THE UNIVERSITY OF SYDNEY

Economics Working Paper Series

2020 - 03

The PCDID Approach: Difference-in-Differences when Trends are Potentially Unparallel and Stochastic

Marc K. Chan and Simon Kwok

March 2020

The PCDID Approach: Difference-in-Differences when Trends are Potentially Unparallel and Stochastic

Marc K. Chan
University of Melbourne

Simon Kwok
The University of Sydney*

This version: March 2020

Abstract

We develop a class of regression-based estimators, called Principal Components Difference-in-Differences estimators (PCDID), for treatment effect estimation. Analogous to a control function approach, PCDID uses factor proxies constructed from control units to control for unobserved trends, assuming that the unobservables follow an interactive effects structure. We clarify the conditions under which the estimands in this regression-based approach represent useful causal parameters of interest. We establish consistency and asymptotic normality results of PCDID estimators under minimal assumptions on the specification of time trends. We show how PCDID can be extended to micro/group-level data and be used for testing parallel trends under the interactive effects structure. The PCDID approach is illustrated in an empirical exercise that examines the effects of welfare waiver programs on welfare caseloads in the US.

Keywords: principal components difference-in-differences, interactive fixed effects, factor-augmented regressions, treatment effects, parallel trends

*We would like to thank Manuel Arellano, John Ham, Rosa Matzkin, Tim Moore, Morten Nielsen, Tim Robinson, Chris Skeels, Yongcheol Shin, Katrien Stevens, Chris Taber, and participants of various seminars and conferences for their valuable comments. Special thanks to James Ziliak who provided the key years of welfare caseload data to us. Chan: Department of Economics, Faculty of Business and Economics, University of Melbourne, Parkville VIC 3010, Australia. Email: marc.chan@unimelb.edu.au. Kwok: School of Economics, The University of Sydney, NSW 2006 Australia; Email: simon.kwok@sydney.edu.au.

1 Introduction

The difference-in-differences (DID) method is a workhorse for policy evaluation in empirical economics and other disciplines. Its key underlying assumption is that trends are parallel among the control and treated units. In this paper, we develop a class of regression-based estimators, called Principal Components Difference-in-Differences Estimators (PCDID), that can be applied to scenarios in which trends are potentially unparallel and stochastic among control and treated units. Unlike existing approaches such as synthetic control, unconfoundedness and matrix completion estimators, PCDID uses factor-augmented regressions to estimate specific parameters related to treatment effects. Specifically, PCDID does the following: (1) use a data-driven method (principal component analysis (PCA)) on data from control units to form factor proxies that capture the endogeneity arising from unparallel trends; (2) among treated unit(s), run regressions using the factor proxies as extra covariates. Our method is analogous to the control function approach in the microeconometric literature, in the sense that the factor proxies play the same role as control functions.

Our main theoretical findings are as follows. First, we clarify the conditions under which the estimands in this regression-based approach represent useful causal parameters of interest. This is useful because the recent literature has shown that, under treatment effect heterogeneity, a standard DID two-way fixed effect (2wfe) regression may yield an estimate that no longer represents a useful causal parameter of interest, but instead a weighted average of treatment effects where some of the weights can be negative (e.g., Borusyak and Jaravel (2017), Abraham and Sun (2018), Athey and Imbens (2018), de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018)). Because our factor-augmented regression extends the 2wfe regression by incorporating unparallel trends, it is important to show that it targets meaningful causal parameters under reasonable assumptions. Second, based on the various ways the factor-augmented regression can be carried out, we consider three different PCDID estimators (basic, mean-group, pooled), and establish consistency and asymptotic normality results for each estimator with respect to its target causal parameter. Their differences in identification conditions and rates of convergence are clarified. Importantly, we show that asymptotic normality holds under minimal assumptions on the specification of trends, e.g., it encompasses nonstationary trends. Thus standard inference (e.g., t-tests with standard critical values) is valid and does not depend on the unknown nature of the time trends. Third, based on the PCDID approach, we develop a test of parallel trends under the functional form specification of the interactive effects model. We establish consistency and asymptotic normality results of the test statistic and, as in PCDID estimators, standard inference is valid and does not depend on the unknown nature of the time trends. We discuss how the test is related to the parallel trend assumptions in the fully nonparametric DID framework of Callaway and Sant'Anna (2018). Extending the above theoretical results, we consider micro/group-level data where an aggregated PCDID estimator is introduced. We then consider Monte Carlo simulations, a placebo exercise based on Bertrand et al. (2004), and an empirical illustration that examines the effects of welfare waiver programs on welfare caseloads in the US.

The PCDID approach is related to the large literature on factor-augmented regression models (e.g., Stock and Watson (2002), Bernanke et al. (2005), Bai and Ng (2006)), which forecast one or several time series using a large number of predictor series. We extend this literature to treatment effect estimation. Our stepwise implementation is also related to common correlated effects (CCE) estimators pioneered by Pesaran (2006). The key difference is that PCDID specifically exploits the DID data structure when it constructs factor proxies.[1] Other estimators in interactive effects models, such as Bai (2009) or Moon and Weidner (2015), are typically based on the assumption of homogeneous parameters, and they do not exploit the specific data structure as in our approach. Bai (2009) has recently been used or adapted for treatment effect estimation, e.g., Kim and Oka (2014), Gobillon and Magnac (2016), Xu (2017). These approaches use Bai (2009)'s iterative procedure to attain numerical convergence of at least some of the coefficients, whereas PCDID does not require numerical iterations.

The PCDID approach is related to the rapidly growing literature on synthetic control (SC), unconfoundedness and matrix completion estimators (e.g., Abadie et al. (2010), Hsiao et al. (2012), Xu (2017), Ferman and Pinto (2018), Ben-Michael et al. (2018), Athey et al. (2018), Chernozhukov et al. (2018), Arkhangelsky et al. (2019)). In this literature, both the control panel and pre-intervention data of treated units (an "L-shaped" matrix, as in

[1] In addition, CCE estimators, which are chiefly based on exogenously weighted cross-sectional averages in generic panel data, have a different theoretical framework to that based on principal components; see Westerlund and Urbain (2015) for a detailed analysis. Greenaway-McGrevy et al. (2012) extends Pesaran's estimator by constructing factor proxies from principal components based on the whole panel data set, whereas the PCDID approach and extracts factor proxies from data on control units only. This important difference explains the different asymptotic properties of our estimator from theirs.

Athey et al. (2018)) are used to impute the counterfactual outcomes of treated units in post-intervention periods; the treatment effect is then estimated as the difference between these counterfactual outcomes and observed outcomes. Athey et al. (2018) show that the original SC method (Abadie et al. (2010)), vertical regression (SC methods) and horizontal regression (unconfoundedness methods) all belong to the same class of matrix completion methods based on matrix factorization (i.e., low-rank matrices) with different restrictions/regularizations. Gobillon and Magnac (2016) and the generalized SC estimator (GSC) of Xu (2017) also consider this step-wise approach, by first applying Bai's estimator on the control panel and pre-intervention data of treated units to form counterfactual outcomes of treated units in post-intervention periods. Unlike the above approaches, PCDID exploits the data structure in the form of two rectangular-shaped matrices. It first constructs control functions (factor proxies) from the control panel exclusively. Then, it runs a factor-augmented regression using all (pre- and post-intervention) data of treated units, with the control functions as additional covariates. We show that PCDID is at least as efficient as GSC, and the difference in efficiency is large when the control functions are correlated with the intervention status variable and when post-intervention periods take up a large proportion of the sample periods.

We note that all our results are derived under the specification assumption of interactive effects, i.e., the unobservables can be factorized as described in Athey et al. (2018). Interactive effects impose a structure on inferring how the unparallel trends behave, based on observed variations in outcomes across units and time. They are less general than many other structures for the unobservables. In fact, Callaway and Sant'Anna (2018) show that, in a fully nonparametric DID setting, the parallel trend assumption is untestable but a stronger, augmented version of this assumption is testable (see also Section 4.4). Interactive effects are more general than two-way fixed effects but it comes with costs; namely, large-T asymptotics are typically required, and the asymptotic distributions and valid inference procedures are non-trivial under weak assumptions on time trends.[2] Whether interactive effects are reasonable and important depend on the specific application.[3] For instance, in the welfare caseload

[2]Quasi-differencing methods have been developed for fixed-T models assuming that the factors are unknown (e.g., Holtz-Eakin et al. (1988), Ahn et al. (2001), Ahn et al. (2013)). These methods require stronger assumptions on idiosyncratic errors (e.g., iid) or require exclusion restrictions.

[3]Such models have deep roots in economics; early analysis include Goldberger (1972), Jöreskog and Goldberger (1975), MaCurdy (1982), Holtz-Eakin et al. (1988) and Ahn et al. (2001) with many applications in micro- and macroeconomics, e.g., Chamberlain and Rothschild (1983), Altug and Miller (1990), Townsend (1994), Cawley et al. (1997), Carneiro et al. (2003).

example, each state's pre-existing welfare program differs in generosity and structure and the potential recipient populations are different. This implies that variations in trends that represent macroeconomic conditions or national sentiment on welfare policy (e.g., changes in welfare stigma) will have heterogeneous effects on welfare participation in different states, resulting in more elastic responses in some states than others.[4]

The paper is organized as follows. Section 2 describes the basic framework. Section 3 discusses the full model. Section 4 discusses the PCDID estimators and the parallel trend test. Section 5 discusses extensions to micro-level data. Section 6 reports results from Monte Carlo simulations. Sections 7 and 8 discuss the empirical analysis. Section 9 concludes. Supplementary materials are included in a separate appendix.

2 Basic Framework

We first illustrate the basic framework using potential outcomes. There are N units and T periods. Let $y_{it}(1)$ and $y_{it}(0)$ be the potential outcomes of unit i in period t with and without treatment, respectively. The treatment effect (TE) is denoted by $\Delta_{it} := y_{it}(1) - y_{it}(0)$. Let $D_{it} = 1$ if unit i receives treatment in period t, $D_{it} = 0$ otherwise. The observed outcome is $y_{it} = D_{it}y_{it}(1) + (1 - D_{it})y_{it}(0)$ (we cannot observe both $y_{it}(1)$ and $y_{it}(0)$). Denote the set of never-treated units as C (called "control units"; there are N_C such units), and the rest of units as E (called "treated units"; there are N_E such units). Suppose once a unit receives treatment, it remains so thereafter, i.e., $D_{it} = 1$ implies $D_{i,t+k} = 1$ for all $k > 0$. Let a treated unit $i \in E$ have T_{0i} pre-intervention periods and T_{1i} post-intervention periods, i.e., its first treated period is $T_{0i} + 1$. We can then express D_{it} as a product of two indicator functions: $D_{it} = 1_{\{i \in E\}} 1_{\{t > T_{0i}\}}$. The observed outcome can be rewritten as $y_{it} = \Delta_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + y_{it}(0)$.

We now impose a key functional form specification on potential outcome: $y_{it}(0) = \varsigma_i + \mu_i^{'} f_t + \tilde{\epsilon}_{it}$, where ς_i, μ_i, f_t, and $\tilde{\epsilon}_{it}$ are all individually unobserved (see the next section for covariates). The term $\mu_i^{'} f_t$ is known as a factor structure or interactive effects, which contains

[4] Similarly, labor market conditions may react differently to time-varying aggregate forces (e.g., structural or technological change) in states that are economic/population centers. See also Meyer (1995) and Blundell and Dias (2009) for a critique on DID.

an $\ell \times 1$ vector of factor loadings μ_i and time-varying factors f_t. The observed outcome is:

$$y_{it} = \Delta_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \varsigma_i + \mu_i' f_t + \tilde{\epsilon}_{it}, \tag{1}$$

The factor structure is used in various literatures such as the SC and matrix completion literatures (e.g., Xu (2017), Ferman and Pinto (2018), Ben-Michael et al. (2018), Athey et al. (2018), Chernozhukov et al. (2018), Arkhangelsky et al. (2019)). In the latter literatures, $\mu_i' f_t$ is seen as a low-rank (less complex) matrix which, together with idiosyncratic noise $\tilde{\epsilon}_{it}$, generates $y_{it}(0)$.

Equation (1) reduces to the two-way fixed effects (2wfe) model when factor loadings are homogeneous across units: $\mu_i = \mu_0$ for all i. The factor structure is then $\mu_0' f_t$, which can be re-expressed as a scalar time fixed effect τ_t, yielding $y_{it} = \Delta_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \varsigma_i + \tau_t + \tilde{\epsilon}_{it}$. In this case, regardless of the specification of f_t, the trend τ_t can be eliminated in a 2wfe regression (when $\Delta_{it} \equiv \Delta_0$). Recent work on this model have also shown that, when Δ_{it} is heterogeneous, a standard 2wfe regression yields an estimate that may no longer represent a useful causal parameter of interest, but instead a weighted average of treatment effects where some of the weights can be negative (e.g., Borusyak and Jaravel (2017), Abraham and Sun (2018), Athey and Imbens (2018), de Chaisemartin and D'Haultfœuille (2018), Goodman-Bacon (2018)).

We assume factor loadings are heterogeneous, which permits trends to be potentially unparallel across units, albeit parsimoniously via the factor structure. We allow for endogeneity of D_{it}. Some examples include correlation between: $1_{\{i \in E\}}$ and μ_i (loadings differ between control and treated units), $1_{\{t > T_{0i}\}}$ and f_t (e.g., factors increase over time), T_{0i} and μ_i (policy intervention date correlated with loadings).[5] We also examine what we call "weak parallel trend (PTW)" under the functional form specification in (1):

$$\textbf{PTW}: \qquad E(\mu_i | i \in E) = E(\mu_i | i \in C), \tag{2}$$

where the average factor loadings are the same between control and treated units. Section 4.4 provides a formal analysis of PTW when covariates are present, and discusses its rela-

[5] As in the 2wfe literature, $1_{\{i \in E\}}$ and T_{0i} may also be correlated with ς_i. In addition, μ_i may be correlated with ς_i.

tionship with the parallel trend assumption in a fully nonparametric framework (Callaway and Sant'Anna (2018)).[6] Nevertheless, whether PTW holds or not, it can be shown that the 2wfe regression is not sufficient to eliminate the nuisance terms related to the trends. These nuisance terms may create serious issues depending on the specification of f_t (e.g., nonstationarity).[7] Several representative numerical examples are discussed in Section 6.1.

Unlike synthetic-control type estimators, PCDID estimators are regression-based. It is therefore important to show that the PCDID regression coefficients converge to meaningful causal parameters of interest. Our key results show that the key causal parameters identified and estimated are ITET and ATET, defined as

$$\textbf{ITET}: \quad \bar{\Delta}_i := E(\Delta_{it}|t > T_{0i}) \text{ for fixed } i \in E, \tag{3}$$

$$\textbf{ATET}: \quad \bar{\bar{\Delta}} := E(\bar{\Delta}_i|i \in E). \tag{4}$$

The basic building block is the ITET $\bar{\Delta}_i$, which is the treatment effect of a unit $i \in E$ averaged over post-intervention periods. The ATET $\bar{\bar{\Delta}}$ is the average of the ITET across units $i \in E$. Under weak regularity conditions, the basic PCDID estimator converges to $\bar{\Delta}_i$ when $N_C, T \to \infty$ (Theorem 1), whereas the simple mean-group PCDID estimator (PCDID-MG) converges to $\bar{\bar{\Delta}}$ when $N_C, N_E, T \to \infty$ (Theorem 3). Moreover, the t-statistic from these estimators converges to a standard normal distribution and does not depend on the specification of f_t.

There are many alternative definitions of causal parameters, some of which are more natural under certain settings. For example, Callaway and Sant'Anna (2018) consider DID with multiple time periods in a fully nonparametric setting. They introduce *group-time average treatment effects* (in our notations, $ATT(g, t) := E(\Delta_{it}|T_{0i} + 1 = g)$), which is

[6] In a nonparametric framework, parallel trend holds when, given each $t > T_0$, the potential outcome $y_{it}(0)$ satisfies conditions akin to $E(y_{it}(0) - y_{i,t-1}(0)|i \in E) = E(y_{it}(0) - y_{i,t-1}(0)|i \in C)$; see Callaway and Sant'Anna (2018) for details. Given t and $t-1$, the factor structure implies $E(y_{it}(0) - y_{i,t-1}(0)|i \in E) = E(\mu_i^{'}|i \in E)(f_t - f_{t-1})$ and $E(y_{it}(0) - y_{i,t-1}(0)|i \in C) = E(\mu_i^{'}|i \in C)(f_t - f_{t-1})$, hence the interest in PTW.

[7] This can be illustrated by decomposing the simple DID estimator $DID_y := (\bar{y}_{E,post} - \bar{y}_{E,pre}) - (\bar{y}_{C,post} - \bar{y}_{C,pre})$ with self-explanatory notations, e.g., $\bar{y}_{E,post} = \frac{1}{N_E T_1}\sum_{i\in E}\sum_{t>T_0} y_{it}$, and assuming homogeneous $\Delta_{it} = \Delta_0$ and $T_{0i} = T_0$. Without covariates, DID_y is equivalent to a 2wfe regression. By simple arithmetic, $DID_y = \Delta_0 + (\bar{\mu}_E - \bar{\mu}_C)\left(\bar{f}_{post} - \bar{f}_{pre}\right) + (\bar{\epsilon}_{E,post} - \bar{\epsilon}_{E,pre}) - (\bar{\epsilon}_{C,post} - \bar{\epsilon}_{C,pre})$. Consider the multiplicative term $(\bar{\mu}_E - \bar{\mu}_C)\left(\bar{f}_{post} - \bar{f}_{pre}\right)$. If f_t follows a driftless random walk process (nonstationary) $f_t = f_{t-1} + \xi_t$, where ξ_t is an *iid* noise with mean zero, then $\bar{f}_{post} - \bar{f}_{pre}$ does not converge to any constants as $T \to \infty$; specifically, $\bar{f}_{post} = O_p(\sqrt{T_1})$ and $\bar{f}_{pre} = O_p(\sqrt{T_0})$, both growing without bound as T increases. When PTW does not hold, the multiplicative term is divergent as $N_C, N_E, T \to \infty$. When PTW holds, the limit depends on the convergence (and divergence) rate of $\bar{\mu}_E - \bar{\mu}_C$ versus $\bar{f}_{post} - \bar{f}_{pre}$.

identified under the assumption that the potential outcome $y_{it}(0)$ satisfies a parallel trend assumption for all $t \geq g$ given observed covariates (note that this assumption is violated in our context because the factor structure generates potentially unparallel trends). They show how this definition can be used to examine general dynamic treatment effects and aggregated into various summary measures of causal effects in the related literature. Reflecting the nature of our estimation method, our causal parameters consider aggregation over the time dimension first (ITET) before the cross-sectional dimension (ATET), an approach that has been considered by some studies, e.g., Goodman-Bacon (2018). As in most regression-based approaches, we need to impose some restrictions on treatment effect heterogeneity (see Section 4). For instance, Δ_{it} can be arbitrarily correlated with $1_{\{i \in E\}}$, $1_{\{t > T_{0i}\}}$, ς_i and μ_i, but some important treatment effect dynamics are ruled out, e.g., Δ_{it} cannot be driven directly by f_t.

Although some of the modeling assumptions that we introduce are rather technical, the key intuition that underlies our method is analogous to the control function approach in the microeconometric literature. We use data from control units ($i \in C$) to form $\ell \geq 1$ control functions that capture the unobserved factor structure, $\mu_i^{'} f_t$, which creates the endogeneity problem as illustrated earlier. Then, in a factor-augmented regression for $i \in E$, the endogeneity is corrected by including the control functions as extra covariates in the regression.

3 The Model

We consider an extension of equation (1), assuming the potential outcome $y_{it}(0)$ is a linear function of covariates:

$$y_{it} = \Delta_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \beta_i' x_{it} + \varsigma_i + \mu_i' f_t + \tilde{\epsilon}_{it}, \tag{5}$$

where the $k \times 1$ vector x_{it} stores the time-varying covariates with unit-specific parameters β_i.[8] The TE is assumed to take the following decomposition

$$\Delta_{it} = \bar{\Delta}_i + \tilde{\Delta}_{it}, \tag{6}$$

where $\bar{\Delta}_i := E(\Delta_{it}|t > T_{0i})$ is the ITET of unit i, a key estimand in this paper. The term $\tilde{\Delta}_{it}$ represents the deviation of Δ_{it} from the ITET, and it is the demeaned, time-varying idiosyncratic component of Δ_{it} for $t > T_{0i}$; by construction, $E(\tilde{\Delta}_{it}|t > T_{0i}) = 0$ for each $i \in E$. Substituting $\Delta_{it} = \bar{\Delta}_i + \tilde{\Delta}_{it}$ into (5) yields a reduced-form model, which is the main model that we examine:

$$y_{it} = \bar{\Delta}_i 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \beta_i' x_{it} + \varsigma_i + \mu_i' f_t + \epsilon_{it}, \tag{7}$$

where $\epsilon_{it} := \tilde{\Delta}_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \tilde{\epsilon}_{it}$ is the idiosyncratic error in the reduced-form model. The composite term $\tilde{\Delta}_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}}$ equals $\tilde{\Delta}_{it}$ when $i \in E$ and $t > T_{0i}$, and it equals zero otherwise. It does not contain $\tilde{\Delta}_{it}$ for any $i \in C$ or $t \leq T_{0i}$. The decomposition in equation (6) implies that the composite term has zero mean. To identify ITET and ATET, the errors ϵ_{it} need to satisfy further assumptions to be stated formally below. The first key assumption of our model is:

Assumption E (predeterminedness, treatment and intervention statuses):

(i) $E(\epsilon_{it}|\bar{\Delta}_i, 1_{\{i \in E\}}, 1_{\{t > T_{0i}\}}, \beta_i, \varsigma_i, \mu_i, f_t, x_{it}) = 0$ for each i and t.

(ii) $0 < E(1_{\{i \in E\}}) < 1$ for each i.

(iii) for each $i \in E$, $T_{1i}/T \xrightarrow{p} \kappa_i$ as $T, T_{1i} \to \infty$, where $0 < \kappa_i < 1$.

Assumption E(i) is a predeterminedness condition on idiosyncratic errors in the reduced-form model. It exerts orthogonality between ϵ_{it} and the conditioning components $\bar{\Delta}_i$, $1_{\{i \in E\}}$, $1_{\{t > T_{0i}\}}$, β_i ς_i, μ_i, f_t, x_{it}, but it does not preclude the conditioning components from being correlated with one another; see the examples of endogeneity in Section 2.[9] By the definition

[8] As in Wooldridge (2005) and Pesaran (2006), β_i is unit-specific and allowed to be correlated with the covariates x_{it} (see also Assumption E(i)). While Wooldridge shows that a fixed-effect estimator may consistently estimate the population average of the β_i, he does not consider a factor structure, a key feature in Pesaran (2006) and our model. As in the 2wfe literature, time-invariant covariates (e.g., demographic characteristics that do not vary over time) are subsumed into fixed effect ς_i. We do not attempt to estimate the parameters of these covariates.

[9] We do not impose the time-homogeneity condition on period-specific disturbances ($\mu_i f_t + \epsilon_{it}$ in our

of ϵ_{it} $(= \tilde{\epsilon}_{it} + \tilde{\Delta}_{it})$, E(i) imposes restrictions on treatment effect dynamics. Generally speaking, given each $i \in E$, the idiosyncratic TE component $\tilde{\Delta}_{it}$ cannot exhibit time-series correlation with f_t and x_{it} over $t > T_{0i}$, e.g., TE dynamics cannot be driven by time-varying factors and covariates directly. Assumption E(ii) is the same as Gobillon and Magnac (2016). It allows for the presence of control (i.e., never-treated) units, ruling out more extreme forms of staggered adoption in which all units receive treatment eventually; the literature has shown that, in such cases, the DID-2wfe estimator may represent uninterpretable weighted averages of treatment effects (see Section 2). Assumption E(iii) assumes the proportions of pre-intervention and post-intervention periods do not vanish to zero in the limit.

Next, we discuss the assumptions on factors and factor loadings. Denote $F := [f_1, \ldots, f_T]'$ and $X_i := \begin{bmatrix} 1 & ,\ldots, & 1 \\ x_{i1} & ,\ldots, & x_{iT} \end{bmatrix}'$, and let $G_i := (F, X_i)$. Denote $\mu_C := [\mu_{N_E+1}, \ldots, \mu_N]'$. As a convention, the norm of a matrix A is given by $\|A\| = [trace(A'A)]^{1/2}$, and "*a.s.*" stands for "almost surely".

Assumption F (factors and covariates): Let $\Gamma := diag(T^{r_1}, \ldots, T^{r_\ell}, T^{0.5}, \ldots, T^{0.5})$ where $r_1, \ldots, r_\ell \geq 0.5$. For each i, the following conditions are satisfied:
(i) For all T, $E \left\| \Gamma^{-1} G_i' G_i \Gamma^{-1} \right\|^2 \leq c$ for some constant $c > 0$.
(ii) $\text{plim}_{T\to\infty} \Gamma^{-1} G_i' G_i \Gamma^{-1}$ is positive definite *a.s.*.

Assumption FL$_C$ (factor loadings of control units): The following conditions are satisfied:
(i) For all $i \in C$, $E \|\mu_i\|^2 \leq c$ for some constant $c > 0$.
(ii) $\text{plim}_{N_C\to\infty} \frac{1}{N_C} \mu_C' \mu_C$ is positive definite.

Assumption F(i) accommodates a wide range of factor dynamics (deterministic and random, stationary and nonstationary) with possibly heterogeneous normalization orders.[10] The

context) as in the semi-/non-parametric panel data models of Chernozhukov et al. (2013). This also highlights the importance of imposing a (factor) structure on the disturbances. Nonetheless, the unobserved μ_i and f_t involve minimal assumptions; see Assumptions F and FL_C.

[10]The *normalization order* reflects the degree of complexity of the factor dynamics. It is defined as the maximum value of r such that the partial sum $\sum_{t=1}^T f_t^2$ is $O(T^{2r})$ if the factor f_t is deterministic, and is $O_p(T^{2r})$ if it is random. If f_t is deterministic, r is related to the *maximal polynomial order* d by $d = (2r-1)/2$. E.g., linear trends ($d = 1$; $r = 1.5$) and quadratic trends ($d = 2$; $r = 2.5$). If f_t is random, r is linked to the *integration order* d by $r = 0.5$ for $d \leq 0.5$, and $r = d$ for $d > 0.5$. E.g., stationary ARMA process ($d = 0$; $r = 0.5$), stationary long memory process ($0 < d < 0.5$; $r = 0.5$), and unit root process ($d = 1$; $r = 1$). Factors with structural breaks are accommodated as long as the integration order is well-defined. Since it is assumed that $r \geq 0.5$, weak factors are ruled out in our model (Onatski (2012)).

factors may contain a unit root (e.g., Bai (2004)), display long-range dependence (e.g., Ergemen and Velasco (2017), Ergemen (2019)), or exhibit structural breaks (e.g., Chen et al. (2014)). Assumption F(ii) ensures sufficient variability in factors and covariates and assumes linear independence, ruling out multicollinearity. Assumption FL_C(i) is a standard moment condition on the factor loadings of control units. Assumption FL_C(ii) ensures sufficient variability and assumes linear independence in factor loadings among control units as the number of control units grow large. Both Assumptions F and FL_C(ii) are crucial for asymptotic identification of the linear factor space (of dimension ℓ). Given the assumptions, the model given by (5)-(7) rules out the possibility that some treated units are exposed to factors that do not affect control unit outcomes. The factors and factor loadings may be deterministic or random – in the former case, the probability limits in the assumptions reduce to deterministic limits.

We want to highlight that identifying ITET and ATET requires identification of $\mu_i' f_t$ but not μ_i and f_t separately. This is evident from (5). Although μ_i and f_t are only separately identified up to a rotation matrix, we are not interested in, nor required to, identify and estimate them separately. Because $\mu_i' f_t = \mu_i' R R^{-1} f_t$ for all invertible rotation matrix R, it follows that both ITET and ATET are uniquely determined regardless of the choice of R.

4 Estimation Procedure and Key Results

4.1 Construction of Factor Proxies

The PCDID approach uses the following procedure to construct factor proxies, which serve as control functions in factor-augmented regression:

1. For each $i \in C$, perform a linear projection of y_{it} on x_{it} using data from $t = 1, \ldots, T$. Obtain the residuals $\hat{u}_{it}$ and form the $N_C \times T$ residual matrix $\hat{u}_C$.[11]

2. Apply PCA on the $N_C \times N_C$ sample covariance matrix $S := \hat{u}_C \hat{u}_C' / T$. With a prespecified p, obtain p factor proxies as the first p principal components of S. Specifically, first construct the $N_C \times p$ eigenvector matrix W with columns being the p eigenvectors associated with the p largest eigenvalues of S. Then obtain the $T \times p$ factor proxy

[11] This can be done by running a time-series regression (with intercept) of y_{it} on x_{it}. Other projections can also be used if we assume $\beta_i = \beta_0$ for all i; for example, run a fixed-effect panel regression of y_{it} on x_{it} using the control panel, and then obtain residuals (net of the fixed effects).

matrix $\hat{F} := \frac{\hat{u}_C' W}{N_C}$.

The factor proxies (columns of $\hat{F}$) can be viewed as weighted averages of the residuals from step 1 over the control panel. The sets of weights are determined by PCA such that the p linear combinations of the residuals jointly explain the most time-variation of control unit outcomes after partialling out the covariates. The residuals are, by construction of projection, orthogonal to the covariates. As long as the covariates are not linear combinations of the factors in model (7) (as implied by Assumption F(ii)), the residuals will preserve the rank and other information about the factor space. Note that the covariates and factors can be orthogonal or correlated. For instance, in Appendix 7, we consider a model in which the covariates are a linear function of factors plus idiosyncratic errors: $X_i = F\Pi_i + V_i$, which is contained as a special case studied by Pesaran (2006) and Bai (2009). The PCDID procedure remains valid for this example.

The total number of factors ℓ is assumed finite and unknown to users. This is an assumption commonly adopted in the interactive effects model literature (e.g., Bai (2003), Bai (2009), Pesaran (2006)). Moon and Weidner (2015) gives the theoretical result that, in these models, the least squares estimator of the regression coefficient is consistent for the true parameter as long as the number of factors used in estimation is at least ℓ. Our framework fits into theirs and has similar properties.[12] In practice, there are statistical methods that provide guidance on the number of factors p in finite samples, e.g., growth ratio (GR) test of Ahn and Horenstein (2013). Section 6.1 considers the performance of PCDID when the number of factors is chosen by the GR test.[13] We also consider conservative versions of the GR test, motivated by Moon and Weidner (2015)'s result that overestimation of the number of factors is less problematic than underestimation.

[12]In our model, the DID regressor is an indicator variable and, in the absence of other covariates, the PCs extracted beyond the ℓ-th one will come from reduced-form idiosyncratic errors ϵ_{it}. Although ϵ_{it} contains an idiosyncratic component $\tilde{\Delta}_{it}$, it is, by construction of the decomposition of Δ_{it}, orthogonal to the DID regressor. Asymptotic irrelevance of the extra PCs would hold as a special case of Moon and Weidner (2015).

[13]When the complexity of factor dynamics differ vastly across factors, we may rely on a systematic procedure that detects and extracts factors of different dynamical complexity in a recursive manner. Appendix 8 presents a procedure that works for an unknown number of factors and heterogeneous normalization orders that are integer-valued.

4.2 ITET Estimation

We now discuss the intuition behind the identification and estimation of ITET $\bar{\Delta}_i := E(\Delta_{it}|t > T_{0i})$, which is our casual parameter of interest for a given treated unit $i \in E$. The factor proxies $\hat{F}$ span the same linear space as the true factors F in the limit as the control panel grows in both the time and cross-sectional dimensions.[14] This large-sample result justifies the use of $\hat{F}$ as an approximation for F, and motivates the following factor-augmented time series regression (*aka PCDID regression*) for treated unit $i \in E$ using data from $t = 1, .., T$:

$$y_{it} = b_{0i} + \delta_i 1_{\{t>T_{0i}\}} + a_i' \hat{f}_t + b_{1i}' x_{it} + e_{it}, \tag{8}$$

where the $p \times 1$ vector $\hat{f}_t$ is the transpose of the t^{th} row of $\hat{F}$.

We now present the asymptotic theory of estimating ITET $\bar{\Delta}_i$. We rewrite the PCDID regression (8) in vector form. Let $1_{post,i}$ be the $T \times 1$ vector consisting of T_{0i} zeros followed by T_{1i} ones. For treated unit $i \in E$,

$$y_i = \delta_i 1_{post,i} + \hat{F} a_i + X_i b_i + e_i, \tag{9}$$

where $y_i = [y_{i1}, \ldots, y_{iT}]'$, $e_i = [e_{i1}, \ldots, e_{iT}]'$, $b_i = [b_{0i}, b_{1i}']'$, and $\hat{F} := \hat{u}_C' W / N_C$, the $T \times \ell$ factor proxy matrix. The PCDID estimator of $\bar{\Delta}_i$ is $\hat{\delta}_i$, the least squares estimator of δ_i:

$$\hat{\delta}_i = (1_{post,i}' M_{[\hat{F},X_i]} 1_{post,i})^{-1} 1_{post,i}' M_{[\hat{F},X_i]} y_i, \tag{10}$$

where $M_A = I - A(A'A)^{-1}A'$ for a given matrix A.

In the following, we list and discuss the key assumptions (**Assumption AI**$_i$ and **ES**) for the asymptotic analysis of PCDID estimation. Define $\tilde{G}_i = \{1_{post,i}, G_i\} = \{1_{post,i}, F, X_i\}$.

Assumption AI$_i$ (asymptotic identification, PCDID estimator): For each $i \in E$,
(i) $\rho_i := \text{plim}_{T\to\infty} \frac{1}{T} 1_{post,i}' M_{G_i} 1_{post,i}$ exists and is strictly positive *a.s.*.
(ii) $\xi_i^2 := \text{plim}_{T\to\infty} E\left(\left\| \frac{1}{\sqrt{T}} 1_{post,i}' M_{G_i} \epsilon_i \right\|^2 \middle| \tilde{G}_i \right)$ exists and is strictly positive *a.s.*.

Assumption ES (strict exogeneity in time): For each i, $E(\epsilon_{it}|\tilde{G}_i) = 0$ *a.s.* for all t.

[14] More precisely, we show that estimation error of the sample cross moments (over t) of $\hat{F}$ and other model components (e.g., covariates, idiosyncratic errors) vanishes as $N_C, T \to \infty$. See lemma A4(d)-(g) for technical details.

The key assumption for asymptotic identification of $\bar{\Delta}_i$ is **Assumption AI$_i$(i)**. This assumption rules out multi-collinearity of the intervention status $1_{post,i}$ with X_i and F in the limit. This is a relatively weak assumption; multi-collinearity occurs when X_i or F is a step function that jumps at $T_{0i}+1$ but constant otherwise, which is not common in practice. Gobillon and Magnac (2016) imposes an analogous condition under finite T in the absence of covariates, but that is insufficient for asymptotic identification of $\bar{\Delta}_i$ in our model.[15] **Assumption AI$_i$(ii)** ensures that $\hat{\delta}_i$ has a non-degenerate distribution after normalization. **Assumption ES** strengthens **Assumption E(i)** to a strict exogeneity condition on the time series of factors, intervention status and covariates. This is crucial for the conditioning argument that leads to our asymptotic normality result.

More technical assumptions (**Assumptions IE, M, D$_i$**) are detailed in Appendix 1 due to space limitation. These are regularity conditions that are relatively standard in the interactive effects and time series literatures. We briefly summarize them below and highlight the key properties and restrictions. **Assumption IE** accommodates heteroskedasticity and weak dependence (e.g., cross-sectional and serial correlations) in the idiosyncratic errors ϵ_{it} $(= \tilde{\epsilon}_{it} + \tilde{\Delta}_{it})$. It assumes the higher-order moments are bounded, which rules out unbounded treatment effect dynamics. **Assumption M** exerts control on the dependence among various model components (idiosyncratic errors, factors, factor loadings, covariates and their cross-products). For analytical convenience, covariates and factors are assumed to be orthogonal (**Assumption MX(iii)**), although this is not necessary for the PCDID approach to deliver valid results; see Section 4.1 and Appendix 7. **Assumption D$_i$** governs the dynamical properties of regressors, factors and idiosyncratic errors by allowing them to be mixing and heteroskedastic over time. This enables us to apply the functional central limit theorem of DeJong and Davidson (2000).[16]

Theorem 1 states the consistency and asymptotic normality results related to $\hat{\delta}_i$:

Theorem 1 *Suppose Assumptions E, F, FL$_C$, AI$_i$, IE and M hold. Then, as* $T, N_C \to \infty$

[15] Requiring that $1_{post,i}$ and F are not collinear for all finite T is not enough, because (i) the estimation error of $\hat{F}$ does not vanish in finite samples, and (ii) this is not equivalent to non-collinearity in the limit. A counterexample is given by $f_t = c_0 + c_1 1_{\{t>T_{0i}\}} + u_t$ where u_t is *iid* $N(0, T^{-1})$. Then, for all finite T, $1_{\{t>T_{0i}\}}$ and f_t satisfy the non-collinear condition in equation (15) of Gobillon and Magnac (2016), but **Assumption AI$_i$(i)** is clearly violated as $T \to \infty$.

[16] While fractionally integrated processes are ruled out in **Assumption D$_i$**, there exist more general central limit theorems that accommodate factors with fractional normalization order (e.g., Davidson and DeJong (2000)). To avoid further complicating our analysis, we do not proceed with such level of generality.

jointly and $\frac{\sqrt{T}}{N_C} \to 0$*, we have for each* $i \in E$*:*

(a) $\hat{\delta}_i \xrightarrow{p} \bar{\Delta}_i$.

(b) $\sqrt{T}\sigma_{Ti}^{-1}(\hat{\delta}_i - \bar{\Delta}_i) \xrightarrow{d} N(0,1)$ if additionally Assumptions ES and D_i hold, where $\sigma_{Ti}^2 := Var[\sqrt{T}(\hat{\delta}_i - \bar{\Delta}_i)|\tilde{G}_i]$.

The consistency and asymptotic normality result is robust to a wide range of factor dynamics, including nonstationary processes. This may seem surprising given that the DID estimator is potentially inconsistent when the factors contain a unit root (see Section 2). The intuition behind the result is that the outcomes and factors form a cointegrating relationship in our model. Since the intervention dummy $1_{post,i}$ is a bounded time series, its coefficient δ_i can be estimated consistently by least squares method at the $\sqrt{T}$-rate regardless of the factor dynamics.

The factor proxies contain not only F but also idiosyncratic errors of control units. The "measurement error" thus introduced creates an endogeneity issue when the idiosyncratic errors of treated and control units are correlated. The asymptotic condition $\sqrt{T}/N_C \to 0$ is necessary to remove the asymptotic bias due to the measurement error.

An important implication of Theorem 1(b) is that the limiting distribution of the studentized statistic remains to be $N(0,1)$ under a wide range of factors (stationary and nonstationary) allowed by Assumption F.[17] Inference can be carried out using the studentized form of the PCDID estimate and standard critical values. To compute the t-statistic, the population standard deviation σ_{Ti} may be replaced by a suitable sample analog that reflects the error dependence structure, e.g., the Newey-West HAC standard error is nonparametric and is widely used in the time series literature. An alternative approach is bootstrapping the t-statistic.[18] More details about the bootstrap implementation are given in Appendix 6. The finite sample performance of the inference procedures is studied in Section 6.2.

Remark: The PCDID method resembles SC, unconfoundedness and matrix completion approaches in that they allow for unparallel trends via factor models. PCDID uses all periods

[17]This is a non-trivial result. For example, Chernozhukov et al. (2018) state that the key assumption underlying their inference procedure is stationarity of the data.

[18]Although we are unable to provide formal proofs, the asymptotic normality result provides support for bootstrapping the t-statistic. By contrast, the limiting distribution of the non-studentized statistic may vary discontinuously with the serial dependence properties of the factors (e.g., stationary, near unit-root and unit-root processes). Hence we do not recommend bootstrapping the PCDID estimate directly.

of data from treated units in a factor-augmented regression to estimate its target parameter (e.g., equation (8)). The factor proxies are constructed from data on control units ($i \in C$) only. The other approaches first estimate the missing potential outcome ($\hat{y}_{it}(0)$) of treated units after intervention ($i \in E$ and $t > T_{0i}$) using data from treated units before intervention ($i \in E$ and $t \leq T_{0i}$) and control units ($i \in C$); then, they compute $\hat{\Delta}_{it} := y_{it}(1) - \hat{y}_{it}(0)$.[19] The regression-based approach of PCDID suggests that assumptions on treatment effect dynamics are needed to ensure that the estimator converges to a well-defined causal parameter of interest (in this case, ITET). Xu (2017)'s GSC estimator is particularly relevant to ours, as it: (1) estimates factors and factor loadings using the control observations; (2) uses these estimates and the pre-intervention periods of treated units to estimate the factor loadings of treated units, and then (3) use all the above estimates to predict $\hat{y}_{it}(0)$ of treated units after intervention. Appendix 3.1 provides a full comparison of identification and efficiency of PCDID and GSC. We show that, given the estimands are aligned, both approaches are numerically different (even in the absence of covariates), GSC uses stronger assumption on identification, PCDID is at least as efficient as GSC, and the former is more efficient when the factor structure is correlated with D_{it}. The intuition is that GSC discards data in the post-intervention subsample which contain useful information about the correlation between factors and intervention dummy (see #2 above). This leads to efficiency loss, which magnifies as $\frac{T_{1i}}{T}$ becomes higher. Their finite-sample performance is compared in Section 6.

4.3 ATET Estimation

We now discuss the identification and estimation of ATET $\bar{\bar{\Delta}} := E(\bar{\Delta}_i | i \in E)$, the key estimand in this section.[20] As discussed in Section 2, this estimand is built on the ITET $\bar{\Delta}_i$. To facilitate the ensuing discussion, we break down the treatment effect Δ_{it} using decomposition

[19] Vertical regressions (SC methods), horizontal regression (unconfoundedness methods), and regularized matrix completion methods have been used to estimate $\hat{y}_{it}(0)$. Doubly weighted methods with both vertical and horizontal regressions have recently been considered, e.g., Arkhangelsky et al. (2019). Their method numerically solves for vertical and horizontal weights with non-negativity constraints; hence it can be viewed as an extension of Abadie et al. (2010) where vertical nonnegative weights are involved. Gobillon and Magnac (2016) derive support conditions implied by nonnegative weights in Abadie et al. (2010). Athey et al. (2018) show that vertical and horizontal regressions belong to the same class of matrix completion methods based on matrix factorization (i.e., low-rank matrices) with different restrictions/regularizations.

[20] Although we only focus on the ATET, other estimands are available, such as the conditional moments and quantiles of ITET among $i \in E$. These estimands can help unveil the distributional features of ITET.

(6) and the definition of ATET, obtaining

$$\Delta_{it} = \bar{\Delta}_i + \tilde{\Delta}_{it} = \bar{\bar{\Delta}} + \upsilon_i + \tilde{\Delta}_{it}, \tag{11}$$

where $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}}$ is the unit-specific deviation of the ITET from the ATET. By construction, $E(\upsilon_i | i \in E) = 0$. We consider the following two cases:

(a) Homogeneous ITET over treated units: $\upsilon_i = 0$ for all $i \in E$; and

(b) Heterogeneous ITET over treated units: υ_i varies over $i \in E$.

Case (a) assumes that $\bar{\Delta}_i := E(\Delta_{it} | t > T_{0i})$ are identical for all $i \in E$. This assumption is weaker than full homogeneity of Δ_{it} (i.e., $\Delta_{it} = \Delta_0$), but stronger than case (b). We will consider two types of estimators, the simple mean-group estimator and the pooled estimator. We will show that in case (a), both estimators identify the ATET under similar assumptions. However, to identify ATET in case (b), the pooled estimator requires stronger assumptions than the simple mean-group estimator. Both estimators have a faster asymptotic convergence rate in case (a) than in case (b). The pooled estimator is analogous to the least square estimator in 2wfe regression, and is included for completeness of our theory. Readers who are not interested in a thorough comparison may skip the section on the pooled estimator.

4.3.1 Simple Mean-group Estimator

The *simple mean-group estimator* (PCDID-MG) is defined as the simple average of the ITET estimates:[21]

$$\hat{\delta}^{mg} = \frac{1}{N_E} \sum_{i \in E} \hat{\delta}_i. \tag{12}$$

The additional assumptions for deriving the large-sample properties of the PCDID-MG estimator are listed below. Define $\tilde{G} := \{\tilde{G}_i\}_{i \in E}$.

Assumption FL (factor loadings): Assumption FL_C holds. In addition,
(i) For all $i \in E$, $E \left\| \mu_i \right\|^2 \leq c$ for some constant $c > 0$.

[21] A more general estimator is the *weighted mean-group estimator*, defined as $\hat{\delta}^{mg,(\omega)} = \sum_{i \in E} \omega_i \hat{\delta}_i$ for a given set of weights $\omega = (\omega_1, \omega_2, \ldots, \omega_{N_E})'$ on the treated group with $\sum_{i \in E} \omega_i = 1$. It encompasses both the simple mean-group and pooled estimators as special cases.

(ii) $\text{plim}_{N_E \to \infty} \frac{1}{N_E} \mu_E' \mu_E$ is positive definite.

Assumption AI$_{mg}$ (asymptotic identification, simple mean-group estimator):
(i) $\text{plim}_{N_E, T \to \infty} \inf_{i \in E} \frac{1}{T} 1_{post,i}' M_{G_i} 1_{post,i}$ exists and is strictly positive $a.s.$.
(ii) $\zeta^2 := \text{plim}_{N_E, T \to \infty} E\left(\left\| \frac{1}{\sqrt{N_E T}} \sum_{i \in E} 1_{post,i}' M_{G_i} \epsilon_i \right\|^2 \middle| \tilde{G} \right)$ exists and is strictly positive $a.s.$.

Assumption ESS (strict exogeneity in panel): $E(\epsilon_{it} | \tilde{G}) = 0$ $a.s.$ for all $i \in E$ and t.

Assumption D (panel dependence): Assumption D$_i$ holds for all $i \in E$. In addition, the following conditions are satisfied:
(i) ϵ_{it} are independent over $i \in E$.
(ii) $\zeta_i^2 := \lim_{T \to \infty} Var(T^{-1/2} 1_{post,i}' M_{G_i} \epsilon_i | \tilde{G})$ satisfies $\frac{\max_{i \in E} \zeta_i^2}{\sum_{i \in E} \zeta_i^2} = O_p\left(\frac{1}{N_E}\right)$ as $N_E \to \infty$.

Assumption RT$_{mg}$ (treatment effects, simple mean-group estimator): Let $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}}$. The following conditions are satisfied:
(i) For some $p > 2$, there exists $0 < c < \infty$ such that $E\,|\upsilon_i|^p \leq c$ for all $i \in E$.
(ii) υ_i is a mixing process with mixing coefficient ϕ of size $-p/2(p-1)$ for $p \geq 2$, or α of size $-p/(p-2)$, $p > 2$.
(iii) $\lim_{N_E \to \infty} Var(N_E^{-1/2} \sum_{i \in E} \upsilon_i)$ exists and is strictly positive.

We highlight the key restrictions below. To account for multiple treated units, Assumptions **FL, AI$_{mg}$, ESS and D** involve stronger (but still relatively standard) regularity conditions than their previous counterparts (Assumptions FL$_C$, ES and D$_i$ in Theorem 1). This also applies to the technical assumption **MM** (relative to Assumption M in Theorem 1) in Appendix 1. The key asymptotic identification assumption, **AI$_{mg}$(i)**, requires that every single $\hat{\delta}_i$ be well defined.

