_{-r} \\ \mathbf{0}_{(m+1-r)\times r} & N^{-1/2}\mathbf{I}_{m+1-r} \end{array}\right] \left[\begin{array}{c} \overline{\mathbf{\Lambda}}_r^{-1} \\ \mathbf{0}_{(m+1-r)\times r} \end{array}\right] = \left[\begin{array}{c} \mathbf{I}_r \\ \mathbf{0}_{(m+1-r)\times r} \end{array}\right]. \quad \text{(A.15)}$$

This implies

$$(\overline{\mathbf{H}}\mathbf{D}_N)^{-1}\mathbf{a}_i = \left[\begin{array}{c} \boldsymbol{\alpha}_i \\ \mathbf{0}_{(m+1-r)\times 1} \end{array}\right] = \boldsymbol{\alpha}_i^0, \quad \text{(A.16)}$$

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

$$\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0 = (\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}^{0\prime}[-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i]. \quad \text{(A.17)}$$

We similarly have

$$\widehat{\mathbf{x}}_{i,t}(\infty) = \widehat{\boldsymbol{\lambda}}_i'\widehat{\mathbf{f}}_t = \mathbf{x}_i'\widehat{\mathbf{f}}(\widehat{\mathbf{f}}'\widehat{\mathbf{f}})^{-1}\widehat{\mathbf{f}}_t = \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0, \quad \text{(A.18)}$$

from which it follows that

$$\widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] = \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0]. \quad \text{(A.19)}$$

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

$$\begin{aligned} \widehat{\Delta}_{i,g,t} &= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i^{0\prime}(\widehat{\mathbf{f}}_t^0 - \mathbf{f}_t^0) - (\widehat{\mathbf{a}}_i^0 - \boldsymbol{\alpha}_i^0)'\widehat{\mathbf{f}}_t^0 + \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \widehat{\mathbf{x}}_{i,t}(\infty)] + \varepsilon_{i,t} \\ &= \eta_{i,g,t} - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i'\overline{\mathbf{e}}_{r,t}^0 - [-\mathbf{x}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_i) - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i]'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0 \\ &+ \widehat{\boldsymbol{\beta}}'[\mathbf{x}_{i,t} - \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0] + \varepsilon_{i,t} \\ &= \eta_{i,g,t} + \boldsymbol{\beta}_i'\mathbf{x}_{i,t} - \boldsymbol{\alpha}_i'\overline{\mathbf{e}}_{r,t}^0 + \varepsilon_{i,t} - (\mathbf{x}_i\boldsymbol{\beta}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i)'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0. \quad \text{(A.20)} \end{aligned}$$

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

$$\mathbf{x}_i = \mathbf{f}\boldsymbol{\lambda}_i + \mathbf{v}_i = \widehat{\mathbf{f}}\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i - (\widehat{\mathbf{f}} - \mathbf{f}\overline{\mathbf{\Lambda}})\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i + \mathbf{v}_i = \widehat{\mathbf{f}}^0\mathbf{D}_N^{-1}\overline{\mathbf{H}}^{-1}\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\lambda}_i + \mathbf{v}_i, \quad \text{(A.21)}$$

for $t < g_{\min}$. If, on the other hand, $t \geq g_{\min}$, then

$$\mathbf{x}_{i,t} = \boldsymbol{\tau}_{i,g,t} + \boldsymbol{\lambda}_i'\mathbf{f}_t + \mathbf{v}_{i,t} = \boldsymbol{\tau}_{i,g,t} + \boldsymbol{\lambda}_i'\overline{\mathbf{H}}_r'\overline{\mathbf{H}}^{-1\prime}\mathbf{D}_N^{-1}\widehat{\mathbf{f}}_t^0 - \boldsymbol{\lambda}_i'\overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t}. \quad \text{(A.22)}$$

These two last results imply

$$\begin{aligned} &\mathbf{x}_{i,t} - \mathbf{x}_i'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0 \\ &= \boldsymbol{\tau}_{i,g,t} + \boldsymbol{\lambda}_i'\overline{\mathbf{H}}_r'\overline{\mathbf{H}}^{-1\prime}\mathbf{D}_N^{-1}\widehat{\mathbf{f}}_t^0 - \boldsymbol{\lambda}_i'\overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t} - (\widehat{\mathbf{f}}^0\mathbf{D}_N^{-1}\overline{\mathbf{H}}^{-1}\overline{\mathbf{H}}_r\boldsymbol{\lambda}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\lambda}_i + \mathbf{v}_i)'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0 \\ &= \boldsymbol{\tau}_{i,g,t} - \boldsymbol{\lambda}_i'\overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t} - (-\overline{\mathbf{e}}_r^0\boldsymbol{\lambda}_i + \mathbf{v}_i)'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0, \quad \text{(A.23)} \end{aligned}$$