Assumptions ESS and **D** are only needed for deriving asymptotic normality related to $\hat{\delta}^{mg}$ *under homogeneous ITET (case (a)).* **Assumption ESS** requires that the regressors are strictly exogeneous in the panel setting. This is necessary for pursuing a conditioning argument that yields asymptotic normality. **Assumption D(i)** restricts the idiosyncratic errors to be independent over treated units. **Assumption D(ii)** is an asymptotic negligibility condition on the cross-sectional variation of the treatment panel. While stronger than weak dependence (**Assumption IE**), the restrictions enable us to invoke the joint central limit theorem of Phillips and Moon (1999) and obtain asymptotic normality. It is possible to relax the cross-sectional independence at the expense of more technical assumptions.

By contrast, under heterogeneous ITET (case (b)), we need another assumption, $\mathbf{RT}_{mg}$, to derive consistency and asymptotic normality. It is a set of weak regularity conditions that place control on the cross-sectional variation and higher-order moments of the term υ_i $(:= \bar{\Delta}_i - \bar{\bar{\Delta}})$. No parametric assumptions are involved. Importantly, neither $\mathbf{RT}_{mg}$ nor the other assumptions preclude $\bar{\Delta}_i$ (or Δ_{it}) from being cross-sectionally correlated with $1_{\{i \in E\}}$, T_{0i}, x_{it}, ς_i and μ_i (see Section 3). This relaxes Gobillon and Magnac (2016)'s assumption that covariates are uncorrelated with treatment effects for the treated, for example.

The asymptotic result for PCDID-MG is presented below.

Theorem 2 (simple mean-group estimator) *Suppose Assumptions E, F, FL, AI_{mg}, IE and MM hold. As $T, N_E, N_C \to \infty$ jointly and $\frac{T}{N_C} \to 0$, we have the following results:*

(a) (homogeneous ITET) Suppose $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}} \equiv 0$. Then,

(i) $\hat{\delta}^{mg} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E T}\bar{\varsigma}^{-1}_{N_E,T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$ if additionally Assumptions ESS and D hold, where $\bar{\varsigma}^2_{N_E,T} := Var[\sqrt{N_E T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})|\tilde{G}]$.

(b) (heterogeneous ITET) Suppose υ_i satisfies Assumption RT_{mg}. Then,

(i) $\hat{\delta}^{mg} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E}\bar{\bar{\varsigma}}^{-1}_{N_E,T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$, where $\bar{\bar{\varsigma}}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})]$.

A number of remarks are in order. First, the asymptotic efficiency of the PCDID-MG estimator depends on the extent of treatment effect heterogeneity in the model. PCDID-MG is $\sqrt{N_E}$-consistent under heterogeneous ITET, and achieves a faster $\sqrt{N_E T}$ rate of convergence under homogeneous ITET.[22] The intuition behind this difference is that, under heterogeneous ITET, υ_i dominates the idiosyncratic errors ϵ_{it} in the limit. Second, compared to ITET estimation, a stronger asymptotic condition $T/N_C \to 0$ is required to remove the asymptotic bias due to the estimation error of factor proxies. This is necessary when idiosyncratic errors in the treated and control panels are correlated. Third, the consistency and asymptotic normality results in Theorem 2(a) are comparable to the interactive effects estimator of Bai (2009)

[22] In addition, it can be easily shown using the same technique that when the treatment effect is fully homogeneous (i.e., $\Delta_{it} = \Delta_0$), the rate of convergence is also $\sqrt{N_E T}$.

(Theorem 2), in which the analysis is confined to the case of homogeneous slope parameters (corresponding to full TE homogeneity ($\Delta_{it} = \Delta_0$) in our context) and stationary factors.

As in Theorem 1, the limiting distribution of the studentized PCDID-MG statistic remains to be $N(0,1)$ under a wide range of factors (stationary and nonstationary), which greatly facilitates inference procedures. In particular, under heterogeneous ITET, the idiosyncratic errors are dominated by the variation in υ_i in the limit (see previous paragraph). This justifies a nonparametric variance estimator as in Pesaran (2006): $\widehat{var}(\hat{\delta}^{mg}) = \frac{1}{N_E(N_E-1)} \sum_{i\in E}(\hat{\delta}_i - \hat{\delta}^{mg})^2$, which provides a convenient way to form the t-statistic. More details regarding inference can be found in Appendix 6 and Section 6.2.

4.3.2 Pooled Estimator

We consider the following factor-augmented panel regression using data from the treated panel: for $i \in E$ and $t = 1, ..T$,

$$y_{it} = b_{0i} + \delta 1_{\{t>T_{0i}\}} + a_i' \hat{f}_t + b_{1i}' x_{it} + e_{it}. \tag{13}$$

The *pooled estimator* is the least squares estimator of δ, given by:

$$\hat{\delta}^{pl} = \left(\sum_{i\in E} 1_{post,i}' M_{[\hat{F},X_i]} 1_{post,i} \right)^{-1} \sum_{i\in E} 1_{post,i}' M_{[\hat{F},X_i]} y_i. \tag{14}$$

We first list below the assumptions that are different from those for PCDID-MG:

Assumption AI$_{pl}$ (asymptotic identification, pooled estimator): Same as Assumption AI$_{mg}$, except replacing AI$_{mg}$(i) by:
(i) $\rho := \text{plim}_{N_E,T\to\infty} \frac{1}{N_E T} \sum_{i\in E} 1_{post,i}' M_{G_i} 1_{post,i}$ exists and is strictly positive $a.s.$.

Assumption RT$_{pl}$ (treatment effects, pooled estimator): Let $R_{Ti} := \frac{1}{T} 1_{post,i}' M_{G_i} 1_{post,i}$, $MSR_T := \{R_{Ti}\}_{i\in E}$ and $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}}$. The following conditions are satisfied:
(i) For each $i \in E$, $E(\upsilon_i | MSR_T) = 0$ $a.s.$.
(ii) For some $p > 2$, there exists $0 < c < \infty$ such that $E(|\upsilon_i|^p \, | MSR_T) \le c$ for all $i \in E$ and T.
(iii) Conditional on MSR_T, υ_i is a mixing process with mixing coefficient ϕ of size $-p/2(p-1)$ for $p \ge 2$, or α of size $-p/(p-2)$, $p > 2$.

(iv) $\text{plim}_{N_E,T\to\infty} Var(N_E^{-1/2}\sum_{i\in E} R_{Ti}\upsilon_i|MSR_T)$ exists and is strictly positive *a.s.*.

Assumption AI$_{pl}$ is the key identification condition for $\hat{\delta}^{pl}$ under homogeneous ITET. It is weaker than AI$_{mg}$ due to the different manner that the estimator pools information from the treated units, as is evident from the summations in formula (14). Unlike PCDID-MG, $\hat{\delta}^{pl}$ does not require $\hat{\delta}_i$ to be well-defined for all $i \in E$, i.e., it allows time-series multicollinearity to occur in a finite number of treated units as long as there is sufficient variation in intervention status in the treated panel.

Assumption RT$_{pl}$ is the *extra* key identification condition for $\hat{\delta}^{pl}$ under heterogeneous ITET. It is stronger than Assumption RT$_{mg}$ due to the presence of a conditioning variable R_{Ti}, which represents unit i's residual variance of the intervention status $1_{post,i}$ after it is projected on X_i and F. In the absence of covariates and factors, R_{Ti} is the variance of demeaned $1_{post,i}$, equalling $\frac{T_{0i}(T-T_{0i})}{T^2}$, which is quadratic in T_{0i} and maximized when $T_{0i} = \frac{T}{2}$ (i.e., variance is largest when the intervention date is $\frac{T}{2}$). **RT**$_{pl}$**(i)** implies that units with T_{0i} farther away from $\frac{T}{2}$ have similar ITET as units with T_{0i} closer to $\frac{T}{2}$. This assumption is crucial in ensuring the pooled estimator identifies the ATET, consisting of ITETs with equal weights across $i \in E$. Otherwise, $\hat{\delta}^{pl}$ identifies a weighted function of ITETs.[23] A similar condition can be found in Goodman-Bacon (2018) for DID models. Note that when $T_{0i} = T_0$ (homogeneous intervention date) or T_{0i} is randomly assigned as in an experiment, **RT**$_{pl}$ is satisfied.

The asymptotic results below are parallel to those for PCDID-MG:

Theorem 3 (pooled estimator) *Suppose Assumptions E, F, FL, AI*$_{pl}$*, IE and MM hold. As* $T, N_E, N_C \to \infty$ *jointly and* $\frac{T}{N_C} \to 0$*, we have the following results:*

(a) (homogeneous ITET) Suppose $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}} \equiv 0$*. Then,*

(i) $\hat{\delta}^{pl} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E T}\bar{\sigma}^{-1}_{N_E,T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$ *if additionally Assumptions ESS and D hold, where* $\bar{\sigma}^2_{N_E,T} := Var[\sqrt{N_E T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|\tilde{G}]$.

[23] The pooled estimator can be expressed as $\hat{\delta}^{pl} = \sum_{i\in E}\omega_i\hat{\delta}_i = \bar{\bar{\Delta}} + \sum_{i\in E}\omega_i\upsilon_i + o_p(1)$, where $\omega_i = R_{Ti}/\sum_{j\in E} R_{Tj}$. It follows that $\hat{\delta}^{pl}$ is consistent if R_{Ti} and υ_i are uncorrelated in the limit. By contrast, it is clear from the decomposition $\hat{\delta}^{mg} = N_E^{-1}\sum_{i\in E}\hat{\delta}_i = \bar{\bar{\Delta}} + N_E^{-1}\sum_{i\in E}\upsilon_i + o_p(1)$ that the simple mean-group estimator is consistent under milder conditions on υ_i without involving R_{Ti} (Assumption RT$_{mg}$).

(b) (heterogeneous ITET) Suppose υ_i satisfies Assumption RT_{pl}. Then,

(i) $\hat{\delta}^{pl} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E}\bar{\tilde{\sigma}}^{-1}_{N_E,T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$*, where* $\bar{\tilde{\sigma}}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|MSR_T]$.

4.3.3 Efficiency Comparison of Estimators

Theroems 2 and 3 indicate that the simple mean-group and pooled PCDID estimators share the same asymptotic rate of convergence, but they do not provide further details about their asymptotic variances. Under the simple setting in which the idiosyncratic errors are *iid*, it is possible to carry out a more refined efficiency analysis that would shed light on the preferred estimator to use (provided all the assumptions are satisfied).[24]

The asymptotic result below shows that: (a) under homogeneous ITET, the pooled estimator is more efficient than the simple mean-group estimator; (b) under heterogeneous ITET, the reverse is true. Here, σ^2 and ς^2 denote the asymptotic variances of the pooled and simple mean-group estimators, respectively, when ITET are homogeneous. Similarly, $\tilde{\sigma}^2$ and $\tilde{\varsigma}^2$ denote the associated quantities under the heterogeneous ITET case.

Theorem 4 (asymptotic efficiency) *Suppose Assumptions E, F, FL, IE, D, AI_{pl}, AI_{mg} and MM hold. For each $i \in E$, define $\rho_i := \text{plim}_{T\to\infty} R_{Ti}$, where $R_{Ti} = 1'_{post,i} M_{G_i} 1_{post,i}/T$. The following results are valid a.s.:*

(a) (homogeneous ITET) Suppose ϵ_{it} are iid over i and t with mean 0 and variance σ^2_ϵ. Then $\sigma^2 \leq \varsigma^2$ a.s.. Equality holds iff ρ_i are identical over $i \in E$.

(b) (heterogeneous ITET) Suppose υ_i are iid with mean 0 and variance σ^2_υ, and that υ_i are independent of R_{Tj} for all $i, j \in E$. Then $\tilde{\varsigma}^2 \leq \tilde{\sigma}^2$. Equality holds iff ρ_i are identical over $i \in E$.

[24] Note that the optimality result of Theorem 4 is sensitive to the error dependence structure, which affects the asymptotic variance of the estimators. Nonetheless, the argument leading to this result is constructive in the search for the asymptotically efficient estimator under more general error dependence structure. More generally, the *weighted mean-group estimator* achieves asymptotic efficiency when the weights are chosen to be proportional to the reciprocal of the (conditional) variance of the associated ITET estimator.

4.4 Parallel Trend Test under the Factor Structure

We develop a test of what we call "weak parallel trends (PTW)" under the functional form specification in equation (5):

PTW: $E(\mu_i|i \in C) = E(\mu_i|i \in E) =: \mu_0$ for some finite and non-zero vector μ_0,

which posits that the expected factor loadings are the same between control and treated units. To motivate the test, we first note that any $\ell \times 1$ factor loading of treated unit j can be uniquely represented by:

$$\mu_j = \alpha_j E(\mu_i|i \in C) + v_j, \tag{15}$$

where α_j is a scalar and v_j is an $\ell \times 1$ vector.[25] In other words, the factor loading vector of treated unit j is proportional to the mean factor loading vector over control group, except for the deviation v_j.

Next, we rewrite the reduced-form model (7) for $j \in E$ as:

$$y_{jt} = \varsigma_j + \beta_j' x_{jt} + \bar{\Delta}_j 1_{\{t>T_{0j}\}} + \alpha_j E(\mu_i'|i \in C) f_t + v_j' f_t + \epsilon_{jt}.$$

The linear combination $E(\mu_i'|i \in C) f_t$ is a time-varying *scalar* variable that captures the cross-sectional average of the factor structure among control units. Since $E(\mu_i'|i \in C) f_t = E(y_{it} - \varsigma_i - \beta_i' x_{it}|i \in C)$ by (7) for $i \in C$ and Assumption E(i), it is approximated by the cross-sectional average $\bar{u}_{Ct} := \frac{1}{N_C} \sum_{i \in C} \hat{u}_{it}$, where $\hat{u}_{it}$ is the control panel residual from the linear projection of y_{it} on x_{it} (with intercept); see step 1 of the PCDID procedure in Section 4.1. This motivates a factor-augmented time-series regression for each $j \in E$:

$$y_{jt} = b_{0j} + b_{1j}' x_{jt} + \delta_j 1_{\{t>T_{0j}\}} + a_j \bar{u}_{Ct} + e_{jt}. \tag{16}$$

from which we obtain the OLS estimator $\hat{a}_j$ of a_j. A single factor proxy $\bar{u}_{Ct}$ in equation (16) is adequate as control for $E(\mu_i'|i \in C) f_t$ although there may be multiple factors in the model (unlike PCDID-MG in Section 4.3.1).

[25] To ensure uniqueness, v_j must satisfy normalization restrictions, e.g., the sum of the elements of v_j is zero and that $v_j \neq \gamma_j E(\mu_i|i \in C)$ for any non-zero scalars γ_j. See Appendix 4.1 for details.

We define the *Alpha statistic* as the simple mean-group estimator

$$\hat{a}^{mg} := \frac{1}{N_E} \sum_{j \in E} \hat{a}_j.$$

From (15), we see that $\alpha := E(\alpha_j | j \in E) = 1$ and $E(v_j | j \in E) = 0$ (a zero vector) under PTW. The converse is also true.[26] The asymptotic result that $\hat{a}^{mg}$ estimates $\alpha = 1$ under PTW forms the basis of *Alpha test* for weak parallel trends.

The Alpha test performance (including test size and power) is not affected by the specific normalization applied to the factors and factor loadings. This follows from the rotational invariance of the factor structure $\mu_i' f_t$ (i.e., for any $\ell \times \ell$ invertible matrix R, we have $\mu_i' f_t = \mu_i' R R^{-1} f_t$). This holds for all $i \in C \cup E$ as long as a common normalization is applied to all treated and control units (see also the end of Section 3).

To prepare for the asymptotic analysis, we list the assumptions below.

Assumption AI$_\alpha$ (asymptotic identification, alpha test): Let r be the normalization order of $F\mu_0$ such that $\|F\mu_0\|^2 / T^{2r} = O_p(1)$ as $T \to \infty$. The following conditions hold:
(i) $\text{plim}_{N_E, T \to \infty} \inf_{i \in E} \frac{1}{T^{2r}} \mu_0' F' M_{[1_{post,i}, X_i]} F \mu_0$ exists and is strictly positive *a.s.*.
(ii) $\text{plim}_{N_E, T \to \infty} E\left(\left\| \frac{1}{\sqrt{N_E} T^r} \sum_{i \in E} \mu_0' F' M_{[1_{post,i}, X_i]} F(\mu_i - \mu_0) \right\|^2 \middle| \tilde{G} \right)$ exists and is strictly positive *a.s.*.

Assumption FLM (mixing factor loadings): The following conditions are satisfied:
(i) For some $p > 1$, there exists $0 < c < \infty$ such that $E(\|\mu_i\|^p) \leq c$ for all $i \in C \cup E$.
(ii) $\{\mu_i : i \in C\}$ and $\{\mu_i : i \in E\}$ are mixing sequences with mixing coefficients ϕ of size $-p/(2p-1)$ for $p \geq 1$, or α of size $-p/(p-1)$ for $p > 1$.

Assumption FLM2 (conditional mixing factor loadings, treated units): The following conditions are satisfied:
(ii) For each $i \in E$, $E(\mu_i | \tilde{G}) = \mu_0$ *a.s.*.
(ii) For some $p > 2$, there exists $0 < c < \infty$ such that $E(\|\mu_i\|^p | \tilde{G}) \leq c$ for all $i \in E$.
(iii) Conditional on $\tilde{G}$, $\{\mu_i : i \in E\}$ is a mixing sequence with mixing coefficients ϕ of size $-p/2(p-1)$ for $p \geq 2$, or α of size $-p/(p-2)$ for $p > 2$.

[26] Suppose $\alpha = 1$ and $E(v_j | j \in E) = 0$. It follows from (15) that $E(\mu_j' | j \in E) f_t = E(\mu_i' | i \in C) f_t$ for all t, or in vector form: $F[E(\mu_j | j \in E) - E(\mu_i | i \in C)] = 0$. That F has full column rank ℓ (Assumption F(ii)) implies $E(\mu_j | j \in E) = E(\mu_i | i \in C) = \mu_0$ where $\mu_0 \neq 0$, and hence PTW is satisfied.

Assumption $\mathbf{AI}_\alpha$**(i)** rules out multicollinearity and ensures the Alpha statistic is well-defined. **Assumption** $\mathbf{AI}_\alpha$**(ii)** ensures that the Alpha statistic is nondegenerate after suitable normalization. To obtain consistency of the Alpha statistic under PTW, **Assumption FLM** is required to control the cross-sectional dependence of factor loadings over treated and control units. Coupled with conditional mixing and moment conditions on the factor loadings of treated units in **Assumption FLM2**, we achieve asymptotic normality. The asymptotic result of Alpha test is stated formally below.

Theorem 5 (Alpha test) *Suppose Assumptions E, F, FLM,* AI_α*, IE and MM hold. Then, under Assumption PTW, we have the following results as* $T, N_E, N_C \to \infty$ *jointly and* $\frac{T}{N_C} \to 0$*:*

(a) $\hat{a}^{mg} \xrightarrow{p} 1$.

(b) $\sqrt{N_E}\bar{\varphi}^{-1}_{N_E,T}(\hat{a}^{mg} - 1) \xrightarrow{d} N(0,1)$ *if additionally Assumption FLM2 holds, where* $\bar{\varphi}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{a}^{mg} - 1)|\tilde{G}]$.

The above theorem shows that, under the null hypothesis of PTW, the Alpha statistic is consistent for unity, and the normalized statistic has an asymptotic standard normal distribution. As in previous theorems, the consistency and asymptotic normality result and the asymptotic $\sqrt{N_E}$-rate are preserved under a wide range of factor specifications (stationary and nonstationary). As in PCDID-MG, a nonparametric variance estimator is available: $\widehat{var}(\hat{a}^{mg}) = \frac{1}{N_E(N_E-1)}\sum_{i\in E}(\hat{a}_i - \hat{a}^{mg})^2$, which provides a convenient way to form the t-statistic, $\frac{\hat{a}^{mg}-1}{\sqrt{\widehat{var}(\hat{a}^{mg})}}$ (more details regarding inference can be found in Appendix 6 and Section 6.2). It readily follows from the asymptotic normality result that the power of Alpha test approaches one in the limit under local alternatives, e.g., $\alpha = 1 + O(1/\sqrt{N_E})$ as $N_E \to \infty$, in which the "average factor" for the treated group, $E(\mu_j'|j \in E)f_t$ remains proportional to that for the control group, $E(\mu_i'|i \in C)f_t$, but the proportionality constant is different from (but converges to) one.[27]

The Alpha test is directly built on the PCDID estimation approach, which allows for potentially unparallel trends via the factor structure of the interactive effects model; therefore,

[27]The Alpha test is powerful against more general alternatives with nonparallel trends. Suppose $E(\mu_j'|j \in E) = [1 + 1/\sqrt{N_E}, 1]'$ and $E(\mu_i'|i \in C) = [1,1]'$. This creates a time-varying gap between the "average factors" for the treated and control groups, leading to asymptotic bias (when factors are stationary and/or deterministic) or explosion (when stochastic trends are present) of $\hat{a}^{mg}$. Appendix 4.1 contains more details.

a rejection may imply an incorrect model specification. In a fully nonparametric setting, Callaway and Sant'Anna (2018) show that the parallel trend assumption (PTA) in DID analysis is, in general, untestable. Importantly, they also show that a stronger, augmented version of this assumption is testable. In our notations, their augmented PTA assumes that the potential outcome $y_{it}(0)$ satisfies conditions akin to $E(y_{it}(0) - y_{i,t-1}(0)|X, i \in E) = E(y_{it}(0) - y_{i,t-1}(0)|X, i \in C)$ *for each* $t = 2, .., T$, whereas the PTA assumes this for each $t > T_0$ only. Our PTW is testable in the sense that under the factor structure, it can be viewed as a special case of the *augmented PTA* (see also footnote 6).[28]

We also attempted an alternative test approach based on a comparison between 2wfe and PCDID-MG estimators. The intuition is given as follows. Assume fully homogeneous TE ($\Delta_{it} \equiv \Delta_0$), homogeneous intervention dates ($T_{0i} \equiv T_0$), *iid* idiosyncratic errors ($\epsilon_{it} \sim iid(0, \sigma_\epsilon^2)$), and the absence of covariates. Under the null hypothesis of *homogeneous* factor loadings, i.e., $\mu_i \equiv \mu_0$ for all $i = 1, \ldots, N$, we have "strong" parallel trends in which the 2wfe estimator is consistent for Δ_0 regardless of factor specification (see Section 2). Under this null hypothesis, the PCDID-MG estimator is consistent for Δ_0 but less efficient than 2wfe.[29] **Proposition H** in Appendix 4.2, which assumes the null and other restrictions above, reports the finite-sample variances of 2wfe and PCDID-MG estimators and shows that the Hausman equality holds, i.e., $Var(\hat{\delta}^{mg} - \hat{\delta}^{2wfe}) = Var(\hat{\delta}^{mg}) - Var(\hat{\delta}^{2wfe}) > 0$. This motivates a feasible Durbin-Wu-Hausman test statistic $t^2_{Haus} := \frac{(\hat{\delta}^{mg} - \hat{\delta}^{DID})^2}{\widehat{Var}(\hat{\delta}^{mg}) - \widehat{Var}(\hat{\delta}^{DID})}$, which has a limiting chi-square distribution with one degree of freedom (provided the regularity conditions in such tests hold). This approach is quite restrictive for two reasons. First, under treatment effect heterogeneity, the estimand of 2wfe may be different from that of PCDID, making the estimates incomparable. Second, the approach is only powerful against those departures from the null in which the 2wfe estimator is biased and inconsistent. Unfortunately, the behavior of 2wfe is sensitive to the factor structure specification (see the nuisance terms in Section 2). By focussing on PCDID, the Alpha test circumvents these limitations.

[28] Given t and $t-1$, we have $E(y_{it}(0) - y_{i,t-1}(0)|i \in E) - E(y_{it}(0) - y_{i,t-1}(0)|i \in C) = E(\beta_i^{'}(x_{it} - x_{i,t-1})|i \in E) - E(\beta_i^{'}(x_{it} - x_{i,t-1})|i \in C) + E(\mu_i^{'}|i \in E)(f_t - f_{t-1}) - E(\mu_i^{'}|i \in C)(f_t - f_{t-1}) = E(\beta_i^{'}(x_{it} - x_{i,t-1})|i \in E) - E(\beta_i^{'}(x_{it} - x_{i,t-1})|i \in C)$. This term is nonzero but it is parametric in x and can be identified (e.g., using pre-intervention data). Importantly, this result holds for every $t = 2, .., T$, hence the link to augmented PTA.

[29] However, when factor loadings are heterogeneous, PCDID-MG is not necessarily less efficient than 2wfe (provided that both estimators are consistent for Δ_0). This is because PCDID takes the factor structure into account (see also Section 2). Indeed, the simulations in Section 6.1 find that PCDID has a lower empirical SD than 2wfe (in scenerios where both estimators are valid).

5 Extension to Micro-Level Data

The existing DID literature on micro-level data assumes that the data consist of groups (clusters) (e.g., Bertrand et al. (2004), Donald and Lang (2007), Cameron et al. (2008), Conley and Taber (2011), Mackinnon and Webb (2017)). The workhorse in this literature is a 2wfe model resembling $y_{igt} = \Delta_{igt} 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \beta_g' x_{igt} + \varsigma_g + \tau_t + \tilde{\epsilon}_{igt}$, where y_{igt} is the observed outcome of unit i in group g at time t, the indicator functions $1_{\{g \in \mathscr{E}\}}$ and $1_{\{t > T_{0g}\}}$ depend on group g, and ς_g and τ_t are group and time fixed effects (parallel trends), respectively. The error term $\tilde{\epsilon}_{igt}$ is assumed independent across groups but can exhibit within-group correlation. Moulton (1990) and subsequent work (e.g., Carter et al. (2017)) showed that the asymptotic properties of estimators depend on the number of groups.

To fix notation, denote the set of treated and control groups in script form, $\mathscr{E}$ and $\mathscr{C}$. There are $N_{\mathscr{E}}$ treated groups and $N_{\mathscr{C}}$ control groups. There are $N = N_E + N_C$ units, with a total of N_E units in the treated groups and N_C units in the control groups. The number of units in group g is denoted by N_g. There are T time periods. For treated group $g \in \mathscr{E}$, there are T_{0g} pre-intervention periods and T_{1g} post-intervention periods.

Note that $\tilde{\epsilon}_{igt}$ above cannot be independent across groups if it has an interactive effects structure. Relatively few studies in the interactive effects or SC literature have examined this micro-level context. This motivates a micro-level model with interactive effects $y_{igt} = \Delta_{igt} 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \beta_{ig}' x_{igt} + \varsigma_{ig} + \mu_{ig}' f_t + \tilde{\epsilon}_{igt}$. Similar to Section 3, the TE is assumed to take the decomposition $\Delta_{igt} = \bar{\Delta}_i + \tilde{\Delta}_{igt}$, where $\bar{\Delta}_i := E(\Delta_{igt} | t > T_{0g})$ is the ITET of unit i in group g (drop subscript g for notational simplicity). This yields a reduced-form model that we examine:

$$y_{igt} = \bar{\Delta}_i 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \beta_{ig}' x_{igt} + \varsigma_{ig} + \mu_{ig}' f_t + \epsilon_{igt}, \tag{17}$$

where $\epsilon_{igt} := \tilde{\Delta}_{igt} 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \tilde{\epsilon}_{igt}$. Theorems 1,2 and 3 are applicable to this model given the underlying assumptions are satisfied. In such a case, the asymptotic properties of PCDID estimation do not depend on the number of groups nor the number of units in each group. For example, PCDID remains valid when there are many control groups ($N_{\mathscr{C}} \to \infty$) or few control groups ($N_{\mathscr{C}} \geq 1$) as long as $N_C \to \infty$; the case of few control groups has been tricky in the existing literature.[30] Assumption FL_C allows factor loadings to be arbitrarily clustered, and Assumption IE allows idiosyncratic errors to have weak cross-sectional dependence as well as general heteroskedasticity and serial correlation. However, stronger forms of clustering are potentially problematic. For example, in an error-components model $\epsilon_{igt} = \check{\epsilon}_{gt} + \check{\epsilon}_{igt}$ where the group-specific component $\check{\epsilon}_{gt}$ is independent of the unit-specific component $\check{\epsilon}_{igt}$, the original PCDID will be invalid when $N_{\mathscr{C}} \to \infty$, but it remains valid when $N_{\mathscr{C}}$ is fixed and finite but $N_C \to \infty$ (see **Corollary 4** in Appendix 5).[31]

In what follows, we consider an aggregated PCDID estimator that is applicable to more general forms of micro-level data. This estimator utilizes the group structure, and we consider its asymptotic properties in the conventional setting of many control groups ($N_{\mathscr{C}} \to \infty$). As an illustration, we will focus on the baseline case of ITET estimation. The estimand is $\bar{\Delta}_i := E(\Delta_{igt}|t > T_{0g})$ given g and $i \in g$. The spirit is the same for ATET estimation with aggregated mean-group and pooled estimators, where we focus on the estimand $\bar{\bar{\Delta}} := E(\bar{\Delta}_i|i \in E)$ (see **Corollaries 2 and 3** in Appendix 5). We first introduce a group structure assumption:

Assumption G (group structure): The micro-level model with interactive effects has at least one treated group and one control group, i.e., $N_{\mathscr{E}} \geq 1$ and $N_{\mathscr{C}} \geq 1$. The factor loading μ_{ig} and idiosyncratic error ϵ_{igt} are independent across groups.

We consider the aggregated PCDID estimator for $\bar{\Delta}_i$, which is denoted by $\hat{\delta}_i^{AGG}$. This estimator uses the same procedure as in Section 4, except that we perform PCA on the *aggregated* control panel $\{\bar{\hat{u}}_{gt}\}_{t=1,\dots,T}^{g=1,\dots,N_{\mathscr{C}}}$, which consists of $N_{\mathscr{C}} \times T$ observations, where $\bar{\hat{u}}_{gt} :=$

[30]Existing approaches require many control groups, or at least a moderately large number of control and treated groups. Recent methods such as Conley and Taber (2011) are applicable to the case of many control groups and few treated groups, allowing very short T. Under micro-level data, their estimator requires the number of units in a group to grow uniformly at the same rate as the number of groups. Relaxing Conley and Taber's assumption that the error is *iid* across groups, Ferman and Pinto (2019) extend their work and propose an estimator (allowing very short T) for aggregate-level data, which permit specific forms of heteroskedastic errors across groups. They assume that the heteroskedasticity in the aggregate-level data is implicitly generated by micro-level data in which the number of units differs across groups. Canay et al. (2018) consider wild bootstrap with a small number of large groups. They require the distribution of covariates to be homogeneous across groups.

[31]Suppose we designate the group-specific component as part of the interactive effects structure. Express ϵ_{igt} for $i \in C$ as $\epsilon_{igt} = [\mu_{ig}' \;\; e_g'][f_t' \;\; \check{\epsilon}_{1t} \;\; \check{\epsilon}_{2t} \;\; \dots \;\; \check{\epsilon}_{N_{\mathscr{C}}t}]' + \check{\epsilon}_{igt}$, where e_g is a $N_{\mathscr{C}} \times 1$ vector with the value of 1 in the g-th element and zero in all other elements. As $N_{\mathscr{C}} \to \infty$, this "enlarged" interactive effects model violates the assumption that the number of factors is fixed. Note that this is no longer an issue when $N_{\mathscr{C}}$ is fixed and finite.

$N_g^{-1}\sum_{i\in g}\hat{u}_{igt}$ is the cross-sectional within-group average of the first-stage residuals $\hat{u}_{igt}$ obtained from linear projections in Section 4.1, step 1 for each $i \in C$. The main result is:

Corollary 1 (PCDID-AGG): Suppose Assumptions G, EG, F, FLG$_C$, AI$_i$, IEG and MG hold in the micro-level model with interactive effects. Then, as $T, N_{\mathscr{C}} \to \infty$ jointly, we have for each $i \in E$:

(a) $\hat{\delta}_i^{AGG} \xrightarrow{p} \bar{\Delta}_i$.

(b) $\sqrt{T}\breve{\sigma}_{Ti}^{-1}(\hat{\delta}_i^{AGG} - \bar{\Delta}_i) \xrightarrow{d} N(0,1)$ if additionally Assumptions ES and D$_i$ hold, where $\breve{\sigma}_{Ti}^2 := Var[\sqrt{T}(\hat{\delta}_i^{AGG} - \bar{\Delta}_i)|\tilde{G}_i]$.

The overall intuition is the same as in Section 4. **Assumption EG** is essentially the same as Assumption E; **Assumptions FLG$_C$**, **IEG** and **MG** list standard regularity conditions on components in the *aggregated* control panel from which factor proxies are constructed (see Appendix 5).[32] No assumptions are imposed on N_g (number of units in group g) and, within each control group, relatively arbitrary forms of clustering in factor loadings and idiosyncratic errors are allowed.[33] Due to group-level independence (**Assumption G**), we do not need the condition $\frac{\sqrt{T}}{N_{\mathscr{C}}} \to 0$ to bound the cross-sectional correlation between the estimation error of factor proxies and idiosyncratic errors of the treated unit. This assumption is stronger than necessary for Corollaries 1-3 and can be relaxed.

6 Small Sample Properties of Estimators via Simulations

6.1 Baseline specifications and results

All DGPs follow the general form $y_{it} = \Delta_{it}1_{\{i\in E\}}1_{\{t>T_{0i}\}} + \beta_i^{'}x_{it} + \varsigma_i + \mu_i^{'}f_t + \tilde{\epsilon}_{it}$. We first discuss the DGPs for ITET estimation, and then the DGPs for ATET estimation. The DGPs for ITET estimation set $N_E = 1$, $T_{0i} = \frac{T}{2}$, $\Delta_{it} = 3$, $\varsigma_i = 0$ and $x_{it} = 0$ (relaxed below). The idiosyncratic error exhibits serial correlation as well as heteroscedasticity across

[32] The method can be readily applied to repeated cross-sectional data. Provided that extra regularity conditions for cross-sectional data are satisfied, an aggregated control panel can be formed, wherein each observation is constructed from all units observed in that particular g and t. Under repeated cross-sectional data, it is more convenient to assume that β_{ig} is homogeneous or group-specific, e.g., pooled regressions of y_{igt} on x_{igt} can be used to obtain first-stage residuals.

[33] Assumption IEG, which is weaker than Assumption IE, accommodates weak dependence of $\bar{\epsilon}_{gt} := N_g^{-1}\sum_{i=1}^{N_g}\epsilon_{igt}$ across control groups, across control groups and treated units, and across time, but unlike IE, it does not necessarily require weak dependence of ϵ_{igt} across units within a control group. Assumption FLG states that the aggregated panel preserves key variations in factor loadings that identify the factor space.

i; specifically, $\tilde{\epsilon}_{it} = \rho_\epsilon \tilde{\epsilon}_{i,t-1} + \nu_{it} h_i$ where $\rho_\epsilon = 0.1$, $\nu_{it} \sim N(0, 0.01(1-\rho_\epsilon^2))$ is iid across i and t, and $h_i \sim unif(0.5, 1.5)$ is iid across i. Three interactive effect scenarios are considered:[34]

1. Stationary factors: For each of the three factors $j = 1,2,3$, $f_{jt} = \phi_j + \eta_j 1_{\{t>\frac{T}{2}\}} + \rho_j f_{jt-1} + u_{jt}$ with i.i.d. $u_{jt} \sim N(0, \sigma_{uj}^2)$, $\phi_1 = \phi_2 = \phi_3 = 0$, $\eta_1 = \eta_2 = \eta_3 = 0$, $\rho_1 = 0.5$, $\rho_2 = 0.7$, $\rho_3 = 0.9$, $\sigma_{u1}^2 = 0.0675$, $\sigma_{u2}^2 = 0.0459$ and $\sigma_{u3}^2 = 0.0171$. Hence all three factors are AR(1) and have variance 0.09.

2. Stationary factors with break: same as #1 except that $\eta_2 = 1.2$. Hence factor 2 is AR(1) except for a jump at $t = \frac{T}{2} + 1$. Factor 2 is positively correlated with $1_{\{t>T_{0i}\}}$ but not collinear because it is time-varying for all t (Assumption AI(i) is satisfied).

3. Nonstationary factors with drift: same as #1 except that $\phi_1 = 0.1$, $\rho_1 = \rho_2 = \rho_3 = 1$, $\sigma_{u1} = 0.3$, $\sigma_{u2} = 0.5$, and $\sigma_{u3} = 0.1$. Hence all three factors are I(1), and factor 1 drifts upward by 0.1 per period.

In all scenarios, the factor loading of unit i for factor j is distributed as $\mu_{ij} \sim N(m_{Cj}, \sigma_{\mu_C}^2)$ if $i \in C$, and $\mu_{ij} \sim N(m_{Ej}, \sigma_{\mu_E}^2)$ if $i \in E$. We set $(m_{C1}, m_{C2}, m_{C3}) = (1, 0.9, 0.8)$, $(m_{E1}, m_{E2}, m_{E3}) = (1.2, 1.4, 1.6)$, $\sigma_{\mu_C} = \sigma_{\mu_E} = 0.3$ so that the average factor loadings are higher among treated units than control units (μ_i is positively correlated with $1_{\{i\in E\}}$). Yet in Scenario 1, $\mu_i' f_t$ remains uncorrelated with $1_{\{i\in E\}} 1_{\{t>T_{0i}\}}$ because f_t is uncorrelated with $1_{\{t>T_{0i}\}}$. In scenario 2, $\mu_i' f_t$ is positively correlated with $1_{\{i\in E\}} 1_{\{t>T_{0i}\}}$ due to factor 2. In scenario 3, $\mu_i' f_t$ tends to be larger among treated units and in later time periods (due to factor 1), but its correlation with $1_{\{i\in E\}} 1_{\{t>T_{0i}\}}$ is undefined due to unit roots in f_t.

We compare seven estimators: (i) PCDID, (ii) DID with two-way FE ("DID-2wfe"), (iii) DID with FE and unit-specific cubic time trend ("DID-trend"; e.g., Wooldridge (2005)), (iv) Bai (2009), (v) GSC (Xu (2017)), (vi) stepwise GM (Gobillon and Magnac (2016)), (vii) nuclear norm matrix completion ("MC-NNM"; Athey et al. (2018)). Details of the last three estimators are in Appendix 3.[35] PCDID, BAI, GSC and GM assume there are three factors,

[34] Existing studies have typically considered factors in simpler forms, e.g., iid factor or deterministic sinusoid functions (Bai (2009), Gobillon and Magnac (2016)). For space reasons, we do not report scenarios that involve deterministic trends; such cases are trivial and are encompassed by PCDID as a special case. We also do not report scenarios that involve a mix of nonstationary and stationary factors because the results look similar to the scenario with nonstationary factors only.

[35] DID-2wfe performs regression $y_{it} = \delta 1_{\{i\in E\}} 1_{\{t>T_{0i}\}} + b_{0i} + \tau_t + e_{it}$ on the full sample. DID-trend also performs regression on the full sample: $y_{it} = \delta 1_{\{i\in E\}} 1_{\{t>T_0\}} + b_{0i} + b_1 1_{\{t>T_0\}} + \sum_{m=1}^{3} a_{im} t^m + e_{it}$ if $T_{0i} = T_0$; $y_{it} = \delta 1_{\{i\in E\}} 1_{\{t>T_{0i}\}} + b_{0i} + \sum_{m=1}^{3} a_{im} t^m + e_{it}$ under staggered adoption. In Bai (2009), the cth iteration

whereas MC-NNM computes the rank (or number of factors) automatically. We consider 10 different (N, T) combinations, and compute the bias and standard deviation (SD) of each estimator based on 1000 replications.

Table I reports the ITET estimation results. Overall, PCDID yields the best performance in terms of bias and SD. Other methods have relatively uneven performance across scenarios and sample size. DID-2wfe performs worst, as expected, followed by DID-trend. MC-NNM outperforms DID methods in scenario A, which is consistent with Athey et al. (2018). It is biased heavily upward in B and C, even though the estimated rank is close to the true number of factors when sample size is large. Both GSC and GM have a high SD when T is small. When T is large, their performance are similar to PCDID in A, but still worse in B and C. Compared to PCDID, BAI has worse bias overall but slightly better SD in B. The average number of iterations required for convergence is nontrivial, ranging from 8 to 120 depending on scenario and sample size, and non-convergence is quite common.[36]

In the DGPs for ATET estimation, we set $N_E = N_C = \frac{N}{2}$. In addition, we set $\Delta_{it} = 3 + \tilde{\Delta}_i + 0.25(\mu_{i1} - 1.2) + \frac{5}{T}[|T_{0i} - \frac{T}{2}| - E(|T_{0i} - \frac{T}{2}|)] + \rho_\Delta \Delta_{i,t-1} + u_{\Delta,it}$, where $\tilde{\Delta}_i \sim N(0, 1)$, $\rho_\Delta = 0.1$ and $u_{\Delta,it} \sim N(0, 0.01(1 - \rho_\Delta^2))$. Therefore, the treatment effect varies cross-sectionally and over time, and it is correlated with both the unobserved loading of the first factor μ_{i1} and the policy intervention date $(T_{0i}+1)$. The treatment effect is smallest when T_{0i} equals $\frac{T}{2}$, and it is larger when T_{0i} is closer to the beginning or the end of the sample period. We compare the simple mean-group estimator (PCDID-MG) with the other estimators.

Table II reports the ATET estimation results when the policy intervention date is homogeneous ($T_{0i} = \frac{T}{2}$). The relative advantage of PCDID is larger compared to ITET estimation – PCDID-MG outperforms the other estimators in all scenarios and sample sizes. The next-best performers are GSC and GM, both having a large SD when T is small. Somewhat

contains two estimation sub-procedures: (i) based on the panel of residuals $y_{it} - \hat{\delta}^{(c-1)} 1_{\{i\in E\}} 1_{\{t>T_{0i}\}} - \hat{b_0}^{(c-1)}$ where $(\hat{\delta}^{(c-1)}, \hat{b_0}^{(c-1)})$ are estimates from the *(c-1)*th iteration, use PCA to estimate the interactive effects structure $\widehat{\mu_i' f_t}^{(c)}$, and (ii) subtract $\widehat{\mu_i' f_t}^{(c)}$ from y_{it}, and obtain $(\hat{\delta}^{(c)}, \hat{b_0}^{(c)})$ from the regression $y_{it} - \widehat{\mu_i' f_t}^{(c)} = \delta 1_{\{i\in E\}} 1_{\{t>T_{0i}\}} + b_0 + e_{it}$ on the full sample. Similar iterative procedures are found in Ahn et al. (2001) and Moon and Weidner (2015); because the numerical properties are similar in our DGPs, we focus on Bai's procedure.

[36]There has been limited attention on the numerical convergence properties of Bai (2009). The choice of initial values, maximum/actual number of iterations and convergence criterion vary widely across studies. We draw the initial value of δ from $N(2.7, 0.09)$ and the initial value of intercept b_0 is zero. We stop iterating when the Euclidean distance between the estimates in the cth and *(c+1)*th iterations is smaller than 10^{-3}, or when the number of iterations exceeds 500. We flag the result if all 1000 replications attain convergence.

surprisingly, numerical convergence in BAI becomes strenuous in all scenarios.[37] DID-2wfe, DID-trend and MC-NNM are satisfactory in scenario A only.

Table III reports the ATET results under staggered adoption, which draws T_{0i} from a uniform discrete distribution: $T_{0i} \sim Unif\{\lceil \frac{T}{4} \rceil, \lfloor \frac{3T}{4} \rfloor\}$. PCDID-MG remains the best performer. GSC and GM have very large SDs when T is small, as some treated units have short pre-intervention periods under staggered adoption; when T_{0i} is equal to or smaller than the number of PCs/factors, under-identification occurs and unit i is dropped, yielding a negative bias.[38] BAI improves due to staggered adoption, although some bias remain. DID-2wfe has bias, even in scenario A, which is consistent with Goodman-Bacon (2018) that it assigns different weights to treated units with middle intervention dates versus those with early or late dates. MC-NNM has similar performance to DID-2wfe. DID-trend has better performance due to staggered adoption, although it is still dominated by PCDID-MG.

We examine the performance of PCDID-MG when the number of factor proxies is chosen by the growth ratio (GR) test in Ahn and Horenstein (2013).[39] Consistent with their findings, the GR test underestimates the number of factors when the sample size is small. Nevertheless, this worsens PCDID only slightly, even in scenarios B and C (see Appendix Table A1, compared with Table II). As in Moon and Weidner (2015), the limiting distribution of PCDID is not affected by the number of factor proxies as long as this number is not underestimated. Therefore, overestimation of number of factors is less problematic than underestimation. To examine this, we consider (1) a conservative version of the GR test that may overestimate the number of factors,[40] and (2) setting the number of factor proxies to be 5. Both result in very small deterioration of PCDID performance.

We also examine the performance of PCDID-MG when exogenous or endogenous covariates are present (see Section 4.1 for the estimation algorithm). The DGP assumes

[37]We also tried a tighter convergence criterion (10^{-4}) and higher maximum iterations (2,000), but the results remain similar.

[38]For each treated unit, both GSC and GM use T_{0i} periods to estimate the unit-specific intercept and the p unit-specific factor loadings. There are insufficient degrees of freedom if $T_{0i} < 1 + p$.

[39]They develop two tests, the eigenvalue ratio (ER) test and the GR test. Both are easy to compute. In the ER test, the estimated number of factors is obtained by maximizing the ratio of two adjacent eigenvalues (from the sample covariance matrix of control units) arranged in descending order. The GR test is a smoothed version of the ER test. See Appendix 8 for details. In our simulations, the GR test tends to perform better.

[40]Ahn and Horenstein (2013) note a one-factor bias problem in these tests, i.e., in finite samples, the estimated number of factors is often one. In our conservative approach, whenever this occurs, we reset the estimated number of factors to be equal to half of the maximum possible number of factors in the model ($\frac{kmax}{2}$). In all simulations, we set $kmax = 6$.

$\varsigma_i \sim N(1 + 1_{\{i \in E\}}0.5, 0.09)$, $x_{it} = 1_{\{i\in E\}}0.3 + \rho_x \mu_{i3} f_{3t} + u_{x,it}$ where $u_{x,it} \sim N(0,1)$, and the slope $\beta_i = 1 + 1_{\{i\in E\}}0.2$. Therefore, ς_i, x_{it}, and β_i are positively correlated with $1_{\{i\in E\}}$. We consider $\rho_x = 0$ and $\rho_x = 0.3$; in the latter case, x_{it} is correlated with the interactive effects of the third factor ($\mu_{i3} f_{3t}$). PCDID performs well in both cases (see Appendix Table A1). The results are similar when x is exogenous or endogenous (to the factor structure), lending support to the robustness of our approach.

6.2 Inference on ITET/ATET and parallel trend test

We examine the finite-sample performance of PCDID inference procedures for ITET and ATET. The DGPs are the same as in Table I and II, respectively (except we set $\rho_\epsilon = 0$ in the DGP for ITET inference). The null hypothesis is set at the DGPs' true value ($\delta_0 = 3$). We use 1000 replications and a nominal size of 5%. We compare four procedures. The first two are based on the full sample only, while the other two involve 199 bootstrap samples per replication: (1) **TrueF**: assume factors are observed in PCDID estimation (infeasible) and compute t-statistic, reject if $|t| \geq 1.96$. (2) **Asym**: use factor proxies in PCDID estimation and compute t-statistic, reject if $|t| \geq 1.96$. (3) **b-t**: same as Asym, but reject if $t \leq c_{0.025}$ or $t \geq c_{0.975}$ where $c_{0.025}, c_{0.975}$ are percentiles of the bootstrap distribution of t-statistics. (4) **b-se**: same as Asym, but the standard error in the t-statistic formula is obtained from bootstrap samples.

Comparing Asym with TrueF will give the relative performance when factors are estimated instead of known (e.g., Gonçalves and Perron (2014)). Because Asym is asymptotically valid (see Theorems 1,2,3), comparing b-t/b-se with Asym will show whether bootstrapping yields reasonable results relative to this baseline approach. For ITET inference, a mix of wild and stationary bootstrap is used; for ATET inference, wild bootstrap is used.[41] See Appendix 6 for details on bootstrap sample construction. In TrueF, Asym and b-t, we compute t-statistics based on analytical standard errors. For ITET inference, it is $\frac{\hat{\delta}_i - \delta_0}{se(\hat{\delta}_i)}$ where $se(\hat{\delta}_i)$ is obtained from the classical standard error formula based on the time-series regression for unit i.[42] For ATET inference, it is $\frac{\hat{\delta}^{mg} - \delta_0}{se(\hat{\delta}^{mg})}$ with the nonparametric estimator

[41]Stationary bootstrap is performed on residuals; it does not require parametric assumptions but requires a tuning parameter. We set the tuning parameter at $\frac{1}{T}$ so that the "contiguous" block length increases with T. See Politis and Romano (1994) for details.

[42]As mentioned above, the DGP for ITET inference sets $\rho_\epsilon = 0$; see Gonçalves and Perron (2014) for a similar setup. The formula can be replaced by the Newey-West HAC estimator if $\rho_\epsilon \neq 0$.

$se(\hat{\delta}^{mg}) = \sqrt{\frac{1}{N_E(N_E-1)} \sum_{i \in E}(\hat{\delta}_i - \hat{\delta}^{mg})^2}$ (see discussion after theorem 3).

Table IV reports the simulation results. In ITET inference (left panel), TrueF has a rejection rate close to the nominal size of 5%, as expected. Asym's rejection rate is similar to TrueF's, especially when the sample is large; however, it tends to overreject when $N << T$. B-t performs similarly to Asym, and has better performance when $N << T$. B-se has a lower rejection rate (i.e., more conservative) compared to b-t. In ATET inference (right panel), the inference procedures perform better. TrueF, Asym, b-t and b-se all yield similar results in general, with rejection rates close to 5% when the sample is large, and Asym and b-t having less severe overrejection when $N << T$.

We then examine the performance of the parallel trend alpha test. The DGP assumes the following factor scenarios: (i) one AR(1) factor ($\phi = 0$, $\eta = 0$, $\rho = 0.5$, $\sigma_u^2 = 0.0675$); (ii) one AR(1) factor with break (same as (i) except $\eta = 2$); (iii) one I(1) factor with drift ($\phi = 0.1$, $\eta = 0$, $\rho = 1$, $\sigma_u = 0.3$). In addition, it assumes $m_C = m_E = 1$ (PTW holds) and $\sigma_{\mu_C} = \sigma_{\mu_E} = 0.1$. The DGP is otherwise the same as in Table II. We use 1000 replications and a nominal size of 5%. We consider TrueF, Asym and b-t procedures, which compute t-statistics based on analytical standard errors: $\frac{\hat{a}^{mg}-1}{se(\hat{a}^{mg})}$ where $se(\hat{a}^{mg}) = \sqrt{\frac{1}{N_E(N_E-1)} \sum_{i \in E}(\hat{a}_i - \hat{a}^{mg})^2}$ and $\hat{a}^{mg} := \frac{1}{N_E} \sum_{i \in E} \hat{a}_i$.

Table V reports the rejection rates of the alpha test. In light of the asymptotic ratio in Theorem 5, the sample size combinations focus on increasing N_C while keeping T fixed at various levels. As expected, TrueF has a rejection rate close to 5%. Asym's rejection rate converges to 5% when N_C becomes large relative to T; however, it overrejects when $N_C << T$. B-t performs similarly to Asym, and has better performance when $N_C << T$.

We also examine the power and rotational invariance of the alpha test. The DGP is the same as in Table V, except $m_C = 2$ and $m_E \in [1.8, 2.2]$. Note that the alpha test is based on identification of the relative instead of absolute levels of m_C and m_E. Due to this feature and test consistency of t-tests, its asymptotic rejection rate equals the nominal size when $\frac{m_E}{m_C} = \frac{2}{2} = 1$, and equals 100% otherwise. This is confirmed in the simulations; Appendix Figure A2 also compares the rejection rates of TrueF, Asym and b-t procedures under various values of $\frac{m_E}{m_C}$, and show that they are similar when the sample is large.

6.3 Micro-level data

The DGPs largely follow the baseline in Section 6.1. The general form is $y_{igt} = \Delta_{igt} 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \varsigma_{ig} + \mu_{ig}^{'} f_t + \tilde{\epsilon}_{igt}$ where i, g, t represent unit, group, and time, respectively. We set $\varsigma_{ig} = 0$, $T_{0g} = \frac{T}{2} = 25$ and the same f_t specifications as in the baseline. We assume mutually independent group and unit error components: $\mu_{ig} = \breve{\mu}_g + \breve{\mu}_{ig}$ and $\tilde{\epsilon}_{igt} = \breve{\epsilon}_{gt} + \breve{\epsilon}_{igt}$, and set the within-cluster correlation of factor loading r_μ and idiosyncratic error r_ϵ at $(r_\mu, r_\epsilon) = (0.2, 0.2)$.[43] In the DGPs for ITET estimation/inference, we set $\Delta_{igt} = 3$; in those for ATET, we set $\Delta_{igt} = 3 + \tilde{\Delta}_{ig} + 0.25(\mu_{ig1} - 1.2)$ where $\tilde{\Delta}_{ig} \sim N(0, 1)$.

In Table VI, the results on the left are based on DGPs for ITET estimation/inference with one treated group/unit only ($N_{\mathscr{E}} = N_E = 1$). The results on the right are based on DGPs for ATET estimation/inference with one treated group consisting of 25 units ($N_{\mathscr{E}} = 1$, $N_E = 25$). We consider four sample structures in the controls: (A) Multiple control groups ($N_{\mathscr{C}} = 5, 10, 25, 50$), each consisting of 25 units. (B) One control group ($N_{\mathscr{C}} = 1$) consisting of $5, 10, 25, 50$ units. (C) Multiple control groups ($N_{\mathscr{C}} = 10, 50$), each consisting of different numbers of units (average=25).[44] (D) Multiple control groups ($N_{\mathscr{C}} = 10, 50$), each consisting of 35 units. We use the aggregated PCDID/PCDID-MG (mean-group across $i \in E$) estimator in Panels A, C and D, and use the original estimator in Panel B. All reported outcomes – empirical bias and SD, as well as the rejection rates of the TrueF, Asym, b-t and b-se inference procedures, are defined in the same way as in Tables I, II and IV. The results are similar to those tables, providing support for the use of PCDID in various forms of micro-level data.[45]

[43] The factor loadings follow $\breve{\mu}_g \sim iidN((1, 0.9, 0.8)^{'}, \sigma^2_{\mu_g} I)$ for $g \in \mathscr{C}$ and $\breve{\mu}_g \sim iidN((1.2, 1.4, 1.6)^{'}, \sigma^2_{\mu_g} I)$ for $g \in \mathscr{E}$, $\breve{\mu}_{ig} \sim iidN((0, 0, 0)^{'}, \sigma^2_{\mu_{ig}} I)$, and $\sigma^2_{\mu_g} + \sigma^2_{\mu_{ig}} = 0.09$. The idiosyncratic errors follow $\breve{\epsilon}_{gt} = 0.1\breve{\epsilon}_{g,t-1} + \breve{\nu}_{gt} h_g$, where $\breve{\nu}_{gt} \sim iidN(0, \sigma^2_{\breve{\nu}_g})$, $\breve{\epsilon}_{igt} = 0.1\breve{\epsilon}_{ig,t-1} + \breve{\nu}_{igt} h_{ig}$ where $\breve{\nu}_{igt} \sim iidN(0, \sigma^2_{\breve{\nu}_{ig}})$, and $h_g, h_{ig} \sim iidU(0.5, 1.5)$, $\sigma^2_{\breve{\nu}_g} + \sigma^2_{\breve{\nu}_{ig}} = 0.0099$. $r_\mu := \sigma^2_{\mu_g} / (\sigma^2_{\mu_g} + \sigma^2_{\mu_{ig}})$ and $r_\epsilon := \sigma^2_{\breve{\nu}_{gt}} / (\sigma^2_{\breve{\nu}_{gt}} + \sigma^2_{\breve{\nu}_{igt}})$, respectively. Note that even when $(r_\mu, r_\epsilon) = (0, 0)$, there is still cross-sectional dependence due to the factor structure. The literature's benchmark is closer to $r_\epsilon = 0.1$, which is sufficient to cause a severe downward bias of OLS standard errors (e.g., Moulton (1990), Ferman and Pinto (2019)).

[44] When $N_{\mathscr{C}} = 10$, each control group has 20,21,...24,26,..,30 units, respectively. When $N_{\mathscr{C}} = 50$, each control group has 13,13,14,14,...,37,37, respectively.

[45] Existing methods, which allow for short T, emphasize the impact on estimator performance from the number of units in a group (N_g) and its heterogeneity, or from small $N_{\mathscr{C}}$ (e.g., Donald and Lang (2007), Cameron et al. (2008), Conley and Taber (2011), Carter et al. (2017), Mackinnon and Webb (2017)).

7 Placebo Designs

The "placebo law" approach by Bertrand, Duflo and Mullainathan (2004) (henceforth BDM) has been used as a benchmark in subsequent studies, e.g., Cameron et al. (2008) and Mackinnon and Webb (2017), to evaluate the performance of hypothesis tests related to DID-2wfe. We use it to compare the overall performance of DID-2wfe and PCDID estimators. As in BDM, the data come from the Current Population Survey (CPS) between 1979 and 1999 with the benchmark model being $y_{igt} = \Delta I_{gt} + \beta' x_{igt} + \varsigma_g + \tau_t + u_{igt}$, where y_{igt} is the log earnings of individual i (women aged 25-50) in state g and year t, I_{gt} is a dummy variable equaling one if the placebo law is in place in state g and year t (explained below), x_{igt} is a set of individual characteristics including age and education, ς_g is the state fixed effect, τ_t is the year fixed effect, and u_{igt} is the disturbance term.

As Mackinnon and Webb (2017) argue, there is limited loss of power by aggregating the data to a state-year panel because the key variable, I_{gt}, does not vary at the individual level. Nevertheless, we present results from both aggregate data (1,071 observations) and micro-level data (549,735 observations). Following BDM, the dependent variable in the aggregate data is the state-by-year average log earnings after partialing out the individual characteristics.[46] We consider 1000 replications. Following BDM, each replication draws a year at random between 1985-95 ($T_0 \sim Unif\{1985, 1995\}$); then, select a subset of states at random and designate them as treated states ($g \in \mathscr{E}$). Thus the policy variable is $I_{gt} := \mathbf{1}\{g \in \mathscr{E}\}\mathbf{1}\{t > T_0\}$.

We consider two scenarios. Scenario A serves as the original benchmark, by selecting 10 states at random from the pool of 51 states. Because randomization on I_{gt} is carried out unconditionally (which is rarely true empirically), the DID-2wfe estimator for Δ will, trivially, yield an empirical mean close to zero; in fact, a between-group estimator is already sufficient. As BDM pointed out, DID-2wfe is appropriate when the interventions are as good as random, conditional on time and group fixed effects. Motivated by this, we add Scenario B, which selects 10 states at random from the pool of the largest 25 states as defined by the state's population in 1990. The spirit remains the same if we construct the pool differently;[47] for

[46] BDM constructs this variable by regressing y_{igt} on x_{igt}, obtain the residuals $\hat{\nu}_{igt}$, and compute the simple average of the residuals across all i within the same g and t, i.e., $\bar{\hat{\nu}}_{gt} = \frac{1}{N_{gt}} \sum_{i=1}^{N_{gt}} \hat{\nu}_{igt}$. They then regress this variable $\bar{\hat{\nu}}_{gt}$ on I_{gt} allowing for fixed effects ς_g and τ_t. To allow for direct comparison, in the micro-level data we use $\hat{\nu}_{igt}$ as the dependent variable. Our data are extracted from the archive of Mackinnon and Webb (2017).

[47] For example, select 10 states at random from the pool of Eastern and Midwest states.

DID-2wfe to be valid, the difference between pool and non-pool states ought to be adequately captured by state fixed effects (i.e., parallel trends). We can view Scenario B as carrying out randomization on I_{gt} conditional on state fixed effects. This is true if, plausibly, no systematic relationship exists between changes in average earnings and changes in the state population.

Table VII reports the results. We first discuss aggregate data (left side of the table). In scenario A, both DID-2wfe and PCDID yield an empirical mean close to zero, as expected. Interestingly, the empirical SD of DID-2wfe is larger than that of PCDID (see also Table II), which suggests an efficiency gain by accounting for the factor structure. To facilitate comparison, we present rejection rates from bootstrap b-p and b-se procedures in DID-2wfe and PCDID (see also Appendix 6).[48] Consistent with earlier studies, the rejection rate for DID-2wfe is close to the nominal size of 5%. For PCDID, the test under-rejects when 1 or 3 PCs are used and it converges to the nominal size when 5 or 7 PCs are used, hinting at the complexity of the unobserved trends. The results are different in scenario B. DID-2wfe is biased upwards with an empirical mean of 0.020. This magnitude is non-trivial; for example, Cameron et al. (2008) investigate the power of their tests against the alternative hypothesis that $\delta_0 = 0.02$. PCDID has an empirical mean of close to zero. The bias in DID-2wfe causes the test to over-reject the null hypothesis. For PCDID, the rejection rates exhibit a similar pattern to scenario A.

The right side of the table considers micro-level data. In scenario A, the results are similar to those when aggregate data are used. Despite the huge increase in sample size, DID-2wfe has a small efficiency gain only (empirical SD drops to 0.173 from 0.193), while there is no efficiency gain for PCDID. In scenario B, DID-2wfe is still biased upward with an empirical mean of 0.008 – the bias becomes smaller possibly because the estimate is implicitly weighted by state population, which is the selection criterion of our pool. In a different context of studying the effects of divorce law reforms using an aggregate state-year panel, Kim and Oka (2014) also find that DID-2wfe estimates are sensitive to the weighting scheme. For PCDID, the results are qualitatively similar to that in scenario A. Overall, PCDID demonstrates superior performance to DID-2wfe as far as robustness is concerned.

[48] Our wild bootstrap for DID-2wfe is similar to Cameron et al. (2008) and Mackinnon and Webb (2017). The b-se procedure for DID-2wfe corresponds to the "wild cluster bootstrap-se" method in Cameron et al. (2008). Because the analytical standard errors are different, we do not compare b-t procedures between DID-2wfe and PCDID. The B-p procedure is otherwise the same as b-t except it uses the original instead of t statistic. Note that b-p and b-se are both valid when the outcome is stationary – a plausible assumption in this application.

8 Effect of Waiver Programs on Welfare Caseloads

DID regressions have been widely used for examining the effects of various welfare reforms on labor market outcomes (see e.g., Chan and Moffitt (2018) for a recent survey). In this illustration, we examine the effects of waiver programs on caseloads in the Aid to Families with Dependent Children (AFDC) program. Providing cash assistance to low-income female-headed families, AFDC had been one of the largest means-tested transfer programs in the US. In the 1990s, many states sought waivers from the federal government, which allowed them to deviate from federal AFDC rules. The majority of the states had a waiver in place when AFDC was replaced by the Temporary Assistance for Needy Families (TANF) program in 1997.

Our analysis builds on Ziliak et al. (2000), who found mixed evidence regarding the effects of waivers. Their model is similar in spirit to a DID regression with state fixed effects and state-specific cubic time trends, based on a first-differenced monthly panel of state-level caseloads and heteroscedasticity-robust standard errors. Our focus is to apply both PCDID and DID regressions (taking into account of clustering) and compare the results. Our data set covers 50 states plus the District of Columbia for 117 months from Oct1986 to Jun1996.[49] Because PCDID is robust to the presence of nonstationarity trends, we do not first-difference the data. As in Ziliak et al. (2000), we use the waiver approval date to define policy intervention. Specifically, we define T_{0i} as the approval date of state i's work requirement waiver or time limit waiver, whichever is earlier.[50] The policy intervention exhibits staggered adoption with T_{0i} as early as mid-1992 (see Figure A4). There are 20 control states, which have neither of the waivers by the end of the sample period.

We perform PCDID estimation separately in four different samples, all of which have $T = 117$ periods: (1) Control plus all treated states $(N_C, N_E) = (20, 31)$; (2) Control plus 10 Southern treated states $(N_C, N_E) = (20, 10)$; (3) Control plus 21 non-Southern treated states $(N_C, N_E) = (20, 21)$; (4) Control plus Wyoming $(N_C, N_E) = (20, 1)$. Summary statistics

[49]The data came from Quarterly Public Assistance Statistics published by the Office of Family Assistance of the U.S. Department of Health and Human Services. We would like to thank James Ziliak for generously providing the data for analysis.

[50]We follow the definitions of variables in Ziliak et al. (2000). Work requirements and time limits are the key components of welfare reform and they unambiguously reduce welfare participation. For simplicity, we do not consider earnings disregards and parental responsibility waivers, which yield mixed evidence in the literature, e.g. see Chan (2013) and Chan and Moffitt (2018) for details.

indicate that the average characteristics of treated and control states are generally similar, while individual states, e.g., Wyoming, can be quite different from the average (see Table A2). The Southern states are defined as states in the South Census region (18 in total); we will examine whether the policy had larger effects in the South, where the welfare system was relatively stringent.

We apply PCDID to the model $y_{it} = \Delta_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \beta_i' x_{it} + \varsigma_i + \mu_i' f_t + \tilde{\epsilon}_{it}$, where y_{it} is the log of per-capita welfare caseload in state i at time t, Δ_{it} is the treatment effect, T_{0i} is the waiver approval date, x_{it} is a vector of six time-varying covariates including the maximum combined real AFDC/Food Stamp benefits for a family of three, state unemployment rate, state log employment-to-population ratio, and calendar quarter dummies; ς_i is the state fixed effect, $\mu_i' f_t$ is the interactive effects structure, and $\tilde{\epsilon}_{it}$ is the idiosyncratic error. We use PCDID ($\hat{\delta}_i$) for ITET ($\bar{\Delta}_i := E(\Delta_{it} | t > T_{0i})$) estimation in sample 4, and PCDID-MG ($\hat{\delta}^{mg} := \frac{1}{N_E} \sum_{i \in E} \hat{\delta}_i$) for ATET ($\bar{\bar{\Delta}} := E(\bar{\Delta}_i | i \in E)$) estimation in samples 1,2,3. The estimation and inference algorithms are the same as in Section 4.[51] We report analytical standard errors (Asym), bootstrapped standard errors (b-se), and p-values from bootstrapped t-statistics (b-t); note that b-se tends to be more conservative than the other two (see Section 5). For comparison, we also use sample 1 to estimate: (1) DID-2wfe regression $y_{it} = \delta 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + b_1' x_{it} + b_{0i} + \tau_t + e_{it}$; (2) DID regression with state-specific time trends $y_{it} = \delta 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + b_1' x_{it} + b_{0i} + \sum_{m=1}^{M} a_{im} t^m + e_{it}$ where $M = 3$ (cubic) or $M = 4$ (quartic) (e.g., Wooldridge (2005)). For DID regressions, we report cluster wild bootstrapped standard errors, as in Section 6.

Table VIII reports the estimates from PCDID (panels A, B) and DID regressions (panel C). Panel A reports the preferred specification of PCDID with 4 PCs (GR test discussed below). In sample 1 (all treated states), the PCDID-MG coefficient on policy intervention is -0.017, implying that waivers reduced per-capita welfare caseload by 1.7% on average.[52] It is statistically significant at the 5% level by Asym and b-t, and 10% level by b-se. The effects are larger in the South – in sample 2 (Southern treated states), the coefficient is -0.024 and

[51] We include some covariates that are potentially endogenous to the factor structure, e.g., unemployment rate. PCDID is robust to such covariates (see Section 4.1, Table A1 and Appendix 7). We perform fixed effect estimation of y_{it} on x_{it} in the control panel (2,340 observations) to obtain residuals, which are then used for constructing factor proxies $\hat{f}_t$ by PCA.

[52] Among the other covariates, the PCDID-MG coefficient on unemployment rate is positive and statistically significant at the 1% level; the welfare maximum benefit has a statistically significant positive coefficient in some specifications; the employment-to-population ratio is statistically insignificant.

statistically significant at the 1% (Asym and b-t) or 5% (b-se) level, whereas in sample 3 (non-Southern treated states), the coefficient is -0.013 and statistically insignificant at the 10% level. Waivers reduced Wyoming's per-capita caseload by 11.4% on average – in sample 4 (Wyoming), the PCDID coefficient is -0.114 and statistically significant at the 1% (Asym and b-t) or 5% (b-se) level.[53]

Table VIII and Figure 1 also report the fraction of the change in per-capita caseload in treated states between Jan93 (t=76) and Jun96 (t=117) that can be explained by the policy intervention. Specifically, we compute $1 - \frac{\bar{y}^{cf}_{E,117} - \bar{y}^{cf}_{E,76}}{\bar{y}^{pred}_{E,117} - \bar{y}^{pred}_{E,76}}$ where $\bar{y}^{cf}_{E,t}$ denotes the counterfactual average per-capita caseload in treated states at period t assuming no reform at all (i.e., set $T_{0i} = \infty \ \forall i$) and $\bar{y}^{pred}_{E,t}$ denotes the predicted caseload when T_{0i} is the same as in the data. Among all treated states, the waivers explained 6.88% of the drop in caseloads between Jan93 and Jun96. Among southern treated states, this proportion is 10.41%. In Wyoming, this proportion is 24.86%. Figure 1 plots the actual and predicted caseloads. The model fits the treated state caseloads well, even though the factor proxies are extracted from control states only. Caseloads peaked slightly earlier among treated states than control states. In samples 1, 2 and 3, the trajectories appear similar between control and treated states; the t-statistic from the Alpha test is not statistically significant from zero, therefore the null of weak parallel trend cannot be rejected. In sample 4, Wyoming had an abrupt drop in caseload at the policy intervention date and the model predicts a smoother reduction if the policy was absent.

Our 4-PC specification is conservative, given that overestimation of number of factors is less problematic than underestimation (see Section 4.1). The original GR test and a recursive version of the test yield two and three factors, respectively.[54] Panel B of Table VIII reports the 3-PC specification as a robustness check, with similar results. The coefficients stabilize when 3 or more PCs are used. Our plots of PCs/factor proxies also show that the underlying trends, which are highly nonlinear, are largely captured by the first 4 PCs (see Figure A5).

Panel C of Table VIII reports the DID regressions using sample 1. The policy coefficient is close to zero and statistically insignificant when state-specific cubic or quartic trends are used. These parametric specifications are restrictive relative to PCDID. The 2wfe regression yields a policy coefficient of -0.054, implying that waivers explained 24% of the drop in caseloads

[53]In sample 4, the analytical standard errors in Asym and b-t are obtained from the Newey-West HAC estimator with 3 lags.

[54]See Appendix 8 for details. We set $jmax = 0$ (original test), $jmax = 1$ (recursive version) and $kmax = 10$.

between Jan93 and Jun96. This estimate is implausibly large and it is similar to the 1-PC PCDID specification (-0.056, not shown in table).

9 Conclusions

In this paper, we developed a class of factor-augmented regression estimators (PCDID) for treatment effect estimation. PCDID was similar in spirit to the control function approach and it used factor proxies constructed from control units to control for unobserved trends, assuming that the unobservables followed an interactive effects structure. The estimation procedure was relatively straightforward. After defining the key causal parameters of interest, ITET and ATET, we showed that the basic PCDID estimator targeted the ITET, whereas the simple mean-group (PCDID-MG) and pooled estimators targeted the ATET. We showed consistency and asymptotic normality of these estimators under minimal assumptions on the trend specification. We also showed that, when treatment effects were more heterogeneous (i.e., ITET was heterogeneous across treated units), PCDID-MG required weaker identification conditions than the pooled estimator and, given that both estimators satisfied their identification conditions, PCDID-MG was more efficient. We provided inference procedures based on the asymptotic normality results. We developed a parallel trend test (assuming an interactive effects structure), called the "Alpha" test, based on the PCDID approach. We also introduced aggregated PCDID estimators (PCDID-AGG) for micro/group-level data.

In Monte Carlo simulations, we considered scenarios of unparallel trends and nonstationary trends. We compared PCDID with DID-2wfe, DID with unit-specific cubic trends (e.g., Wooldridge (2005)), Bai (2009)'s iterative estimator, Xu (2017)'s GSC estimator, Gobillon and Magnac (2016)'s stepwise estimator, and Athey et al. (2018)'s MC-NNM estimator in ITET and ATET estimation, with and without staggered adoption. We found that PCDID had better finite sample performance in terms of bias and empirical SD, and it was robust across various scenarios. We examined PCDID's performance when the number of factors was unknown (using various auxiliary procedures to determine the number of factors), and when the covariates were correlated with the factor structure. We compared various inference procedures (analytical and bootstrap) for PCDID in finite samples, examined the size and power of the Alpha test, and examined estimation and inference based on PCDID in micro-level data.

In a placebo exercise, we showed that PCDID could be more efficient than DID-2wfe in scenarios where both approaches were valid, and it was less biased than DID-2wfe otherwise. In the welfare caseload analysis, we found that waivers programs (work requirements and time limits) reduced welfare caseload per capita by an average of 1.7% among all treated states, explaining 6.88% of the overall reduction in caseload between 1993Q1 and 1996Q2. We found that the effects were larger among Southern states (-2.4%), explaining 10.41% of the reduction between 93Q1-96Q2. The above estimates were based on the PCDID-MG estimator. In Wyoming, using the basic PCDID estimator, we found that waivers reduced welfare caseload per capita by an average of 11.4%, explaining 24.86% of the reduction between 93Q1-96Q2. All the above estimates were statistically significant at the 5% or 1% levels based on our preferred inference procedures. We plotted actual and predicted caseloads as well as the factor proxies over time. We found that the Alpha test did not reject the null of PTW. We examined the sensitivity of estimates to the number of factor proxies, and considered auxiliary procedures to determine the number of factors. We found that DID-2wfe tended to overestimate the effects, whereas DID with state-specific time polynomials failed to find any statistically significant effects.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505.

Abraham, S. and L. Sun (2018): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Working Paper.*

Ahn, S. and A. Horenstein (2013): "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203–1227.

Ahn, S., Y. Lee, and P. Schmidt (2001): "GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects," *Journal of Econometrics*, 101, 219–255.

——— (2013): "Panel Data Models with Multiple Time-Varying Individual Effects," *Journal of Econometrics*, 174, 1–14.

Altug, S. and R. Miller (1990): "Household Choices in Equilibrium," *Econometrica*, 58, 543–570.

Arkhangelsky, D., S. Athey, D. Hirshberg, G. Imbens, and S. Wager (2019): "Synthetic Difference in Differences," *Working Paper.*

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018): "Matrix Completion Methods for Causal Panel Data Models," *Working Paper.*

Athey, S. and G. Imbens (2018): "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption," *Working Paper.*

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171.

——— (2004): "Estimating Cross-Section Common Stochastic Trends in Nonstationary Panel Data," *Journal of Econometrics*, 122, 137–183.

——— (2009): "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Bai, J. and S. Ng (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74, 1133–1150.

Ben-Michael, E., A. Feller, and J. Rothstein (2018): "The Augmented Synthetic Control Method," *Working Paper.*

Bernanke, B., J. Boivin, and P. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, 120, 387–422.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics*, 119, 249–275.

Blundell, R. and M. C. Dias (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 44, 565–640.

Borusyak, K. and X. Jaravel (2017): "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume," *Working Paper.*

Callaway, B. and P. Sant'Anna (2018): "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment," Working Paper.

Cameron, A., J. Gelbach, and D. Miller (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 90, 414–427.

Canay, I., A. Santos, and A. Shaikh (2018): "The Wild Bootstrap with a "Small" Number of "Large" Clusters," *Working Paper.*

Carneiro, P., K. Hansen, and J. Heckman (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44, 362–422.

Carter, A., K. Schnepel, and D. Steigerwald (2017): "Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity," *Review of Economics and Statistics*, 99, 698–709.

Cawley, J., K. Connelly, J. Heckman, and E. Vytlacil (1997): "Cognitive Ability, Wages, and Meritocracy," in *Intelligence Genes, and Success: Scientists Respond to the Bell Curve*, ed. by B. Devlin, S. Feinberg, D. Resnick, and K. Roeder, Berlin: Springer-Verlag, 179–192.

Chamberlain, G. and M. Rothschild (1983): "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets," *Econometrica*, 51, 1281–1304.

Chan, M. K. (2013): "A Dynamic Model of Welfare Reform," *Econometrica*, 81, 941–1001.

Chan, M. K. and R. A. Moffitt (2018): "Welfare Reform and the Labor Market," *Annual Review of Economics*, 10, 347–381.

Chen, L., J. Dolado, and J. Gonzalo (2014): "Detecting Big Structural Breaks in Large Factor Models," *Journal of Econometrics*, 180, 30–48.

Chernozhukov, V., I. Fernandez-Val, J. Han, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2018): "Inference on Average Treatment Effects in Aggregate Panel Data Settings," *Working Paper*.

Conley, T. G. and C. R. Taber (2011): "Inference with "Difference in Differences" with a Small Number of Policy Changes," *The Review of Economics and Statistics*, 93, 113–125.

Davidson, J. and R. DeJong (2000): "The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals II: Fractionally Integrated Processes," *Econometric Theory*, 16, 643–666.

de Chaisemartin, C. and X. D'Haultfœuille (2018): "Two-way fixed effects estimators with heterogeneous treatment effects," *Working Paper*.

DeJong, R. and J. Davidson (2000): "The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals I: Weakly Dependent Processes," *Econometric Theory*, 16, 621–642.

Donald, S. G. and K. Lang (2007): "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 89, 221–233.

Ergemen, Y. (2019): "System Estimation of Panel Data Models Under Long-Range Dependence," *Journal of Business & Economic Statistics*, 37, 13–26.

Ergemen, Y. and C. Velasco (2017): "Estimation of Fractionally Integrated Panels with Fixed Effects and Cross-Section Dependence," *Journal of Econometrics*, 196, 248–258.

Ferman, B. and C. Pinto (2018): "Synthetic Controls with Imperfect Pre-Treatment Fit," *Working Paper*.

——— (2019): "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *Review of Economics and Statistics*, 101, 452–467.

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

Goldberger, A. (1972): "Stuctural Equations Methods in the Social Sciences," *Econometrica*, 40, 979–1001.

Gonçalves, S. and B. Perron (2014): "Bootstrapping Factor-Augmented Regression Models," *Journal of Econometrics*, 182, 156–173.

Goodman-Bacon, A. (2018): "Difference-in-Differences with Variation in Treatment Timing," Working Paper.

Greenaway-McGrevy, R., C. Han, and D. Sul (2012): "Asymptotic Distribution of Factor Augmented Estimators for Panel Regression," *Journal of Econometrics*, 169, 48–53.

Holtz-Eakin, D., W. Newey, and H. Rosen (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371–1395.

Hsiao, C., H. S. Ching, and S. K. Wan (2012): "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China," *Journal of Applied Econometrics*, 27, 705–740.

Jöreskog and A. Goldberger (1975): "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *Journal of the American Statistical Association*, 70, 631–639.

Kim, D. and T. Oka (2014): "Divorce Law Reforms and Divorce Rates in the USA: An Interactive Fixed-Effects Approach," *Journal of Applied Econometrics*, 29, 231–245.

Mackinnon, J. and M. Webb (2017): "Wild Bootstrap Inference for Wildly Different Cluster Sizes," *Journal of Applied Econometrics*, 32, 233–254.

MaCurdy, T. (1982): "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics*, 122, 81–126.

Meyer, B. D. (1995): "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics*, 13, 151–161.

Moon, H. R. and M. Weidner (2015): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83, 1543–1579.

Moulton, B. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 72, 334–338.

Onatski, A. (2012): "Asymptotics of the Principal Components Estimator of Large Factor Models with Weakly Influential Factors," *Journal of Econometrics*, 168, 244–258.

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

Phillips, P. and H. Moon (1999): "Linear Regression Limit Theorems for Nonstationary Panel Data," *Econometrica*, 67, 1057–1111.

Politis, D. and J. Romano (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

Stock, J. and M. Watson (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

Townsend, R. (1994): "Risk and Insurance in Village India," *Econometrica*, 62, 539–592.

Westerlund, J. and J. Urbain (2015): "Cross-Sectional Averages Versus Principal Components," *Journal of Econometrics*, 185, 372–377.

Wooldridge, J. (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *The Review of Economics and Statistics*, 87, 385–390.

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

Ziliak, J. P., D. N. Figlio, E. E. Davis, and L. S. Connolly (2000): "Accounting for the Decline in AFDC Caseloads: Welfare Reform or the Economy?" *Journal of Human Resources*, 35, 570–586.

TABLE I: SMALL SAMPLE PROPERTIES OF ESTIMATORS, ITET ESTIMATION[a]

		PCDID		DID-2wfe		DID-trend		BAI			GSC		GM		MC-NNM		
N_E+N_C	T	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Avg #Iter	Bias	SD	Bias	SD	Bias	SD	Avg Rank
Panel A: Stationary factors																	
51	10	0.00	0.15	0.00	0.47	-0.01	0.33	-0.02	0.19	*23*	0.01	0.59	0.01	0.57	0.01	0.33	*1.7*
51	20	0.00	0.09	-0.01	0.46	-0.02	0.31	-0.01	0.07	*13* [b]	0.00	0.14	0.00	0.10	-0.01	0.26	*1.5*
51	50	0.00	0.05	-0.02	0.41	0.00	0.36	-0.02	0.19	*11*	0.00	0.06	0.00	0.04	-0.01	0.13	*3.0*
6	100	0.00	0.09	0.00	0.27	0.02	0.37	-0.10	0.24	*40*	0.00	0.09	0.00	0.08	0.01	0.18	*1.1*
11	100	0.00	0.05	0.00	0.28	0.03	0.37	-0.05	0.22	*26*	0.00	0.05	0.00	0.04	0.00	0.16	*1.3*
26	100	0.00	0.03	0.00	0.27	0.02	0.36	-0.03	0.23	*13*	0.00	0.03	0.00	0.02	0.00	0.14	*1.7*
6	10	0.01	0.24	0.02	0.48	0.01	0.34	-0.17	0.39	*81*	0.01	0.55	0.01	0.55	0.01	0.34	*2.2*
11	20	0.00	0.13	-0.02	0.47	0.00	0.33	-0.07	0.29	*42*	0.00	0.17	0.00	0.16	-0.01	0.28	*1.5*
26	50	-0.01	0.05	-0.02	0.42	-0.01	0.37	-0.03	0.23	*15*	-0.01	0.06	0.00	0.04	0.00	0.15	*2.6*
51	100	0.00	0.03	0.00	0.27	0.02	0.34	-0.01	0.18	*8*	0.00	0.03	0.00	0.02	0.00	0.11	*2.8*
Panel B: Stationary factors with break																	
51	10	-0.01	0.23	1.27	0.87	0.43	0.42	-0.02	0.20	*49* [b]	-0.01	2.25	-0.01	2.25	2.00	1.19	*1.9*
51	20	0.00	0.16	1.56	1.08	0.59	0.49	-0.02	0.15	*49* [b]	-0.01	0.68	-0.01	0.68	2.63	1.22	*2.0*
51	50	0.00	0.12	1.77	1.19	1.05	0.77	-0.02	0.11	*64* [b]	0.01	0.36	0.01	0.36	3.22	1.18	*2.3*
6	100	0.06	0.34	1.85	1.28	1.43	1.04	-0.10	0.23	*105*	0.09	0.68	0.09	0.67	4.25	1.20	*1.6*
11	100	0.02	0.18	1.94	1.27	1.49	1.01	-0.06	0.14	*97* [b]	0.05	0.39	0.05	0.39	4.07	1.21	*1.7*
26	100	0.01	0.12	2.00	1.20	1.52	0.97	-0.03	0.11	*84* [b]	0.01	0.27	0.01	0.27	3.99	1.21	*1.7*
6	10	0.03	0.32	1.23	0.96	0.42	0.43	-0.09	0.42	*120*	0.05	2.65	0.05	2.65	2.16	1.20	*2.7*
11	20	0.01	0.20	1.53	1.08	0.59	0.50	-0.02	0.21	*76*	0.07	0.82	0.07	0.82	2.78	1.19	*2.2*
26	50	0.00	0.14	1.89	1.17	1.11	0.74	-0.03	0.13	*69* [b]	0.02	0.38	0.02	0.38	3.44	1.17	*2.3*
51	100	0.01	0.11	1.96	1.21	1.50	0.95	-0.03	0.10	*78* [b]	0.01	0.24	0.01	0.24	3.41	1.10	*2.9*
Panel C: Nonstationary factors with drift																	
51	10	0.01	0.17	0.13	0.66	-0.03	0.45	0.00	0.17	*34*	0.03	0.66	0.03	0.66	0.17	0.80	*2.2*
51	20	-0.01	0.11	0.18	0.93	-0.04	0.49	-0.01	0.11	*29* [b]	0.00	0.22	0.00	0.22	0.27	1.09	*2.1*
51	50	0.00	0.07	0.53	1.53	0.02	0.67	-0.01	0.07	*31* [b]	0.00	0.15	0.00	0.15	0.70	1.39	*2.5*
6	100	-0.01	0.19	0.98	2.44	-0.03	0.99	-0.03	0.14	*67* [b]	-0.03	0.33	-0.03	0.33	2.94	3.50	*2.4*
11	100	-0.01	0.09	0.98	2.43	-0.02	0.98	-0.01	0.08	*46* [b]	-0.02	0.21	-0.02	0.20	2.01	2.74	*2.4*
26	100	0.00	0.06	0.87	2.39	-0.02	0.99	-0.01	0.06	*35* [b]	-0.01	0.13	-0.01	0.13	1.40	2.08	*2.4*
6	10	-0.02	0.22	0.13	0.63	0.00	0.44	-0.12	0.42	*94*	0.01	1.23	0.01	1.23	0.29	1.07	*2.6*
11	20	-0.01	0.15	0.20	0.94	-0.04	0.48	-0.04	0.26	*50*	-0.01	0.26	-0.01	0.25	0.32	1.16	*2.4*
26	50	-0.01	0.09	0.51	1.52	0.00	0.72	-0.01	0.07	*34* [b]	-0.02	0.17	-0.02	0.17	0.74	1.49	*2.5*
51	100	0.00	0.05	0.95	2.41	-0.01	1.02	-0.01	0.05	*32* [b]	-0.01	0.12	-0.01	0.11	1.31	1.98	*2.4*

[a] See Section 6.1 for details of DGPs and estimators. PCDID: basic PCDID estimator. DID-2wfe: Two way fixed effects estimator. DID-trend: DID with unit-specific cubic trend. BAI: Bai (2009)'s iterative estimator. GSC: Xu (2017)'s generalized synthetic control estimator. GM: Gobillon and Magnac (2016)'s stepwise estimator. MC-NNM: Athey et al. (2018)'s nuclear norm matrix completion estimator. SD: empirical standard deviation of estimator. Avg #Iter: average number of iterations used. Avg Rank: Average matrix rank computed. Panels A, B and C refer to DGPs 1, 2 and 3 in Section 6.1. Number of replications=1000. In all specifications, N_E=1 and T_0=T_1=T/2. PCDID, BAI, GSC and GM assume 3 factors.

[b] Numerical convergence is attained in all replications. See Section 6.1 for numerical convergence criteria.

TABLE II: SMALL SAMPLE PROPERTIES OF ESTIMATORS, ATET ESTIMATION[a]

		PCDID		DID-2wfe		DID trend		BAI			GSC		GM		MC-NNM		
N_E+N_C	T	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Avg #Iter	Bias	SD	Bias	SD	Bias	SD	Avg Rank
Panel A: Stationary factors																	
100	10	-0.01	0.15	0.00	0.44	-0.01	0.27	-0.34	0.49	*82*	0.00	0.26	0.00	0.26	0.00	0.48	*1.5*
100	20	0.00	0.15	-0.02	0.45	-0.01	0.27	-0.36	0.57	*113*	0.00	0.15	0.00	0.15	-0.02	0.41	*1.1*
100	50	0.00	0.15	-0.02	0.40	-0.01	0.34	-0.35	0.52	*97*	0.00	0.15	0.00	0.15	-0.04	0.25	*1.7*
10	100	0.01	0.45	0.01	0.51	0.02	0.56	-0.36	0.69	*179*	0.01	0.45	0.01	0.45	0.01	0.48	*1.1*
20	100	0.02	0.32	0.01	0.41	0.04	0.45	-0.36	0.66	*142*	0.02	0.32	0.02	0.32	0.02	0.38	*1.1*
50	100	0.00	0.21	0.00	0.32	0.02	0.37	-0.34	0.57	*107*	0.00	0.21	0.00	0.21	-0.01	0.27	*1.1*
10	10	0.02	0.48	0.02	0.62	0.01	0.53	-0.32	0.62	*181*	0.02	0.62	0.02	0.62	0.02	0.64	*2.2*
20	20	0.01	0.35	-0.02	0.55	-0.01	0.42	-0.34	0.66	*169*	0.01	0.36	0.01	0.36	-0.01	0.46	*1.6*
50	50	0.01	0.19	0.00	0.42	0.01	0.36	-0.33	0.57	*116*	0.01	0.20	0.01	0.20	-0.03	0.26	*2.4*
100	100	0.01	0.14	0.01	0.28	0.03	0.34	-0.37	0.47	*72*	0.01	0.14	0.01	0.14	-0.01	0.23	*1.1*
Panel B: Stationary factors with break																	
100	10	0.00	0.16	1.24	0.46	0.41	0.27	-0.34	0.69	*203*	0.03	1.13	0.03	1.13	2.61	0.95	*2.2*
100	20	0.00	0.15	1.54	0.50	0.59	0.29	-0.35	0.76	*238*	0.02	0.27	0.02	0.27	3.37	0.69	*2.2*
100	50	0.00	0.15	1.80	0.44	1.06	0.36	-0.26	0.70	*263*	0.01	0.18	0.01	0.18	3.82	0.51	*2.3*
10	100	0.07	0.52	1.86	0.89	1.42	0.78	-0.24	0.66	*203*	0.13	0.69	0.13	0.69	4.62	0.74	*2.0*
20	100	0.04	0.33	1.91	0.63	1.47	0.58	-0.20	0.57	*196*	0.06	0.41	0.06	0.41	4.40	0.59	*2.3*
50	100	0.00	0.21	1.90	0.47	1.45	0.45	-0.20	0.51	*186*	0.02	0.24	0.02	0.24	4.15	0.47	*2.2*
10	10	0.06	0.50	1.28	0.76	0.45	0.55	-0.29	0.74	*223*	0.15	2.61	0.15	2.61	2.69	1.10	*2.7*
20	20	0.02	0.37	1.53	0.70	0.59	0.45	-0.32	0.76	*235*	0.07	0.63	0.07	0.63	3.54	0.80	*2.1*
50	50	0.02	0.21	1.81	0.53	1.07	0.40	-0.23	0.68	*245*	0.03	0.27	0.03	0.27	4.05	0.54	*2.2*
100	100	0.01	0.14	1.91	0.37	1.46	0.38	-0.19	0.48	*176*	0.01	0.16	0.01	0.16	3.39	0.39	*3.0*
Panel C: Nonstationary factors with drift																	
100	10	-0.01	0.15	0.12	0.54	-0.02	0.38	-0.27	0.64	*143*	-0.03	0.41	-0.03	0.41	0.54	1.68	*2.1*
100	20	0.00	0.15	0.18	0.75	-0.03	0.43	-0.21	0.94	*239*	0.00	0.16	0.00	0.16	0.78	1.99	*2.5*
100	50	0.00	0.15	0.51	1.12	0.02	0.57	-0.17	1.30	*342*	0.00	0.15	0.00	0.15	1.98	2.60	*3.1*
10	100	-0.01	0.48	0.92	1.90	0.00	0.98	-0.23	1.36	*299*	-0.02	0.53	-0.02	0.53	5.34	4.80	*2.7*
20	100	0.01	0.32	0.98	1.82	0.00	0.91	-0.22	1.42	*335*	0.00	0.35	0.00	0.35	5.01	4.52	*3.0*
50	100	0.00	0.21	0.95	1.66	-0.01	0.84	-0.23	1.48	*359*	-0.01	0.22	-0.01	0.22	4.82	4.33	*2.9*
10	10	0.00	0.49	0.15	0.75	0.01	0.60	-0.28	0.74	*212*	0.00	0.67	0.00	0.67	0.59	1.85	*2.7*
20	20	-0.01	0.35	0.19	0.81	-0.03	0.54	-0.23	0.96	*249*	0.00	0.38	0.00	0.38	0.86	2.17	*2.7*
50	50	0.01	0.20	0.53	1.14	0.03	0.59	-0.22	1.27	*332*	0.00	0.21	0.00	0.21	2.28	2.91	*3.1*
100	100	0.01	0.14	0.98	1.60	-0.01	0.83	-0.22	1.42	*361*	0.01	0.15	0.01	0.15	3.96	3.70	*3.1*

[a] See Section 6.1 for details of DGPs and estimators. PCDID: PCDID simple mean-group (PCDID-MG) estimator. DID-2wfe: Two way fixed effects estimator. DID-trend: DID with unit-specific cubic trend. BAI: Bai (2009)'s iterative estimator. GSC: Xu (2017)'s generalized synthetic control estimator. GM: Gobillon and Magnac (2016)'s stepwise estimator. MC-NNM: Athey et al. (2018)'s nuclear norm matrix completion estimator. SD: empirical standard deviation of estimator. Avg #Iter: average number of iterations used. Avg Rank: Average matrix rank computed. Panels A, B and C refer to DGPs 1, 2 and 3 in Section 6.1. Number of replications=1000. In all specifications, N_E=N_c=N/2 and T_0=T_1=T/2. PCDID, BAI, GSC and GM assume 3 factors. In BAI, numerical convergence is attained in some replications only. See Section 6.1 for numerical convergence criteria.