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

$$\begin{aligned}
\widehat{\Delta}_{i,g,t} &= \eta_{i,g,t} + \boldsymbol{\beta}_i'(\boldsymbol{\tau}_{i,g,t} - \boldsymbol{\lambda}_i'\overline{\mathbf{e}}_{r,t}^0 + \mathbf{v}_{i,t}) - \boldsymbol{\alpha}_i'\overline{\mathbf{e}}_{r,t}^0 + \varepsilon_{i,t} \\
&\quad - [(-\overline{\mathbf{e}}_r^0\boldsymbol{\lambda}_i + \mathbf{v}_i)\boldsymbol{\beta}_i - \overline{\mathbf{e}}_r^0\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i]'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0 \\
&= \Delta_{i,g,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\overline{\mathbf{e}}_{r,t}^0 + \boldsymbol{\beta}_i'\mathbf{v}_{i,t} + \varepsilon_{i,t} \\
&\quad - [-\overline{\mathbf{e}}_r^0(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i]'\widehat{\mathbf{f}}^0(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}_t^0. \qquad \text{(A.24)}
\end{aligned}$$

where $\Delta_{i,g,t} = \eta_{i,g,t} + \boldsymbol{\beta}_i'\boldsymbol{\tau}_{i,g,t}$ as in the main paper.

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 $\mathbf{f}^+ = [\mathbf{f}_1^+, ...\mathbf{f}_{g_{\min}-1}^+]' = [\mathbf{f}, \overline{\mathbf{e}}_{-r}^0]$, a $(g_{\min} - 1) \times (m + 1)$ 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,

$$\|(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1} - (\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\| = O_p(N^{-1/2}), \qquad \text{(A.25)}$$

where

$$\mathbf{f}^{+\prime}\mathbf{f}^+ = \begin{bmatrix} \mathbf{f}'\mathbf{f} & \mathbf{f}'\overline{\mathbf{e}}_{-r}^0 \\ \overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{f} & \overline{\mathbf{e}}_{-r}^{0\prime}\overline{\mathbf{e}}_{-r}^0 \end{bmatrix}, \qquad \text{(A.26)}$$

$$(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1} = \begin{bmatrix} \begin{matrix} (\mathbf{f}'\mathbf{f})^{-1} + (\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'\overline{\mathbf{e}}_{-r}^0(\overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}_{-r}^0)^{-1}\overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{f}(\mathbf{f}'\mathbf{f})^{-1} \\ -(\overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}_{-r}^0)^{-1}\overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{f}(\mathbf{f}'\mathbf{f})^{-1} \end{matrix} \\ \begin{matrix} -(\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'\overline{\mathbf{e}}_{-r}^0(\overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}_{-r}^0)^{-1} \\ (\overline{\mathbf{e}}_{-r}^{0\prime}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}_{-r}^0)^{-1} \end{matrix} \end{bmatrix}. \qquad \text{(A.27)}$$

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

$$\begin{aligned}
\widehat{\mathbf{f}}_t^{0\prime}(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1}\widehat{\mathbf{f}}^{0\prime} &= \widehat{\mathbf{f}}_t^{0\prime}[(\widehat{\mathbf{f}}^{0\prime}\widehat{\mathbf{f}}^0)^{-1} - (\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}]\widehat{\mathbf{f}}^{0\prime} + \widehat{\mathbf{f}}_t^{0\prime}(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\widehat{\mathbf{f}}^{0\prime} \\
&= \widehat{\mathbf{f}}_t^{0\prime}(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\widehat{\mathbf{f}}^{0\prime} + O_p(N^{-1/2}) \\
&= \mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^+)^{-1}\mathbf{f}^{+\prime} + O_p(N^{-1/2}). \qquad \text{(A.28)}
\end{aligned}$$