TABLE III: SMALL SAMPLE PROPERTIES OF ESTIMATORS, ATET ESTIMATION (STAGGERED ADOPTION)[a]

		PCDID		DID-2wfe		DID trend		BAI			GSC		GM		MC-NNM		
N_E+N_C	T	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Avg #Iter	Bias	SD	Bias	SD	Bias	SD	Avg Rank
Panel A: Stationary factors																	
100	10	-0.01	0.17	0.02	0.36	-0.04	0.20	-0.01	0.21	*156*	-0.10	4.20	-0.19	3.47	0.00	0.45	*1.3*
100	20	0.00	0.16	0.01	0.37	-0.01	0.20	-0.03	0.18	*147*	0.00	0.18	0.00	0.17	-0.02	0.38	*1.1*
100	50	0.00	0.16	0.01	0.33	-0.01	0.21	-0.04	0.17	*99*	0.00	0.16	0.00	0.16	-0.03	0.24	*1.6*
10	100	0.01	0.48	0.03	0.52	0.01	0.60	-0.04	0.62	*93*	0.01	0.48	0.01	0.48	0.01	0.50	*1.2*
20	100	0.01	0.34	0.03	0.39	0.01	0.43	-0.02	0.43	*77*	0.01	0.34	0.01	0.34	0.01	0.37	*1.7*
50	100	0.00	0.22	0.02	0.29	-0.01	0.28	-0.04	0.25	*74* [b]	0.00	0.22	0.00	0.22	-0.01	0.27	*1.0*
10	10	0.01	0.52	0.04	0.60	0.00	0.60	-0.10	0.71	*138*	0.10	2.95	-0.03	2.81	0.02	0.61	*2.1*
20	20	0.00	0.37	0.02	0.50	-0.01	0.45	-0.04	0.46	*140*	0.00	0.38	0.00	0.38	0.00	0.46	*1.4*
50	50	0.01	0.20	0.02	0.36	-0.01	0.27	-0.05	0.23	*99*	0.01	0.21	0.01	0.21	-0.02	0.25	*2.3*
100	100	0.00	0.15	0.03	0.24	-0.01	0.19	-0.04	0.17	*73* [b]	0.00	0.15	0.00	0.15	-0.01	0.22	*1.1*
Panel B: Stationary factors with break																	
100	10	-0.01	0.17	0.94	0.38	-0.06	0.26	-0.04	0.20	*116*	-0.88	32.94	-0.35	27.60	2.08	0.70	*2.1*
100	20	0.00	0.16	1.12	0.41	-0.03	0.28	-0.06	0.17	*121*	0.00	0.26	0.00	0.26	2.58	0.53	*2.1*
100	50	0.00	0.16	1.28	0.36	0.04	0.33	-0.10	0.18	*89*	0.01	0.18	0.01	0.18	2.31	0.40	*2.8*
10	100	0.03	0.49	1.33	0.77	0.09	1.14	-0.07	0.63	*91*	0.08	0.56	0.08	0.56	3.25	0.86	*1.9*
20	100	0.02	0.34	1.33	0.53	0.05	0.81	-0.08	0.44	*77*	0.04	0.37	0.04	0.37	2.96	0.63	*2.2*
50	100	0.00	0.22	1.31	0.39	0.03	0.53	-0.10	0.27	*71* [b]	0.01	0.23	0.01	0.23	2.38	0.44	*2.8*
10	10	0.04	0.52	1.01	0.71	-0.01	0.77	-0.12	0.70	*141*	-0.47	17.08	-0.01	14.11	2.15	0.93	*3.1*
20	20	0.01	0.37	1.13	0.62	-0.01	0.61	-0.04	0.44	*126*	0.06	0.55	0.06	0.55	2.70	0.69	*2.3*
50	50	0.01	0.21	1.29	0.44	0.04	0.45	-0.10	0.24	*94*	0.02	0.23	0.02	0.23	2.73	0.47	*2.3*
100	100	0.01	0.15	1.30	0.31	0.02	0.36	-0.11	0.17	*69* [b]	0.01	0.16	0.01	0.16	2.49	0.33	*2.1*
Panel C: Nonstationary factors with drift																	
100	10	-0.01	0.17	0.12	0.43	-0.04	0.23	-0.03	0.22	*109*	0.15	0.68	0.18	0.82	0.39	1.24	*1.8*
100	20	0.00	0.16	0.17	0.58	-0.02	0.26	-0.03	0.18	*86*	0.00	0.19	0.00	0.19	0.53	1.36	*2.4*
100	50	0.00	0.15	0.42	0.87	-0.02	0.32	-0.03	0.18	*63*	0.00	0.16	0.00	0.16	1.41	1.80	*2.7*
10	100	-0.01	0.50	0.78	1.58	-0.04	1.31	-0.02	0.59	*95*	-0.03	0.55	-0.03	0.55	3.80	3.44	*2.7*
20	100	0.01	0.34	0.82	1.47	-0.02	0.94	-0.01	0.38	*72*	0.00	0.36	0.00	0.36	3.21	2.96	*3.2*
50	100	0.00	0.22	0.78	1.29	-0.05	0.63	-0.02	0.24	*56*	-0.01	0.23	-0.01	0.23	2.69	2.43	*3.1*
10	10	0.00	0.52	0.15	0.69	-0.02	0.65	-0.08	0.72	*136*	0.17	2.39	0.21	2.17	0.42	1.36	*2.8*
20	20	-0.01	0.37	0.18	0.69	-0.02	0.54	-0.01	0.44	*106*	-0.01	0.41	-0.01	0.41	0.64	1.65	*2.3*
50	50	0.01	0.21	0.44	0.90	0.00	0.48	-0.02	0.23	*67*	0.00	0.22	0.00	0.22	1.50	1.92	*2.8*
100	100	0.00	0.15	0.80	1.24	-0.03	0.46	-0.02	0.17	*49*	0.00	0.15	0.00	0.15	2.24	2.06	*3.1*

[a] All DGPs and estimators are the same as in Table II (ATET estimation), except that T_0 is drawn from a discrete uniform distribution: $T_0 \sim Unif\{\lceil T/4 \rceil, \lfloor 3T/4 \rfloor\}$. See Section 6.1 for details. PCDID: PCDID simple mean-group (PCDID-MG) estimator. DID-2wfe: Two way fixed effects estimator. DID-trend: DID with unit-specific cubic trend. BAI: Bai (2009)'s iterative estimator. GSC: Xu (2017)'s generalized synthetic control estimator. GM: Gobillon and Magnac (2016)'s stepwise estimator. MC-NNM: Athey et al. (2018)'s nuclear norm matrix completion estimator. Number of replications=1000. In all specifications, N_E=N_c=N/2 and T_1=T-T_0. PCDID, BAI, GSC and GM assume 3 factors.

[b] Numerical convergence is attained in all replications. See Section 6.1 for numerical convergence criteria.

TABLE IV: PCDID INFERENCE PROCEDURES, REJECTION RATE (%)[a]

DGPs for ITET estimation (Table I) Basic PCDID estimator, ITET inference (H_0: ITET=3)						DGPs for ATET estimation (Table II) PCDID-MG estimator, ATET inference (H_0: ATET=3)					
N_E+N_C	T	TrueF	Asym	b-t	b-se	N_E+N_C	T	TrueF	Asym	b-t	b-se
Panel A: Stationary factors											
51	10	4.4	5.5	3.5	3.7	100	10	5.4	6.5	5.3	6.5
51	20	4.2	6.2	6.1	4.5	100	20	5.0	5.3	5.2	5.6
51	50	4.1	5.0	6.9	3.3	100	50	5.7	5.9	6.6	5.6
6	100	4.6	22.6	10.6	3.8	10	100	5.1	5.4	7.3	3.5
11	100	5.4	14.3	6.8	1.0	20	100	3.9	3.9	4.9	4.3
26	100	4.8	6.3	3.5	1.1	50	100	5.1	5.3	5.4	4.9
6	10	4.8	8.8	6.6	6.4	10	10	5.6	6.5	10.4	5.2
11	20	4.8	9.5	7.3	4.6	20	20	5.5	6.7	7.9	8.9
26	50	5.0	7.8	7.0	3.3	50	50	4.8	4.7	4.9	5.6
51	100	5.1	4.7	4.7	3.1	100	100	4.5	4.8	4.5	4.9
Panel B: Stationary factors with break											
51	10	4.9	5.9	4.7	5.5	100	10	5.5	7.6	5.3	5.4
51	20	6.2	7.2	7.0	5.4	100	20	5.0	6.4	4.8	5.0
51	50	5.1	4.9	5.3	3.6	100	50	6.0	6.0	4.9	4.8
6	100	5.0	32.4	13.6	9.7	10	100	5.2	7.2	7.8	2.3
11	100	5.8	13.9	6.5	2.9	20	100	4.9	5.9	5.5	3.9
26	100	6.7	8.1	7.0	3.5	50	100	5.2	5.3	5.1	5.1
6	10	4.7	5.3	4.2	3.6	10	10	5.9	7.6	9.4	4.0
11	20	4.0	7.1	6.4	3.2	20	20	6.2	7.9	7.7	7.0
26	50	5.8	6.3	7.2	4.5	50	50	4.6	5.2	4.3	4.1
51	100	4.9	5.0	5.1	3.7	100	100	4.7	5.5	5.2	4.7
Panel C: Nonstationary factors with drift											
51	10	4.1	6.0	4.5	5.7	100	10	5.4	6.6	5.9	7.3
51	20	4.8	5.6	6.5	5.3	100	20	4.7	6.2	5.8	5.9
51	50	6.6	6.2	6.7	5.5	100	50	5.8	5.7	5.2	5.3
6	100	3.6	26.6	7.9	3.3	10	100	5.0	6.2	7.8	2.4
11	100	5.3	15.0	9.8	3.6	20	100	4.0	4.0	4.8	4.0
26	100	4.8	5.9	7.2	2.5	50	100	4.8	5.1	5.4	5.0
6	10	5.3	6.8	5.5	3.7	10	10	5.6	6.9	9.5	5.1
11	20	3.2	8.4	6.4	4.3	20	20	6.5	7.6	8.1	7.9
26	50	4.8	6.6	8.0	4.6	50	50	5.0	5.4	4.8	5.7
51	100	4.8	4.9	5.9	3.8	100	100	4.6	4.6	4.9	4.6

[a] See Section 6.2 for details of inference procedures. All DGPs are the same as in Table I and II, respectively (except setting ρ_ϵ=0 in the DGPs for ITET inference). The null hypothesis is set at the DGPs' true value. TrueF: assume factors are observed in PCDID estimation (infeasible) and compute t-statistic, reject if $|t|>=1.96$. Asym: use 3 factor proxies in PCDID estimation and compute t-statistic, reject if $|t|>=1.96$. b-t: same as Asym, but reject if $t<=c_{0.025}$ or $t>=c_{0.975}$ where $c_{0.025}$, $c_{0.975}$ are percentiles of the bootstrap distribution of t-statistics. b-se: same as Asym, but the standard error in the t-statistics formula is obtained from bootstrap samples. See Section 6.2 for analytical standard errrors used in TrueF, Asym and b-t. See Appendix 6 for details on bootstrap sample construction. Number of replications=1000. Bootstrap repetitions = 199. $T_0=T_1=T/2$. In ITET inference, N_E=1; in ATET inference, $N_E=N_C=N/2$. The nominal size is 5%.

TABLE V: PARALLEL TREND ALPHA TEST, REJECTION RATE (%)[a]

		DGP1: Stationary Factor			DGP2: Stationary Factor with break			DGP3: Nonstationary Factor with drift		
N_C	T	TrueF	asym	b-t	TrueF	asym	b-t	TrueF	asym	b-t
Panel A: N_E =5										
5	10	4.6	12.3	7.3	6.3	12.0	6.7	5.1	9.6	5.1
10	10	4.6	8.2	5.2	4.6	8.2	6.2	4.6	8.6	5.9
25	10	5.3	6.0	4.4	4.8	6.0	4.4	3.9	5.5	4.1
50	10	3.7	5.5	5.8	5.4	6.1	6.1	4.9	5.5	5.2
5	20	5.3	10.9	6.4	5.3	12.1	7.7	4.3	12.5	8.0
10	20	4.7	8.9	6.5	5.7	8.4	5.3	3.8	7.8	5.0
25	20	4.3	5.2	4.0	4.5	5.5	3.8	5.2	6.6	5.3
50	20	4.4	4.9	5.2	5.3	5.7	3.8	4.1	4.9	4.3
5	50	4.6	13.5	7.6	3.8	11.5	7.3	5.0	12.3	7.3
10	50	4.8	8.3	5.2	3.4	5.9	4.0	4.8	7.6	5.2
25	50	3.9	4.9	4.2	4.6	5.9	4.9	4.9	5.9	4.6
50	50	4.6	5.4	3.9	5.7	6.0	5.3	5.0	5.0	4.8
Panel B: N_E =10										
5	10	4.9	20.8	10.7	5.6	20.6	11.0	4.7	21.0	11.0
10	10	4.8	12.7	6.3	4.6	12.8	6.9	4.5	11.6	6.4
25	10	5.0	8.4	5.8	5.0	8.6	6.5	6.1	8.3	6.0
50	10	5.6	6.4	5.6	5.7	6.9	5.7	5.6	7.2	5.6
5	20	4.5	22.2	11.7	5.2	20.5	10.6	5.0	21.2	11.3
10	20	4.8	13.5	6.9	3.7	14.4	7.4	5.0	13.1	7.4
25	20	6.1	9.4	6.0	5.4	8.8	6.2	6.0	7.9	6.2
50	20	6.1	6.6	5.1	3.0	5.6	5.5	3.4	4.9	4.4
5	50	4.2	23.6	12.6	4.3	22.2	12.6	4.7	21.6	11.2
10	50	4.7	15.7	8.7	4.6	14.8	8.1	4.7	13.7	7.8
25	50	4.4	7.5	4.6	4.7	7.9	5.7	5.0	8.6	5.8
50	50	4.4	7.0	5.2	4.7	6.3	5.2	5.0	6.7	5.1

[a] See Section 6.2 for details of DGPs and inference procedures. All DGPs assume weak parallel trend (PTW) holds, with m_C=m_E=1. DGP1: one AR(1) factor. DGP2: one AR(1) factor with break. DGP3: one I(1) factor with drift. The DGPs are otherwise the same as in Table II (ATET estimation). TrueF: assume the factor is observed in PCDID estimation (infeasible) and compute t-statistic, reject if |t|>=1.96. Asym: use a factor proxy in PCDID estimation and compute t-statistic, reject if |t|>=1.96. b-t: same as Asym, but reject if t<=$c_{0.025}$ or t>=$c_{0.975}$ where $c_{0.025}$, $c_{0.975}$ are percentiles of the bootstrap distribution of t-statistics. See Section 6.2 for analytical standard errrors used in TrueF, Asym and b-t. See Appendix 6 for details on bootstrap sample construction. Number of replications=1000. Bootstrap repetitions = 199. T_0=T_1=T/2. The nominal size is 5%.

TABLE VI: PCDID ESTIMATION AND INFERENCE IN MICRO-LEVEL DATA[a,b,c]

	DGPs for ITET estimation and inference (H_0: ITET=3)						DGPs for ATET estimation and inference (H_0: ATET=3)					
	Bias	SD	TrueF	Asym	b-t	b-se	Bias	SD	TrueF	Asym	b-t	b-se
Panel A: Multiple control groups (N_c=5,10,25,50), each has 25 units:												
DGP1: Stationary factors												
N_c=5	0.00	0.11	5.0	24.7	13.1	6.8	0.00	0.23	5.8	8.6	8.0	8.9
10	0.00	0.07	5.1	13.3	7.7	3.4	0.00	0.21	5.5	6.2	6.0	6.2
25	0.00	0.05	5.0	7.1	6.5	3.8	0.01	0.21	5.4	6.0	6.1	6.2
50	0.00	0.05	4.9	6.4	6.4	5.5	0.00	0.19	4.6	4.5	5.1	4.7
DGP2: Stationary factors with break												
N_c=5	0.02	0.26	5.1	18.5	9.1	5.8	0.04	0.29	6.5	15.3	10.4	10.2
10	0.00	0.17	4.8	9.2	5.7	2.9	0.01	0.23	6.7	9.4	6.6	6.5
25	0.00	0.12	4.3	5.3	5.7	3.3	0.01	0.22	5.9	7.6	7.2	5.8
50	0.00	0.12	5.8	6.3	8.0	6.0	0.00	0.20	5.4	5.7	6.1	5.7
DGP3: Nonstationary factors with drift												
N_c=5	-0.03	0.17	4.4	20.6	8.4	4.1	-0.02	0.26	6.1	12.0	6.5	7.2
10	-0.02	0.11	5.2	15.2	9.3	5.2	-0.01	0.22	5.8	6.6	6.2	6.5
25	0.00	0.08	5.2	6.4	7.7	5.7	0.00	0.21	6.2	6.5	6.0	6.2
50	0.00	0.07	5.5	5.4	6.8	5.4	0.00	0.20	5.0	5.0	5.5	5.2
Panel B: One control group (N_c=1) with 5, 10, 25, 50 units:												
DGP1: Stationary factors												
#units=5	0.01	0.14	5.0	24.4	12.0	5.1	-0.01	0.24	4.7	8.8	7.1	8.0
10	0.00	0.09	4.3	15.4	9.1	3.8	-0.02	0.21	5.3	6.4	5.7	6.9
25	-0.01	0.06	4.4	6.8	5.9	3.2	0.01	0.20	5.5	5.2	5.1	5.4
50	0.00	0.05	5.4	7.0	7.3	5.9	0.00	0.21	6.1	6.7	7.3	7.5
DGP2: Stationary factors with break												
#units=5	0.08	0.32	4.8	19.4	7.7	4.5	0.07	0.35	5.6	22.2	10.0	10.3
10	0.03	0.19	5.3	8.7	4.9	1.8	0.01	0.26	5.1	11.8	6.4	6.7
25	0.01	0.15	6.2	7.0	8.3	5.6	0.03	0.23	5.7	8.4	7.7	7.4
50	0.01	0.14	5.4	5.4	6.3	5.1	0.01	0.24	7.5	8.7	8.8	8.6
DGP3: Nonstationary factors with drift												
#units=5	-0.02	0.20	5.2	20.5	8.0	3.6	-0.03	0.26	4.9	11.4	5.6	6.2
10	-0.01	0.13	4.6	10.9	7.0	4.0	-0.02	0.23	5.3	8.6	6.8	7.5
25	-0.01	0.09	4.0	7.2	8.0	6.0	0.01	0.21	5.2	5.7	5.7	5.6
50	0.00	0.08	7.1	6.4	7.9	6.6	0.01	0.22	6.4	7.5	7.8	7.6
Panel C: Multiple control groups (N_c=10, 50), unbalanced number of units (average=25):												
N_c=10	0.00	0.07	5.1	13.4	6.7	3.4	0.00	0.21	5.5	6.4	6.4	6.7
50	0.00	0.05	4.9	6.7	6.6	5.4	0.00	0.20	4.6	4.7	5.0	4.9
N_c=10	0.01	0.16	4.8	9.0	5.9	2.6	0.01	0.23	6.7	8.8	6.7	7.0
50	0.00	0.12	5.8	5.6	7.1	5.9	0.00	0.20	5.4	5.6	5.9	5.8
N_c=10	-0.02	0.11	5.2	14.9	10.3	6.1	-0.01	0.22	5.8	6.7	6.3	6.4
50	0.00	0.07	5.5	5.4	7.8	6.0	0.00	0.20	5.0	5.2	5.3	5.4
Panel D: Multiple control groups (N_c=10, 50), each has 35 units:												
N_c=10	0.00	0.07	4.6	13.2	8.0	3.5	0.00	0.21	4.2	5.1	5.2	5.9
50	0.00	0.05	5.6	6.1	5.8	5.0	0.01	0.20	5.2	5.6	5.8	5.6
N_c=10	0.01	0.16	3.7	8.9	5.3	3.5	0.00	0.23	5.5	7.0	4.8	5.1
50	0.00	0.11	4.7	5.2	5.2	4.7	0.01	0.21	6.4	5.7	6.0	5.6
N_c=10	-0.01	0.11	4.4	12.2	8.7	4.8	-0.01	0.22	4.2	5.3	5.3	5.2
50	-0.01	0.07	5.5	5.2	6.7	5.8	0.01	0.21	5.2	5.6	5.5	5.4

[a] See Section 6.3 for details of DGPs and estimation/inference procedures. The DGPs on the left (ITET) have 1 treated group with 1 treated unit. The DGPs on the right (ATET) have 1 treated group with 25 treated units. In Panels C and D, rows 1 and 2 are based on DGP1, rows 3 and 4 are based on DGP2, rows 5 and 6 are based on DGP3.

[b] Panels A, C and D use the basic PCDID-AGG estimator (ITET) and PCDID-MG-AGG estimator (ATET). Panel B uses the basic PCDID estimator (ITET) and PCDID-MG estimator (ATET). SD: empirical standard deviation of estimator.

[c] In inference procedures, the null hypothesis is set at the DGPs' true value and rejection rates (%) are reported. TrueF: assume factors are observed in PCDID estimation (infeasible) and compute t-statistic, reject if $|t|>=1.96$. Asym: use 3 factor proxies in PCDID estimation and compute t-statistic, reject if $|t|>=1.96$. b-t: same as Asym, but reject if $t<=c_{0.025}$ or $t>=c_{0.975}$ where $c_{0.025}$, $c_{0.975}$ are percentiles of the bootstrap distribution of t-statistics. b-se: same as Asym, but the standard error in the t-statistics formula is obtained from bootstrap samples. See Section 6.2 for analytical standard errrors used in TrueF, Asym and b-t. See Appendix 6 for details on bootstrap sample construction. Number of replications=1000. Bootstrap repetitions = 199. $T_0=T_1=T/2=25$. The nominal size is 5%.

TABLE VII: RESULTS FROM PLACEBO DESIGN[a]

	Aggregated State-Year Panel (# obs=1071)					Micro-Level Data (#obs=549,735)				
	DID-	PCDID-MG				DID-	PCDID-MG-AGG			
	2wfe	1PC	3PC	5PC	7PC	2wfe	1PC	3PC	5PC	7PC
Scenario A: The treated states come from the pool of all states										
Empirical mean and SD of estimators:										
Mean	*-0.0004*	-0.0008	-0.0011	-0.0004	-0.0001	*0.0003*	-0.0002	-0.0002	-0.0001	-0.0001
SD	*0.0193*	0.0167	0.0138	0.0163	0.0178	*0.0173*	0.0173	0.0147	0.0164	0.0181
Inference procedures (H_0: ATET=0), rejection rate (%):										
b-p	*5.4* [b]	2.3	2.2	5.1	4.9	*4.7* [b]	1.8	2.1	4.2	4.8
b-se	*3.8* [b]	2.1	1.8	4.7	5.4	*3.7* [b]	1.4	1.4	3.4	4.9
Scenario B: The treated states come from the pool of the largest 25 states										
Empirical mean and SD of estimators:										
Mean	*0.0200*	0.0011	0.0006	0.0012	0.0013	*0.0080*	0.0012	0.0005	0.0011	0.0013
SD	*0.0130*	0.0160	0.0129	0.0149	0.0165	*0.0139*	0.0156	0.0129	0.0134	0.0151
Inference procedures (H_0: ATET=0), rejection rate (%):										
b-p	*15.2* [b]	2.2	1.8	4.6	4.7	*6.7* [b]	2.5	1.8	2.4	4.0
b-se	*13.3* [b]	1.7	1.6	3.8	4.3	*5.4* [b]	1.8	1.3	2.2	3.0

[a] See Section 7 for details of the placebo design. Number of replications=1000. In each replication, there are 10 treated states and 41 control states. PC: number of factor proxies used. b-p: reject if the estimate<=$c_{0.025}$ or >=$c_{0.975}$ where $c_{0.025}$, $c_{0.975}$ are percentiles of the bootstrap distribution of the estimator. It is assumed that the outcomes are stationary. b-se: compute t-statistic where the standard error is obtained from bootstrap samples, reject if |t|>=1.96. Bootstrap repetitions = 199. See Appendix 6 for details on bootstrap sample construction.

[b] Wild cluster bootstrap imposing H_0 (no effect) is used. See Appendix 6 for details.

TABLE VIII: EFFECTS OF WELFARE WAIVERS ON WELFARE CASELOADS[a]

	Panel A: PCDID (#PC=4)[b]				Panel B: PCDID (#PC=3)[b]				Panel C: DID regressions[c]		
									Unit-specific time polynomial		
	ALL treated states	Southern treated states	Non-southern treated states	Wyoming (ITET)[d]	ALL treated states	Southern treated states	Non-southern treated states	Wyoming (ITET)[d]	Quartic	Cubic	Two-way fixed effect
Policy intervention dummy	-0.017	-0.024	-0.013	-0.114	-0.018	-0.021	-0.016	-0.138	-0.007	-0.008	-0.054
asym-se	(0.007)	(0.007)	(0.010)	(0.029)	(0.008)	(0.009)	(0.011)	(0.039)			
b-se	(0.009)	(0.012)	(0.011)	(0.051)	(0.010)	(0.014)	(0.011)	(0.053)	(0.010)	(0.013)	(0.032)
b-t-pval	{0.030}	{0.000}	{0.241}	{0.010}	{0.050}	{0.040}	{0.161}	{0.010}			
Max monthly welfare ben. ($100)	0.014	0.046	-0.001	-0.010	0.035	0.092	0.008	0.014	-0.003	-0.003	0.020
asym-se	(0.008)	(0.019)	(0.004)	(0.020)	(0.016)	(0.044)	(0.004)	(0.017)			
b-se	(0.016)	(0.033)	(0.009)	(0.027)	(0.016)	(0.046)	(0.008)	(0.027)	(0.002)	(0.002)	(0.014)
b-t-pval	{0.121}	{0.000}	{0.834}	{0.754}	{0.010}	{0.020}	{0.221}	{0.673}			
State unemployment rate (%)	0.021	0.016	0.023	-0.030	0.029	0.022	0.032	-0.017	0.007	0.011	0.023
asym-se	(0.004)	(0.006)	(0.004)	(0.008)	(0.005)	(0.008)	(0.006)	(0.009)			
b-se	(0.007)	(0.007)	(0.008)	(0.014)	(0.007)	(0.010)	(0.009)	(0.013)	(0.002)	(0.003)	(0.008)
b-t-pval	{0.000}	{0.000}	{0.000}	{0.020}	{0.000}	{0.010}	{0.000}	{0.261}			
Ln(state empl. to popn ratio)	0.058	0.070	0.052	-0.622	-0.128	-0.291	-0.050	-0.754	-0.338	-0.198	-1.346
asym-se	(0.129)	(0.236)	(0.157)	(0.357)	(0.134)	(0.306)	(0.138)	(0.352)			
b-se	(0.164)	(0.241)	(0.167)	(0.502)	(0.170)	(0.282)	(0.169)	(0.549)	(0.087)	(0.110)	(0.450)
b-t-pval	{0.724}	{0.744}	{0.764}	{0.231}	{0.412}	{0.281}	{0.704}	{0.161}			
Percent of predicted change in caseload explained by reform (Jan93-Jun96)	6.88%	10.41%	5.11%	24.86%	6.89%	8.89%	5.90%	30.18%	2.88%	3.29%	24.00%
Alpha statistic (raw)	0.992	1.189	0.898	-	0.992	1.189	0.898	-	-	-	-
asym-se	(0.138)	(0.183)	(0.183)	-	(0.138)	(0.183)	(0.183)	-	-	-	-
b-t-pval	{0.824}	{0.362}	{0.864}	-	{0.824}	{0.362}	{0.864}	-	-	-	-
#factors, GR test (original, recursive)	2, 3	2, 3	2, 3	2, 3	2, 3	2, 3	2, 3	2, 3	-	-	-
Number of treated states	31	10	21	1	31	10	21	1	31	31	31

[a] The sample period is from Oct86 to Jun96. There are 20 control states. Policy intervention dummy: =1 if a work requirement or time limit waiver is approved/implemented in the state, =0 otherwise. Max monthly welfare ben.: maximum combined real AFDC/Food Stamp benefits for a family of three (in $100). Except the 2wfe specifications, covariates include calendar quarter dummies. GR test: growth-ratio test in Ahn and Horenstein (2013); see Appendix 8 for details.

[b] Standard errors from the asymptotic formula (asym-se) and bootstrapping the coefficient (b-se) are reported in parentheses. P-values from bootstrapping the t-statistic (b-t-pval) are reported in curly brackets. See Appendix 6 for details of bootstrap sample construction. Bootstrap repetitions=199.

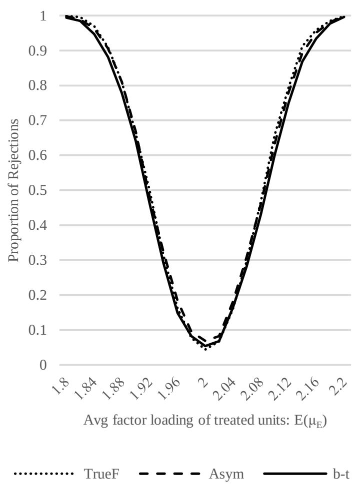
[c] Wild cluster bootstrapped standard errors are in parentheses. See Appendix 6 for details. Bootstrap repetitions=199.

[d] For ITET, the asymptotic standard error is obtained from the Newey-West HAC estimator with $T^{1/4}$ ~= 3 lags.

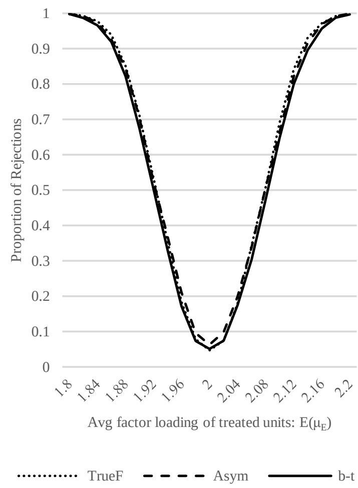
Appendix Table A1: PCDID-MG estimator, sensitivity to number of factors and covariates[a]

		Decide #PC by GR test					Decide #PC by GR test (conservative)					Use 5PCs		Exogenous covariate		Endogenous covariate	
N_E+N_C	T	Bias	SD	#PC mean	#PC sd	#PC max	Bias	SD	#PC mean	#PC sd	#PC max	Bias	SD	Bias	SD	Bias	SD
Panel A: Stationary factors																	
100	10	-0.01	0.19	*1.6*	*0.7*	*3*	-0.01	0.16	*2.7*	*0.5*	*3*	-0.01	0.16	-0.01	0.16	-0.01	0.16
100	20	0.00	0.18	*1.9*	*0.9*	*3*	0.00	0.15	*2.8*	*0.4*	*3*	0.00	0.15	0.00	0.15	0.00	0.15
100	50	0.00	0.16	*2.3*	*0.9*	*3*	0.00	0.15	*3.0*	*0.1*	*3*	0.00	0.15	0.00	0.15	0.00	0.15
10	100	0.01	0.46	*1.1*	*0.4*	*3*	0.01	0.45	*2.9*	*0.3*	*3*	0.01	0.45	0.01	0.45	0.01	0.45
20	100	0.01	0.34	*1.2*	*0.6*	*3*	0.02	0.32	*2.9*	*0.3*	*3*	0.02	0.32	0.02	0.32	0.02	0.32
50	100	0.00	0.22	*1.9*	*1.0*	*3*	0.00	0.21	*3.0*	*0.1*	*3*	0.00	0.21	0.00	0.21	0.00	0.21
10	10	0.02	0.49	*1.3*	*0.5*	*3*	0.02	0.48	*2.8*	*0.4*	*3*	0.01	0.50	0.02	0.50	0.03	0.50
20	20	0.00	0.37	*1.4*	*0.7*	*6*	0.00	0.35	*2.8*	*0.4*	*6*	0.01	0.35	0.01	0.35	0.00	0.35
50	50	0.01	0.22	*1.8*	*1.0*	*3*	0.01	0.20	*2.9*	*0.2*	*3*	0.01	0.20	0.01	0.19	0.01	0.20
100	100	0.01	0.14	*2.7*	*0.7*	*3*	0.01	0.14	*3.0*	*0.0*	*3*	0.01	0.14	0.01	0.14	0.01	0.14
Panel B: Stationary factors with break																	
100	10	0.01	0.20	*1.9*	*0.8*	*3*	-0.01	0.18	*2.6*	*0.5*	*3*	-0.01	0.18	-0.01	0.21	-0.01	0.22
100	20	0.02	0.21	*2.0*	*0.9*	*3*	0.01	0.17	*2.8*	*0.4*	*3*	0.00	0.16	0.00	0.19	0.00	0.21
100	50	0.02	0.20	*2.2*	*1.0*	*3*	0.01	0.15	*3.0*	*0.2*	*3*	0.00	0.15	0.00	0.15	0.01	0.16
10	100	0.04	0.71	*1.1*	*0.2*	*2*	0.07	0.52	*2.9*	*0.2*	*3*	0.06	0.51	0.10	0.53	0.12	0.55
20	100	-0.02	0.51	*1.2*	*0.5*	*3*	0.04	0.34	*2.9*	*0.3*	*3*	0.04	0.33	0.05	0.34	0.06	0.35
50	100	-0.03	0.32	*1.7*	*0.9*	*3*	0.00	0.21	*3.0*	*0.1*	*3*	0.00	0.21	0.00	0.21	0.01	0.22
10	10	0.12	0.55	*1.4*	*0.6*	*3*	0.06	0.51	*2.7*	*0.5*	*3*	0.05	0.56	0.16	0.63	0.17	0.64
20	20	0.06	0.43	*1.5*	*0.7*	*3*	0.02	0.37	*2.7*	*0.4*	*3*	0.01	0.37	0.04	0.40	0.05	0.41
50	50	0.02	0.29	*1.8*	*0.9*	*3*	0.02	0.21	*2.9*	*0.3*	*3*	0.02	0.20	0.02	0.21	0.02	0.22
100	100	0.00	0.18	*2.5*	*0.9*	*3*	0.01	0.14	*3.0*	*0.1*	*3*	0.01	0.14	0.01	0.14	0.01	0.15
Panel C: Nonstationary factors with drift																	
100	10	-0.04	0.18	*1.8*	*0.7*	*3*	-0.01	0.16	*2.5*	*0.5*	*3*	-0.01	0.16	-0.01	0.17	-0.02	0.17
100	20	-0.01	0.17	*2.3*	*0.7*	*3*	0.00	0.16	*2.6*	*0.5*	*3*	0.00	0.15	0.00	0.16	0.00	0.17
100	50	-0.01	0.16	*2.7*	*0.5*	*3*	-0.01	0.16	*2.7*	*0.5*	*3*	0.00	0.15	0.00	0.15	0.00	0.17
10	100	-0.13	0.69	*1.8*	*0.7*	*3*	-0.02	0.50	*2.5*	*0.5*	*3*	-0.01	0.48	-0.02	0.51	-0.01	0.55
20	100	-0.01	0.38	*2.4*	*0.7*	*3*	0.01	0.34	*2.6*	*0.5*	*3*	0.01	0.32	0.01	0.34	0.02	0.38
50	100	0.00	0.23	*2.7*	*0.5*	*3*	0.00	0.22	*2.8*	*0.4*	*3*	0.00	0.21	0.00	0.21	0.01	0.25
10	10	-0.04	0.50	*1.4*	*0.6*	*3*	-0.01	0.49	*2.6*	*0.5*	*3*	0.00	0.51	-0.01	0.52	-0.01	0.52
20	20	-0.04	0.38	*1.8*	*0.7*	*3*	-0.01	0.35	*2.5*	*0.5*	*3*	0.00	0.35	-0.01	0.36	-0.01	0.36
50	50	0.00	0.22	*2.5*	*0.6*	*3*	0.01	0.21	*2.6*	*0.5*	*3*	0.01	0.20	0.01	0.21	0.01	0.22
100	100	0.01	0.15	*2.9*	*0.4*	*3*	0.01	0.15	*2.9*	*0.3*	*3*	0.01	0.14	0.01	0.14	0.01	0.18

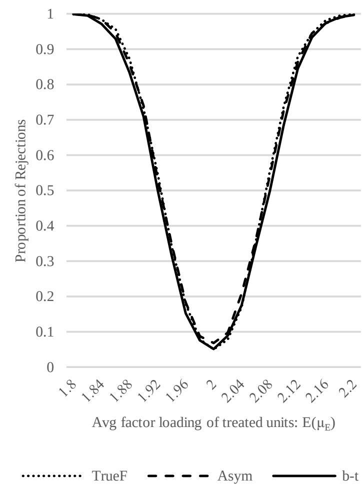
[a] See Section 6.1 for details of DGPs and estimator. All DGPs are the same as in Table II (ATET), except that in "exogenous covariate" and "endogenous covariate", ς_i, β_i, and x_{it} follow different DGPs. Exogenous covariate: x is not correlated with the factor structure. Endogenous covariate: x is correlated with the factor structure. Decide #PC by GR test: use the growth ratio test in Ahn and Horestein (2013) to determine the number of factor proxies. Decide #PC by GR test (conservative): if the GR test yield 1 factor, adjust it to kmax/2 = 3 factors. Use 5 PCs: use 5 factor proxies. #PC mean: average number of factor proxies. #PC sd: empircal standard deviation of the number of factor proxies. #PC max: maximum number of factor proxies. Number of replications=1000. In all specifications, $N_E=N_c=N/2$ and $T_0=T_1=T/2$.



(a) Stationary factor.



(b) Stationary factor with break.



(c) Non-stationary factor with drift.

Appendix Figure A1. -- power of parallel trend Alpha test.

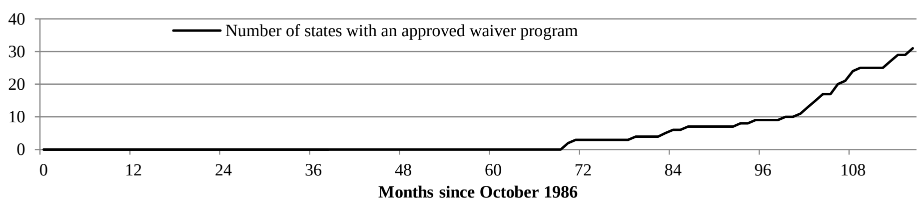
Note: See Section 6.2 for details of DGPs and inference procedures. TrueF: assume the factor is observed in PCDID estimation (infeasible) and compute t-statistic, reject if $|t|>=1.96$. Asym: use a factor proxy in PCDID estimation and compute t-statistic, reject if $|t|>=1.96$. b-t: same as Asym, but reject if $t<=c_{0.025}$ or $t>=c_{0.975}$ where $c_{0.025}$, $c_{0.975}$ are percentiles of the bootstrap distribution of t-statistics. See Section 6.2 for analytical standard errrors used in TrueF, Asym and b-t. See Appendix 6 for details on bootstrap sample construction. Number of replications=1000. Bootstrap repetitions = 199. Average factor loading of control units is set to be $E(\mu_C)=2$ in all scenarios. Average factor loading of treated units is set between 1.8 and 2.2. The DGPs are otherwise the same as in Table V. $T_0=T_1=25$, $N_C=50$, $N_E=10$.

Appendix Table A2: Characteristics of control and treated states, Oct86[a]

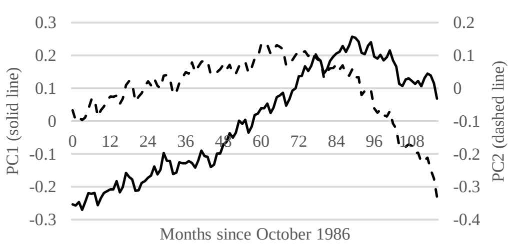
	Control states (N=20)		ALL Treated states (N=31)	
Characteristics	Mean	SD	Mean	SD
Population ('000)	4259.37	4530.29	5067.02	5651.37
Welfare caseload ('000)	63.53	83.04	78.77	113.40
Ln(caseload per capita)	-4.32	0.42	-4.39	0.37
Max combined real AFDC/Food Stamps benefits for family of 3 (in $100)	4.67	1.47	5.00	0.72
Unemployment rate (%)	6.93	2.00	6.18	2.35
Employment to population ratio (%)	45.61	3.19	46.63	3.49

Among treated states:	10 Southern treated states		21 Nonsouthern treated states		Wyoming (1 state)
Characteristics	Mean	SD	Mean	SD	Mean
Population ('000)	5365.22	4386.88	4925.02	6259.49	476.9
Welfare caseload ('000)	63.15	38.59	86.21	135.77	4.16
Ln(caseload per capita)	-4.37	0.28	-4.41	0.41	-4.74
Max combined real AFDC/Food Stamps benefits for family of 3 (in $100)	5.12	0.48	4.93	0.81	4.88
Unemployment rate (%)	7.16	3.00	5.72	1.88	8.80
Employment to population ratio (%)	44.72	4.12	47.55	2.81	47.95

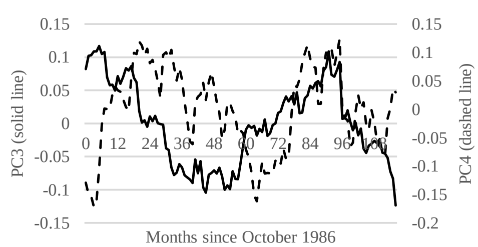
[a] Control states consist of Alabama*, Alaska, Arkansas*, Colorado, DC*, Florida*, Idaho, Kansas, Kentucky*, Maine, Maryland*, Minnesota, Nevada, New Hampshire, New Mexico, New York, Pennsylvania, Rhode Island, South Carolina*, Tennessee*. States in the south are noted by an asterisk.



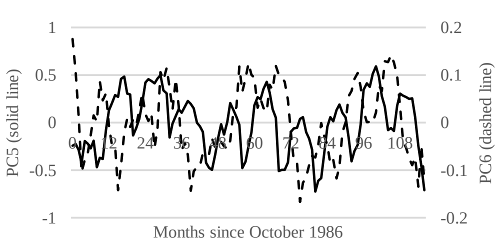
Appendix Figure A2. -- Number of states with an approved waiver program.



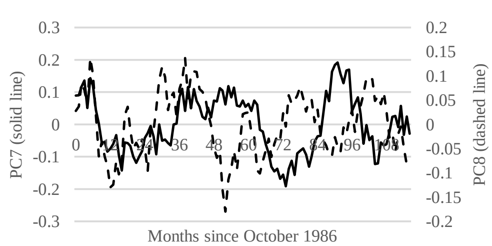
(a) PC1 and PC2.



(b) PC3 and PC4.



(c) PC5 and PC6.



(d) PC7 and PC8.

Appendix Figure A3. -- Factor proxies from the control panel (number of observations = 2340).

Supplementary Materials

The PCDID Approach: Difference-in-Differences when Trends are Potentially Unparallel and Stochastic

Contents

1 Main Text Results

1.1 Preliminaries

Key notations:

- Treatment effect (TE): Δ_{it}.
- ITET for treated unit i: $\bar{\Delta}_i := E(\Delta_{it}|t > T_{0i})$.
- ATET: $\bar{\bar{\Delta}} := E(\bar{\Delta}_i|i \in E)$.
- Idiosyncratic part of TE: $\tilde{\Delta}_{it} := \Delta_{it} - \bar{\Delta}_i = \Delta_{it} - \bar{\bar{\Delta}} - \upsilon_i$ (where $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}}$).
- Vector of intervention dummies for treated unit i: $1_{post,i} = [0, \dots, 0, 1, \dots, 1]'$ (T_{0i} zeros followed by T_{1i} ones).
- Vector of composite idiosyncratic errors: $\epsilon_i = \tilde{\epsilon}_i + \tilde{\Delta}_i 1_{\{i \in E\}} 1_{post,i}$.
- Variable sets: $G_i = \{F, X_i\}$, $\tilde{G}_i = \{1_{post,i}, G_i\}$, $\hat{G}_i = \{\hat{F}, X_i\}$, $\tilde{G} := \{\tilde{G}_i\}_{i \in E}$.
- Mean-squared residuals of regressing $1_{post,i}$ on 1, F and X_i: $R_{Ti} := \frac{1}{T} 1'_{post,i} M_{G_i} 1_{post,i} = \frac{1}{T}[1'_{post,i} 1_{post,i} - 1'_{post,i} G_i (G_i' G_i)^{-1} G_i' 1_{post,i}]$.
- Set of all mean-squared residuals in treated group: $MSR_T := \{R_{Ti}\}_{i \in E}$.
- Residual vector from linear projection of y_i on X_i: $\hat{u}_i := M_{X_i} y_i$.
- Sample mean of $\hat{u}_i$ over control group: $\bar{u}_C := \frac{1}{N_C} \sum_{i \in C} \hat{u}_i$.

Table of estimators/statistics and their conditional variance:

	Estimator/statistic	Conditional variance
PCDID	$\hat{\delta}_i := \frac{1'_{post,i} M_{\hat{G}_i} y_i}{1'_{post,i} M_{\hat{G}_i} 1_{post,i}}$	$\sigma^2_{Ti} := Var[\sqrt{T}(\hat{\delta}_i - \bar{\Delta}_i)\|\tilde{G}_i]$
mean gp	$\hat{\delta}^{mg} := \frac{1}{N_E}\sum_{i\in E}\hat{\delta}_i$	$\bar{\varsigma}^2_{N_E,T} := Var[\sqrt{N_E T}(\hat{\delta}^{mg} - \bar{\Delta})\|\tilde{G}]$ (homo ITET) $\bar{\bar{\varsigma}}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})]$ (hetero ITET)
pooled	$\hat{\delta}^{pl} := \frac{\sum_{i\in E} 1'_{post,i} M_{\hat{G}_i} y_i}{\sum_{i\in E} 1'_{post,i} M_{\hat{G}_i} 1_{post,i}}$	$\bar{\sigma}^2_{N_E,T} := Var[\sqrt{N_E T}(\hat{\delta}^{pl} - \bar{\Delta})\|\tilde{G}]$ (homo ITET) $\bar{\bar{\sigma}}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})\|MSR_T]$ (hetero ITET)
Alpha test	$\hat{a}^{mg} := \frac{1}{N_E}\sum_{i\in E}\hat{a}_i$, $\hat{a}_i = \frac{\bar{u}'_C M_{[1_{post,i},X_i]} y_i}{\bar{u}'_C M_{[1_{post,i},X_i]}\bar{u}_C}$	$\bar{\varphi}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{a}^{mg} - 1)\|\tilde{G}]$

Here are a few notations on the characterization of probabilistic statements and stochastic order. The phrase "$a.s.$" stands for "almost surely" or "with probability one," while "$w.p.a.1$" stands for "with probability approaching one."

For sequences of random variables U_n and V_n, we say that $U_n = O_p(V_n)$ as $n \to \infty$ if, for any $\varepsilon > 0$, there exists $C = C(\varepsilon) < \infty$ and $N = N(\varepsilon) > 0$ such that for all $n > N$, $P(|U_n/V_n| < C) > 1 - \varepsilon$. This is sometimes expressed in short form as $|U_n/V_n| \le C \quad w.p.a.1$ as $n \to \infty$ for some $C < \infty$, or in words as "U_n/V_n is bounded in probability". We say that $U_n = o_p(V_n)$ as $n \to \infty$ if, for any $\varepsilon > 0$ and $\delta > 0$, there exists $N = N(\varepsilon, \delta) > 0$ such that for all $n > N$, $P(|U_n/V_n| < \delta) > 1 - \varepsilon$. In words, this is the same as saying "U_n/V_n converges in probability to zero." The big-O and small-o notations are sometimes applied to represent the relationship between vector/matrix sequences, with the understanding that the relationship is valid elementwise.

As a convention on matrices and vectors, the norm of a matrix/vector A is given by $\|A\| = [trace(A'A)]^{1/2}$. For any matrix/vector A of full column rank, we denote the projection matrix as $P_A = A(A'A)^{-1}A'$ and the annihilator matrix as $M_A = I - P_A$.

1.2 Assumptions

1.2.1 Exogeneity, Treatment and Intervention Dummies

Assumption E (predeterminedness, treatment and intervention dummies):
(i) $E(\epsilon_{it}|\bar{\Delta}_i, 1_{\{i\in E\}}, 1_{\{t>T_{0i}\}}, \beta_i, \varsigma_i, \mu_i, f_t, x_{it}) = 0$ for each i and t.
(ii) $0 < E(1_{\{i\in E\}}) < 1$ for each i.
(iii) for each $i \in E$, $T_{1i}/T \xrightarrow{p} \kappa_i$ as $T, T_{1i} \to \infty$, where $0 < \kappa_i < 1$.

Assumption ES (strict exogeneity in time): For each i, $E(\epsilon_{it}|\tilde{G}_i) = 0$ $a.s.$ for all t.

Assumption ESS (strict exogeneity in panel): $E(\epsilon_{it}|\tilde{G}) = 0$ $a.s.$ for all $i \in E$ and t.

Remark: Assumption ES strengthens Assumption E(i) to a strict exogeneity condition on the time series of factors, intervention status and covariates. This is crucial for the conditioning argument that leads to the asymptotic normality of PCDID estimator. Assumption ESS requires that the regressors are strictly exogeneous in the panel setting. This is necessary for deriving the asymptotic normality of the pooled and mean-group estimators. Assumption ESS implies Assumption ES by the law of iterated expectations.

1.2.2 Factors and Factor Loadings

Denote $F := [f_1, \ldots, f_T]'$ the $T \times \ell$ matrix of factors with normalization orders $r_1, \ldots, r_\ell \ge 0.5$. For unit i, let $X_i := \begin{bmatrix} 1 & ,\ldots, & 1 \\ x_{i1} & ,\ldots, & x_{iT} \end{bmatrix}'$ be the $T \times k$ matrix of stationary covariates, and let $G_i := (F, X_i)$, a $T \times (\ell + k)$ matrix. Define $\mu_E := [\mu_1, \ldots, \mu_{N_E}]'$ and $\mu_C := [\mu_{N_E+1}, \ldots, \mu_N]'$, the $N_E \times \ell$ and $N_C \times \ell$ matrices of factor loadings for the treated and control groups, respectively. Define $\Upsilon := diag(T^{r_1}, \ldots, T^{r_\ell})$, an $\ell \times \ell$ diagonal matrix for normalizing the factors, and $\Gamma :=$

$diag(T^{r_1}, \dots, T^{r_\ell}, T^{0.5}, \dots, T^{0.5})$, an $(\ell+k)\times(\ell+k)$ diagonal matrix for normalizing the factors and covariates.

Assumption F (factors and covariates): Let $\Gamma := diag(T^{r_1}, \dots, T^{r_\ell}, T^{0.5}, \dots, T^{0.5})$ where $r_1, \dots, r_\ell \geq 0.5$. For each i, the following conditions are satisfied:
(i) For all T, $E\left\|\Gamma^{-1}G_i'G_i\Gamma^{-1}\right\|^2 \leq c$ for some constant $c > 0$.
(ii) $\text{plim}_{T\to\infty}\,\Gamma^{-1}G_i'G_i\Gamma^{-1}$ is positive definite *a.s.*.

Assumption FL$_C$ (factor loadings of control units):
(i) For all $i \in C$, $E\left\|\mu_i\right\|^2 \leq c$ for some constant $c > 0$.
(ii) $\text{plim}_{N_C\to\infty}\,\frac{1}{N_C}\mu_C'\mu_C$ is positive definite.

Assumption FL (factor loadings): Assumption FL$_C$ holds. In addition,
(i) For all $i \in E$, $E\left\|\mu_i\right\|^2 \leq c$ for some constant $c > 0$.
(ii) $\text{plim}_{N_E\to\infty}\,\frac{1}{N_E}\mu_E'\mu_E$ is positive definite.

Assumption FLM (mixing factor loadings): The following conditions are satisfied:
(i) For some $p > 1$, there exists $0 < c < \infty$ such that $E(\left\|\mu_i\right\|^p) \leq c$ for all $i \in C \cup E$.
(ii) $\{\mu_i : i \in C\}$ and $\{\mu_i : i \in E\}$ are mixing sequences with mixing coefficients ϕ of size $-p/(2p-1)$ for $p \geq 1$, or α of size $-p/(p-1)$ for $p > 1$.

Assumption FLM2 (conditional mixing factor loadings, treated units): The following conditions are satisfied:
(ii) For each $i \in E$, $E(\mu_i|\tilde{G}) = \mu_0$ *a.s.*.
(ii) For some $p > 2$, there exists $0 < c < \infty$ such that $E(\left\|\mu_i\right\|^p|\tilde{G}) \leq c$ for all $i \in E$.
(iii) Conditional on $\tilde{G}$, $\{\mu_i : i \in E\}$ is a mixing sequence with mixing coefficients ϕ of size $-p/2(p-1)$ for $p \geq 2$, or α of size $-p/(p-2)$ for $p > 2$.

Remark: Assumptions F, FL$_C$ and FL are customized versions of those in the standard interactive-effects model literature (e.g., Assumptions A and B of Bai (2003), Assumption B of Bai (2009)). We customize the assumptions to allow for a broader range of factor dynamics (Assumption F) and to accommodate the specific data feature in program evaluation (Assumptions FL$_C$ and FL).

Remark: The weak form of parallel trend hypothesis allows for heterogeneous factor loadings. To obtain consistency of the Alpha statistic under the weak parallel trend hypothesis, Assumption FLM is required to control the cross-sectional dependence of factor loadings over treated and control units. Coupled with the conditional mixing and moment conditions on the factor loadings of treated units in Assumption FLM2, we achieve asymptotic normality.

1.2.3 Asymptotic Identification

Assumption AI$_i$ (asymptotic identification, PCDID estimator): For each $i \in E$,
(i) $\rho_i := \text{plim}_{T\to\infty}\,\frac{1}{T}1_{post,i}'M_{G_i}1_{post,i}$ exists and is strictly positive *a.s.*.
(ii) $\xi_i^2 := \text{plim}_{T\to\infty}\,E\left(\left\|\frac{1}{\sqrt{T}}1_{post,i}'M_{G_i}\epsilon_i\right\|^2\middle|\tilde{G}_i\right)$ exists and is strictly positive *a.s.*.

Assumption AI$_{mg}$ (asymptotic identification, simple mean-group estimator):
(i) $\text{plim}_{N_E,T\to\infty}\inf_{i\in E}\frac{1}{T}1_{post,i}'M_{G_i}1_{post,i}$ exists and is strictly positive *a.s.*.
(ii) $\zeta^2 := \text{plim}_{N_E,T\to\infty}\,E\left(\left\|\frac{1}{\sqrt{N_E T}}\sum_{i\in E}1_{post,i}'M_{G_i}\epsilon_i\right\|^2\middle|\tilde{G}\right)$ exists and is strictly positive *a.s.*.

Assumption AI$_{pl}$ (asymptotic identification, pooled estimator): Same as Assumption AI$_{mg}$, except replacing AI$_{mg}$(i) by:

(i) $\rho := \text{plim}_{N_E,T\to\infty} \frac{1}{N_E T} \sum_{i\in E} 1'_{post,i} M_{G_i} 1_{post,i}$ exists and is strictly positive $a.s.$.

Assumption AI$_\alpha$ (asymptotic identification, Alpha test): Let r be the normalization order of $F\mu_0$ such that $\|F\mu_0\|^2 / T^{2r} = O_p(1)$ as $T \to \infty$. The following conditions hold:
(i) $\text{plim}_{N_E,T\to\infty} \inf_{i\in E} \frac{1}{T^{2r}} \mu_0' F' M_{[1_{post,i},X_i]} F\mu_0$ exists and is strictly positive $a.s.$.
(ii) $\text{plim}_{N_E,T\to\infty} E\left(\left\| \frac{1}{\sqrt{N_E}T^r} \sum_{i\in E} \mu_0' F' M_{[1_{post,i},X_i]} F(\mu_i - \mu_0) \right\|^2 \middle| \tilde{G} \right)$ exists and is strictly positive $a.s.$.

Remark: It is possible to asymptotically identify ATET with the pooled estimator even if Assumption AI$_i$(i) breaks down for a finite number of treated units, as long as Assumption AI$_{pl}$(i) holds. By contrast, asymptotically identifying ATET with the mean-group estimator requires that *every* treated unit is well identified as the treatment panel grows, i.e., $\rho_i > 0$ uniformly over all $i \in E$ as $N_E \to \infty$ (Assumption AI$_{mg}$(i)). This is crucial for a well-defined mean-group estimator, defined as the average of PCDID estimators over all $i \in E$.

1.2.4 Heterogeneous Treatment Effects

Assumption RT$_{mg}$ (treatment effects, simple mean-group estimator): Let $\tilde{X}_i$ be the matrix of the covariates (excluding constant and deterministic ones) stacked over time for unit i, and $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}}$. The following conditions are satisfied:
(i) For some $p > 2$, there exists $0 < c < \infty$ such that $E|\upsilon_i|^p \le c$ for all $i \in E$.
(ii) υ_i is a mixing process with mixing coefficient ϕ of size $-p/2(p-1)$ for $p \ge 2$, or α of size $-p/(p-2)$, $p > 2$.
(iii) $\lim_{N_E\to\infty} Var(N_E^{-1/2} \sum_{i\in E} \upsilon_i)$ exists and is strictly positive.

Assumption RT$_{pl}$ (treatment effects, pooled estimator): Let $R_{Ti} := \frac{1}{T} 1'_{post,i} M_{G_i} 1_{post,i}$, $MSR_T := \{R_{Ti}\}_{i\in E}$ and $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}}$. The following conditions are satisfied:
(i) For each $i \in E$, $E(\upsilon_i | MSR_T) = 0$ $a.s.$.
(ii) For some $p > 2$, there exists $0 < c < \infty$ such that $E(|\upsilon_i|^p | MSR_T) \le c$ for all $i \in E$ and T.
(iii) Conditional on MSR_T, υ_i is a mixing process with mixing coefficient ϕ of size $-p/2(p-1)$ for $p \ge 2$, or α of size $-p/(p-2)$, $p > 2$.
(iv) $\text{plim}_{N_E,T\to\infty} Var(N_E^{-1/2} \sum_{i\in E} R_{Ti}\upsilon_i | MSR_T)$ exists and is strictly positive $a.s.$.

Remark: To obtain asymptotic results in a heterogeneous treatment effect environment, we impose Assumptions RT$_{pl}$ and RT$_{mg}$ respectively for the pooled and mean-group estimators. Both of them exert control on the cross-section mean, variation and higher-order moments of the treatment effects over the treated; nevertheless, to obtain desirable asymptotic properties of the pooled estimator we need additional restrictions. Importantly, to ensure consistency of the pooled estimator, it is required that the heterogeneous treatment effect is uncorrelated with the treatment status after partialling out the factors and covariates (Assumption RT$_{pl}$(i)). The mean-group estimator does not require such assumption for consistency. For consistency of the mean-group estimator it is necessary that $E(\upsilon_i | i \in E) = 0$, but this is automatically satisfied by construction.

1.2.5 Weak Dependence of Idiosyncratic Errors

Define $\sigma_{ij,st} := E(\epsilon_{is}\epsilon_{jt})$ and $\gamma_T(i,j) := E\left[T^{-1} \sum_{t=1}^T \epsilon_{it}\epsilon_{jt}\right]$.
Assumption IE (idiosyncratic errors): There exists a positive constant $0 < c < \infty$ such that for all N_E, N_C and T:
(i) $E|\epsilon_{it}|^8 \le c$ for each i and t.
(ii) $|\gamma_T(i,i)| \le c$ and $\sum_{j\in C} |\gamma_T(i,j)| \le c$ for each i.
(iii) $|\sigma_{ii,st}| \le |\sigma_{st}|$ for some σ_{st} and for all i. In addition, $\frac{1}{T} \sum_{s,t=1}^T |\sigma_{st}| \le c$.

(iv) $\frac{1}{N_C T}\sum_{i,j\in C}\sum_{s,t=1}^{T}|\sigma_{ij,st}| \leq c$ and $\frac{1}{N_E T}\sum_{i,j\in E}\sum_{s,t=1}^{T}|\sigma_{ij,st}| \leq c$.
(v) $E\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\epsilon_{it}\epsilon_{jt} - E(\epsilon_{it}\epsilon_{jt})\right]\right|^4 \leq c$ for each (i,j) pair.

Remark: Note that Assumption E(i) implies $E(\epsilon_{it}) = 0$. The conditions in Assumption IE are no stronger than those in Bai and Ng (2002) and Bai (2003). They allow for heteroskedastic and weakly dependent idiosyncratic errors, both over cross-section and over time. Assumption IE(i) is imposed to ensure the existence of those moments of ϵ_{it} appearing in the proofs of all lemmas and theorems (including those in the rest of Assumption IE). Except for Assumption IE(ii) which is a combination of Assumptions C(2) and E(1) of Bai (2003), all the rest are obtained from Bai and Ng (2002) and Bai (2003) by interchanging the subscripts for i and t.

1.2.6 Moment Conditions

Assumption MX (moments on covariates): There exists $0 < c < \infty$ such that the following moment conditions are satisfied for all N_E, N_C and T:
(i) $E\sup_i\left\|\frac{X_i'X_i}{T}\right\| \leq c$.
(ii) $\inf_i \inf_{v\in\mathbb{R}^k\setminus\{0\}} v'\left(\frac{X_i'X_i}{T}\right)v > 0$.
(iii) $E\sup_i\left\|X_i'F\Upsilon^{-1}\right\|^2 \leq c$.
(iv) $E\sup_{i,j}\left\|\frac{1}{\sqrt{T}}X_i'\epsilon_j\right\|^2 \leq c$.
(v) $\frac{1}{TN_C}\sum_{i\in C}\sum_{s=1}^{T}\sum_{t=1}^{T}\|Cov(x_{is}\epsilon_{is}, x_{it}\epsilon_{it})\| \leq c$.
(vi) $E\left\|\frac{1}{\sqrt{N_C T}}\sum_{i\in C}\sum_{t=1}^{T}x_{it}\epsilon_{it}\mu_i\right\|^2 \leq c$.
(vii) $E\sup_i\left\|\frac{1}{\sqrt{N_C T}}\sum_{j\in C}X_i'\epsilon_j\right\|^2 \leq c$ and $E\sup_i\left\|\frac{1}{\sqrt{N_E T}}\sum_{j\in E}X_i'\epsilon_j\right\|^2 \leq c$.

Remark: MX(i) states that the second (self- and cross-) moments of covariates are uniformly bounded over all i. In particular, this implies that all covariates are temporally stationary. MX(ii) means that the matrices $X_i'X_i$ are positive definite uniformly over all i. MX(iii) states that x_{it} and f_t are orthogonal in the limit. When all factors are I(0), the orthogonality condition implies that all elements of $T^{-1}\sum_{t=1}^{T}x_{it}f_t'$ are $O_p(T^{-0.5}) = o_p(1)$ uniformly over all i. This assumption is imposed for analytical convenience when deriving some of the theoretical results. MX(iv) and (vii) are standard assumptions that guarantee the existence of second moments of $x_{it}\epsilon_{jt}$ while allowing for weak dependence across time and units. MX(v) is the weak dependence assumption on $x_{it}\epsilon_{it}$.

Assumption M (moments): Assumption MX holds. In addition, there exists $0 < c < \infty$ such that the following conditions are satisfied for all N_C and T:
(i) $E\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{\sqrt{N_C}}\sum_{i\in C}\mu_i\epsilon_{it}\right\|^2\right) \leq c$.
(ii) $E\sup_i\left\|\Upsilon^{-1}\sum_{t=1}^{T}f_t\epsilon_{it}\right\|^2 \leq c$.
(iii) $E\left\|\frac{1}{\sqrt{N_C}}\Upsilon^{-1}\sum_{t=1}^{T}\sum_{i\in C}\epsilon_{it}\mu_i f_t'\right\|^2 \leq c$.
(iv) $E\sup_{j\in E}\left\|\frac{1}{\sqrt{N_C T}}\sum_{t=1}^{T}\sum_{i\in C}\mu_i[\epsilon_{it}\epsilon_{jt} - E(\epsilon_{it}\epsilon_{jt})]\right\|^2 \leq c$.

Remark: Assumptions M(i)-(iv) are the same as Assumptions D and F(1)-(3) in Bai (2003) after interchanging the roles of factors and factor loadings. In particular, Assumption M(i) is used in the proof of lemmas A1, A2 and A3(a). Assumptions M(ii)-(iv) are required for obtaining the consistency of PCDID estimator in Theorem 1. Assumption M(iii) is a moment condition analogous to Assumption MX(vi) with factors in place of covariates.

Assumption MM (moments): Assumption M holds. In addition, there exists $0 < c < \infty$ such that the following conditions are satisfied for all N_E, N_C and T:
(i) $\frac{1}{N_C}\sum_{i,j\in C}|\gamma_T(i,j)| \leq c$ and $\frac{1}{\sqrt{N_C N_E}}\sum_{i\in C}\sum_{j\in E}|\gamma_T(i,j)| \leq c$.
(ii) $E\left\|\Upsilon^{-1}\frac{1}{\sqrt{N_E}}\sum_{j\in E}\sum_{t=1}^T f_t\epsilon_{jt}\right\|^2 \leq c$ and $E\left\|\Upsilon^{-1}\frac{1}{\sqrt{N_C}}\sum_{i\in C}\sum_{t=1}^T f_t\epsilon_{it}\right\|^2 \leq c$.
(iii) $E\left\|\frac{1}{\sqrt{N_E N_C T}}\sum_{t=1}^T\sum_{j\in E}\sum_{i\in C}\mu_i[\epsilon_{it}\epsilon_{jt} - E(\epsilon_{it}\epsilon_{jt})]\right\|^2 \leq c$.

Remark: Assumption MM lists the conditions necessary for the consistency of pooled and mean-group estimators and alpha test. Assumption MM(i) is the customized version of Bai (2003)'s Assumption C(2), and is weaker than our Assumption IE(ii). Assumptions MM(ii) and (iii) generalize Assumptions M(ii) and (iv) to the large panel setting.

1.2.7 Time Series and Cross-Sectional Dependence

Denote $\|v\|_r = [E(\sum_{j=1}^k |v_j|^r)]^{1/r}$ be the r-th norm of a k-dimensional vector v. We say that a vector sequence v_t is L_2-*NED* (near-epoch dependent) of size $-p$ on another process h_t with respect to scaling constants π_t if $\left\|v_t - E(v_t|\mathcal{F}_{t-m}^{t+m})\right\|_2 \leq \pi_t a_m$, where $\mathcal{F}_s^t = \sigma(h_s,\ldots,h_t)$ for $s \leq t$, π_t is positive for all t, and $a_m = O(m^{-p'})$ for some $p' > p$ as $m \to \infty$ (e.g., Definition 1 of DeJong and Davidson (2000)). The definitions of ϕ- and α-mixing processes are standard in the literature (e.g., Section 3.4 of White (2001)).

Assumption D$_i$ (time series dependence): Denote $u_t = (u_{1t},\ldots,u_{\ell t})'$. Let $\tilde{x}_{it}$ be the vector of covariates excluding the constant intercept and deterministic variables.
Each factor f_{jt} ($j = 1,\ldots,\ell$) takes any one of the following forms:
(1) a deterministic trend with maximal polynomial order $d_j \geq 0$,
(2) an L_2-*NED* process $f_{jt} = u_{jt}$, or
(3) an integrated process with integer integration order $d_j = 1, 2\ldots$ and representation given by $f_{jt} = \sum_{s_{d_j-1}=1}^t \sum_{s_{d_j-2}=1}^{s_{d_j-1}} \cdots \sum_{s_1=1}^{s_2}\sum_{s=1}^{s_1} u_{js}$.

For each $i \in E$, the vector sequence $v_{it} = (1_{\{t>T_{0i}\}}, \tilde{x}_{it}', u_t', \epsilon_{it})'$ satisfies all of the following conditions:
(i) v_{it} is L_2-*NED* of size $-\frac{1}{2}$ on a process z_{it} with respect to π_{it}, where z_{it} is ϕ-mixing of size $-\frac{p}{2(p-1)}$ for $p \geq 2$, or α-mixing of size $-\frac{p}{p-2}$ for $p > 2$, and π_{it} satisfies $0 < \pi_{it} \leq \|v_{it}\|_p$ for all t.
(ii) $\sup_t \|v_{it} - E(v_{it})\|_p < \infty$ for p defined in (i), and if $p = 2$ then each component of v_{it} is uniformly integrable.
(iii) $\lim_{T\to\infty} Var(T^{-1/2}\sum_{t=1}^T v_{it})$ exists and is positive definite.

Assumption D (panel dependence): Assumption D$_i$ holds for all $i \in E$. In addition, the following conditions are satisfied:
(i) ϵ_{it} are independent over $i \in E$.
(ii) $\zeta_i^2 := \lim_{T\to\infty} Var(T^{-1/2}1_{post,i}' M_{G_i}\epsilon_i|\tilde{G})$ satisfies $\frac{\max_{i\in E}\zeta_i^2}{\sum_{i\in E}\zeta_i^2} = O_p\left(\frac{1}{N_E}\right)$ as $N_E \to \infty$.

Remark: To derive the asymptotic normality of PCDID estimator we need to strengthen existing assumptions on idiosyncratic errors. Assumption D$_i$ puts restrictions on the time series dynamics of regressors, factors and idiosyncratic errors by allowing for weak dependence and heteroskedasticity over time. This enables us to apply the FCLT of DeJong and Davidson (2000). While fractionally integrated processes are ruled out in Assumption D$_i$, there exist more general assumptions and FCLT results that accommodate factors with fractional normalization order (e.g., Davidson and DeJong (2000)). To avoid further complicating our analysis we do not consider such level of generality.

Remark: Assumption D is the basis for obtaining the asymptotic distribution of the pooled and mean-group estimators. Assumption D(i) restricts the idiosyncratic errors to be independent over treated units. Assumption D(ii) is an asymptotic negligibility condition on the cross-sectional variation of the treatment panel. While stronger than weak dependence (Assumption IE), the restrictions enable us to invoke the joint CLT of Phillips and Moon (1999) and obtain asymptotic normality. It is possible to relax the cross-sectional independence at the expense of more technical assumptions.

1.2.8 Parallel Trend Hypothesis

Assumption PTW (weak parallel trends): $E(\mu_i|i \in C) = E(\mu_i|i \in E) = \mu_0$ for some finite and non-zero vector μ_0.

1.3 Result on PCDID Estimator

Theorem 1 (PCDID estimator): Suppose Assumptions E, F, FL$_C$, AI$_i$, IE and M hold. Then, as $T, N_C \to \infty$ jointly and $\frac{\sqrt{T}}{N_C} \to 0$, we have for each $i \in E$:

(a) $\hat{\delta}_i \xrightarrow{p} \bar{\Delta}_i$.

(b) $\sqrt{T}\sigma_{Ti}^{-1}(\hat{\delta}_i - \bar{\Delta}_i) \xrightarrow{d} N(0,1)$ if additionally Assumptions ES and D$_i$ hold, where $\sigma_{Ti}^2 := Var[\sqrt{T}(\hat{\delta}_i - \bar{\Delta}_i)|\tilde{G}_i]$.

1.4 Result on Simple Mean-Group Estimator

Theorem 2 (simple mean-group estimator): Suppose Assumptions E, F, FL, AI$_{mg}$, IE and MM hold. As $T, N_E, N_C \to \infty$ jointly and $\frac{T}{N_C} \to 0$, we have the following results:

(a) (homogeneous ITET) Suppose $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}} \equiv 0$. Then,
(i) $\hat{\delta}^{mg} \xrightarrow{p} \bar{\bar{\Delta}}$.
(ii) $\sqrt{N_E T}\varsigma_{N_E,T}^{-1}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$ if additionally Assumptions ESS and D hold, where $\bar{\varsigma}_{N_E,T}^2 := Var[\sqrt{N_E T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})|\tilde{G}]$.

(b) (heterogeneous ITET) Suppose υ_i satisfies Assumption RT$_{mg}$. Then,
(i) $\hat{\delta}^{mg} \xrightarrow{p} \bar{\bar{\Delta}}$.
(ii) $\sqrt{N_E}\bar{\bar{\varsigma}}_{N_E,T}^{-1}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$, where $\bar{\bar{\varsigma}}_{N_E,T}^2 := Var[\sqrt{N_E}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})]$.

1.5 Result on Pooled Estimator

Theorem 3 (pooled estimator): Suppose Assumptions E, F, FL, AI$_{pl}$, IE and MM hold. As $T, N_E, N_C \to \infty$ jointly and $\frac{T}{N_C} \to 0$, we have the following results:

(a) (homogeneous ITET) Suppose $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}} \equiv 0$. Then,
(i) $\hat{\delta}^{pl} \xrightarrow{p} \bar{\bar{\Delta}}$.
(ii) $\sqrt{N_E T}\bar{\sigma}_{N_E,T}^{-1}(\hat{\delta}^{pl} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$ if additionally Assumptions ESS and D hold, where $\bar{\sigma}_{N_E,T}^2 := Var[\sqrt{N_E T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|\tilde{G}]$.

(b) (heterogeneous ITET) Suppose υ_i satisfies Assumption RT$_{pl}$. Then,
(i) $\hat{\delta}^{pl} \xrightarrow{p} \bar{\bar{\Delta}}$.
(ii) $\sqrt{N_E}\bar{\bar{\sigma}}_{N_E,T}^{-1}(\hat{\delta}^{pl} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$, where $\bar{\bar{\sigma}}_{N_E,T}^2 := Var[\sqrt{N_E}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|MSR_T]$.

1.6 Result on Asymptotic Efficiency

Theorem 4 (asymptotic efficiency): Suppose Assumptions E, F, FL, IE, D, AI_{pl}, AI_{mg} and MM hold. For each $i \in E$, define $\rho_i := \text{plim}_{T\to\infty} 1'_{post,i} M_{G_i} 1_{post,i}/T$, where $R_{Tj} := 1'_{post,j} M_{G_j} 1_{post,j}/T$. The following results are valid *a.s.*:

(a) (homogeneous ITET) Suppose ϵ_{it} are *iid* over i and t with mean 0 and variance σ_ϵ^2. Then $\sigma^2 \leq \varsigma^2$ *a.s.*. Equality holds iff ρ_i are identical over $i \in E$.

(b) (heterogeneous ITET) Suppose υ_i are *iid* with mean 0 and variance σ_υ^2, and that υ_i are independent of R_{Tj} for all $i, j \in E$. Then $\tilde{\varsigma}^2 \leq \tilde{\sigma}^2$. Equality holds iff ρ_i are identical over $i \in E$.

1.7 Result on Alpha Test

Theorem 5 (Alpha test): Suppose Assumptions E, F, FLM, AI_α, IE and MM hold. Then, under Assumption PTW, we have the following results as $T, N_E, N_C \to \infty$ jointly and $\frac{T}{N_C} \to 0$:

(a) $\hat{a}^{mg} \xrightarrow{p} 1$.

(b) $\sqrt{N_E}\bar{\varphi}^{-1}_{N_E,T}(\hat{a}^{mg} - 1) \xrightarrow{d} N(0,1)$ if additionally Assumption FLM2 holds, where $\bar{\varphi}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{a}^{mg} - 1)|\tilde{G}]$.

2 Intermediate Results

Let V be the $\ell \times \ell$ eigenvalue matrix V, and W be the associated $N_C \times \ell$ eigenvector matrix for the principal component analysis. The following lemma shows that both V and W have full (column) rank in the limit as the control panel grows in size.

Lemma A1: Suppose Assumptions F, FL_C, IE, M(i) and MX(i)-(iii) hold. Then we have $rank(W) = rank(V) \xrightarrow{p} \ell$ as $N_C, T \to \infty$.

The factor space can be identified up to a rotation. The following lemma constructs a rotation matrix which has full rank for large N_C and T. This construction is necessary for the asymptotic arguments in the main theorems. Define $\Upsilon = diag(T^{r_1}, \ldots, T^{r_\ell})$.

Lemma A2: Suppose Assumptions F, FL_C, IE, M(i) and MX(i)-(iii) hold. Then there exists an $\ell \times \ell$ block-diagonal rotation matrix H defined by $H := \Upsilon^{-1} F' F \frac{\mu_C' W}{N_C} V^{-1} \Upsilon^{-1}$ such that $rank(H) = \ell$ almost surely for large N_C and T.

The following lemma states a number of asymptotic results which will be useful for the proofs of the main theorems. Let w_i be the i^{th} row of the eigenvector matrix W. H is as constructed in lemma A2.

Lemma A3: Under Assumptions E, F, FL_C, IE and M, the following results hold as $N_C, T \to \infty$:
(a) $\frac{1}{N_C} \sum_{i \in C} \|w_i - H\mu_i\|^2 = O_p\left(\frac{1}{\min(N_C, T)}\right)$.
(b) $\left\|\frac{W'(W - \mu_C H)}{N_C}\right\| = O_p\left(\frac{1}{\min(\sqrt{N_C}, \sqrt{T})}\right)$.
(c) $\left\|\frac{(W - \mu_C H)' \epsilon_C \epsilon_C' (W - \mu_C H)}{T N_C^2}\right\| = O_p\left(\frac{1}{\min(N_C, T)}\right)$.
(d) $\left\|\frac{W' \epsilon_C \epsilon_C' W}{T N_C^2}\right\| = O_p\left(\frac{1}{\min(N_C, T)}\right)$.

(e) $\left\| \Upsilon^{-1} \frac{F'\epsilon_C' W}{N_C} \right\| = O_p\left(\frac{1}{\min(\sqrt{N_C}, \sqrt{T})}\right)$.

The following lemmas are the key to the main theorems (lemmas A4 and A5 for Theorem 1, lemma A6 for Theorem 2, lemma A7 for Theorem 3). The important message from the lemmas is that the estimation error of the principal components does not contribute to the asymptotic limit of the treatment effect estimators. Recall that M_A is the orthogonal projection matrix of A defined as $M_A = I - A(A'A)^{-1}A'$.

Lemma A4: Suppose Assumptions E, F, FL$_C$, IE, M and MX hold. Then, as $N_C, T \to \infty$,
(a) $\sup_{j\in E} \left\| \frac{1}{N_C\sqrt{T}} \sum_{i\in C} 1'_{post,j} M_{X_i} F\mu_i w_i \Upsilon^{-1} - \frac{1}{\sqrt{T}} 1'_{post,j} F\Upsilon^{-1} H'^{-1} \right\| = o_p(1)$.
(b) $\left\| \frac{1}{N_C\sqrt{T}} \sum_{i\in C} X_i'\epsilon_i w_i \right\| = O_p\left(\frac{1}{\min(\sqrt{T}, \sqrt{N_C})}\right)$.
(c) $\sup_{j\in E} \left\| \frac{1}{N_C\sqrt{T}} \sum_{i\in C} 1'_{post,j} M_{X_i}\epsilon_i w_i \Upsilon^{-1} \right\| = o_p(1)$.
(d) $\sup_{j\in E} \left\| \frac{1'_{post,j}\hat{F}}{\sqrt{T}} \Upsilon^{-1} - \frac{1'_{post,j}F}{\sqrt{T}} \Upsilon^{-1} H'^{-1} \right\| = o_p(1)$.
(e) $\sup_{j\in E} \left\| \frac{X_j'\hat{F}}{\sqrt{T}} \Upsilon^{-1} \right\| = o_p(1)$.
(f) $\left\| \Upsilon^{-1}\hat{F}'\hat{F}\Upsilon^{-1} - H^{-1}\Upsilon^{-1}F'F\Upsilon^{-1}H'^{-1} \right\| = o_p(1)$.
(g) $\sup_{j\in E} \left\| \epsilon_j'\hat{F}\Upsilon^{-1} - \epsilon_j' F\Upsilon^{-1}H'^{-1} \right\| = o_p(1)$ if in addition $\frac{\sqrt{T}}{N_C} \to 0$.

Lemma A5: Suppose Assumptions E, F, FL$_C$, IE, M and MX hold. Then, as $N_C, T \to \infty$,
(a) $\sup_{j\in E} \left| \frac{1}{T} 1'_{post,j} M_{[\hat{F}, X_j]} 1_{post,j} - \frac{1}{T} 1'_{post,j} M_{[F, X_j]} 1_{post,j} \right| = o_p(1)$.
(b) $\sup_{j\in E} \left| \frac{1}{\sqrt{T}} 1'_{post,j} M_{[\hat{F}, X_j]} (F\mu_j + X_j\beta_j + \epsilon_j) - \frac{1}{\sqrt{T}} 1'_{post,j} M_{[F, X_j]}\epsilon_j \right| = o_p(1)$ if in addition $\frac{\sqrt{T}}{N_C} \to 0$.

Lemma A6: Suppose Assumptions E, F, FL, IE, MM and MX hold. Then, as $N_E, N_C, T \to \infty$ and $\frac{T}{N_C} \to 0$,

$$\left| \frac{1}{\sqrt{N_E T}} \sum_{j\in E} 1'_{post,j} M_{[\hat{F}, X_j]} (F\mu_j + X_j\beta_j + \epsilon_j) - \frac{1}{\sqrt{N_E T}} \sum_{j\in E} 1'_{post,j} M_{[F, X_j]}\epsilon_j \right| = o_p(1).$$

Lemma A7: Suppose Assumptions E, F, FLM, IE, MM and MX hold. Suppose further that $E(\mu_i | i \in C) =: \mu_0$, a finite and non-zero vector. Let r be the normalization order of $F\mu_0$ such that $\|F\mu_0/T^r\| = O_p(1)$ as $T \to \infty$. Then, as $N_C, T \to \infty$,
(a) $\sup_{j\in E} \left| \frac{\bar{u}_C' M_{[1_{post,j}, X_j]} \bar{u}_C}{T^{2r}} - \mu_0' \frac{F' M_{[1_{post,j}, X_j]} F}{T^{2r}} \mu_0 \right| = o_p(1)$.
(b) $\sup_{j\in E} \left| \frac{\bar{u}_C' M_{[1_{post,j}, X_j]} f\mu_j}{T^{2r}} - \mu_0' \frac{F' M_{[1_{post,j}, X_j]} F\mu_j}{T^{2r}} \right| = o_p(1)$.
(c) $\left| \sum_{j\in E} \frac{\bar{u}_C' M_{[1_{post,j}, X_j]}\epsilon_j}{\sqrt{N_E} T^r} - \mu_0' \sum_{j\in E} \frac{F' M_{[1_{post,j}, X_j]}\epsilon_j}{\sqrt{N_E} T^r} \right| = o_p(1)$ if in addition $N_E \to \infty$ and $T/N_C \to 0$.

Remark: When some factors are $I(0)$ and the idiosyncratic errors display heteroskedasticity and weak dependence, the asymptotic condition $\frac{\sqrt{T}}{N_C} \to 0$ is necessary to achieve consistency of the individual treatment effect estimator, and the asymptotic condition $\frac{T}{N_C} \to 0$ is necessary to achieve consistency of the pooled estimator and the alpha test.

Lemma A8: For each $i \in E$, define $S_{\tilde{G}\epsilon} := \tilde{\Gamma}^{-1} \sum_{t=1}^{T} \tilde{G}_{it}\epsilon_{it}$. Suppose Assumptions ES and D$_i$ hold, and that the asymptotic variance $\Omega_{\tilde{G}\epsilon} := \text{plim}_{T\to\infty} Var(S_{\tilde{G}\epsilon} | \tilde{G}_i)$ is $O_p(1)$ and positive definite *a.s.*. Then we have $S_{\tilde{G}\epsilon}|_{\tilde{G}_i} \xrightarrow{d} N(0, \Omega_{\tilde{G}\epsilon})$ as $T \to \infty$.

3 Comparison with Other Methods

In this section, we study the generalized synthetic control (GSC) estimator of Xu (2017). A formal comparison with the PCDID estimator (defined in Section 4) is made in terms of identification and efficiency.

3.1 The Generalized Synthetic Control Estimator

The DGP is the same as Section 3, equation (5): $y_{it} = \Delta_{it} 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \beta_i' x_{it} + \varsigma_i + \mu_i' f_t + \tilde{\epsilon}_{it}$. The estimator is obtained from the following procedure:

1. Use Bai (2009)'s estimator on the control panel ($N_C \times T$ observations) to obtain $\hat{\beta}_i^{bai}$, $\hat{\varsigma}_i^{bai}$, $\hat{\mu}_i^{bai}$ and $\hat{f}_t^{bai}$ for $i \in C$ and $t = 1, .., T$. The cth iteration contains two estimation sub-procedures: (i) based on the control panel of residuals $y_{it} - \hat{\beta}_i^{(c-1)'} x_{it} - \hat{\varsigma}_i^{(c-1)}$ where $\{\hat{\beta}_i^{(c-1)}, \hat{\varsigma}_i^{(c-1)}\}_{i \in C}$ are estimates from the *(c-1)*th iteration, use PCA to estimate $\hat{\mu}_i^{(c)}$ and $\hat{f}_t^{(c)}$, and (ii) subtract $\hat{\mu}_i^{(c)'} \hat{f}_t^{(c)}$ from y_{it}, and obtain $\{\hat{\beta}_i^{(c)}, \hat{\varsigma}_i^{(c)}\}_{\{i \in C\}}$ from the regression $y_{it} - \hat{\mu}_i^{(c)'} \hat{f}_t^{(c)} = b_{0i} + b_{1i}' x_{it} + e_{it}$ using the control panel.

2. For each treated unit ($i \in E$), use data from *pre-intervention* periods ($t = 1, .., T_{0i}$) to obtain $\hat{\varsigma}_i^{gsc}$, $\hat{\mu}_i^{gsc}$ from the regression $y_{it} - \hat{\beta}_i^{bai'} x_{it} = b_{0i} + b_{1i}' \hat{f}_t^{bai} + e_{it}$. The least squares formula is

$$[\hat{\varsigma}_i^{gsc}, \hat{\mu}_i^{gsc}]' = (\tilde{\hat{G}}_{0i}' \tilde{\hat{G}}_{0i})^{-1} \tilde{\hat{G}}_{0i}' (\tilde{\hat{Y}}_i), \tag{1}$$

where $\tilde{\hat{G}}_{0i} := \begin{bmatrix} 1 & ,\dots, & 1 \\ \hat{f}_1^{bai} & ,\dots, & \hat{f}_{T_{0i}}^{bai} \end{bmatrix}'$ and $\tilde{\hat{Y}}_i := \begin{bmatrix} y_{i1} - \hat{\beta}_i^{bai'} x_{i1} & ,\dots, & y_{iT_{0i}} - \hat{\beta}_i^{bai'} x_{iT_{0i}} \end{bmatrix}'$.

3. For each $i \in E$ and $t = T_{0i} + 1, .., T$, compute

$$\hat{\delta}_{it}^{gsc} := y_{it} - \hat{\beta}_i^{bai'} x_{it} - \hat{\varsigma}_i^{gsc} - \hat{\mu}_i^{gsc'} \hat{f}_t^{bai}.$$

As in most synthetic control or matrix completion approaches, GSC uses the control panel and *pre-intervention* periods of treated units to construct the counterfactual outcome. Then, using the *post-intervention* periods of treated units, it computes the treatment effect by subtracting the counterfactual outcome from the actual outcome.

Without losing the key insights, the ensuing analysis abstracts from the presence of covariates and assume that all factors are stationary processes. The lengths of pre- and post-intervention periods (T_0 and T_1) are assumed to be homogeneous. The idiosyncratic errors $\tilde{\epsilon}_{it}$ are assumed to be *iid* with mean zero and independent of the factors, factor loadings and treatment effects. The number of factors ℓ is assumed fixed.

3.1.1 Identification

A key identification condition of the GSC estimator is:[1]

$$G_0 \text{ has full column rank,} \tag{2}$$

where $G_0 = [1_{pre}, F_0]$, 1_{pre} is a $T \times 1$ vector storing T_0 ones followed by T_1 zeros, and F_0 is a $T \times \ell$ matrix which is the same as F except that the last T_1 rows are set to zero. This is evident from (1), which requires $(\tilde{\hat{G}}_0' \tilde{\hat{G}}_0)^{-1}$ to exist ($\tilde{\hat{G}}_{0i} \equiv \tilde{\hat{G}}_0$). In practice, it means that there is sufficient time variation in factors during pre-intervention periods that can identify the factor loadings among all treated units. For the PCDID estimator, the identification condition is:

$$1_{post}' M_G 1_{post} > 0, \tag{3}$$

[1] This condition is analogous to Assumption 4.4 of Xu (2017b) after switching the roles of t and i.

where $G = [1, F]$ and 1_{post} is a $T \times 1$ vector storing T_0 zeros followed by T_1 ones. It turns out that condition (3) is stronger than (4). This is stated in the following proposition. The proof is found at the end of this appendix section.

Proposition S1: (2) $\Longrightarrow$ (3).

Geometrically, condition (2) implies that 1_{post} does not lie on the linear subspace spanned by the columns of F. The full rank condition is crucial for GSC estimator to be well-defined, but it is optional for PCDID estimator. Some examples illustrating this point are given below.

Example 1a*:* Set $T_0 = T_1 = T/2$, $f_t = 1_{\{t>0.75T\}}$ (single factor). GSC estimator is not well defined due to lack of variation in f_t for $t \le T_0$. PCDID estimator is well defined because the intervention dummy $1_{\{t>T_0\}}$ is not collinear with f_t over $t = 1, .., T$. ($\frac{1'_{post}M_G 1_{post}}{T} > 0$ *uniformly* over all T, T_1, and hence Assumption AI(i) holds.).

Example 1b*:* Let u_t be an *iid* sequence with mean 0 and variance $\sigma_u^2 > 0$. Set $T_0 = T_1 = T/2$, $f_t = u_t 1_{\{t \le T_0\}} + 1_{\{t>T_0\}}$ (single factor). GSC estimator is well-defined due to variation in f_t for $t \le T_0$. PCDID estimator is also well-defined because the intervention dummy $1_{\{t>T_0\}}$ is not collinear with f_t for $t = 1, .., T$.

Due to treatment effect heterogeneity, we must also align the estimands between GSC and PCDID when discussing identification. That said, an important benchmark case in the synthetic control or matrix completion literature is $T_1 = N_E = 1$ but T_0 and N_C are large. As in those literatures, assume that the treatment effect Δ_{it} as given once the sample is drawn, i.e., the estimand is Δ_{it} instead of some population moments of Δ_{it}. Then, in this case, the estimands are identical and, clearly, PCDID has a weaker identification condition than GSC according to Proposition S1.

More generally, the building block of PCDID's estimand is the ITET $\bar{\Delta}_i := E(\Delta_{it}|t > T_0)$ for each $i \in E$, which is identified when T_1 is large. The asymptotic analog of (3) and (4) are, respectively,

$$\operatorname*{plim}_{T_0 \to \infty} \frac{G_0' G_0}{T_0} \text{ is invertible,} \qquad \text{and} \tag{4}$$

$$\operatorname*{plim}_{T_1, T \to \infty} \frac{1'_{post} M_G 1_{post}}{T} > 0, \tag{5}$$

which is Assumption AI_i(i). The proposition below is the asymptotic analog of Proposition S1.

Proposition S2: Suppose $0 < \kappa < 1$, where $\kappa := \lim_{T, T_1 \to \infty} T_1/T$. Then (4) $\Longrightarrow$ (5).

In practice, $0 < \kappa < 1$ means that the proportions of pre- and post-intervention periods do not vanish in the limit (see also Assumption E(iii)). This makes the comparison of pre- and post-intervention periods meaningful when T becomes large, as illustrated by the example below.

Example 2: Set $\ell = 1$ (single factor), $f_t = 1_{\{t>T_0+0.5T_1\}}$. GSC estimator is not well defined due to lack of variation in f_t for $t \le T_0$. As long as $0 < \kappa < 1$, PCDID estimator is well-defined because the intervention dummy $1_{\{t>T_0\}}$ is not collinear with f_t for $t = 1, .., T$. When $\kappa = 0$ or $\kappa = 1$, PCDID estimator is not well defined because there is no variation in the intervention dummy in the limit.

3.1.2 Efficiency

To compare the efficiency of GSC and PCDID estimators, we first align the estimands. Without losing the main intuition, we consider the ITET as the estimand in large samples.[2] The relevant estimators

[2] The results still hold even when $T_1 = 1$, by replacing $\bar{\Delta}_i$ with Δ_{it}. This can then be used to construct $\frac{1}{N_E}\sum_{i \in E} \hat{\delta}_{it}^{gsc}$ found in Xu (2017) and PCDID-MG estimators, using $\frac{1}{N_E}\sum_{i \in E} \Delta_{it}$ as the estimand, with similar results.

are then $\hat{\delta}_i^{gsc} := \frac{1}{T_1}\sum_{t>T_0}\hat{\delta}_{it}^{gsc}$ and $\hat{\delta}_i^{pcdid}$. To facilitate our analysis, it is helpful to decompose $\hat{\delta}_i^{gsc}$ as follows. Recall that the vector of composite idiosyncratic errors is $\epsilon_i = \tilde{\epsilon}_i + \tilde{\Delta}_i 1_{\{i\in E\}} 1_{post}$.

Lemma S1: As $N_C, T \to \infty$ and $T/N_C \to 0$, we have

$$\hat{\delta}_i^{gsc} - \bar{\Delta}_i = \frac{1}{T_1} 1'_{post}\epsilon_i - \frac{1}{T_1} 1'_{post} G(G'_0 G_0)^{-1} G'_0 \epsilon_i + o_p(1).$$

Suppose ϵ_{it} are iid with mean zero and variance σ_ϵ^2. Then, by the above lemma, the asymptotic variance of $\hat{\delta}_i^{gsc}$ is given by:

$$V^{gsc} := Var(\sqrt{T}\hat{\delta}_i^{gsc}|1_{post}, G) = \frac{\sigma_\epsilon^2}{T_1^2}(1'_{post}1_{post} + 1'_{post}G(G'_0G_0)^{-1}G'1_{post}) + o_p(1). \tag{6}$$

On the other hand, the asymptotic variance of $\hat{\delta}_i^{pcdid}$ is (c.f. Theorem 1 and its proof):

$$V^{pcdid} := Var(\sqrt{T}\hat{\delta}_i^{pcdid}|1_{post}, G) = \frac{T\sigma_\epsilon^2}{1'_{post}M_G 1_{post}} + o_p(1). \tag{7}$$

To analyze the variances in a simple setting, we consider the DGP

$$y_{it} = \Delta_{it} 1_{\{i\in E\}} 1_{\{t>T_0\}} + \varsigma_i + \mu'_i f_t + \tilde{\epsilon}_{it}, \tag{8}$$

where f_t consists of a single factor that is a stationary, mean-zero process possibly correlated with the intervention dummy $1_{\{t>T_0\}}$. Let $b_0 := E(f_t|t \le T_0)$ and $b_1 := E(f_t|t > T_0)$. Suppose further that $Var(f_t|t \le T_0) = Var(f_t|t > T_0) =: v_f > 0$. The following proposition compares the variances of the two estimators under this DGP.

Proposition S3: Suppose the DGP is given by (8), and that $0 < \kappa < 1$ and (4) holds. Then $\text{plim}\, V^{gsc} \ge \text{plim}\, V^{pcdid}$ as $N_C, T \to \infty$ and $T/N_C \to 0$. Equality holds iff $b_0 = b_1 = 0$.

The implication is that PCDID estimator is at least as efficient as GSC estimator in large samples, and both estimators are equally efficient if and only if the factor is uncorrelated with the intervention dummy. To explain the rationale, we note that GSC estimator discards data in the post-intervention subsample which contain useful information about the correlation between factors and intervention dummy. This leads to efficiency loss, which becomes bigger as the proportion of post-intervention periods to be discarded is higher. This is illustrated in Example 3 below.