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

$$\begin{aligned}
&\mathbf{f}_t^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}^{+\prime} \\
&= [\mathbf{f}_t', \overline{\mathbf{e}}^{0\prime}_{-r,t}] \begin{bmatrix} (\mathbf{f}'\mathbf{f})^{-1} + (\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'\overline{\mathbf{e}}^{0}_{-r}(\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}^{0}_{-r})^{-1}\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{f}(\mathbf{f}'\mathbf{f})^{-1} & -(\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'\overline{\mathbf{e}}^{0}_{-r}(\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}^{0}_{-r})^{-1} \\ -(\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}^{0}_{-r})^{-1}\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{f}(\mathbf{f}'\mathbf{f})^{-1} & (\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}^{0}_{-r})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{f}' \\ \overline{\mathbf{e}}^{0\prime}_{-r} \end{bmatrix} \\
&= \mathbf{f}_t'(\mathbf{f}'\mathbf{f})^{-1}\mathbf{f}'[\mathbf{I}_{g_{\min}-1} - \overline{\mathbf{e}}^{0}_{-r}(\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}^{0}_{-r})^{-1}\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}] + \overline{\mathbf{e}}^{0\prime}_{-r,t}(\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}\overline{\mathbf{e}}^{0}_{-r})^{-1}\overline{\mathbf{e}}^{0\prime}_{-r}\mathbf{M}_{\mathbf{f}}. \qquad (A.29)
\end{aligned}$$

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

$$\begin{aligned}
\widehat{\Delta}_{i,g,t} &= \Delta_{i,g,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0}_{r,t} + \boldsymbol{\beta}_i'\mathbf{v}_{i,t} + \varepsilon_{i,t} \\
&\quad - [-\overline{\mathbf{e}}^{0}_{r}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i]'\mathbf{f}^{+}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}_t^{+} + O_p(N^{-1/2}) \\
&= \Delta_{i,g,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0*}_{r,t} + \boldsymbol{\beta}_i'\mathbf{v}^{*}_{i,t} + \varepsilon^{*}_{i,t} + O_p(N^{-1/2}), \qquad (A.30)
\end{aligned}$$

where

$$\mathbf{a}^{*}_{i,t} = \mathbf{a}_{i,t} - \mathbf{a}_i'\mathbf{f}^{+}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}_t^{+} = \mathbf{a}_{i,t} - \sum_{s=1}^{g_{\min}-1}\mathbf{a}_{i,s}\mathbf{f}_s^{+\prime}(\mathbf{f}^{+\prime}\mathbf{f}^{+})^{-1}\mathbf{f}_t^{+} \qquad (A.31)$$

for any vector $\mathbf{a}_{i,t}$ with $(g_{\min}-1)$-rowed stack $\mathbf{a}_i = [\mathbf{a}_{i,1}, ..., \mathbf{a}_{i,g_{\min}-1}]'$. 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 $\sqrt{|\mathcal{I}_g|}$. In order to appreciate this, we make use of the fact that $\|\sqrt{|\mathcal{I}_g|}\overline{\mathbf{e}}^{0}_{r}\| = O_p(1)$, 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 $\|(|\mathcal{I}_g|)^{-1/2}\sum_{i\in\mathcal{I}_g}\mathbf{v}_i\boldsymbol{\beta}_i\| = O_p(1)$. It follows that

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{|\mathcal{I}_g|}}\sum_{i\in\mathcal{I}_g}[-\overline{\mathbf{e}}^{0}_{r}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i] \right\| \\
&\leq \|\sqrt{|\mathcal{I}_g|}\overline{\mathbf{e}}^{0}_{r}\| \left\| \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g}(\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) \right\| + \left\| \frac{1}{\sqrt{|\mathcal{I}_g|}}\sum_{i\in\mathcal{I}_g}\mathbf{v}_i\boldsymbol{\beta}_i \right\| + \left\| \frac{1}{\sqrt{|\mathcal{I}_g|}}\sum_{i\in\mathcal{I}_g}\boldsymbol{\varepsilon}_i \right\| = O_p(1).
\end{aligned}$$

$$(A.32)$$

We can therefore show that

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} [-\overline{\mathbf{e}}_r^0(\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i]' [\mathbf{f}^+ (\mathbf{f}^{+\prime} \mathbf{f}^+)^{-1} \mathbf{f}_t^+ - \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0] \right\| \\
&\leq \left\| \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} [-\overline{\mathbf{e}}_r^0(\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i) + \mathbf{v}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i] \right\| \|\mathbf{f}^+ (\mathbf{f}^{+\prime} \mathbf{f}^+)^{-1} \mathbf{f}_t^+ - \widehat{\mathbf{f}}^0 (\widehat{\mathbf{f}}^{0\prime} \widehat{\mathbf{f}}^0)^{-1} \widehat{\mathbf{f}}_t^0 \| \\
&= O_p(N^{-1/2}),
\end{aligned} \tag{A.33}$$

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 $\sqrt{|\mathcal{I}_g|}$.