Example 3: Set $\sigma_\epsilon^2 = 1$, $T_0 = 0.75T$, $T_1 = 0.25T$, f_t is an *iid* sequence with mean 0 and (unconditional) variance $\sigma_f^2 > 0$. Assume that f_t and $1_{\{t>T_0\}}$ are correlated such that $b_1 = E(f_t|t > T_0) = 0.5$. Further assume that $v_f = Var(f_t|t \le T_0) = Var(f_t|t > T_0) = 1$. Then we have $\kappa = \frac{1}{4}$, $\theta := \frac{\kappa}{1-\kappa} = \frac{1}{3}$, $b_0 = -\theta b_1 = -\frac{1}{6}$, $\sigma_f^2 = v_f + b_0^2(1-\kappa) + b_1^2\kappa = \frac{13}{12}$, $A := \frac{b_1^2}{v_f} = \frac{1}{4}$ and $B := \frac{b_1^2}{\sigma_f^2} = \frac{3}{13}$. The variance of GSC estimator is $Var(\hat{\delta}_i^{gsc}) = \frac{1}{T_1}\sigma_\epsilon^2\left(1 + \theta\left[1 + (\theta+1)^2 A\right]\right) = \frac{40}{27}\frac{1}{T_1} = \frac{1.4815}{T_1}$, which is bigger than that of PCDID estimator: $Var(\hat{\delta}_i^{pcdid}) = \frac{1}{T_1}\sigma_\epsilon^2\frac{1}{1-\kappa(1+B)} = \frac{13}{9}\frac{1}{T_1} = \frac{1.4444}{T_1}$.

Now treat κ and b_1 as free parameters. As illustrated in the table below, the ratio $\frac{\text{plim}\, V^{gsc}}{\text{plim}\, V^{pcdid}} \ge 1$ for all (κ, b_1) pairs. For example, when $\kappa = 0.5$ and $b_1 = 1$ ($A = 1$), $\text{plim}\, V^{gsc}$ is 50 percent larger than $\text{plim}\, V^{pcdid}$.

b_1	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2
$A := \frac{b_1^2}{v_f}$	0.0	0.04	0.16	0.36	0.64	1.0	1.44	1.96	2.56	3.24	4.00
$\kappa = 0.1$	1.00	1.00	1.00	1.00	1.01	1.01	1.02	1.02	1.02	1.03	1.03
$\kappa = 0.2$	1.00	1.00	1.01	1.02	1.03	1.05	1.07	1.08	1.10	1.11	1.13
$\kappa = 0.3$	1.00	1.01	1.03	1.06	1.09	1.13	1.16	1.20	1.22	1.25	1.27
$\kappa = 0.4$	1.00	1.02	1.06	1.13	1.20	1.27	1.33	1.38	1.42	1.46	1.48
$\kappa = 0.5$	1.00	1.04	1.14	1.26	1.39	1.50	1.59	1.66	1.72	1.76	1.80
$\kappa = 0.6$	1.00	1.08	1.29	1.53	1.73	1.90	2.03	2.12	2.19	2.24	2.29
$\kappa = 0.7$	1.00	1.20	1.63	2.07	2.40	2.63	2.80	2.91	3.00	3.06	3.11
$\kappa = 0.8$	1.00	1.55	2.56	3.36	3.88	4.20	4.41	4.55	4.64	4.71	4.76
$\kappa = 0.9$	1.00	3.38	6.31	7.88	8.67	9.10	9.36	9.52	9.63	9.70	9.76

3.2 The Step-wise GM Estimator

Gobillon and Magnac (2016) examine identification and estimation of treatment effects under the *original* Bai (2009) approach, which uses the entire panel for estimation. They also consider a step-wise estimator ("GM") as follows. This estimator uses the same #1 and #2 procedure as the GSC estimator. Then, do the following:

3. Use the Bai (2009)'s estimator on the full panel ($N \times T$ observations) where the outcomes of treated units during post-intervention periods (y_{it} for all $i \in E$ and $t = T_{0i}+1, ..., T$) are replaced by the counterfactual outcome $\hat{\beta}_i^{bai'} x_{it} + \hat{\varsigma}_i^{gsc} + \hat{\mu}_i^{gsc'} \hat{f}_t^{bai}$. Denote the estimates by $\hat{\beta}_i^{bbai}$, $\hat{\varsigma}_i^{bbai}$, $\hat{\mu}_i^{bbai}$ and $\hat{f}_t^{bbai}$ for $i \in C, E$ and $t = 1, .., T$.

4. For each $i \in E$ and $t = T_{0i} + 1, .., T$, compute $\hat{\delta}_{it}^{gm} := y_{it} - \hat{\beta}_i^{bbai'} x_{it} - \hat{\varsigma}_i^{bbai} - \hat{\mu}_i^{bbai'} \hat{f}_t^{bbai}$.

From the #2 procedure, the identification condition is the same as GSC. In addition, we expect GM to have similar efficiency performance as GSC because in #3, information of the actual outcomes of treated units during post-intervention periods are not used for estimating the factor structure and treatment effects.

3.3 The MC-NNM Estimator

The Matrix Completion with Nuclear Norm Minimization (MC-NNM) estimator (Athey et al. (2018)) assumes the potential outcome without treatment is given by $Y(0) = L^* + \epsilon$, where $Y(0), L^*, \epsilon$ are $N \times T$ matrices, L^* is a low-rank matrix of rank R (e.g., factor structure with R factors), and the measurement error ϵ satisfies $E(\epsilon|L^*) = 0$. See their paper for the full list of assumptions. They show that MC-NNM, vertical regression (synthetic control methods), horizontal regression (unconfoundedness methods), and the original synthetic control approach (Abadie et al. (2010)) can all be viewed as matrix completion methods based on matrix factorization with different restrictions/regularizations.

Let Y be a $N \times T$ matrix of actual outcomes, and $\mathcal{O} := \{(i, j)|(i \in C \cap 1 \leq t \leq T) \cup (i \in E \cap 1 \leq t \leq T_{0i})\}$ be a set of indices consisting of all periods of control units and pre-intervention periods of treated units. Define the $N \times T$ matrix $P_{\mathcal{O}}(Y)$ with elements as follows: $P_{\mathcal{O}}(Y)_{it} = Y_{it}$ if $(i, t) \in \mathcal{O}$, $P_{\mathcal{O}}(Y)_{it} = 0$ otherwise. Their estimator for L^* is

$$\hat{L} = argmin_L \{ \frac{1}{|\mathcal{O}|} ||P_{\mathcal{O}}(Y - L)||_F^2 + \lambda ||L||_* \}, \tag{9}$$

where $||.||_F^2$ is the Fröbenius norm, λ is a penalty factor for the regularization problem, and $||L||_* := \sum_i \sigma_i(L)$ is the nuclear norm given by the sum of all singular values of L.[3] This estimator finds a low-rank matrix $\hat{L}$ that best fits the observed entries of $Y(0)$, where the optimal rank is determined by the relative size of the penalty term $\lambda ||L||_*$.

The minimization problem can be solved by a recursive algorithm as follows. Initialize $L^{(1)} := P_{\mathcal{O}}(Y)$ (i.e., the initial L fits observed $Y(0)$ perfectly). The cth iteration uses the following sub-procedures to update $L^{(c)}$:

1. Given $L^{(c-1)}$, compute $A^{(c)} = P_{\mathcal{O}}(Y) + P_{\mathcal{O}}^{\perp}(L^{(c-1)})$ where $P_{\mathcal{O}}^{\perp}(L^{(c-1)}) := L^{(c-1)} - P_{\mathcal{O}}(L^{(c-1)})$.

2. Perform singular value decomposition (SVD) on $A^{(c)}$, yielding $S^{(c)} \Sigma^{(c)} R^{(c)'}$ where $\Sigma^{(c)}$ is a $N \times T$ diagonal matrix with ordered diagonal elements $\sigma_{ii}^{(c)}$ for $i = 1, .., \min\{N, T\}$ as singular values.

3. Perform shrinkage operation on $\Sigma^{(c)}$ by replacing each diagonal element $\sigma_{ii}^{(c)}$ with $\max\{\sigma_{ii}^{(c)} - \frac{\lambda|\mathcal{O}|}{2}, 0\}$ for $i = 1, .., \min\{N, T\}$. Denote the new matrix by $\tilde{\Sigma}^{(c)}$.

[3]The singular value decomposition (SVD) $L_{N \times T} = S_{N \times N} \Sigma_{N \times T} R'_{T \times T}$ yields singular values $\sigma_i(L), i = 1, .., min(N, T)$ as the ordered diagonal elements of $\Sigma_{N \times T}$. The rank of L is equal to the number of non-zero singular values of L.

4. Compute $L^{(c)} = S^{(c)}\tilde{\Sigma}^{(c)}R^{(c)\prime}$.

The optimal penalty factor λ is chosen by cross-validation within $\mathcal{O}$. Consider a random subset $\tilde{\mathcal{O}} \subset \mathcal{O}$ as the training set and its complement, $\mathcal{O}\backslash\tilde{\mathcal{O}}$, as the testing set. Given each λ, apply the recursive algorithm to the training set and compute the mean squared prediction error (MSPE) in the testing set. The optimal λ is the one that minimizes the MSPE.

4 Parallel Trend Tests: More Details

4.1 Alpha Test

The factor loading μ_j for each treated unit $j \in E$ can be represented as

$$\mu_j = \alpha_j E(\mu_i | i \in C) + v_j, \tag{10}$$

where α_j is a scalar and v_j is an $\ell \times 1$ vector. The representation exists for the factor loading of any treated unit j.[4] To ensure uniqueness of the above representation, we impose the restrictions that (i) the sum of the elements of v_j is zero, and (ii) $v_j \neq \gamma_j E(\mu_i | i \in C)$ for all non-zero scalars γ_j (see the proposition below and its proof in Section 9).

Proposition A (Alpha test): Restrictions (i) and (ii) hold *iff* any $\ell \times 1$ factor loading can be uniquely represented by (10).

The Alpha test involves the following steps: (1) for each unit $i \in C \cup E$, run a simple regression of y_{it} on x_{it} (with intercept) and obtain the residuals $\hat{u}_{it}$; (2) compute the simple average of the residuals over $i \in C$, and obtain the factor proxy $\bar{u}_{Ct} = \frac{1}{N_C}\sum_{i\in C}\hat{u}_{it}$ which approximates $E(\mu_i' | i \in C)f_t$; (3) for each treated unit $j \in E$, run the time series regression $\hat{u}_{jt} = \delta_j 1_{\{t>T_{0j}\}} + a_j \bar{u}_{Ct} + w_{jt}$, and obtain the OLS estimator $\hat{a}_j$ of a_j; (4) compute the simple mean-group estimator, given by $\hat{a}^{mg} := \frac{1}{N_E}\sum_{j\in E}\hat{a}_j$.

Let us discuss what happens under PTW. Define $\alpha := E(\alpha_j | j \in E)$. By (10), the reduced-form DGP of y_{jt} for $j \in E$ can be rewritten as

$$y_{jt} = \varsigma_i + \beta_i' x_{it} + \bar{\Delta}_j 1_{\{t>T_{0j}\}} + \alpha E(\mu_i' | i \in C)f_t + [(\alpha_j - \alpha)E(\mu_i' | i \in C)f_t + v_j' f_t + \epsilon_{jt}]. \tag{11}$$

On the other hand, $\hat{a}^{mg}$ is the simple mean-group estimator based on the regression

$$y_{jt} = b_{0j} + b_{1j}' x_{jt} + \delta_j 1_{\{t>T_{0j}\}} + a\bar{u}_{Ct} + q_{jt}. \tag{12}$$

Matching the terms, we see that the regression error q_{jt} in (12) is given by

$$q_{jt} = \alpha[E(\mu_i' | i \in C)f_t - \bar{u}_{Ct}] + (\alpha_j - \alpha)E(\mu_i' | i \in C)f_t + v_j' f_t + \epsilon_{jt}.$$

Under PTW, we see that $\alpha = 1$ after taking conditional expectation of μ_j in (10). On the other hand, PTW entails that the linear combination $E(\mu_i' | i \in C)f_t$ becomes $\mu_0' f_t$, which may be viewed as a single factor $\tilde{f}_t$ common to all units i. According to the above matching, the regression coefficient a of the factor proxy $\bar{u}_{Ct}$ in (12) corresponds to the loading of $\tilde{f}_t := E(\mu_i' | i \in C)f_t$ in (11), and so we expect that the estimate of a is close to unity under PTW.

The error of the reduced-form DGP is given by the square-bracketed expression in (11). Each of the error components is described below. The first term $\alpha[E(\mu_i' | i \in C)f_t - \bar{u}_{Ct}]$ represents the estimation error due to the use of factor proxy $\bar{u}_{Ct}$. It vanishes in the limit by the asymptotic theory of PCDID estimation (see Lemma A7). The second term $(\alpha_j - \alpha)E(\mu_i' | i \in C)f_t$ measures the portion of the treated-group factor structure that departs from parallel trends but remains proportional to the "average factor" for the control group. Its conditional mean over treated group is non-zero,

[4] If μ_j takes the form of $\gamma_j E(\mu_i | i \in C)$ for some non-zero scalar γ_j, then $\alpha_j = \gamma_j$ and $v_j = 0$ (the zero vector); otherwise, $\alpha_j = 0$ and $v_j = \mu_j$.

except when Assumption PTW holds (in which case $\alpha = E(\alpha_j|j \in E) = 1$). The third term $v_j' f_t = [\mu_j - \alpha_j E(\mu_i'|i \in C)]' f_t$ (by (10)) captures the portion of the treated-group factor structure that is not proportional to the "average factor" for the control group. Its conditional mean over treated group is non-zero, except when Assumption PTW holds (in which case $E(v_j|j \in E) = 0$ by (10)). The last term ϵ_{jt} is the idiosyncratic errors in the reduced-form DGP.

There are several ways by which PTW is violated in the DGP. We expect that the Alpha test is powerful against the following departures from PTW.

(A) $\alpha \neq 1$ and $E(v_j|j \in E) = 0$. In other words, the "average factor" for the treated group $E(\mu_j'|j \in E) f_t$ remains proportional to that for the control group $E(\mu_i'|i \in C) f_t$, but the proportionality constant is different from one. The Alpha test has power as $\hat{a}^{mg}$ approaches α, which is different from unity.

(B) $E(v_j|j \in E) \neq 0$ and f_t is covariance stationary (possibly with deterministic trends). In this case, the "average factors" for the treated and control groups are not proportional to each other. This gap is time-varying in general. This is captured by the third term $v_j' f_t$ in the regression error ϵ_{jt}. Since the gap may be correlated with the factor proxy (with covariance c), the Alpha test has power as $\hat{a}^{mg}$ approaches $\alpha + c$, which is different from unity in general.

(C) $E(v_j|j \in E) \neq 0$ and f_t contains stochastic trends (e.g., integrated processes). This is similar to (B) in that the gap between the "average factors" is time-varying, but the gap contains stochastic trends over time. The Alpha test has power as $\hat{a}^{mg}$ explodes in the limit.

4.2 Hausman Test

In the context of Hausman test, we suppose that the factors are $I(0)$, the treatment effects are homogeneous, i.e., $\bar{\Delta}_i \equiv \bar{\bar{\Delta}}$ for all $i \in E$, and the intervention dates are identical across treated units (so that $1_{post,i} \equiv 1_{post}$ for all $i \in E$). Without losing the main insight, we assume *iid* errors ($\epsilon_{it} \sim iid(0, \sigma_\epsilon^2)$) and the absence of covariates. Suppose that the *strong parallel trend hypothesis* holds: $\mu_i \equiv \mu_0$ for all $i \in C \cup E$. We then see that multiple factors combine into a single factor $f := F\mu_0$ common to all units. The DGP thus reduces to

$$y_{it} = \varsigma_i + f_t + \bar{\bar{\Delta}} 1_{\{i \in E\}} 1_{\{t > T_0\}} + \epsilon_{it}. \tag{13}$$

Under the strong parallel trend hypothesis and the above DGP, it suffices to use a single factor proxy in simple mean-group estimation. Let $\hat{\delta}^{mg}$ be the simple mean-group estimator that uses the simple cross-sectional average of control-group outcomes $\bar{y}_{Ct} := N_C^{-1} \sum_{i \in C} y_{it}$ as the factor proxy. Let $\hat{\delta}^{2wfe}$ be the OLS estimator of $\bar{\bar{\Delta}}$ in the two-way fixed-effects regression associated with (13). The following proposition compares the variances of the two estimators, providing motivation for the construction of Hausman test. The proof is found at the end of the Appendix.

Proposition H (Hausman test): Suppose the *strong parallel trend hypothesis* holds: $\mu_i \equiv \mu_0$ for all $i \in C \cup E$, and that the DGP is given by (13). We then have:
(a) $Var(\hat{\delta}^{2wfe}) = \sigma_\epsilon^2 \left(\frac{1}{T_0} + \frac{1}{T_1}\right)\left(\frac{1}{N_E} + \frac{1}{N_C}\right) = \sigma_\epsilon^2 \left(1_{post}' M_{[1]} 1_{post}\right)^{-1} \left(\frac{1}{N_E} + \frac{1}{N_C}\right)$,
(b) $Var(\hat{\delta}^{mg}) = \sigma_\epsilon^2 (1_{post}' M_{[1,\bar{y}_C]} 1_{post})^{-1} \left(\frac{1}{N_E} + \frac{1}{N_C}\right)$,
(c) $Var(\hat{\delta}^{mg} - \hat{\delta}^{2wfe}) = Var(\hat{\delta}^{mg}) - Var(\hat{\delta}^{2wfe})$.

5 Micro-Level Analysis

5.1 Assumptions

To simplify the notation, the subscript g is suppressed wherever it is redundant, e.g, write $\bar{\Delta}_{ig}$ ($:= E(\Delta_{igt}|t > T_{0g})$) as $\bar{\Delta}_i$, x_{igt} as x_{it}, μ_{ig} as μ_i, and ϵ_{igt} ($:= \tilde{\Delta}_{igt} 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \tilde{\epsilon}_{igt}$) as ϵ_{it}, with the understanding that unit i is in group g (which has N_g units). Let $\bar{\mu}_g := N_g^{-1} \sum_{i=1}^{N_g} \mu_{ig}$ and $\bar{\epsilon}_{gt} := N_g^{-1} \sum_{i=1}^{N_g} \epsilon_{igt}$. Denote $\mathscr{C}$ the collection of $N_{\mathscr{C}}$ control groups, and $\mathscr{E}$ the collection of $N_{\mathscr{E}}$ treated groups.

Assumption G (group structure): The micro-level model with interactive effects has at least one treated group and one control group, i.e., $N_{\mathscr{E}} \geq 1$ and $N_{\mathscr{C}} \geq 1$. The factor loading μ_{ig} and idiosyncratic error ϵ_{igt} are independent across groups.

Assumption EG (predeterminedness, treatment and intervention dummies):
(i) $E(\epsilon_{it}|\bar{\Delta}_i, 1_{\{g\in\mathscr{E}\}}, 1_{\{t>T_{0g}\}}, \beta_i, \varsigma_i, \mu_i, f_t, x_{it}) = 0$ for each i in group g and for each t.
(ii) $0 < E(1_{\{g\in\mathscr{E}\}}) < 1$ for each g.
(iii) for each $g \in \mathscr{E}$, $T_{1g}/T \xrightarrow{p} \kappa_g$ as $T, T_{1g} \to \infty$, where $0 < \kappa_g < 1$.

Assumption FLG$_C$ (factor loadings of control groups):
(i) For all $g \in \mathscr{C}$, $E\left\|\bar{\mu}_g\right\|^2 \leq c$ for some constant $c > 0$.
(ii) $\operatorname{plim}_{N_{\mathscr{C}}\to\infty} \frac{1}{N_{\mathscr{C}}} \sum_{g\in\mathscr{C}} \bar{\mu}_g \bar{\mu}_g'$ is positive definite.

Assumption FLG (factor loadings): Assumption FLG$_C$ holds. In addition, Assumption FL(i)-(ii) hold.

Define $\bar{\sigma}_{gh,st} := E(\bar{\epsilon}_{gs}\bar{\epsilon}_{ht})$ and $\bar{\gamma}_T(g,h) := E\left[T^{-1}\sum_{t=1}^T \bar{\epsilon}_{gt}\bar{\epsilon}_{ht}\right]$.
Assumption IEG (idiosyncratic errors): Assumptions IE(i)-(iii) and (v) hold for all $i, j \in E$. In addition, there exists a positive constant $0 < c < \infty$ such that for all N_E, $N_{\mathscr{C}}$ and T:
(i) $E\left|\bar{\epsilon}_{gt}\right|^8 \leq c$ for each $g \in \mathscr{C}$ and t.
(ii) $|\bar{\gamma}_T(g,g)| \leq c$ and $\sum_{h\in\mathscr{C}} |\bar{\gamma}_T(g,h)| \leq c$ for each $g \in \mathscr{C}$.
(iii) $|\bar{\sigma}_{gg,st}| \leq |\bar{\sigma}_{st}|$ for some $\bar{\sigma}_{st}$ and for all $g \in \mathscr{C}$. In addition, $\frac{1}{T}\sum_{s,t=1}^T |\bar{\sigma}_{st}| \leq c$.
(iv) $\frac{1}{N_{\mathscr{C}}T}\sum_{g,h\in\mathscr{C}}\sum_{s,t=1}^T |\bar{\sigma}_{gh,st}| \leq c$.
(v) $E\left|\frac{1}{\sqrt{T}}\sum_{t=1}^T [\bar{\epsilon}_{gt}\bar{\epsilon}_{ht} - E(\bar{\epsilon}_{gt}\bar{\epsilon}_{ht})]\right|^4 \leq c$ for all $g, h \in \mathscr{C}$.

Assumption IEDG (decomposition of idiosyncratic errors): Suppose $1 \leq N_{\mathscr{C}} < \infty$. Without loss of generality, let $g \in \mathscr{C}' := \{1, \ldots, N_{\mathscr{C}}'\}$ (where $1 \leq N_{\mathscr{C}}' \leq N_{\mathscr{C}}$) be the control groups with $N_g \to \infty$. For each $g \in \mathscr{C}'$, the idiosyncratic errors can be decomposed as $\epsilon_{igt} = \check{\gamma}_{ig}'\check{\epsilon}_{gt} + \check{\epsilon}_{igt}$, so that the interactive-effect structure of the micro-level model can be rewritten as $\mu_{ig}'f_t + \epsilon_{igt} = \check{\mu}_{ig}'\check{f}_t + \check{\epsilon}_{igt}$, where $\check{\mu}_{ig}' := [\begin{array}{cccc} \mu_{ig}' & c_{i1}\check{\gamma}_{i1}' & \cdots & c_{iN_{\mathscr{C}}'}\check{\gamma}_{iN_{\mathscr{C}}'}' \end{array}]'$, c_{ig} is a scalar equalling 1 if individual i is in group g and 0 otherwise, and $\check{f}_t = [\begin{array}{cccc} f_t' & \check{\epsilon}_{1t}' & \cdots & \check{\epsilon}_{N_{\mathscr{C}}'t}' \end{array}]'$.

Remark: Note that Assumption IEDG is *not* needed by aggregated (AGG) PCDID estimators.

Assumption MG (moments): Assumption MX holds. In addition, there exists $0 < c < \infty$ such that the following conditions are satisfied for all $N_{\mathscr{C}}$ and T:
(i) $E\left(\frac{1}{T}\sum_{t=1}^T \left\|\frac{1}{\sqrt{N_{\mathscr{C}}}}\sum_{g\in\mathscr{C}} \bar{\mu}_g\bar{\epsilon}_{gt}\right\|^2\right) \leq c$.
(ii) $E\sup_i \left\|\Upsilon^{-1}\sum_{t=1}^T f_t\epsilon_{it}\right\|^2 \leq c$.
(iii) $E\left\|\frac{1}{\sqrt{N_{\mathscr{C}}}}\Upsilon^{-1}\sum_{t=1}^T\sum_{g\in\mathscr{C}} \bar{\epsilon}_{gt}\bar{\mu}_g f_t'\right\|^2 \leq c$.
(iv) $E\sup_{j\in E}\left\|\frac{1}{\sqrt{N_{\mathscr{C}}T}}\sum_{t=1}^T\sum_{g\in\mathscr{C}} \bar{\mu}_g[\bar{\epsilon}_{gt}\epsilon_{jt} - E(\bar{\epsilon}_{gt}\epsilon_{jt})]\right\|^2 \leq c$.

Assumption MMG (moments): Assumption MG holds. In addition, there exists $0 < c < \infty$ such that the following conditions are satisfied for all N_E, $N_{\mathscr{C}}$ and T:

(i) $\frac{1}{N_{\mathscr{C}}}\sum_{g,h\in\mathscr{C}}|\bar{\gamma}_T(g,h)| \le c$ and $\frac{1}{\sqrt{N_{\mathscr{C}}N_E}}\sum_{g\in\mathscr{C}}\sum_{j\in E}\left|E[T^{-1}\sum_{t=1}^{T}\bar{\epsilon}_{gt}\epsilon_{jt}]\right| \le c$.

(ii) $E\left\|\Upsilon^{-1}\frac{1}{\sqrt{N_E}}\sum_{j\in E}\sum_{t=1}^{T} f_t\epsilon_{jt}\right\|^2 \le c$ and $E\left\|\Upsilon^{-1}\frac{1}{\sqrt{N_{\mathscr{C}}}}\sum_{g\in\mathscr{C}}\sum_{t=1}^{T} f_t\bar{\epsilon}_{gt}\right\|^2 \le c$.

(iii) $E\left\|\frac{1}{\sqrt{N_E N_{\mathscr{C}} T}}\sum_{t=1}^{T}\sum_{j\in E}\sum_{g\in\mathscr{C}}\bar{\mu}_g[\bar{\epsilon}_{gt}\epsilon_{jt} - E(\bar{\epsilon}_{gt}\epsilon_{jt})]\right\|^2 \le c$.

Assumptions ES, ESS, **AI**$_i$, **AI**$_{pl}$, **AI**$_{mg}$**, RT**$_{pl}$**, RT**$_{mg}$, **MX**, **MM**, **D**$_i$ and **D**: same as before.

5.2 Key Results

Corollary 1 (PCDID-AGG): Suppose Assumptions G, EG, F, FLG$_C$, AI$_i$, IEG and MG hold in the micro-level model with interactive effects. Then, as $T, N_{\mathscr{C}} \to \infty$ jointly, we have for each $i \in E$:

(a) $\hat{\delta}_i^{AGG} \xrightarrow{p} \bar{\Delta}_i$.

(b) $\sqrt{T}\breve{\sigma}_{Ti}^{-1}(\hat{\delta}_i^{AGG} - \bar{\Delta}_i) \xrightarrow{d} N(0,1)$ if additionally Assumptions ES and D$_i$ hold, where $\breve{\sigma}_{Ti}^2 := Var[\sqrt{T}(\hat{\delta}_i^{AGG} - \bar{\Delta}_i)|\tilde{G}_i]$.

Similar to Sections 4.3 and 5, we decompose the treatment effect $\Delta_{igt} = \bar{\Delta}_{ig} + \tilde{\Delta}_{igt} = \bar{\bar{\Delta}} + \upsilon_{ig} + \tilde{\Delta}_{igt}$, where $\bar{\Delta}_{ig}$ is the ITET of unit i in group g, $\bar{\bar{\Delta}} := E(\bar{\Delta}_{ig}|i \in E)$ is the ATET of treated units, and $\upsilon_{ig} := \bar{\Delta}_{ig} - \bar{\bar{\Delta}}$ is the unit-specific deviation of the ITET from the ATET. To simplify the notation, write $\bar{\Delta}_{ig}$ as $\bar{\Delta}_i$ and write υ_{ig} as υ_i, with the understanding that unit i is in group g.

Corollary 2 (PCDID-MG-AGG): Suppose Assumptions G, EG, F, FLG, AI$_{mg}$, IEG and MMG hold in the micro-level model with interactive effects. As $T, N_E, N_{\mathscr{C}} \to \infty$ jointly, we have the following results:

(a) (homogeneous ITET) Suppose $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}} \equiv 0$. Then,

(i) $\hat{\delta}^{mg,AGG} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E T}\breve{\varsigma}_{N_E,T}^{-1}(\hat{\delta}^{mg,AGG} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$ if additionally Assumptions ESS and D hold, where $\breve{\varsigma}_{N_E,T}^2 := Var[\sqrt{N_E T}(\hat{\delta}^{mg,AGG} - \bar{\bar{\Delta}})|\tilde{G}]$.

(b) (heterogeneous ITET) Suppose υ_i satisfies Assumption RT$_{mg}$. Then,

(i) $\hat{\delta}^{mg,AGG} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E}\check{\breve{\varsigma}}_{N_E,T}^{-1}(\hat{\delta}^{mg,AGG} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$, where $\check{\breve{\varsigma}}_{N_E,T}^2 := Var[\sqrt{N_E}(\hat{\delta}^{mg,AGG} - \bar{\bar{\Delta}})]$.

Corollary 3 (PCDID pooled AGG): Suppose Assumptions G, EG, F, FLG, AI$_{pl}$, IEG and MMG hold in the micro-level model with interactive effects. As $T, N_E, N_{\mathscr{C}} \to \infty$ jointly, we have the following results:

(a) (homogeneous ITET) Suppose $\upsilon_i := \bar{\Delta}_i - \bar{\bar{\Delta}} \equiv 0$. Then,

(i) $\hat{\delta}^{pl,AGG} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E T}\breve{\sigma}_{N_E,T}^{-1}(\hat{\delta}^{pl,AGG} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$ if additionally Assumptions ESS and D hold, where $\breve{\sigma}_{N_E,T}^2 := Var[\sqrt{N_E T}(\hat{\delta}^{pl,AGG} - \bar{\bar{\Delta}})|\tilde{G}]$.

(b) (heterogeneous ITET) Suppose υ_i satisfies Assumption RT$_{pl}$. Then,

(i) $\hat{\delta}^{pl,AGG} \xrightarrow{p} \bar{\bar{\Delta}}$.

(ii) $\sqrt{N_E}\check{\breve{\sigma}}_{N_E,T}^{-1}(\hat{\delta}^{pl,AGG} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$, where $\check{\breve{\sigma}}_{N_E,T}^2 := Var[\sqrt{N_E}(\hat{\delta}^{pl,AGG} - \bar{\bar{\Delta}})|MSR_T]$.

As discussed in Section 5, Assumption IE in theorems 1, 2 and 3 may be violated if the micro-level data exhibit strong clustering in its idiosyncratic errors. For completeness, we present below alternative assumptions for these theorems that allow strong clustering in idiosyncratic errors. The key alternative assumptions are: (i) group structure, (ii) the number of control groups is fixed and finite, and (iii) the idiosyncratic errors can be decomposed into a strong clustering component and a weak clustering component, via an error-components model or sub factor structure.

Corollary 4 (PCDID estimators, alternative assumptions): Suppose $1 \leq N_{\mathscr{C}} < \infty$ and Assumptions G and IEDG hold. Consider the decomposed micro-level model as in IEDG, where $\check{\mu}_{ig}$ and $\check{\epsilon}_{igt}$ are the factor loadings and idiosyncratic error, respectively. Then, the statements in Theorems 1, 2 and 3 hold for this model.

6 Bootstrap Procedures

The bootstrap procedures below are motivated by the property that the *normalized* PCDID and Alpha statistics have an asymptotic standard normal distribution, regardless of the factor specification (e.g., even when some or all factors are nonstationary). Overall, the procedures mimic the two-step nature of PCDID. Step 1 performs a wild cluster bootstrap on the residuals obtained from an auxiliary regression based on control units to construct the bootstrap factor proxies. Gonçalves and Perron (2014) propose a similar first-step procedure to obtain bootstrap factors in factor-augmented regression models. Our second step is different to that in Gonçalves and Perron (2014). For hypothesis tests related to the ITET (e.g., unit-specific PCDID), step 2 performs the "stationary bootstrap" based on Politis and Romano (1994), by resampling blocks of observations of random length from the time series of the *residuals* of the treated unit.[5] For hypothesis tests related to the ATET (e.g., PCDID-MG), step 2 performs a wild cluster bootstrap among treated units. In both cases, we combine the bootstrap treated sample and bootstrap factor proxies to obtain a bootstrap estimate.

6.1 PCDID: Hypothesis tests on ITET and ATET

1. (Construct residuals) Let $\hat{u}_{it}$ be the residual from stage 1 of PCDID using the *original* sample, and $\hat{f}_t$ be the factor proxies obtained by PCA on $\{\hat{u}_{it}\}_{i\in C,t=1,\ldots,T}$. Obtain residuals $\tilde{v}_{it}$ from the following auxiliary regressions:

$$\begin{aligned} &\text{for } i \in C, \;\; \hat{u}_{it} = \tilde{c}_i + \tilde{a}_i'\hat{f}_t + \tilde{v}_{it}; \\ &\text{for } i \in E, \;\; y_{it} = \tilde{c}_i + \delta 1_{\{t>T_{0i}\}} + \tilde{b}_i' x_{it} + \tilde{\alpha}_i'\hat{f}_t + \tilde{v}_{it}, \end{aligned}$$

   where $(\tilde{c}_i, \tilde{a}_i')$ are the OLS estimates of the regression of $\hat{u}_{it}$ on $\hat{f}_t$, and $(\tilde{c}_i, \tilde{b}_i', \tilde{\alpha}_i')$ are the OLS estimates of the regression of y_{it} on $1_{\{t>T_{0i}\}}$, x_{it} and $\hat{f}_t$ with H_0 imposed (i.e., restricting $\delta = \delta_0$).

2. Do B iterations of this step. On the b^{th} iteration:

   (a) (Wild and pairs cluster bootstrap, control units) For $i \in C$, let $r_i^{(b)}$ be iid variables with a Rademacher distribution (i.e., takes values 1 and -1 with equal probability; see Davidson and Flachaire (2008)). Construct:

$$\hat{u}_{it}^{(b)} = \tilde{c}_i + \tilde{a}_i'\hat{f}_t + \tilde{v}_{it} r_i^{(b)}.$$

   Denote $\hat{\mathbf{u}}_i^{(b)} := (\hat{u}_{i1}^{(b)}, \ldots, \hat{u}_{iT}^{(b)})'$. From the sample of $\{\hat{\mathbf{u}}_i^{(b)}\}_{i\in C}$, resample with replacement N_C times to obtain the bootstrap sample $\{\hat{\mathbf{u}}_i^{*(b)}\}_{i\in C}$.

   (b) Do one of the following:

[5] This method is nonparametric and simple to implement, and it can be interpreted as a weighted average version of moving block bootstrap in which the block length is fixed. By resampling the residuals only, the method also ensures that the key asymptotic identification condition of PCDID (Assumption AI$_i$(i)) is met in the bootstrap sample. An alternative resampling approach is subsampling, which require weaker assumptions than bootstrap.

i. (**ITET**: stationary bootstrap) Consider a treated unit j ($j \in E$). Construct the vector $(\tilde{v}_{j1}^{(b)}, \ldots, \tilde{v}_{jT}^{(b)})$ from $(\tilde{v}_{j1}, \ldots, \tilde{v}_{jT})$ using the following algorithm:[6]

- let $I_{j1}^{(b)}, ..., I_{jT}^{(b)}$ be a sequence of iid variables with a discrete uniform distribution on $\{1, ..., T\}$. Construct the index sequence $k_{j1}^{(b)}, ..., k_{jT}^{(b)}$ as follows: $k_{j1}^{(b)} = I_{j1}^{(b)}$; for $t = 2, ..., T$, let $k_{jt}^{(b)} = \mod(k_{j,t-1}^{(b)}, T) + 1$ with probability $1-p$ and $k_{jt}^{(b)} = I_{jt}^{(b)}$ with probability p (note: p is the *tuning parameter*). Define $(\tilde{v}_{j1}^{(b)}, \ldots, \tilde{v}_{jT}^{(b)}) = (\tilde{v}_{j,k_{j1}^{(b)}}, \ldots, \tilde{v}_{j,k_{jT}^{(b)}})$.

Then compute:

$$y_{jt}^{(b)} = \tilde{c}_j + \delta_0 1_{\{t>T_{0j}\}} + \tilde{b}_j' x_{jt} + \tilde{\alpha}_j' \hat{f}_t + \tilde{v}_{jt}^{(b)}.$$

Compute the PCDID estimate (for ITET) $\hat{\delta}_j^{(b)}$ from the bootstrap sample $\{\hat{\mathbf{u}}_i^{*(b)}\}_{i\in C}$ and $\{(y_{jt}^{(b)}, 1_{\{t>T_{0j}\}}, x_{jt})\}_{t=1,..,T}$.

(c) (**ATET**: wild cluster bootstrap) For $i \in E$, let $r_i^{(b)}$ be iid variables with a Rademacher distribution. Construct:

$$y_{it}^{(b)} = \tilde{c}_i + \delta_0 1_{\{t>T_{0i}\}} + \tilde{b}_i' x_{it} + \tilde{\alpha}_i' \hat{f}_t + \tilde{v}_{it} r_i^{(b)}.$$

Compute the PCDID estimate (for ATET) $\hat{\delta}^{(b)}$ from the bootstrap sample $\{\hat{\mathbf{u}}_i^{*(b)}\}_{i\in C}$ and $\{(y_{it}^{(b)}, 1_{\{t>T_{0i}\}}, x_{it})\}_{i\in E; t=1,..,T}$.

Micro-level data. The bootstrap procedure is as follows:

1. (Construct residuals) Let $\hat{u}_{igt}$ be the residual from stage 1 of PCDID using the *original* sample, $\bar{\hat{u}}_{gt} = \frac{1}{N_g}\sum_{i=1}^{N_g} \hat{u}_{igt}$ be the within-group average value of residuals at time t, and $\hat{f}_t$ be the factor proxies obtained by PCA on $\{\bar{\hat{u}}_{gt}\}_{g\in\mathscr{C}; t=1,...,T}$. Obtain residuals $\tilde{v}_{gt}$ for $g \in \mathscr{C}$ and $\tilde{v}_{igt}$ for $g \in \mathscr{E}$ from the following auxiliary regressions:

$$\begin{aligned} &\text{for } g \in \mathscr{C}, \;\; \bar{\hat{u}}_{gt} = \tilde{c}_g + \tilde{a}_g' \hat{f}_t + \tilde{v}_{gt}; \\ &\text{for } g \in \mathscr{E}, \;\; y_{igt} = \tilde{c}_{ig} + \delta 1_{\{t>T_{0g}\}} + \tilde{b}_{ig}' x_{igt} + \tilde{\alpha}_{ig}' \hat{f}_t + \tilde{v}_{igt}, \end{aligned}$$

where $(\tilde{c}_g, \tilde{a}_g')$ are the OLS estimates of the regression of $\bar{\hat{u}}_{gt}$ on $\hat{f}_t$, and $(\tilde{c}_{ig}, \tilde{b}_{ig}', \tilde{\alpha}_{ig}')$ are the OLS estimates of the regression of y_{igt} on $1_{\{t>T_{0g}\}}$, x_{igt} and $\hat{f}_t$ with H_0 imposed (i.e., restricting $\delta = \delta_0$).

2. Do B iterations of this step. On the b^{th} iteration:

(a) (Wild and pairs cluster bootstrap, control groups) For $g \in \mathscr{C}$, let $r_g^{(b)}$ be iid variables with a Rademacher distribution. Construct:

$$\bar{\hat{u}}_{gt}^{(b)} = \tilde{c}_g + \tilde{a}_g' \hat{f}_t + \tilde{v}_{gt} r_g^{(b)}.$$

Denote $\bar{\hat{\mathbf{u}}}_g^{(b)} := (\bar{\hat{u}}_{g1}^{(b)}, \ldots, \bar{\hat{u}}_{gT}^{(b)})'$. From the sample of $\{\bar{\hat{\mathbf{u}}}_g^{(b)}\}_{g\in\mathscr{C}}$, resample with replacement G_C times to obtain the bootstrap sample $\{\bar{\hat{\mathbf{u}}}_g^{*(b)}\}_{g\in\mathscr{C}}$.

[6] Here is an informal description. Draw the initial index $k_{j1}^{(b)}$ at random from $\{1, ..., T\}$. With probability $1-p$, set the next index as $k_{j2}^{(b)} = k_{j1}^{(b)} + 1$ or if $k_{j1} = T$, set $k_{j2}^{(b)} = 1$; with probability p, draw the next index $k_{j2}^{(b)}$ at random from $\{1, ..., T\}$. Repeat this procedure until the vector of indices $(k_{j1}^{(b)}, ..., k_{jT}^{(b)})$ is formed. Draw $(\tilde{v}_{j1}^{(b)}, \ldots, \tilde{v}_{jT}^{(b)})$ from $(\tilde{v}_{j1}, \ldots, \tilde{v}_{jT})$ according to this vector of indices.

(b) Do one of the following:

  i. (**ITET**: stationary bootstrap) Same procedure as in nonmicro-level data for any treated unit j.

  ii. (**ATET**: wild cluster bootstrap) For $g \in \mathscr{E}$, let $r_g^{(b)}$ be iid variables with a Rademacher distribution. Construct:

$$y_{igt}^{(b)} = \tilde{c}_{ig} + \delta_0 1_{\{t>T_{0g}\}} + \tilde{b}_{ig}' x_{igt} + \tilde{\alpha}_{ig}' \hat{f}_t + \tilde{v}_{igt} r_g^{(b)}.$$

  Compute the PCDID estimate (for ATET) $\hat{\delta}^{(b)}$ from the bootstrap sample $\{\bar{\hat{\mathbf{u}}}_g^{*(b)}\}_{g\in\mathscr{C}}$ and $\{(y_{igt}^{(b)}, 1_{\{t>T_{0g}\}}, x_{igt})\}_{i=1,..,N_E;g\in\mathscr{E};t=1,..,T}$.

6.2 Parallel Trend Alpha Test

1. (Construct residuals) Let $\hat{u}_{it}$ be the residual from stage 1 of PCDID using the *original* sample, and $\bar{\hat{u}}_{Ct} = \frac{1}{N_C}\sum_{i\in C}\hat{u}_{it}$ be the simple average of $\hat{u}_{it}$. Obtain residuals $\tilde{v}_{it}$ from the following auxiliary regressions:

$$\begin{aligned} &\text{for } i \in C, \quad \hat{u}_{it} = \tilde{c}_i + \tilde{a}_i \bar{\hat{u}}_{Ct} + \tilde{v}_{it}; \\ &\text{for } i \in E, \quad y_{it} = \tilde{c}_i + \tilde{\delta}_i 1_{\{t>T_{0i}\}} + \tilde{b}_i' x_{it} + \alpha \bar{\hat{u}}_{Ct} + \tilde{v}_{it}, \end{aligned}$$

where $(\tilde{c}_i, \tilde{a}_i)$ are the OLS estimates of the regression of $\hat{u}_{it}$ on $\bar{\hat{u}}_{Ct}$, and $(\tilde{c}_i, \tilde{\delta}_i, \tilde{b}_i')$ are the OLS estimates of the regression of y_{it} on $1_{\{t>T_{0i}\}}$, x_{it} and $\bar{\hat{u}}_{Ct}$ with H_0 imposed (i.e., restricting $\alpha = 1$).

2. Do B iterations of this step. On the b^{th} iteration:

  (a) (Wild and pairs cluster bootstrap, control units) For $i \in C$, let $r_i^{(b)}$ be iid variables with a Rademacher distribution. Construct:

$$\hat{u}_{it}^{(b)} = \tilde{c}_i + \tilde{a}_i \bar{\hat{u}}_{Ct} + \tilde{v}_{it} r_i^{(b)}.$$

  Denote $\hat{\mathbf{u}}_i^{(b)} := (\hat{u}_{i1}^{(b)}, \ldots, \hat{u}_{iT}^{(b)})'$. From the sample of $\{\hat{\mathbf{u}}_i^{(b)}\}_{i\in C}$, resample with replacement N_C times to obtain the bootstrap sample $\{\hat{\mathbf{u}}_i^{*(b)}\}_{i\in C}$.

  (b) (Wild cluster bootstrap) For $i \in E$, let $r_i^{(b)}$ be iid variables with a Rademacher distribution. Construct:

$$y_{it}^{(b)} = \tilde{c}_i + \tilde{\delta}_i 1_{\{t>T_{0i}\}} + \tilde{b}_i' x_{it} + \bar{\hat{u}}_{Ct} + \tilde{v}_{it} r_i^{(b)}.$$

  Compute the Alpha statistic $\hat{\alpha}^{(b)}$ from the bootstrap sample $\{\hat{\mathbf{u}}_i^{*(b)}\}_{i\in C}$ and $\{(y_{it}^{(b)}, 1_{\{t>T_{0i}\}}, x_{it})\}_{i\in E;t=1,..,T}$.

6.3 Two-Way Fixed-Effects Regressions

For completeness, we describe the bootstrap procedures used for 2wfe regressions in this paper.[7]

1. (Construct residuals) Obtain $\tilde{v}_{it}$ from the following auxiliary regression:

$$y_{it} = \tilde{c}_i + \tilde{\tau}_t + \delta 1_{\{i\in E\}} 1_{\{t>T_{0i}\}} + \tilde{b}' x_{it} + \tilde{v}_{it},$$

where $(\tilde{c}_i, \tilde{\tau}_t, \tilde{b}')$ are the OLS estimates of the regression of y_{it} on $1_{\{i\in E\}} 1_{\{t>T_{0i}\}}$ and x_{it} having unit and time fixed effects, with H_0 imposed (i.e., restricting $\delta = \delta_0$).

[7] For comparability, we try to maintain a symmetry with the PCDID bootstrap procedures. The 2wfe bootstraps are similar to those used in the relevant literature (e.g., Bertrand et al. (2004), Cameron et al. (2008), Mackinnon and Webb (2017)).

2. Do B iterations of this step. On the b^{th} iteration, perform wild cluster bootstrap as follows. For $i \in C \cup E$, let $r_i^{(b)}$ be iid variables with a Rademacher distribution. Construct:

$$y_{it}^{(b)} = \tilde{c}_i + \tilde{\tau}_t + \delta_0 1_{\{i \in E\}} 1_{\{t > T_{0i}\}} + \tilde{b}' x_{it} + \tilde{v}_{it} r_i^{(b)},$$

   Compute the 2wfe estimate $\hat{\delta}^{(b)}$ from the bootstrap sample $\{(y_{it}^{(b)}, 1_{\{i \in E\}} 1_{\{t > T_{0i}\}}, x_{it})\}_{i \in C \cup E; t=1,..,T}$.

For micro-level data, first obtain residuals $\tilde{\nu}_{igt}$ from the auxiliary regression $y_{igt} = \tilde{c}_g + \tilde{\tau}_t + \delta 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \tilde{b}' x_{igt} + \tilde{v}_{igt}$, where $(\tilde{c}_g, \tilde{\tau}_t, \tilde{b}')$ are the OLS estimates of the regression of y_{igt} on $1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}}$ and x_{igt} having group and time fixed effects, with H_0 imposed (i.e., restricting $\delta = \delta_0$). Then, do B iteration of the following step. On the b^{th} iteration, perform wild cluster bootstrap: for $g \in \mathscr{C} \cup \mathscr{E}$, let $r_g^{(b)}$ be iid variables with a Rademacher distribution. Construct $y_{igt}^{(b)} = \tilde{c}_g + \tilde{\tau}_t + \delta_0 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}} + \tilde{b}' x_{igt} + \tilde{v}_{igt} r_g^{(b)}$. Compute the 2wfe estimate $\hat{\delta}^{(b)}$ from the bootstrap sample $\{(y_{igt}^{(b)}, 1_{\{g \in \mathscr{E}\}} 1_{\{t > T_{0g}\}}, x_{igt})\}_{i=1,..,N; g \in \mathscr{C} \cup \mathscr{E}; t=1,..,T}$.

6.4 Rejection Criterion

Let $H_0 : \theta = \theta_0$, $H_a : \theta \neq \theta_0$, $\hat{\theta}$ be the raw test statistic from the full sample, and $\hat{\theta}^{(b)}$ be the raw test statistic from the b^{th} iteration of the bootstrap sample.

1. Bootstrap t-statistic (b-t) criterion: compute the studentized statistic from the full sample $\hat{t} := \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$, where $se(\hat{\theta})$ is obtained from an appropriate standard error formula (see discussions after theorems 1,2,5). Compute the bootstrap studentized statistic $\hat{t}^{(b)} := \frac{\hat{\theta}^{(b)} - \hat{\theta}}{se(\hat{\theta}^{(b)})}$. Let $s_{[q]}$ denote the qth percentile of the bootstrap sample $\{\hat{t}^{(1)}, \ldots, \hat{t}^{(B)}\}$. Reject H_0 at level a if and only if $\hat{t} < s_{[a/2]}$ or $\hat{t} > s_{[1-a/2]}$.
2. Bootstrap standard error (b-se) criterion: compute the bootstrap standard error $b.s.e. := \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \overline{\hat{\theta}} \right)^2}$, where $\overline{\hat{\theta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$. Then compute the studentized statistic from the full sample $\hat{t}_{bse} = \frac{\hat{\theta} - \theta_0}{b.s.e.}$. Obtain the p-value as $p_{bse} = 2[1 - \Phi(|\hat{t}_{bse}|)]$. Reject H_0 if p_{bse} is smaller than the significance level a.

7 Endogenous Covariates

When deriving the asymptotic theory of PCDID estimation, the restriction that covariates are asymptotically orthogonal to factors is imposed for analytical convenience. In this section, we examine some DGP for endogenous covariates under which the proposed PCDID approach delivers valid inference.

As a motivation, we recall that the factor proxies are extracted by PCA from the covariance matrix $\hat{u}_C \hat{u}_C' / T$ of the residuals from the time series regression of control unit's outcomes on the covariates. The factor proxies then act as the regressors (along with the intervention dummy and covariates) in the regression of the treated units' outcomes.

Now suppose that the covariates follow the DGP given below:

$$X_i = F \Pi_i + V_i,$$

where X_i is the $T \times k$ covariate matrix, F is the $T \times \ell$ factor matrix, Π_i is the $\ell \times k$ matrix of factor loadings, and V_i is the $T \times k$ matrix of idiosyncratic errors. The endogeneity of the covariates is due to their dependence on the factors. This DGP is often included as a special case in interactive effects model commonly found in the literature (Pesaran (2006), Bai (2009)).

In vector form, the DGP is expressed as

$$x_{it} = \Pi_i' f_t + v_{it},$$

where x_{it} and v_{it} are $k \times 1$ vectors, and f_t is a $\ell \times 1$ vector.

We assume that v_{it} are *iid* over i and t, $E(v_{it}) = 0$, $E(v_{it} f_t') = O$, v_{it} are independent of ϵ_{js} for all i, j, s, t, and that $E(v_{it} v_{it}')$ has full rank. Furthermore, we assume that $V_i' V_i / T$ has full rank for large enough T. We restrict our analysis to $I(0)$ factors for simplicity.

It follows that $E(x_{it} x_{it}')$ has full rank because $E(x_{it} x_{it}') = \Pi_i' E(f_t f_t') \Pi_i + E(v_{it} v_{it}')$, which is positive definite as $E(v_{it} v_{it}')$ has full rank. Similarly, $X_i' X_i / T$ has full rank for large enough T. Note that the quadratic form $Q_i := \Pi_i' E(f_t f_t') \Pi_i$ is allowed to be positive semidefinite, for example:

- $k > \ell$ (more covariates than factors). In this case, the $k \times k$ quadratic form Q_i does not have full rank (its rank is at most ℓ).
- $k \leq \ell$ and that some covariates are not driven by factors. In this case, some columns of Π_i are zero vectors, and Q_i does not have full rank.
- $k \leq \ell$ and that some covariates are driven by factors in the same way. In this case, some columns of Π_i are linearly dependent, and Q_i does not have full rank.

We want to show that the proposed PCDID procedure works for this DGP. Recall that the first step is to obtain, for each control unit i, the residuals $\hat{u}_i = M_{X_i} y_i$ from the time series regression of y_i on X_i. The factor proxies are then obtained as weighted averages of $\hat{u}_i$ over control units, where the weights are obtained from the eigenvectors from PCA on $\hat{u}_C \hat{u}_C' / T$. More precisely,

$$\begin{aligned} \hat{F} &= \frac{1}{N_C} \sum_{i \in C} \hat{u}_i w_i = \frac{1}{N_C} \sum_{i \in C} M_{X_i} y_i w_i = \frac{1}{N_C} \sum_{i \in C} M_{X_i} \left(F\mu_i + \epsilon_i \right) w_i \\ &= \frac{1}{N_C} \sum_{i \in C} M_{X_i} F \mu_i w_i + \frac{1}{N_C} \sum_{i \in C} M_{X_i} \epsilon_i w_i =: (I) + (II). \end{aligned}$$

By the orthogonality of X_i and ϵ_i (as F is orthogonal to ϵ_i by Assumption E(i), and V_i is independent of ϵ_i by the DGP assumption), the second term (II) reduces to ℓ columns of weighted averages of ϵ_i ($\frac{1}{N_C} \sum_{i \in C} \epsilon_i w_i$) which all vanish to zero for large N_C. The first term (I) consists of ℓ columns of weighted averages of $M_{X_i} F \mu_i$. By the assumption on V_i, we see that X_i and hence $G_i := [F, X_i]$ have full column rank for large T, satisfying Assumption F(i). It follows that the Frisch-Waugh-Lovell theorem ensures the existence of $(F' M_{X_i} F)^{-1}$, which means that $F' M_{X_i} F = F' M_{X_i} M_{X_i} F$ has full rank, i.e., $rank(M_{X_i} F) = rank(F) = \ell$. Since V_i are independent over i, the $T \times \ell$ matrix of weighted averages, $(I) = \frac{1}{N_C} \sum_{i \in C} M_{X_i} F \mu_i w_i$, has full column rank (i.e., $rank(\hat{F}) = \ell$) for large enough N_C. This establishes our main result that the factor proxies have full rank when covariates are endogenous. In addition, since V_i are independent of X_j for $j \in E$, the regressor matrix $G_j := [\hat{F}, X_j]$ in the PCDID regression has full column rank (i.e., $rank(G_j) = \ell + k$) for large enough N_C and T, and so the PCDID estimator is well defined and can identify ITET asymptotically.

8 A Recursive Approach to Factor Extraction

The following recursive procedure constructs factors proxies when the true number of factors is unknown and when the normalization orders of factors are heterogeneous and integer-valued.

1. For each $i \in C$, run the time series regression $y_i = \varsigma_i + X_i \beta_i + u_i$. Obtain the residuals $\hat{u}_i$ and form the $N_C \times T$ residual matrix $\hat{u}_C$.
2. Set the maximum possible number of factors as $kmax < min(N_C, T)$. Set the largest possible normalization order as the integer $jmax$ (set $jmax = 0$ under full stationarity). Initialize $j = jmax$ and $kmax^{(j)} = kmax$. Denote the residuals obtained in step 1 by $u^{(}j)_i$.

3. Using a consistent procedure (e.g., Ahn and Horenstein (2013)), determine $p_j \leq kmax^{(j)}$, the number of orthonormal eigenvectors extracted from $S^{(j)} := \frac{1}{T^{(2j)\vee 1}}\hat{u}_C\hat{u}_C'$. Store them as columns in $W^{(j)}$. Obtain factor proxies $\hat{F}^{(j)} := \frac{\hat{u}_C' W^{(j)}}{N_C}$. If $p_j = kmax^{(j)}$, then STOP.
4. Repeat this step until reaching STOP:
   (a) If $j > 0$, then do the following. Set $j = j-1$. Then set $kmax^{(j)} = kmax^{(j+1)} - p_{j+1}$. For each $i \in C$, run the time series regression $\hat{u}_i^{(j+1)} = \varsigma_i + \hat{F}^{(j+1)}\alpha_i + u_i$. Obtain the residuals $\hat{u}_i^{(j)}$ and form the $N_C \times T$ residual matrix $\hat{u}_C^{(j)}$. Using the same consistent procedure as in step 3, determine $p_j \leq kmax^{(j)}$, the number of orthonormal eigenvectors extracted from $S^{(j)} := \frac{1}{T^{(2j)\vee 1}}\hat{u}_C^{(j)}\hat{u}_C^{(j)\prime}$. Store them as columns in $W^{(j)}$. Obtain factor proxies $\hat{F}^{(j)} := \frac{\hat{u}_C' W^{(j)}}{N_C}$. If $p_j = kmax^{(j)}$, then STOP.
   (b) If $j = 0$, then STOP.

The recursive procedure above generates $p = \sum_{j=0}^{jmax} p_j$ distinct factor proxies given as columns in $\hat{F} := [\hat{F}^{(jmax)}, \hat{F}^{(jmax-1)}, \ldots, \hat{F}^{(0)}]$. This procedure is conservative in that p is overestimated when the researcher sets a $jmax$ that is higher than the maximum normalization order in the DGP.

The Ahn and Horenstein (2013) eigenvalue-ratio (ER) and growth-ratio (GR) tests are implemented as follows. Both tests require the researcher to set $kmax$. Given j, let $s_k^{(j)}$ be the kth largest eigenvalue extracted from $S^{(j)}$. Let $ER^{(j)}(k) \equiv \frac{s_k^{(j)}}{s_{k+1}^{(j)}}$ be the ratio of the kth and $(k{+}1)$th largest eigenvalues. The ER test yields $p_j = \text{argmax}_{1\leq k\leq kmax^{(j)}} ER^{(j)}(k)$. In the GR test, let $V^{(j)}(k) = \sum_{r=k+1}^{\min(N_C,T)} s_r^{(j)}$ be the sum of all eigenvalues smaller than $s_k^{(j)}$. Let $GR^{(j)}(k) \equiv \frac{\ln[V^{(j)}(k-1)/V^{(j)}(k)]}{\ln[V^{(j)}(k)/V^{(j)}(k+1)]}$. The GR test yields $p_j = \text{argmax}_{1\leq k\leq kmax^{(j)}} GR^{(j)}(k)$. To address the one-factor bias in finite samples, a conservative version of the tests involve resetting $p_j = \lceil \frac{kmax^{(j)}}{2} \rceil$ whenever $p_j = 1$.

9 Proofs

9.1 Proof of Lemma A1

Without loss of generality, we reorder the factors (columns of F) in descending order of their normalization orders. Let $R = \{r_{\min}, \ldots, r_{\max}\}$ be the set of all possible normalization orders, and let ℓ_r be the number of factors having the normalization order r. In the proof we will use the same notations as introduced in the section on factor extraction.

We start with the largest normalization order $r = r_{\max}$. The eigenequation is

$$\frac{1}{N_C T^{2r}}\hat{u}_C\hat{u}_C' W^{(r)} = W^{(r)}V^{(r)}, \tag{14}$$

where $V^{(r)}$ is an $\ell_r \times \ell_r$ diagonal matrix storing ℓ_r eigenvalues in descending order from top left to bottom right, and $W^{(r)}$ is an $N_C \times \ell_r$ eigenvector matrix storing the ℓ_r associated eigenvectors as columns. By orthonormality, we have $W^{(r)\prime}W^{(r)}/N_C = I$, so pre-multiplying (14) both sides by $N_C^{-1}W^{(r)\prime}$ gives

$$V^{(r)} = \frac{1}{N_C^2 T^{2r}} W^{(r)\prime}\hat{u}_C\hat{u}_C' W^{(r)}.$$

We want to show that $V^{(r)}$ is a full rank matrix. Let $F^{(r)}$ denote the $T \times \ell_r$ matrix of factors with normalization order r, and let $\mu^{(r)}$ be the associated $N_C \times \ell_r$ factor loading vector. Let $w_i^{(r)}$ be the

i^{th} row of $W^{(r)}$, and $\mu_i^{(r)}$ the i^{th} row of $\mu^{(r)}$. Using the definition of $\hat{u}_C$, we obtain

$$V^{(r)} = \sum_{i\in C}\sum_{j\in C} \frac{w_i^{(r)\prime}\mu_i^{(r)\prime}}{N_C}\frac{F^{(r)\prime}M_{X_i}M_{X_j}F^{(r)}}{T^{2r}}\frac{\mu_j^{(r)}w_j^{(r)}}{N_C} + \sum_{i\in C}\sum_{j\in C}\frac{1}{N_C^2T^{2r}}w_i^{(r)\prime}\epsilon_i M_{X_i}M_{X_j}\epsilon_j' w_j^{(r)}.$$

The first term is

$$\begin{aligned}
&\sum_{i\in C}\sum_{j\in C} \frac{w_i^{(r)\prime}\mu_i^{(r)\prime}}{N_C}\frac{F^{(r)\prime}M_{X_i}M_{X_j}F^{(r)}}{T^{2r}}\frac{\mu_j^{(r)}w_j^{(r)}}{N_C}\\
&= \frac{W^{(r)\prime}\mu_C^{(r)}}{N_C}\frac{F^{(r)\prime}F^{(r)}}{T^{2r}}\frac{\mu_C^{(r)\prime}W^{(r)}}{N_C}\\
&\quad - \sum_{i\in C}\sum_{j\in C}\frac{w_i^{(r)\prime}\mu_i^{(r)\prime}}{N_C}\frac{F^{(r)\prime}X_i}{T^{r+0.5}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'X_j}{T}\left(\frac{X_j'X_j}{T}\right)^{-1}\frac{X_j'F^{(r)}}{T^{r+0.5}}\frac{\mu_j^{(r)}w_j^{(r)}}{N_C}\\
&= \frac{W^{(r)\prime}\mu_C^{(r)}}{N_C}\frac{F^{(r)\prime}F^{(r)}}{T^{2r}}\frac{\mu_C^{(r)\prime}W^{(r)}}{N_C} + o_p(1_\ell 1_\ell'),
\end{aligned}$$

where the last line follows from Assumptions MX(i)-(iii). The second term is

$$\begin{aligned}
&\sum_{i\in C}\sum_{j\in C} w_i^{(r)\prime}\frac{\epsilon_i M_{X_i}M_{X_j}\epsilon_j'}{N_C^2T^{2r}}w_j^{(r)}\\
&= \frac{W^{(r)\prime}\epsilon_C\epsilon_C' W^{(r)}}{N_C^2T^{2r}} - \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C} w_i^{(r)\prime}\frac{\epsilon_i X_i}{T^{r+0.5}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'X_j}{T}\left(\frac{X_j'X_j}{T}\right)^{-1}\frac{X_j'\epsilon_j'}{T^{r+0.5}}w_j^{(r)}\\
&= \frac{W^{(r)\prime}\epsilon_C\epsilon_C' W^{(r)}}{N_C^2T^{2r}} + o_p(1_\ell 1_\ell'),
\end{aligned}$$

where the last line follows from Assumptions MX(i), (ii) and (iv). This implies that

$$V^{(r)} = \frac{W^{(r)\prime}\mu_C^{(r)}}{N_C}\frac{F^{(r)\prime}F^{(r)}}{T^{2r}}\frac{\mu_C^{(r)\prime}W^{(r)}}{N_C} + \frac{W^{(r)\prime}\epsilon_C\epsilon_C' W^{(r)}}{N_C^2T^{2r}} + o_p(1_\ell 1_\ell')$$

as $N_C, T \to \infty$. Note that the two dominating terms of $V^{(r)}$ on the right are clearly non-negative definite. Furthermore, by applying lemma A.3 of Bai (2003) after interchanging the roles of factors and factor loadings, we see that both $V^{(r)}$ and $\frac{W^{(r)\prime}\mu_C^{(r)}}{N_C}\frac{F^{(r)\prime}F^{(r)}}{T^{2r}}\frac{\mu_C^{(r)\prime}W^{(r)}}{N_C}$ converge in probability to a full-rank matrix $\text{plim}_{N_C,T\to\infty} V^{(r)}$. By the continuity of eigenvalues, it follows that $V^{(j)}$ has full rank and $W^{(r)}$ has full column rank for large enough N_C and T.

Provided that $r > r_{\min}$, we apply the same arguments by replacing r by the next largest $r \in R$ and solving for the next ℓ_r eigenvalues that form $V^{(r)}$ and the associated eigenvectors that form $W^{(r)}$. We repeat the above procedures until reaching $r = r_{\min}$. By constructing the diagonal matrix V using all the eigenvalues in $V^{(r_{\max})}, \ldots, V^{(r_{\min})}$ arranged in descending order, and defining $W = [W^{(r_{\max})}, \ldots, W^{(r_{\min})}]$, we deduce that both V and W have full column rank in the limit in the sense that $rank(V) = rank(W) \xrightarrow{p} \ell := \sum_{r\in R}\ell_r$ as $N_C, T \to \infty$.

9.2 Proof of Lemma A2

For each r, we know by lemma A1 and the continuity of eigenvalues that the inverse $(V^{(r)})^{-1}$ exists for large N_C and T. Together with Assumption F and lemma A1, we see that the $\ell_r \times \ell_r$ rotation matrix $H^{(r)}$, $H^{(r)} := \frac{F^{(r)\prime}F^{(r)}}{T^{2r}}\frac{\mu_C^{(r)\prime}W^{(r)}}{N_C}(V^{(r)})^{-1}$ has full rank for large N_C and T. Construct W and V as in lemma A1, and let H be the $\ell \times \ell$ block diagonal matrix with block matrices $H^{(r_{\max})}, \ldots, H^{(r_{\min})}$. We then have $H = \Upsilon^{-1}F'F\frac{\mu_C'W}{N_C}V^{-1}\Upsilon^{-1}$. Since each block matrix has full rank in the limit, we see that $rank(H) = \ell$ for large N_C and T.

9.3 Proof of Lemma A3

Let $z_{iq} := (w_{iq} - \mu_i' h_q)$, the $(i,q)^{th}$ element of $(W - \mu_C H)$.
(a) The proof is obtained by switching the role of factors and factor loadings and adapting the proof of Theorem 1 of Bai and Ng (2002) to the $N_C \times T$ control group panel.
(b) The $(p,q)^{th}$ element of $\frac{W'(W-\mu_C H)}{N_C}$ is

$$\frac{1}{N_C}\sum_{i\in C} w_{ip}(w_{iq} - \mu_i' h_q) \le \left(\frac{1}{N_C}\sum_{i\in C} w_{ip}^2\right)^{1/2}\left(\frac{1}{N_C}\sum_{i\in C} z_{iq}^2\right)^{1/2},$$

which is $O_p\left(\frac{1}{\min(\sqrt{N_C},\sqrt{T})}\right)$ by part (a).
(c) The $(p,q)^{th}$ element of $\frac{(W-\mu_C H)'\epsilon_C\epsilon_C'(W-\mu_C H)}{TN_C^2}$ is

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N_C}\sum_{i\in C}\epsilon_{it} z_{ip}\right)\left(\frac{1}{N_C}\sum_{j\in C}\epsilon_{jt} z_{jq}\right)\\
&\le \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N_C}\sum_{i\in C}\epsilon_{it}^2\right)\left(\frac{1}{N_C}\sum_{i\in C} z_{ip}^2\right)^{1/2}\left(\frac{1}{N_C}\sum_{j\in C} z_{jq}^2\right)^{1/2}.
\end{aligned}$$

By Assumption IE(iv), we see that $\frac{1}{TN_C}\sum_{t=1}^{T}\sum_{i\in C} E(\epsilon_{it}^2) \le \frac{1}{TN_C}\sum_{t=1}^{T}\sum_{i\in C}\sigma_{ii,tt} \le c$. The result follows by applying the result of part (a).
(d) We consider the decomposition

$$\begin{aligned}
W'\epsilon_C\epsilon_C'W &= H'\mu_C\epsilon_C\epsilon_C'\mu_C'H + (W-\mu_C H)'\epsilon_C\epsilon_C'(W-\mu_C H)\\
&\quad + H'\mu_C\epsilon_C\epsilon_C'(W-\mu_C H) + (W-\mu_C H)'\epsilon_C\epsilon_C'\mu_C H.
\end{aligned}$$

The $(p,q)^{th}$ element of $(I) := \frac{\mu_C'\epsilon_C\epsilon_C\mu_C}{TN_C^2}$ is given by

$$\frac{1}{TN_C^2}\sum_{t=1}^{T}\sum_{i\in C}\sum_{j\in C}\mu_{ip}\epsilon_{it}\epsilon_{jt}\mu_{jq} = \frac{1}{TN_C}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{N_C}}\sum_{i\in C}\mu_{ip}\epsilon_{it}\right)\left(\frac{1}{\sqrt{N_C}}\sum_{j\in C}\mu_{jq}\epsilon_{jt}\right)$$

which is $O_p\left(\frac{1}{N_C}\right)$ by Assumption M(i).

By part (c), we know that $(II) := \frac{(W-\mu_C H)'\epsilon_C\epsilon_C'(W-\mu_C H)}{TN_C^2}$ is $O_p\left(\frac{1}{\min(N_C,T)}\right)$.

The $(p,q)^{th}$ element of $(III) := \frac{\mu_C'\epsilon_C\epsilon_C'(W-\mu_C H)}{TN_C^2}$ is

$$\frac{1}{TN_C^2}\sum_{i\in C}\sum_{j\in C}\sum_{t=1}^{T}\mu_{ip}\epsilon_{it}\epsilon_{jt}z_{jq} \le \left[\frac{1}{N_C}\sum_{i\in C}\left(\frac{1}{TN_C}\sum_{i\in C}\sum_{t=1}^{T}\mu_{ip}\epsilon_{it}\epsilon_{jt}\right)^2\right]^{1/2}\left(\frac{1}{N_C}\sum_{j\in C} z_{jq}^2\right)^{1/2}.$$

The second factor on the right is $O_p\left(\frac{1}{\min(\sqrt{N_C},\sqrt{T})}\right)$ by part (a), whereas the first factor on the right is $O_p\left(\frac{1}{\sqrt{N_C}}\right)$. To show the latter, we observe that, for fixed $j \in C$,

$$\begin{aligned}
&\frac{1}{N_C}\sum_{i\in C} E\left(\frac{1}{TN_C}\sum_{t=1}^{T}\sum_{i\in C}\mu_{ip}\epsilon_{it}\epsilon_{jt}\right)^2 \\
&= \frac{1}{N_C}\sum_{i\in C} E\left(\frac{1}{TN_C}\sum_{t=1}^{T}\sum_{i\in C}\mu_{ip}[\epsilon_{it}\epsilon_{jt} - E(\epsilon_{it}\epsilon_{jt})] + \frac{1}{N_C}\sum_{i\in C}\mu_{ip}\gamma_T(i,j)\right)^2 \\
&\leq \frac{2}{N_C}\sum_{i\in C} E\left(\frac{1}{TN_C}\sum_{t=1}^{T}\sum_{i\in C}\mu_{ip}[\epsilon_{it}\epsilon_{jt} - E(\epsilon_{it}\epsilon_{jt})]\right)^2 + \frac{2}{N_C}\sum_{i\in C} E\left(\frac{1}{N_C}\sum_{i\in C}\mu_{ip}\gamma_T(i,j)\right)^2 \\
&\leq \frac{2c_1}{TN_C} + \frac{2}{N_C}\sum_{i\in C}\left(\frac{1}{N_C}\sum_{i\in C}E(\mu_{ip}^2)\right)\left(\frac{1}{N_C}\sum_{i\in C}\gamma_T^2(i,j)\right) \\
&\leq \frac{2c_1}{TN_C} + \frac{2}{N_C}\sum_{i\in C} c_2 \cdot \frac{c_3}{N_C} = \frac{2c_1}{TN_C} + \frac{2c_2c_3}{N_C}.
\end{aligned}$$

The second inequality is by Assumption M(iv) and Cauchy-Schwarz inequality, while the third inequality is by Assumptions FL_C(i) and IE(ii) and the fact that absolute summability implies square summability. We therefore deduce that (III) is $O_p\left(\frac{1}{\sqrt{N_C}\min(\sqrt{N_C},\sqrt{T})}\right)$.

The last term $(IV) := \frac{(W-\mu_C H)'\epsilon_C\epsilon_C'\mu_C H}{TN_C^2}$ is the transpose of (III) and hence of the same order. In summary, we conclude that

$$\begin{aligned}
\left\|\frac{W'\epsilon_C\epsilon_C' W}{TN_C^2}\right\| &= O_p\left(\frac{1}{N_C}\right) + O_p\left(\frac{1}{\min(N_C,T)}\right) + O_p\left(\frac{1}{\sqrt{N_C}\min(\sqrt{N_C},\sqrt{T})}\right) \\
&= O_p\left(\frac{1}{\min(N_C,T)}\right).
\end{aligned}$$

(e) We first decompose $\Upsilon^{-1}\frac{F'\epsilon_C' W}{N_C}$ as follows:

$$\Upsilon^{-1}\frac{F'\epsilon_C' W}{N_C} = \Upsilon^{-1}\frac{F'\epsilon_C'\mu_C}{N_C}H + \Upsilon^{-1}\frac{F\epsilon_C'(W-\mu_C H)}{N_C}.$$

Ignoring H, the term $\Upsilon^{-1}\frac{F'\epsilon_C'\mu_C}{N_C}$ is $O_p\left(\frac{1}{\sqrt{N_C}}\right)$ by Assumption M(iii). The $(p,q)^{th}$ element of $\Upsilon^{-1}\frac{F'\epsilon_C'(W-\mu_C H)}{N_C}$ is bounded from above as follows:

$$\begin{aligned}
\frac{1}{T^{r_p}N_C}\sum_{i\in C}\sum_{t=1}^{T} f_{tp}\epsilon_{it}(w_{iq} - \mu_i' h_q) &\leq \left[\frac{1}{N_C}\sum_{i\in C}\left(\frac{1}{T^{r_p}}\sum_{t=1}^{T} f_{tp}\epsilon_{it}\right)^2\right]^{1/2}\left(\frac{1}{N_C}\sum_{i\in C}(w_{iq} - \mu_i' h_q)^2\right)^{1/2} \\
&\leq O_p(1) O_p\left(\frac{1}{\min(\sqrt{N_C},\sqrt{T})}\right),
\end{aligned}$$

where the last inequality follows from Assumption M(ii) and part (a) of the lemma. It then follows that the elements of $\Upsilon^{-1}\frac{F'\epsilon_C' W}{N_C}$ are $O_p\left(\frac{1}{\sqrt{N_C}}\right) + O_p\left(\frac{1}{\min(\sqrt{N_C},\sqrt{T})}\right) = O_p\left(\frac{1}{\min(\sqrt{N_C},\sqrt{T})}\right)$.

9.4 Proof of Lemma A4

Let 1_ℓ be the $\ell \times 1$ vector of ones.

In several parts of the proof, we will make use of the facts:

$$\left\|\sqrt{T}\Upsilon^{-1}\right\| \leq 1 \text{ as } T \to \infty. \tag{15}$$

$$\left\|\frac{W'\mu_C}{N_C} - H^{-1}\right\| = o_p(1) \text{ as } N_C, T \to \infty. \tag{16}$$

The bound (15) follows from Assumption F, which restricts the minimum normalization order of the factors to be $r_{\min} = 0.5$. To show (16), we consider the decomposition $\mu_C = WH^{-1} - (W - \mu_C H)H^{-1}$. Note the matrix H is invertible by lemma A2. Using the decomposition, we obtain

$$\begin{aligned}\frac{W'\mu_C}{N_C} - H^{-1} &= \frac{1}{N_C}\left[W'W - W'(W - \mu_C H)\right] H^{-1} - H^{-1} \\ &= \left[I - \frac{W'(W - \mu_C H)}{N_C}\right] H^{-1} - H^{-1} \\ &= \frac{W'(W - \mu_C H)}{N_C} H^{-1},\end{aligned}$$

where I is the $N_C \times N_C$ identity matrix. The result (16) follows from $\left\|\frac{W'(W-\mu_C H)}{N_C}\right\| = o_p(1)$ (by lemma A3(b)) and $\|H\| > 0$ (as H is non-singular by lemma A2).

The matrix norm properties will be useful for the computation to be carried out in the proof.[8]

(a) We decompose the expression as follows:

$$\begin{aligned}&\sup_{j\in E}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C} 1'_{post,j} M_{X_i} F\mu_i w_i \Upsilon^{-1} - \frac{1}{\sqrt{T}} 1'_{post,j} F\Upsilon^{-1}H'^{-1}\right\| \\ \leq\ &\sup_{j\in E}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C} 1'_{post,j} M_{X_i} F\mu_i w_i \Upsilon^{-1} - \frac{1}{\sqrt{T}} 1'_{post,j} F\frac{\mu'_C W}{N_C}\Upsilon^{-1}\right\| \\ &+ \sup_{j\in E}\left\|\frac{1}{\sqrt{T}} 1'_{post,j} F\left(\frac{\mu'_C W}{N_C}\Upsilon^{-1} - \Upsilon^{-1}H'^{-1}\right)\right\| \\ =:\ &(I) + (II).\end{aligned}$$

To show that (I) is $o_p(1)$, we observe the following facts: (i) $\sup_{i\in C}\left\|X_i' F\Upsilon^{-1}\right\| = O_p(1)$ by Assumption MX(iii); (ii) $\inf_{i\in C}\frac{X_i'X_i}{T}$ is positive definite by Assumption MX(ii); (iii) $\sup_{i\in C, j\in E}\left\|T^{-1}1'_{post,j}X_i\right\| = O_p(1)$, as

$$E\sup_{i\in C, j\in E}\left\|\frac{1'_{post,j}X_i}{T}\right\|^2 \leq E\sup_{i\in C}\left\|\frac{1'X_i}{T}\right\|^2 \leq E\sup_{i\in C}\left\|\frac{X_i'X_i}{T}\right\| \leq c$$

by Assumption MX(i) and Cauchy-Schwarz inequality; and (iv) $\frac{\mu'_C W}{N_C} = O_p(1)$ as

$$E\left\|\frac{\mu'_C W}{N_C}\right\|^2 \leq E\left\|\frac{\mu'_C\mu_C}{N_C}\right\| \leq c \tag{17}$$

by Assumption FL$_C$(i), Cauchy-Schwarz inequality and that $W'W/N_C = I$. We therefore deduce that

$$\begin{aligned}\|(I)\| &= \sup_{j\in E}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C} 1'_{post,j} P_{X_i} F\mu_i w_i\Upsilon^{-1}\right\| \\ &\leq \frac{1}{\sqrt{T}}\sup_{j\in E}\left\|\frac{1}{N_C}\sum_{i\in C}\frac{1'_{post,j}X_i}{T}\left(\frac{X_i'X_i}{T}\right)^{-1} X_i'F\mu_i w_i\Upsilon^{-1}\right\| \\ &\leq \frac{1}{\sqrt{T}}\sup_{i\in C, j\in E}\left\|\frac{1'_{post,j}X_i}{T}\right\|\cdot\inf_{i\in C}\left\|\frac{X_i'X_i}{T}\right\|^{-1}\cdot\sup_{i\in C}\left\|X_i'F\Upsilon^{-1}\right\|\cdot\left\|\frac{\mu'_C W}{N_C}\right\|,\end{aligned}$$

[8] Given the trace norm $\|A\| := [trace(A'A)]^{1/2}$, the following relationships hold for any $\ell \times \ell$ square matrices A and B: $\|AB\| \leq \|A\|\,\|B\|$, $\|A + B\| \leq \|A\| + \|B\|$, and $\|A\| = \|A'\|$.

which is $O_p\left(\frac{1}{\sqrt{T}}\right)$. Next we turn to (II). Given the block diagonal property of H'^{-1} (lemma A2) and Υ^{-1}, we observe that $\Upsilon^{-1}H'^{-1} = H'^{-1}\Upsilon^{-1}$. Let us compute:

$$\begin{aligned}\left\|\frac{1}{\sqrt{T}}1'_{post,j}F\left(\frac{\mu_C'W}{N_C}\Upsilon^{-1} - \Upsilon^{-1}H'^{-1}\right)\right\|^2 &= \left\|\frac{1}{\sqrt{T}}1'_{post,j}F\left(\frac{\mu_C'W}{N_C}\Upsilon^{-1} - H'^{-1}\Upsilon^{-1}\right)\right\|^2 \\ &\leq \left\|\frac{1}{\sqrt{T}}1'F\right\|^2 \left\|\frac{\mu_C'W}{N_C} - H'^{-1}\right\|^2 \|\Upsilon^{-1}\|^2 \\ &\leq \left\|\frac{1'1}{T}\right\| \|F'F\| \left\|\frac{\mu_C'W}{N_C} - H'^{-1}\right\|^2 \|\Upsilon^{-1}\|^2 \\ &= \left\|\frac{\mu_C'W}{N_C} - H'^{-1}\right\|^2 \\ &= o_p(1),\end{aligned}$$

by (16) and Cauchy-Schwarz inequality. It follows that $\|(II)\|$ is $o_p(1)$. The proof is now completed.
(b) Using the decomposition $w_i = \mu_i'H + (w_i - \mu_i'H)$, we have

$$\begin{aligned}\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i w_i &= \frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i\mu_i'H + \frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i(w_i - \mu_i'H) \\ &=: (I) + (II).\end{aligned}$$

To evaluate (I), we first note that $\frac{\mu_C'W}{N_C} = O_p(1)$ as

$$E\left\|\frac{\mu_C'W}{N_C}\right\|^2 \leq E\left\|\frac{\mu_C'\mu_C}{N_C}\right\| \leq c \tag{18}$$

by Assumption FL$_C$(i), Cauchy-Schwarz inequality and that $W'W/N_C = I$. It follows that

$$\|H\| = \left\|\Upsilon^{-1}F'F\frac{\mu_C'W}{N_C}V^{-1}\Upsilon^{-1}\right\| \leq \|\Upsilon^{-1}F'F\Upsilon^{-1}\| \left\|\frac{\mu_C'W}{N_C}\right\| \|V^{-1}\| = O_p(1), \tag{19}$$

where the last step follows from Assumption F(i), equation (18), and that V is non-singular for large N_C and T by lemma A1. We thus obtain

$$\|(I)\| = \left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i\mu_i'H\right\| = \left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i\mu_i'\right\| \|H\| = O_p\left(\frac{1}{\sqrt{N_C}}\right)O_p(1)$$

by Assumption MX(vi). Now let us turn to (II). By Cauchy-Schwarz inequality, the $(p,q)^{th}$ element of (II) is

$$\frac{1}{N_C\sqrt{T}}\sum_{i\in C}\sum_{t=1}^{T}x_{pit}\epsilon_{it}(w_{iq} - \mu_i'h_q) \leq \left[\frac{1}{N_C}\sum_{i\in C}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{pit}\epsilon_{it}\right)^2\right]^{1/2}\left[\frac{1}{N_C}\sum_{i\in C}(w_i - \mu_i'h_q)^2\right]^{1/2}.$$

Since

$$E\left[\frac{1}{N_C}\sum_{i\in C}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{pit}\epsilon_{it}\right)^2\right] = \frac{1}{TN_C}\sum_{i\in C}\sum_{s=1}^{T}\sum_{t=1}^{T}E(x_{pis}x_{pit}\epsilon_{is}\epsilon_{it}) \leq c$$

by Assumption MX(v), lemma A3(a) implies that the elements of (II) are $O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)$. Combining the results, we conclude that all elements of $\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i w_i$ are $O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)$.

(c) Let us decompose the term

$$\frac{1}{N_C\sqrt{T}}\sum_{i\in C}1'_{post,j}M_{X_i}\epsilon_i w_i\Upsilon^{-1}=\frac{1'_{post,j}\epsilon'_C W}{N_C\sqrt{T}}\Upsilon^{-1}-\frac{1}{N_C\sqrt{T}}\sum_{i\in C}1'_{post,j}P_{X_i}\epsilon_i w_i\Upsilon^{-1}$$
$$=: (I)+(II).$$

First, we evaluate $\sup_{j\in E}\|(I)\|$:

$$\begin{aligned}\sup_{j\in E}\left\|\frac{1'_{post,j}\epsilon'_C W}{N_C\sqrt{T}}\Upsilon^{-1}\right\| &\le \left\|\frac{1'\epsilon'_C W}{N_C\sqrt{T}}\Upsilon^{-1}\right\|=\left\|\frac{1'\epsilon'_C W}{N_C T}\right\|\left\|\sqrt{T}\Upsilon^{-1}\right\|\\ &\le \left\|\frac{1'1}{T}\right\|^{1/2}\left\|\frac{W'\epsilon_C\epsilon'_C W}{TN_C^2}\right\|^{1/2}\left\|\sqrt{T}\Upsilon^{-1}\right\|\\ &\le O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right),\end{aligned}$$

which follows from Cauchy-Schwarz inequality, lemma A3(d) and (15). Next, let us compute $\sup_{j\in E}\|(II)\|$ as follows:

$$\begin{aligned}\sup_{j\in E}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}1'_{post,j}P_{X_i}\epsilon_i w_i\Upsilon^{-1}\right\| &= \left\|\frac{1}{N_C}\sum_{i\in C}\frac{1'X_i}{T}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'\epsilon_i w_i}{\sqrt{T}}\Upsilon^{-1}\right\|\\ &\le \left\|\sup_{i\in C}\frac{1'X_i}{T}\right\|\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i w_i\right\|\left\|\Upsilon^{-1}\right\|\\ &\le \left\|\sup_{i\in C}\frac{X_i'X_i}{T}\right\|^{1/2}\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i w_i\Upsilon^{-1}\right\|\left\|\Upsilon^{-1}\right\|\\ &= O_p(1)O_p(1)O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)O_p\left(\frac{1}{\sqrt{T}}\right),\end{aligned}$$

where the last step is by Assumptions MX(i) and (ii) and part (b). The result follows immediately.
(d) By the decomposition $\hat{F}=\frac{1}{N_C}\sum_{i\in C}M_{X_i}\left(F\mu_i+\epsilon_i\right)w_i$, we obtain

$$\begin{aligned}&\sup_{j\in E}\left\|\frac{1'_{post,j}\hat{F}}{\sqrt{T}}\Upsilon^{-1}-\frac{1}{\sqrt{T}}1'_{post,j}F\Upsilon^{-1}H'^{-1}\right\|\\ &=\sup_{j\in E}\left\|\frac{1}{\sqrt{T}N_C}\sum_{i\in C}1'_{post,j}M_{X_i}\left(F\mu_i+\epsilon_i\right)w_i\Upsilon^{-1}-\frac{1}{\sqrt{T}}1'_{post,j}F\Upsilon^{-1}H'^{-1}\right\|\\ &\le\sup_{j\in E}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}1'_{post,j}M_{X_i}F\mu_i w_i\Upsilon^{-1}-\frac{1}{\sqrt{T}}1'_{post,j}F\Upsilon^{-1}H'^{-1}\right\|\\ &\quad+\sup_{j\in E}\left\|\frac{1}{\sqrt{T}N_C}\sum_{i\in C}1'_{post,j}M_{X_i}\epsilon_i w_i\Upsilon^{-1}\right\|,\end{aligned}$$

which is $o_p(1)$ by parts (a) and (c).
(e) We decompose the term as follows:

$$\frac{X_j'\hat{F}}{\sqrt{T}}\Upsilon^{-1}-\frac{X_j'F}{\sqrt{T}}\Upsilon^{-1}=\left(\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_j'M_{X_i}F\mu_i w_i\Upsilon^{-1}-\frac{X_j'F}{\sqrt{T}}\Upsilon^{-1}\right)+\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_j'M_{X_i}\epsilon_i w_i\Upsilon^{-1}$$
$$=: (I)+(II).$$

We first compute the first expression (I):

$$\begin{aligned}\|(I)\| &= \left\| \frac{1}{N_C\sqrt{T}} \sum_{i\in C} X_j' M_{X_i} F \mu_i w_i \Upsilon^{-1} - \frac{X_j' F}{\sqrt{T}} \Upsilon^{-1} \right\| \\ &\leq \sup_{i\in C, j\in E} \left\| \frac{1}{\sqrt{T}} X_j' M_{X_i} F \right\| \left\| \frac{1}{N_C} \sum_{i\in C} \mu_i w_i \right\| \left\| \Upsilon^{-1} \right\| \\ &= \sup_{i\in C, j\in E} \left\| \frac{1}{\sqrt{T}} X_j' M_{X_i} F \right\| \left\| \frac{\mu_C' W}{N_C} \right\| \left\| \Upsilon^{-1} \right\|.\end{aligned}$$

First, note that $\left\| \frac{\mu_C' W}{N_C} \right\| = \left\| H^{-1} \right\| + o_p(1) = O_p(1)$ by (16) and the fact that $\|H\| > 0$ (as H is non-singular by lemma A2). Furthermore, we have

$$E \sup_{i\in C, j\in E} \left\| \frac{1}{\sqrt{T}} X_j' M_{X_i} F \right\|^2 \leq E \sup_{j\in E} \left\| \frac{1}{\sqrt{T}} X_j' F \right\|^2 \leq E \sup_{j\in E} \left\| \frac{X_j' X_j}{T} \right\| E \left\| F'F \right\| \leq c \left\| \Upsilon \right\|^2$$

by Assumption MX(i). It follows that $\|(I)\| = o_p(1)$.

Simplifying the second expression (II), we have

$$\begin{aligned}(II) &= \frac{1}{N_C\sqrt{T}} \sum_{i\in C} X_j' M_{X_i} \epsilon_i w_i \Upsilon^{-1} \\ &= \frac{1}{N_C\sqrt{T}} \sum_{i\in C} X_j' \epsilon_i w_i \Upsilon^{-1} - \frac{1}{N_C} \sum_{i\in C} \frac{X_j' X_i}{T} \left(\frac{X_i' X_i}{T} \right)^{-1} \frac{X_i' \epsilon_i}{\sqrt{T}} w_i \Upsilon^{-1} \\ &= \frac{1}{N_C\sqrt{T}} \sum_{i\in C} X_j' \epsilon_i \mu_i' H \Upsilon^{-1} + \frac{1}{N_C\sqrt{T}} \sum_{i\in C} X_j' \epsilon_i (w_i - \mu_i' H) \Upsilon^{-1} \\ &\quad - \frac{1}{N_C} \sum_{i\in C} \frac{X_j' X_i}{T} \left(\frac{X_i' X_i}{T} \right)^{-1} \frac{X_i' \epsilon_i}{\sqrt{T}} w_i \Upsilon^{-1} \\ &=: (IIa) + (IIb) + (IIc)\end{aligned}$$

where the last equality follows from the decomposition $w_i = \mu_i' H + (w_i - \mu_i' H)$. First we evaluate $\sup_{j\in E} \|(IIa)\|$. Applying Cauchy-Schwarz inequality, we can bound it as follows:

$$\begin{aligned}\sup_{j\in E} \|(IIa)\| &\leq \frac{1}{\sqrt{N_C}} \sup_{j\in E} \left\| \frac{1}{T} X_j' \left(\frac{1}{\sqrt{N_C}} \sum_{i\in C} \epsilon_i \mu_i' \right) \right\| \|H\| \left\| \sqrt{T} \Upsilon^{-1} \right\| \\ &\leq \frac{1}{\sqrt{N_C}} \sup_{j\in E} \left\| \frac{X_j' X_j}{T} \right\|^{1/2} \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \frac{1}{\sqrt{N_C}} \sum_{i\in C} \epsilon_{it} \mu_i \right\|^2 \right)^{1/2} \|H\| \left\| \sqrt{T} \Upsilon^{-1} \right\| \\ &= O_p\left(\frac{1}{\sqrt{N_C}} \right) O_p(1) O_p(1) O_p(1) O_p(1),\end{aligned}$$

where the last step is obtained by Assumptions MX(i) and M(i), (15) and (19). Then we compute $\sup_{j\in E} \|(IIb)\|$. By Cauchy-Schwarz inequality again, we have

$$\begin{aligned}\sup_{j\in E} \|(IIb)\| &\leq \sup_{j\in E} \left\| \frac{1}{T} X_j' \left[\frac{1}{N_C} \sum_{i\in C} \epsilon_i (w_i - \mu_i' H) \right] \right\| \left\| \sqrt{T} \Upsilon^{-1} \right\| \\ &\leq \sup_{j\in E} \left\| \frac{X_j' X_j}{T} \right\|^{1/2} \left\| \frac{(W - \mu_C H)' \epsilon_C \epsilon_C' (W - \mu_C H)}{T N_C^2} \right\|^{1/2} \left\| \sqrt{T} \Upsilon^{-1} \right\| \\ &= O_p(1) O_p\left(\frac{1}{\min(\sqrt{T}, \sqrt{N_C})} \right) O_p(1),\end{aligned}$$

where the last step is obtained by Assumption MX(i) and lemma A3(c). Next we compute $\sup_{j\in E}\|(IIc)\|$. By Cauchy-Schwarz inequality once again, we obtain

$$\begin{aligned}
\sup_{j\in E}\|(IIc)\| &\le \sup_{j\in E}\left\|\frac{1}{N_C}\sum_{i\in C}\frac{X_j'X_i}{T}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'\epsilon_i}{\sqrt{T}}w_i\Upsilon^{-1}\right\| \\
&\le \sup_{i\in C, j\in E}\left\|\frac{X_j'X_i}{T}\right\|\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C}X_i'\epsilon_i w_i\right\|\left\|\Upsilon^{-1}\right\| \\
&\le \sup_{i\in C}\left\|\frac{X_i'X_i}{T}\right\|^{1/2}\sup_{j\in E}\left\|\frac{X_j'X_j}{T}\right\|^{1/2}\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\left\|\sum_{i\in C}\frac{X_i'\epsilon_i w_i}{N_C\sqrt{T}}\right\|\left\|\Upsilon^{-1}\right\| \\
&= O_p(1)O_p(1)O_p(1)O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)O_p\left(\frac{1}{\sqrt{T}}\right),
\end{aligned}$$

where the last step is obtained by Assumptions MX(i)-(ii), lemma A3(c) and (15). We thus see that the elements of (II) are $O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)$. The result immediately follows after combining all the terms.