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

$$\begin{aligned}
\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t}) &= \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} (\widehat{\Delta}_{i,g,t} - \Delta_{g,t}) \\
&= \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} (\widehat{\Delta}_{i,g,t} - \Delta_{i,g,t} + v_{i,t}) \\
&= \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} [v_{i,t} - (\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)' \overline{\mathbf{e}}_{r,t}^{0*} + \boldsymbol{\beta}_i' \mathbf{v}_{i,t}^* + \varepsilon_{i,t}^*] + O_p(N^{-1/2}).
\end{aligned} \tag{A.34}$$

Moreover, $|\mathcal{I}_g|/N \to_p \tau_g \in (0,1)$ by Assumption 2. Hence, if we in addition define $\mathbf{a}_g = \mathrm{plim}_{N \to \infty} (|\mathcal{I}_g|)^{-1} \sum_{i \in \mathcal{I}_g} (\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)$, the above expression for $\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})$ becomes

$$\begin{aligned}
&\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t}) \\
&= \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} (v_{i,t} + \boldsymbol{\beta}_i' \mathbf{v}_{i,t}^* + \varepsilon_{i,t}^*) - \sqrt{\frac{|\mathcal{I}_g|}{N}} \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} (\boldsymbol{\lambda}_i \boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)' \sqrt{N} \overline{\mathbf{e}}_{r,t}^{0*} + O_p(N^{-1/2}) \\
&= \frac{1}{\sqrt{|\mathcal{I}_g|}} \sum_{i \in \mathcal{I}_g} (v_{i,t} + \boldsymbol{\beta}_i' \mathbf{v}_{i,t}^* + \varepsilon_{i,t}^*) - \sqrt{\tau_g} \mathbf{a}_g' \sqrt{N} \overline{\mathbf{e}}_{r,t}^{0*} + o_p(1).
\end{aligned} \tag{A.35}$$

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 $\sigma^2(\widehat{\Delta}_{g,t}) = \mathrm{var}(\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t}) | \mathcal{C})$, where $\mathcal{C}$ is the sigma-field generated by $\mathbf{f}$. The asymptotic distribution of $\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})$ as $N \to \infty$ can now be stated in the following way:

$$\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t}) \to_d MN(0, \sigma^2(\widehat{\Delta}_{g,t})), \tag{A.36}$$

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

$$\frac{\sqrt{|\mathcal{I}_g|}(\widehat{\Delta}_{g,t} - \Delta_{g,t})}{\sigma(\widehat{\Delta}_{g,t})} \to_d N(0,1), \tag{A.37}$$

as required for part (a).

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

$$\widehat{\Delta}_{i,g,t} = \Delta_{g,t} + v_{i,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0*}_{r,t} + \boldsymbol{\beta}_i'\mathbf{v}^*_{i,t} + \varepsilon^*_{i,t} + O_p(N^{-1/2}), \tag{A.38}$$

$$\frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g}\widehat{\Delta}_{i,g,t} = \Delta_{g,t} + \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g}[v_{i,t} - (\boldsymbol{\lambda}_i\boldsymbol{\beta}_i + \boldsymbol{\alpha}_i)'\overline{\mathbf{e}}^{0*}_{r,t} + \boldsymbol{\beta}_i'\mathbf{v}^*_{i,t} + \varepsilon^*_{i,t}] + O_p(N^{-1/2}). \tag{A.39}$$

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

$$\widehat{\Delta}_{i,g,t} - \frac{1}{|\mathcal{I}_g|}\sum_{j\in\mathcal{I}_g}\widehat{\Delta}_{j,g,t} = z_{i,t} - \frac{1}{|\mathcal{I}_g|}\sum_{j\in\mathcal{I}_g} z_{j,t} + O_p(N^{-1/2}). \tag{A.40}$$

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

$$\begin{aligned}\widehat{\sigma}^2(\widehat{\Delta}_{g,t}) &= \frac{1}{|\mathcal{I}_g|-1}\sum_{i\in\mathcal{I}_g}\left(\widehat{\Delta}_{i,g,t} - \frac{1}{|\mathcal{I}_g|}\sum_{j\in\mathcal{I}_g}\widehat{\Delta}_{j,g,t}\right)^2 \\ &= \frac{1}{|\mathcal{I}_g|-1}\sum_{i\in\mathcal{I}_g}\left(z_{i,t} - \frac{1}{|\mathcal{I}_g|}\sum_{j\in\mathcal{I}_g} z_{j,t}\right)^2 + O_p(N^{-1/2}) \to_p \sigma^2(\widehat{\Delta}_{g,t}) \end{aligned} \tag{A.41}$$

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.

2 Discussion of some of the assumptions

The results reported in the main paper assume that the covariates admits to a common factor representation, which is not needed in pricipal components-based studies such as that of

Chan and Kwok (2022). In this section, we show that the direct effect is estimable even if this assumption fails, although in that case we can no longer identify the overall and indirect ATTs.

Our starting point here is Brown et al. (2022), who consider the same CCE approach as in Pesaran (2006) but under a different set of assumptions. In particular, instead of requiring that $\mathbf{x}_{i,t}$ has factor structure, they assume that $\mathbf{f}_t$ satisfies

$$\mathbf{f}_t = \mathbf{B}'\mathbf{\Psi}_t \tag{B.42}$$

where $\mathbf{\Psi}_t = \mathbb{E}(\mathbf{z}_{i,t})$ is constant in i and $\mathbf{B}$ is an arbitrary $(m+1) \times r$ matrix of constants. Unlike the factors-in-covariates condition, (B.42) is not testable. However, if it holds, $\widehat{\mathbf{f}}_t$ can be used to estimate an arbitrary number of factors. In order to illustrate this last point, note that if (B.42) holds for the untreated potential outcomes,

$$\begin{aligned}
y_{i,t}(\infty) &= \boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) + \boldsymbol{\alpha}_i'\mathbf{f}_t + \epsilon_{i,t} \\
&= \boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) + \mathbf{a}_i'\mathbb{E}(\mathbf{z}_{i,t}|g_i = \infty) + \epsilon_{i,t} \\
&= \boldsymbol{\beta}_i'\mathbf{x}_{i,t}(\infty) + \mathbf{a}_i'\widehat{\mathbf{f}}_t + \mathbf{a}_i'[\mathbb{E}(\mathbf{z}_{i,t}|g_i = \infty) - \widehat{\mathbf{f}}_t] + \epsilon_{i,t}
\end{aligned} \tag{B.43}$$

where $\mathbf{a}_i = \mathbf{B}\boldsymbol{\alpha}_i$ and $\mathbb{E}(\mathbf{z}_{i,t}|g_i = \infty) - \widehat{\mathbf{f}}_t$ is negligible, as $\bar{\mathbf{z}}_t \to_p \mathbb{E}(\mathbf{z}_{i,t})$ as $N \to \infty$ under standard regulatory conditions.

The main advantage of (B.42) is that it leaves the covariates essentially unrestricted. However, because we no longer have a model for the untreated potential covariates, we cannot estimate $\mathbf{x}_{i,t}(\infty)$ in step 3 of the counterfactual estimation procedure. This has two implications; (i) we are unable to identify the effect of the treatment on $\mathbf{x}_{i,t}$, and (ii) we have to use $\mathbf{x}_{i,t}$ as opposed to $\widehat{\mathbf{x}}_{i,t}(\infty)$ when computing $\widehat{y}_{i,t}(\infty)$ in step 4. As a result, similarly to Chan and Kwok (2022), we can only identify the direct ATT. Hence, while we can relax factors-in-covariates condition, this has a price in terms of the estimable ATTs.

References

ABADIR, K. M. AND J. R. MAGNUS (2005): *Matrix algebra*, vol. 1, Cambridge University Press.

BROWN, N. L., P. SCHMIDT, AND J. M. WOOLDRIDGE (2022): "Simple alternatives to the common correlated effects model," ArXiv:2112.01486.

CHAN, M. K. AND S. S. KWOK (2022): "The PCDID approach: difference-in-differences when trends are potentially unparallel and stochastic," *Journal of Business & Economic Statistics*, 40, 1216–1233.

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

WESTERLUND, J., Y. PETROVA, AND M. NORKUTE (2019): "CCE in fixed-T panels," *Journal of Applied Econometrics*, 34, 746–761.