(f) We decompose the expression as follows:

$$\begin{aligned}
\Upsilon^{-1}\hat{F}'\hat{F}\Upsilon^{-1} &= \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'(F\mu_i+\epsilon_i)'M_{X_i}M_{X_j}(F\mu_j+\epsilon_j)w_j\Upsilon^{-1} \\
&= \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'\mu_i'F'M_{X_i}M_{X_j}F\mu_j w_j\Upsilon^{-1} \\
&\quad + \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'\mu_i'F'M_{X_i}M_{X_j}\epsilon_j w_j\Upsilon^{-1} \\
&\quad + \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'\epsilon_i'M_{X_i}M_{X_j}F\mu_j w_j\Upsilon^{-1} \\
&\quad + \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'\epsilon_i'M_{X_i}M_{X_j}\epsilon_j w_j\Upsilon^{-1} \\
&=: (I)+(II)+(III)+(IV).
\end{aligned}$$

Let us further decompose (I) as follows:

$$\begin{aligned}
(I) &= \Upsilon^{-1}\frac{W'\mu_C}{N_C}F'F\frac{\mu_C'W}{N_C}\Upsilon^{-1} \\
&\quad - \frac{1}{TN_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'\mu_i'F'X_i\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'X_j}{T}\left(\frac{X_j'X_j}{T}\right)^{-1}X_j'F\mu_j w_j\Upsilon^{-1} \\
&=: (Ia)+(Ib).
\end{aligned}$$

The norm of (Ia) is given by

$$\|(Ia)\| = \left\|\Upsilon^{-1}F'F\Upsilon^{-1}\right\|\left\|\frac{W'\mu_C}{N_C}\right\|^2 = O_p(1)$$

by (18) and Assumption F(i). By (18) again and Cauchy-Schwarz inequality, the norm of (Ib) can be bounded from above as follows:

$$\begin{aligned}
&\|(Ib)\| \\
&\leq \frac{1}{TN_C^2}\sum_{i\in C}\sum_{j\in C}\left\|\Upsilon^{-1}w_i'\mu_i'\right\|\left\|F'X_i\right\|\left\|\frac{X_i'X_i}{T}\right\|^{-1}\left\|\frac{X_i'X_i}{T}\right\|^{1/2}\left\|\frac{X_j'X_j}{T}\right\|^{1/2}\left\|\frac{X_j'X_j}{T}\right\|^{-1}\left\|X_j'F\right\|\left\|\mu_j w_j\Upsilon^{-1}\right\| \\
&= \frac{1}{TN_C^2}\sum_{i\in C}\sum_{j\in C}\left\|\Upsilon^{-1}w_i'\mu_i'\right\|\left\|F'X_i\right\|\left\|\frac{X_i'X_i}{T}\right\|^{-1/2}\left\|\frac{X_j'X_j}{T}\right\|^{-1/2}\left\|X_j'F\right\|\left\|\mu_j w_j\Upsilon^{-1}\right\| \\
&\leq \frac{1}{T}\left\|\frac{W'\mu_C}{N_C}\right\|\sup_{i\in C}\left\|\Upsilon^{-1}\frac{F'X_i}{\sqrt{T}}\right\|\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\sup_{j\in C}\left\|\frac{X_j'F}{\sqrt{T}}\Upsilon^{-1}\right\|\left\|\frac{\mu_C'W}{N_C}\right\| \\
&= O_p\left(\frac{1}{T}\right)O_p(1)O_p(1)O_p(1)O_p(1)O_p(1),
\end{aligned}$$

where the last step is by (18) and Assumptions MX(ii) and (iii). We thus obtain

$$\|(I)\| = \left\|\Upsilon^{-1}H^{-1}F'FH'^{-1}\Upsilon^{-1}\right\| + O_p\left(\frac{1}{T}\right).$$

The second expression (II) is rewritten as:

$$\begin{aligned}
(II) &= \Upsilon^{-1}\frac{W'\mu_C}{N_C}\frac{F'\epsilon_C'W}{N_C}\Upsilon^{-1} \\
&\quad - \frac{1}{\sqrt{T}N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}w_i'\mu_i'F'X_i\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'X_j}{T}\left(\frac{X_j'X_j}{T}\right)^{-1}\frac{X_j'\epsilon_j w_j}{\sqrt{T}}\Upsilon^{-1} \\
&=: (IIa) + (IIb).
\end{aligned}$$

The norm of (IIa) is

$$\|(IIa)\| \leq \left\|\Upsilon^{-1}\right\|\left\|\frac{W'\mu_C}{N_C}\right\|\left\|\frac{F'\epsilon_C'W}{N_C}\Upsilon^{-1}\right\| = O_p\left(\frac{1}{\sqrt{T}}\right)O_p(1)O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)$$

by (15), (18) and lemma A3(e). By Cauchy-Schwarz inequality, the norm of (IIb) is bounded from above as follows:

$$\begin{aligned}
\|(IIb)\| &\leq \frac{1}{\sqrt{T}}\left\|\frac{W'\mu_C}{N_C}\right\|\sup_{i\in C}\left\|\Upsilon^{-1}F'X_i\right\|\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1/2}\left\|\inf_{j\in C}\frac{X_j'X_j}{T}\right\|^{-1/2}\left\|\sum_{j\in C}\frac{X_j'\epsilon_j w_j}{N_C\sqrt{T}}\right\|\left\|\Upsilon^{-1}\right\| \\
&\leq O_p\left(\frac{1}{\sqrt{T}}\right)O_p(1)O_p(1)O_p(1)O_p(1)O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)O_p\left(\frac{1}{\sqrt{T}}\right) \\
&= O_p\left(\frac{1}{T\min(\sqrt{T},\sqrt{N_C})}\right),
\end{aligned}$$

where the second-to-last step is obtained by (15), (18), Assumptions MX(ii) and (iii), and part (b). We thus see that the elements of (II) are $O_p\left(\frac{1}{\sqrt{T}\min(\sqrt{T},\sqrt{N_C})}\right)$.

The third expression (III), being the transpose of (II), is of the same order.

The fourth expression (IV) is rewritten as:

$$\begin{aligned}
(IV) &= \Upsilon^{-1}\frac{W'\epsilon_C\epsilon_C'W}{N_C^2}\Upsilon^{-1} - \frac{1}{N_C^2}\sum_{i\in C}\sum_{j\in C}\Upsilon^{-1}\frac{w_i'\epsilon_i X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'X_j}{T}\left(\frac{X_j'X_j}{T}\right)^{-1}\frac{X_j'\epsilon_j w_j}{\sqrt{T}}\Upsilon^{-1} \\
&=: (IVa) + (IVb).
\end{aligned}$$

The norm of (IVa) is $\|(IVa)\| = \left\|\Upsilon^{-1}\sqrt{T}\right\| \left\|\frac{W'\epsilon_C\epsilon_C' W}{TN_C^2}\right\| \left\|\sqrt{T}\Upsilon^{-1}\right\| = O_p\left(\frac{1}{\min(T,N_C)}\right)$ by (15) and lemma A3(d). By Cauchy-Schwarz inequality, the norm of (IVb) is bounded from above by:

$$\begin{aligned}
\|(IVb)\| &= \left\|\Upsilon^{-1}\right\|^2 \left\|\sum_{i\in C} \frac{w_i'\epsilon_i' X_i}{N_C\sqrt{T}}\right\| \left\|\inf_{i\in C} \frac{X_i'X_i}{T}\right\|^{-1/2} \left\|\inf_{j\in C} \frac{X_j'X_j}{T}\right\|^{-1/2} \left\|\sum_{j\in C} \frac{X_j'\epsilon_j w_j}{N_C\sqrt{T}}\right\| \\
&\leq O_p\left(\frac{1}{T}\right) O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right) O_p(1)O_p(1)O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right) \\
&= O_p\left(\frac{1}{T\min(T,N_C)}\right),
\end{aligned}$$

where the second-to-last step is obtained by (15), Assumption MX(ii) and part (b). We thus see that the elements of (IV) are $O_p\left(\frac{1}{\min(T,N_C)}\right)$.

Combining all the terms, we have $\left\|\Upsilon^{-1}\hat{F}'\hat{F}\Upsilon^{-1} - \Upsilon^{-1}H^{-1}F'FH'^{-1}\Upsilon^{-1}\right\|^2 = o_p(1)$. The result immediately follows by noting that $H^{-1}\Upsilon^{-1} = \Upsilon^{-1}H^{-1}$ (due to the block-diagonal nature of H) and that

$$\left\|\Upsilon^{-1}\hat{F}'\hat{F}\Upsilon^{-1} - H^{-1}\Upsilon^{-1}F'F\Upsilon^{-1}H'^{-1}\right\|^2 = \left\|\Upsilon^{-1}\hat{F}'\hat{F}\Upsilon^{-1} - \Upsilon^{-1}H^{-1}F'FH'^{-1}\Upsilon^{-1}\right\|^2.$$

(g) We decompose the expression as follows:

$$\begin{aligned}
\epsilon_j'\hat{F}\Upsilon^{-1} &= \frac{1}{N_C}\sum_{i\in C}\epsilon_j' M_{X_i} F\mu_i w_i \Upsilon^{-1} + \frac{1}{N_C}\sum_{i\in C}\epsilon_j' M_{X_i}\epsilon_i w_i\Upsilon^{-1} \\
&=: (I) + (II).
\end{aligned}$$

We further decompose (I) and obtain

$$\begin{aligned}
(I) &= \frac{1}{N_C}\sum_{i\in C}\epsilon_j' M_{X_i} F\mu_i w_i\Upsilon^{-1} \\
&= \frac{1}{N_C}\sum_{i\in C}\epsilon_j' F\mu_i w_i\Upsilon^{-1} - \frac{1}{\sqrt{T}N_C}\sum_{i\in C}\frac{\epsilon_j' X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1} X_i' F\mu_i w_i\Upsilon^{-1} \\
&= \epsilon_j' F\frac{\mu_C' W}{N_C}\Upsilon^{-1} - \frac{1}{\sqrt{T}N_C}\sum_{i\in C}\frac{\epsilon_j' X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1} X_i' F\mu_i w_i\Upsilon^{-1} \\
&=: (Ia) + (Ib).
\end{aligned}$$

We evaluate $\sup_{j\in E}\|(Ia)\|$ as follows:

$$\sup_{j\in E}\|(Ia)\| \leq \sup_{j\in E}\left\|\epsilon_j' F\Upsilon^{-1}\right\| \left\|\frac{\mu_C' W}{N_C}\right\| = O_p(1)O_p(1)$$

by (18) and Assumption M(ii). Next, we evaluate $\sup_{j\in E}\|(Ib)\|$ as follows:

$$\begin{aligned}
\sup_{j\in E}\|(Ib)\| &\leq \frac{1}{\sqrt{T}}\sup_{i\in C, j\in E}\left\|\frac{\epsilon_j' X_i}{\sqrt{T}}\right\| \left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1} \sup_{i\in C, j\in E}\left\|X_i' F\Upsilon^{-1}\right\| \left\|\frac{\mu_C' W}{N_C}\right\| \\
&\leq O_p\left(\frac{1}{\sqrt{T}}\right) O_p(1)O_p(1)O_p(1)O_p(1),
\end{aligned}$$

where the last step is by (18) and Assumption MX(ii), (iii) and (iv). We thus obtain

$$\|(I)\| = \left\|\epsilon_j' F\frac{\mu_C' W}{N_C}\Upsilon^{-1}\right\| + o_p(1).$$

We further decompose (II) and obtain

$$\begin{aligned}
(II) &= \frac{1}{N_C}\sum_{i\in C}\epsilon_j' M_{X_i}\epsilon_i w_i \Upsilon^{-1} \\
&= \frac{1}{N_C}\sum_{i\in C}\epsilon_j'\epsilon_i w_i\Upsilon^{-1} - \frac{1}{N_C}\sum_{i\in C}\frac{\epsilon_j' X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'\epsilon_i}{\sqrt{T}}w_i\Upsilon^{-1} \\
&= \frac{1}{N_C}\sum_{i\in C}[\epsilon_j'\epsilon_i - E(\epsilon_j'\epsilon_i)]\mu_i' H\Upsilon^{-1} + \frac{1}{N_C}\sum_{i\in C}E(\epsilon_j'\epsilon_i)\mu_i' H\Upsilon^{-1} \\
&\quad + \frac{1}{N_C}\sum_{i\in C}\epsilon_j'\epsilon_i(w_i - \mu_i' H)\Upsilon^{-1} - \frac{1}{N_C}\sum_{i\in C}\frac{\epsilon_j' X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'\epsilon_i}{\sqrt{T}}w_i\Upsilon^{-1} \\
&=: (IIa) + (IIb) + (IIc) + (IId).
\end{aligned}$$

We evaluate $\sup_{j\in E}\|(IIa)\|$ as follows:

$$\sup_{j\in E}\|(IIa)\| \le \frac{1}{\sqrt{N_C}}\sup_{j\in E}\left\|\frac{1}{\sqrt{N_C T}}\sum_{i\in C}[\epsilon_j'\epsilon_i - E(\epsilon_j'\epsilon_i)]\mu_i'\right\|\|H\|\left\|\sqrt{T}\Upsilon^{-1}\right\| = O_p\left(\frac{1}{\sqrt{N_C}}\right)$$

by Assumption M(iv). Then, we simplify (IIb) as follows:

$$(IIb) = \frac{\sqrt{T}}{N_C}\sum_{i\in C}\frac{E(\epsilon_j'\epsilon_i)}{T}\mu_i' H(\sqrt{T}\Upsilon^{-1}) = \frac{\sqrt{T}}{N_C}\sum_{i\in C}\gamma_T(j,i)\mu_i' H(\sqrt{T}\Upsilon^{-1}). \tag{20}$$

Let us compute $\sup_{j\in E}\|(IIb)\|$:

$$\begin{aligned}
\sup_{j\in E}\|(IIb)\| &\le \frac{\sqrt{T}}{N_C}\left(\sup_{j\in E}\sum_{i\in C}|\gamma_T(j,i)|\right)\sup_{i\in C}\|\mu_i\|\left\|\sqrt{T}\Upsilon^{-1}\right\|\|H\| \\
&\le O_p\left(\frac{\sqrt{T}}{N_C}\right)O_p(1)O_p(1)O_p(1)O_p(1),
\end{aligned}$$

where the last step follows from (15), (18) and Assumptions FL_C(i) and IE(ii). Next, we evaluate $\sup_{j\in E}\|(IIc)\|$ as follows:

$$\begin{aligned}
\sup_{j\in E}\|(IIc)\| &= \sup_{j\in E}\left\|\sum_{t=1}^{T}\epsilon_{jt}\left[\frac{1}{N_C}\sum_{i\in C}\epsilon_{it}(w_i - \mu_i' H)\right]\Upsilon^{-1}\right\| \\
&\le \sup_t\left\|\frac{1}{N_C}\sum_{i\in C}\epsilon_{it}(w_i - \mu_i' H)\right\|\sup_{j\in E}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\epsilon_{jt}\right|\left\|\sqrt{T}\Upsilon^{-1}\right\|. \qquad (21)
\end{aligned}$$

By Cauchy-Schwarz inequality, the first term in (21) is bounded in probability:

$$\begin{aligned}
\sup_t\left\|\frac{1}{N_C}\sum_{i\in C}\epsilon_{it}(w_i - \mu_i' H)\right\| &\le \sup_t\left(\frac{1}{N_C}\sum_{i\in C}\epsilon_{it}^2\right)^{1/2}\left(\frac{1}{N_C}\sum_{i\in C}\|w_i - \mu_i' H\|^2\right)^{1/2} \\
&= O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right), \qquad (22)
\end{aligned}$$

where the last step follows from lemma A3(a) and Assumption IE(i), which implies that $\frac{1}{N_C}\sum_{i\in C}E(\epsilon_{it}^2) \le c_1 < \infty$ uniformly over t. On the other hand, the second term in (21) is bounded in probability: by Assumption IE(iii), there exists $c_2 < \infty$ such that

$$E\left|\sup_{j\in E}\frac{1}{T}\sum_{s,t=1}^{T}\epsilon_{js}\epsilon_{jt}\right| \le \frac{1}{T}\sum_{s,t=1}^{T}|\sigma_{st}| \le c_2.$$

The last term in (21) is bounded by (15). We therefore have $\sup_{j\in E}\|(IIc)\| = O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)$. At last, we compute $\sup_{j\in E}\|(IId)\|$ as follows:

$$\begin{aligned}\sup_{j\in E}\|(IId)\| &\le \sup_{i\in C, j\in E}\left\|\frac{\epsilon_j' X_i}{\sqrt{T}}\right\| \left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\left\|\frac{1}{N_C\sqrt{T}}\sum_{i\in C} X_i'\epsilon_i w_i\right\| \|\Upsilon^{-1}\| \\ &\le O_p(1)O_p(1)O_p\left(\frac{1}{\min(\sqrt{T},\sqrt{N_C})}\right)O_p\left(\frac{1}{\sqrt{T}}\right),\end{aligned}$$

where the last step follows from (15), part (b), and Assumptions MX(i) and (ii).

Combining all the terms, we obtain

$$\sup_{j\in E}\left\|\epsilon_j'\hat{F}\Upsilon^{-1} - \epsilon_j' F\frac{\mu_C' W}{N_C}\Upsilon^{-1}\right\| = O_p\left(\frac{\sqrt{T}}{N_C}\right).$$

The result follows by noting that

$$\sup_{j\in E}\left\|\epsilon_j'\hat{F}\Upsilon^{-1} - \epsilon_j' F\Upsilon^{-1}H'^{-1}\right\| \le \sup_{j\in E}\left\|\epsilon_j'\hat{F}\Upsilon^{-1} - \epsilon_j' F\frac{\mu_C' W}{N_C}\Upsilon^{-1}\right\| + \sup_{j\in E}\left\|\epsilon_j' F\frac{\mu_C' W}{N_C}\Upsilon^{-1} - \epsilon_j' F\Upsilon^{-1}H'^{-1}\right\|$$

and that

$$\begin{aligned}\sup_{j\in E}\left\|\epsilon_j' F\frac{\mu_C' W}{N_C}\Upsilon^{-1} - \epsilon_j' F\Upsilon^{-1}H'^{-1}\right\|^2 &= \sup_{j\in E}\|\epsilon_j' F\|^2\left\|\frac{\mu_C' W}{N_C}\Upsilon^{-1} - \Upsilon^{-1}H'^{-1}\right\|^2 \\ &= \sup_{j\in E}\|\epsilon_j' F\Upsilon^{-1}\|^2\left\|\frac{\mu_C' W}{N_C} - H'^{-1}\right\|^2 = o_p(1),\end{aligned}$$

where the last step follows from (16) and Assumption M(ii).

9.5 Proof of Lemma A5

Define the $(\ell+k)\times(\ell+k)$ block diagonal matrix $\Gamma := \begin{pmatrix}\Upsilon & O \\ O' & \sqrt{T}I\end{pmatrix}$, where $\Upsilon = diag(T^{r_1},\ldots,T^{r_\ell})$ and I is the $k\times k$ identity matrix. Denote $\tilde{H}^{-1} := \begin{pmatrix}H^{-1} & O \\ O & I\end{pmatrix}$, which is of dimension $(\ell+k)\times(\ell+k)$ and is invertible for large T and N_C by lemma A2.

(a) We compute

$$\begin{aligned}&\frac{1}{T}1_{post,j}' M_{[\hat{F},X_j]}1_{post,j} \\ &= \frac{1}{T}1_{post,j}'1_{post,j} - \left(\frac{1_{post,j}'\hat{F}}{\sqrt{T}}, \frac{1_{post,j}'X_j}{\sqrt{T}}\right)\Gamma^{-1}\left[\Gamma^{-1}\begin{pmatrix}\hat{F}'\hat{F} & \hat{F}'X_j \\ X_j'\hat{F} & X_j'X_j\end{pmatrix}\Gamma^{-1}\right]^{-1}\Gamma^{-1}\begin{pmatrix}\frac{\hat{F}'1_{post,j}}{\sqrt{T}} \\ \frac{X_j'1_{post,j}}{\sqrt{T}}\end{pmatrix} \\ &= \frac{1}{T}1_{post,j}'1_{post,j} - \left(\frac{1_{post,j}'\hat{F}}{\sqrt{T}}\Upsilon^{-1}, \frac{1_{post,j}'X_j}{T}\right)\begin{pmatrix}\Upsilon^{-1}\hat{F}'\hat{F}\Upsilon^{-1} & \Upsilon^{-1}\frac{\hat{F}'X_j}{\sqrt{T}} \\ \frac{X_j'\hat{F}}{\sqrt{T}}\Upsilon^{-1} & \frac{X_j'X_j}{T}\end{pmatrix}^{-1}\begin{pmatrix}\Upsilon^{-1}\frac{\hat{F}'1_{post,j}}{\sqrt{T}} \\ \frac{X_j'1_{post,j}}{T}\end{pmatrix} \\ &= \frac{1_{post,j}'1_{post,j}}{T} - \left(\frac{1_{post,j}'F}{\sqrt{T}}\Upsilon^{-1}, \frac{1_{post,j}'X_j}{T}\right)\tilde{H}'^{-1}\left[\tilde{H}^{-1}\begin{pmatrix}\Upsilon^{-1}F'F\Upsilon^{-1} & \Upsilon^{-1}\frac{F'X_j}{\sqrt{T}} \\ \frac{X_j'F}{\sqrt{T}}\Upsilon^{-1} & \frac{X_j'X_j}{T}\end{pmatrix}\tilde{H}'^{-1}\right]^{-1} \\ &\quad\times\tilde{H}^{-1}\begin{pmatrix}\Upsilon^{-1}\frac{F'1_{post,j}}{\sqrt{T}} \\ \frac{X_j'1_{post,j}}{T}\end{pmatrix} + o_p(1) \\ &= \frac{1}{T}1_{post,j}' M_{[F,X_j]}1_{post,j} + o_p(1).\end{aligned}$$

The third equality follows from lemma A4 (d), (e) and (f). The result holds by noting that the $o_p(1)$ term is uniform over all $j \in E$.

(b) We decompose as follows:

$$
\begin{aligned}
&\frac{1}{\sqrt{T}} 1'_{post,j} M_{[\hat{F},X_j]}(F\mu_j + X_j\beta_j + \epsilon_j) \\
=& \frac{1'_{post,j}(F\mu_j + X_j\beta_j + \epsilon_j)}{\sqrt{T}} - \left(\frac{1'_{post,j}\hat{F}}{\sqrt{T}}, \frac{1'_{post,j}X_j}{\sqrt{T}} \right) \Gamma^{-1} \\
&\times \left[\Gamma^{-1} \begin{pmatrix} \hat{F}'\hat{F} & \hat{F}'X_j \\ X_j'\hat{F} & X_j'X_j \end{pmatrix} \Gamma^{-1} \right]^{-1} \Gamma^{-1} \begin{pmatrix} \hat{F}' \\ X_j' \end{pmatrix} \left[(F, X_j) \begin{pmatrix} \mu_j \\ \beta_j \end{pmatrix} + \epsilon_j \right] \\
=& \frac{1'_{post,j}(F\mu_j + X_j\beta_j + \epsilon_j)}{\sqrt{T}} - \left(\frac{1'_{post,j}F}{\sqrt{T}} \Upsilon^{-1}, \frac{1'_{post,j}X_j}{T} \right) \\
&\times \begin{pmatrix} \Upsilon^{-1}F'F\Upsilon^{-1} & \Upsilon^{-1}\frac{F'X_j}{\sqrt{T}} \\ \frac{X_j'F}{\sqrt{T}}\Upsilon^{-1} & \frac{X_j'X_j}{T} \end{pmatrix}^{-1} \begin{pmatrix} \Upsilon^{-1}F' \\ \frac{1}{\sqrt{T}}X_j' \end{pmatrix} \left[(F, X_j) \begin{pmatrix} \mu_j \\ \beta_j \end{pmatrix} + \epsilon_j \right] + O_p\left(\frac{\sqrt{T}}{N_C} \right) \\
=& \frac{1'_{post,j}\epsilon_j}{\sqrt{T}} - \left(\frac{1'_{post,j}F}{\sqrt{T}} \Upsilon^{-1}, \frac{1'_{post,j}X_j}{T} \right) \begin{pmatrix} \Upsilon^{-1}F'F\Upsilon^{-1} & \Upsilon^{-1}\frac{F'X_j}{\sqrt{T}} \\ \frac{X_j'F}{\sqrt{T}}\Upsilon^{-1} & \frac{X_j'X_j}{T} \end{pmatrix}^{-1} \begin{pmatrix} \Upsilon^{-1}F'\epsilon_j \\ \frac{X_j'\epsilon_j}{\sqrt{T}} \end{pmatrix} + O_p\left(\frac{\sqrt{T}}{N_C} \right) \\
=& \frac{1}{\sqrt{T}} 1'_{post,j} M_{[F,X_j]}\epsilon_j + O_p\left(\frac{\sqrt{T}}{N_C} \right). \qquad (23)
\end{aligned}
$$

The second equality follows from lemma A4(d), (e), (f) and (g). The result holds by noting that the $O_p\left(\frac{\sqrt{T}}{N_C}\right)$ term is uniform over all $j \in E$.

9.6 Proof of Lemma A6

From the proofs of lemmas A5(b) and A4(d)-(g), it suffices to show that

$$
\left\| \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' \hat{F} \Upsilon^{-1} - \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' F \Upsilon^{-1} \right\| = o_p(1). \qquad (24)
$$

To this end, we evaluate the term

$$
\begin{aligned}
\frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' \hat{F} \Upsilon^{-1} =& \frac{1}{\sqrt{N_E}N_C} \sum_{j\in E}\sum_{i\in C} \epsilon_j' M_{X_i} F\mu_i w_i \Upsilon^{-1} + \frac{1}{\sqrt{N_E}N_C} \sum_{j\in E}\sum_{i\in C} \epsilon_j' M_{X_i} \epsilon_i w_i \Upsilon^{-1} \\
=:& (I) + (II).
\end{aligned}
$$

Simplifying the first expression (I), we have

$$
\begin{aligned}
(I) =& \frac{1}{\sqrt{N_E}N_C} \sum_{j\in E}\sum_{i\in C} \epsilon_j' M_{X_i} F\mu_i w_i \Upsilon^{-1} \\
=& \frac{1}{\sqrt{N_E}N_C} \sum_{j\in E}\sum_{i\in C} \epsilon_j' F\mu_i w_i \Upsilon^{-1} - \frac{1}{\sqrt{N_E}N_C} \sum_{j\in E}\sum_{i\in C} \frac{\epsilon_j' X_i}{\sqrt{T}} \left(\frac{X_i'X_i}{T} \right)^{-1} \frac{X_i'F}{\sqrt{T}} \mu_i w_i \Upsilon^{-1} \\
=:& (Ia) + (Ib).
\end{aligned}
$$

The norm of (Ia) is bounded as follows:

$$
\|(Ia)\| = \left\| \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' F \Upsilon^{-1} \right\| \left\| \frac{\mu_C' W}{N_C} \right\| \leq O_p(1)O_p(1)
$$

by (18) and Assumption MM(ii). The norm of (Ib) is bounded as follows:

$$\begin{aligned}
\|(Ib)\| &= \frac{1}{\sqrt{T}N_E N_C}\sum_{j\in E}\sum_{i\in C}\left\|\frac{\epsilon_j' X_i}{\sqrt{T}}\right\|\left\|\frac{X_i'X_i}{T}\right\|^{-1}\|X_i'F\|\,\|\mu_i w_i\|\left\|\Upsilon^{-1}\right\| \\
&\leq \frac{1}{\sqrt{T}}\sup_{i\in C, j\in E}\left\|\sum_{j\in E}\frac{\epsilon_j' X_i}{\sqrt{N_E T}}\right\|\left\|\inf_{i\in C}\frac{X_i'X_i}{T}\right\|^{-1}\sup_{i\in C}\left\|X_i'F\Upsilon^{-1}\right\|\left\|\frac{\mu_C' W}{N_C}\right\| \\
&\leq O_p\left(\frac{1}{\sqrt{T}}\right)O_p(1)O_p(1)O_p(1)O_p(1),
\end{aligned}$$

where the last step holds by (18) and Assumption MX(ii), (iii) and (vii).

Simplifying the second expression (II), we have

$$\begin{aligned}
(II) &= \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}\epsilon_j' M_{X_i}\epsilon_i w_i \Upsilon^{-1} \\
&= \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}\epsilon_j'\epsilon_i w_i\Upsilon^{-1} - \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}\frac{\epsilon_j' X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'\epsilon_i}{\sqrt{T}}w_i\Upsilon^{-1} \\
&= \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}[\epsilon_j'\epsilon_i - E(\epsilon_j'\epsilon_i)]\mu_i' H\Upsilon^{-1} + \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}E(\epsilon_j'\epsilon_i)\mu_i' H\Upsilon^{-1} \\
&\quad + \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}\epsilon_j'\epsilon_i(w_i - \mu_i' H)\Upsilon^{-1} - \frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}\frac{\epsilon_j' X_i}{\sqrt{T}}\left(\frac{X_i'X_i}{T}\right)^{-1}\frac{X_i'\epsilon_i}{\sqrt{T}}w_i\Upsilon^{-1} \\
&= (IIa) + (IIb) + (IIc) + (IId),
\end{aligned}$$

where the second-to-last equality follows from the decomposition $w_i = \mu_i' H + (w_i - \mu_i' H)$. The norm of (IIa) is

$$\begin{aligned}
\|(IIa)\| &\leq \frac{1}{\sqrt{N_C}}\left\|\frac{1}{\sqrt{N_E N_C T}}\sum_{t=1}^{T}\sum_{i\in C}\sum_{j\in E}[\epsilon_{it}\epsilon_{jt} - E(\epsilon_{it}\epsilon_{jt})]\mu_i'\right\|\|H\|\left\|\sqrt{T}\Upsilon^{-1}\right\| \\
&\leq O_p\left(\frac{1}{\sqrt{N_C}}\right)O_p(1)O_p(1)O_p(1)
\end{aligned}$$

by (15), (19) and Assumption MM(iii). The norm of (IIb) is

$$\begin{aligned}
\|(IIb)\| &\leq \sqrt{\frac{T}{N_C}}\frac{1}{\sqrt{N_E N_C}}\sum_{j\in E}\sum_{i\in C}\left|\frac{1}{T}\sum_{t=1}^{T}E(\epsilon_{jt}\epsilon_{it})\right|\|\mu_i\|\,\|H\|\left\|\sqrt{T}\Upsilon^{-1}\right\| \\
&\leq \sqrt{\frac{T}{N_C}}\sup_{i\in C}\|\mu_i\|\frac{1}{\sqrt{N_E N_C}}\sum_{j\in E}\sum_{i\in C}|\gamma_T(j,i)|\,\|H\|\left\|\sqrt{T}\Upsilon^{-1}\right\| \\
&\leq O_p\left(\sqrt{\frac{T}{N_C}}\right)O_p(1)O_p(1)O_p(1)O_p(1)
\end{aligned}$$

by (15), (19) and Assumptions FL_C(i) and MM(i). The norm of (IIc) is

$$\begin{aligned}
\|(IIc)\| &= \left\| \frac{1}{\sqrt{N_E}} \sum_{j\in E} \sum_{t=1}^{T} \epsilon_{jt} \left[\frac{1}{N_C} \sum_{i\in C} \epsilon_{it}(w_i - \mu_i' H) \right] \Upsilon^{-1} \right\| \\
&\leq \sup_t \left\| \frac{1}{N_C} \sum_{i\in C} \epsilon_{it}(w_i - \mu_i' H) \right\| \left| \frac{1}{\sqrt{N_E T}} \sum_{j\in E} \sum_{t=1}^{T} \epsilon_{jt} \right| \left\| \sqrt{T} \Upsilon^{-1} \right\| \\
&\leq O_p\left(\frac{1}{\min(\sqrt{T}, \sqrt{N_C})} \right) O_p(1) O_p(1),
\end{aligned}$$

where the last step follows from (15), (22) and Assumption IE(iv), which implies that

$$\frac{1}{N_E T} \sum_{i,j\in E} \sum_{s,t=1}^{T} |E(\epsilon_{js}\epsilon_{jt})| \leq \frac{1}{N_E T} \sum_{i,j\in E} \sum_{s,t=1}^{T} |\sigma_{ij,st}| \leq c < \infty.$$

The norm of (IId) is

$$\begin{aligned}
\|(IId)\| &\leq \sup_{i\in C} \left\| \sum_{j\in E} \frac{\epsilon_j' X_i}{\sqrt{N_E T}} \right\| \left\| \inf_{i\in C} \frac{X_i' X_i}{T} \right\|^{-1} \left\| \frac{1}{N_C \sqrt{T}} \sum_{i\in C} X_i' \epsilon_i w_i \right\| \|\Upsilon^{-1}\| \\
&\leq O_p(1) O_p(1) O_p\left(\frac{1}{\min(\sqrt{T}, \sqrt{N_C})} \right) O_p\left(\frac{1}{\sqrt{T}} \right) \\
&= O_p\left(\frac{1}{\sqrt{T}\min(\sqrt{T}, \sqrt{N_C})} \right),
\end{aligned}$$

where the second-to-last step follows from (15), lemma A4(b) and Assumptions MX(i) and (vii).

Combining all the above terms, we obtain

$$\left\| \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' \hat{F} \Upsilon^{-1} - \frac{1}{\sqrt{N_E} N_C} \sum_{j\in E} \sum_{i\in C} \epsilon_j' F \mu_i w_i \Upsilon^{-1} \right\| = O_p\left(\sqrt{\frac{T}{N_C}} \right).$$

The upper bound is $o_p(1)$ provided that $T/N_C \to 0$. The claim (24) follows immediately by noting that

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' \hat{F} \Upsilon^{-1} - \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' F \Upsilon^{-1} H'^{-1} \right\| \\
\leq &\left\| \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' \hat{F} \Upsilon^{-1} - \frac{1}{\sqrt{N_E} N_C} \sum_{j\in E} \sum_{i\in C} \epsilon_j' F \mu_i w_i \Upsilon^{-1} \right\| \\
&+ \left\| \frac{1}{\sqrt{N_E} N_C} \sum_{j\in E} \sum_{i\in C} \epsilon_j' F \mu_i w_i \Upsilon^{-1} - \frac{1}{\sqrt{N_E}} \sum_{j\in E} \epsilon_j' F \Upsilon^{-1} H'^{-1} \right\|
\end{aligned}$$

and that

$$\left\|\frac{1}{\sqrt{N_E}N_C}\sum_{j\in E}\sum_{i\in C}\epsilon_j' F\mu_i w_i \Upsilon^{-1} - \frac{1}{\sqrt{N_E}}\sum_{j\in E}\epsilon_j' F\Upsilon^{-1}H'^{-1}\right\|^2$$
$$\leq \left\|\frac{1}{\sqrt{N_E}}\sum_{j\in E}\epsilon_j' F\right\|^2 \left\|\frac{\mu_C' W}{N_C}\Upsilon^{-1} - \Upsilon^{-1}H'^{-1}\right\|^2$$
$$= \left\|\frac{1}{\sqrt{N_E}}\sum_{j\in E}\epsilon_j' F\Upsilon^{-1}\right\|^2 \left\|\frac{\mu_C' W}{N_C} - H'^{-1}\right\|^2 = O_p(1)o_p(1) = o_p(1)$$

by Assumption MM(ii) and (16).

9.7 Proof of Lemma A7

Denote $\bar{\mu}_C := \frac{1}{N_C}\sum_{i\in C}\mu_i$. Assumption FLM enables us to apply the law of large numbers (Corollary 3.48 of White (2001)) on $\bar{\mu}_C$, yielding

$$\|\bar{\mu}_C - \mu_0\| = o_p(1) \tag{25}$$

as $N_C \to \infty$. We will make use of this consistency result in various parts of the proof.

Let us prove (a). For fixed $j \in E$, we have

$$\begin{aligned}
\frac{1}{T^{2r}}\bar{u}_C' M_{[1_{post,j},X_j]}\bar{u}_C &= \frac{1}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)' M_{X_i} M_{[1_{post,j},X_j]} M_{X_{i'}}(F\mu_{i'}+\epsilon_{i'}) \\
&= \frac{1}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)'(F\mu_{i'}+\epsilon_{i'}) \\
&\quad - \frac{2}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)' P_{X_i}(F\mu_{i'}+\epsilon_{i'}) \\
&\quad + \frac{1}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)' P_{X_i}P_{X_{i'}}(F\mu_{i'}+\epsilon_{i'}) \\
&\quad - \frac{1}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)' P_{[1_{post,j},X_j]}(F\mu_{i'}+\epsilon_{i'}) \\
&\quad + \frac{2}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)' P_{X_i}P_{[1_{post,j},X_j]}(F\mu_{i'}+\epsilon_{i'}) \\
&\quad - \frac{1}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}(F\mu_i+\epsilon_i)' P_{X_i}P_{[1_{post,j},X_j]}P_{X_{i'}}(F\mu_{i'}+\epsilon_{i'}) \\
&=: (I) - 2(II) + (III) - (IV) + 2(V) - (VI).
\end{aligned}$$

The first term is

$$\begin{aligned}
(I) &= \bar{\mu}_C'\frac{F'F}{T^{2r}}\bar{\mu}_C + 2\bar{\mu}_C'\frac{1}{N_C T^{2r}}\sum_{i\in C}F'\epsilon_i + \frac{1}{N_C^2 T^{2r}}\sum_{i\in C}\sum_{i'\in C}\epsilon_i'\epsilon_{i'} \\
&= \bar{\mu}_C'\frac{F'F}{T^{2r}}\bar{\mu}_C + \frac{2\bar{\mu}_C'}{\sqrt{N_C}T^r}\frac{1}{\sqrt{N_C}T^r}\sum_{i\in C}f'\epsilon_i + \frac{1}{N_C T^{2r-1}}\frac{1}{N_C}\sum_{i\in C}\sum_{i'\in C}\gamma_T(i,i') \\
&= \mu_0'\frac{F'F}{T^{2r}}\mu_0 + O_p\left(\frac{1}{\sqrt{N_C}T^r}\right) + O_p\left(\frac{1}{N_C T^{2r-1}}\right),
\end{aligned}$$

where the last line follows from (25) and Assumptions MM(i) and (ii). The leading term is $O_p(1)$ by the assumed hypothesis that r is the normalization order of $F\mu_0$. The second expression is

$$
\begin{aligned}
(II) &= \frac{1}{N_C^2 T^{2r}} \sum_{i\in C}\sum_{i'\in C} (F\mu_i + \epsilon_i)' X_i \left(X_i'X_i\right)^{-1} X_i'(F\mu_{i'} + \epsilon_{i'}) \\
&= \frac{1}{N_C^2 T} \sum_{i\in C}\sum_{i'\in C} \mu_i' \frac{F'X_i}{T^r} \left(\frac{X_i'X_i}{T}\right)^{-1} \frac{X_i'F}{T^r}\mu_{i'} \\
&\quad + \frac{2}{N_C^2 T^{r+0.5}} \sum_{i\in C}\sum_{i'\in C} \mu_i' \frac{F'X_i}{T^r} \left(\frac{X_i'X_i}{T}\right)^{-1} \frac{X_i'\epsilon_{i'}}{\sqrt{T}} \\
&\quad + \frac{1}{N_C^2 T^{2r}} \sum_{i\in C}\sum_{i'\in C} \frac{\epsilon_i'X_i}{\sqrt{T}} \left(\frac{X_i'X_i}{T}\right)^{-1} \frac{X_i'\epsilon_{i'}}{\sqrt{T}},
\end{aligned}
$$

which is $O_p\left(\frac{1}{T}\right) = o_p(1)$ by Assumptions MX(ii)-(iv). Similarly, the third expression is $O_p\left(\frac{1}{T}\right) = o_p(1)$ by Assumptions MX(i)-(iv). The fourth expression is simplified into

$$
\begin{aligned}
(IV) &= \frac{1}{N_C^2 T^{2r}} \sum_{i\in C}\sum_{i'\in C} \begin{pmatrix} \mu_i'F'1_{post,j} + \epsilon_i'1_{post,j} & \mu_i'F'X_j + \epsilon_i'X_j \end{pmatrix} \\
&\quad \times \begin{pmatrix} T_{1j} & 1_{post,j}'X_j \\ X_j'1_{post,j} & X_j'X_j \end{pmatrix}^{-1} \begin{pmatrix} 1_{post,j}'F\mu_{i'} + 1_{post,j}'\epsilon_{i'} \\ X_j'F\mu_{i'} + X_j'\epsilon_{i'} \end{pmatrix} \\
&= \begin{pmatrix} \bar{\mu}_C' \frac{F'1_{post,j}}{T^{r+0.5}} + \frac{\sum_{i\in C}\epsilon_i'1_{post,j}}{N_C T^{r+0.5}} & \bar{\mu}_C' \frac{F'X_j}{T^{r+0.5}} + \frac{\sum_{i\in C}\epsilon_i'X_j}{N_C T^{r+0.5}} \end{pmatrix} \\
&\quad \times \begin{pmatrix} \frac{T_{1j}}{T} & \frac{1_{post,j}'X_j}{T} \\ \frac{X_j'1_{post,j}}{T} & \frac{X_j'X_j}{T} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1_{post,j}'F}{T^{r+0.5}}\bar{\mu}_C + \frac{\sum_{i'\in C}1_{post,j}'\epsilon_{i'}}{N_C T^{r+0.5}} \\ \frac{X_j'F}{T^{r+0.5}}\bar{\mu}_C + \frac{\sum_{i'\in C}X_j'\epsilon_{i'}}{N_C T^{r+0.5}} \end{pmatrix} \\
&= \mu_0' \frac{F'P_{[1_{post,j},X_j]}F}{T^{2r}}\mu_0 + O_p\left(\frac{1}{\sqrt{N_C}T^r}\right),
\end{aligned}
$$

where the last line follows from (25) and Assumptions E(iii), IE(iv) and MX(vii). The leading term on the last line is $O_p(1)$ uniformly over $j \in E$ by Assumptions E(iii), F and MX(i)-(iv). The fifth and sixth expressions are $O_p\left(\frac{1}{T}\right) = o_p(1)$ by a similar reasoning. The result follows by combining the above expressions.

Let us prove (b). For fixed $j \in E$, we have

$$
\begin{aligned}
\frac{1}{T^{2r}}\bar{u}_C' M_{[1_{post,j},X_j]} F\mu_j &= \frac{1}{N_C T^{2r}} \sum_{i\in C} (F\mu_i + \epsilon_i)' M_{X_i} M_{[1_{post,j},X_j]} F\mu_j \\
&= \frac{1}{N_C T^{2r}} \sum_{i\in C} (F\mu_i + \epsilon_i)' F\mu_j \\
&\quad - \frac{1}{N_C T^{2r}} \sum_{i\in C} (F\mu_i + \epsilon_i)' P_{X_i} F\mu_j \\
&\quad - \frac{1}{N_C T^{2r}} \sum_{i\in C} (F\mu_i + \epsilon_i)' P_{[1_{post,j},X_j]} F\mu_j \\
&\quad + \frac{1}{N_C T^{2r}} \sum_{i\in C} (F\mu_i + \epsilon_i)' P_{X_i} P_{[1_{post,j},X_j]} F\mu_j \\
&=: (I) - (II) - (III) + (IV).
\end{aligned}
$$

The first expression is

$$
\begin{aligned}
(I) &= \frac{\bar{\mu}_C' F' F}{T^{2r}} \mu_j + \frac{1}{N_C T^{2r}} \sum_{i \in C} \epsilon_i' F \mu_j \\
&= \frac{\mu_0' F' F}{T^{2r}} \mu_j + O_p\left(\frac{1}{\sqrt{N_C} T^r} \right),
\end{aligned}
$$

where the last line follows from (25) and Assumptions F(i) and MM(ii). The leading term is $O_p(1)$ by the assumed hypothesis that r is the normalization order of $F\mu_0$.. The second expression is

$$
\begin{aligned}
(II) &= \frac{1}{N_C T^{2r}} \sum_{i \in C} (F\mu_i + \epsilon_i)' X_i \left(X_i' X_i \right)^{-1} X_i' F \mu_j \\
&= \frac{1}{N_C T} \sum_{i \in C} \mu_i' \left(\frac{F' X_i}{T^r} \right) \left(\frac{X_i' X_i}{T} \right)^{-1} \left(\frac{X_i' F}{T^r} \right) \mu_j \\
&\quad + \frac{1}{N_C T^{r+0.5}} \sum_{i \in C} \left(\frac{\epsilon_i' X_i}{\sqrt{T}} \right) \left(\frac{X_i' X_i}{T} \right)^{-1} \left(\frac{X_i' F}{T^r} \right) \mu_j,
\end{aligned}
$$

which is $O_p\left(\frac{1}{T}\right) = o_p(1)$ by Assumptions MX(ii)-(iv). The third expression is

$$
\begin{aligned}
(III) &= \frac{1}{N_C T^{2r}} \sum_{i \in C} \begin{pmatrix} \mu_i' F' 1_{post,j} + \epsilon_i' 1_{post,j} & \mu_i' F' X_j + \epsilon_i' X_j \end{pmatrix} \\
&\quad \times \begin{pmatrix} T_{1j} & 1_{post,j}' X_j \\ X_j' 1_{post,j} & X_j' X_j \end{pmatrix}^{-1} \begin{pmatrix} 1_{post,j}' F \mu_j \\ X_j' F \mu_j \end{pmatrix} \\
&= \begin{pmatrix} \bar{\mu}_C' \frac{F' 1_{post,j}}{T^{r+0.5}} + \frac{\sum_{i \in C} \epsilon_i' 1_{post,j}}{N_C T^{r+0.5}} & \bar{\mu}_C' \frac{F' X_j}{T^{r+0.5}} + \frac{\sum_{i \in C} \epsilon_i' X_j}{N_C T^{r+0.5}} \end{pmatrix} \\
&\quad \times \begin{pmatrix} \frac{T_{1j}}{T} & \frac{1_{post,j}' X_j}{T} \\ \frac{X_j' 1_{post,j}}{T} & \frac{X_j' X_j}{T} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1_{post,j}' F}{T^{r+0.5}} \mu_j \\ \frac{X_j' F}{T^{r+0.5}} \mu_j \end{pmatrix} \\
&= \mu_0' \frac{F' P_{[1_{post,j}, X_j]} F}{T^{r+0.5}} \mu_j + O_p\left(\frac{1}{\sqrt{N_C} T^r} \right),
\end{aligned}
$$

where the last line follows from (25) and Assumptions E(iii), IE(iv) and MX(vii). The leading term on the last line is $O_p(1)$ by the assumed hypothesis and MX(i)-(iv).The fourth expression is $O_p\left(\frac{1}{T}\right) = o_p(1)$ by a similar reasoning. The result follows by combining the above expressions.

Let us prove (c). We decompose the following term:

$$
\begin{aligned}
\frac{1}{\sqrt{N_E} T^r} \sum_{j \in E} \bar{u}_C' M_{[1_{post,j}, X_j]} \epsilon_j &= \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i \in C} \sum_{j \in E} (F\mu_i + \epsilon_i)' M_{X_i} M_{[1_{post,j}, X_j]} \epsilon_j \\
&= \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i \in C} \sum_{j \in E} (F\mu_i + \epsilon_i)' \epsilon_j \\
&\quad - \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i \in C} \sum_{j \in E} (F\mu_i + \epsilon_i)' P_{X_i} \epsilon_j \\
&\quad - \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i \in C} \sum_{j \in E} (F\mu_i + \epsilon_i)' P_{[1_{post,j}, X_j]} \epsilon_j \\
&\quad + \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i \in C} \sum_{j \in E} (F\mu_i + \epsilon_i)' P_{X_i} P_{[1_{post,j}, X_j]} \epsilon_j \\
&=: (I) - (II) - (III) + (IV).
\end{aligned}
$$

The first expression is

$$
\begin{aligned}
(I) &= \sum_{j\in E} \frac{\bar{\mu}_C' F' \epsilon_j}{\sqrt{N_E} T^r} + \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i\in C} \sum_{j\in E} \epsilon_i' \epsilon_j \\
&= \sum_{j\in E} \frac{\bar{\mu}_C' F' \epsilon_j}{\sqrt{N_E} T^r} + \frac{1}{T^{r-0.5}} \sqrt{\frac{T}{N_C}} \frac{1}{\sqrt{N_C N_E}} \sum_{i\in C} \sum_{j\in E} \gamma_T(i,j) \\
&= \sum_{j\in E} \frac{\mu_0' F' \epsilon_j}{\sqrt{N_E} T^r} + O_p\left(\frac{1}{T^{r-0.5}} \sqrt{\frac{T}{N_C}} \right),
\end{aligned}
$$

where the last line is obtained by (25) and Assumptions MM(ii). The second term on the last line is $o_p(1)$ when the factor is $I(0)$ and $T/N_C \to 0$. The leading term on the last line is $O_p(1)$ by the assumed hypothesis and Assumption MM(ii). The second expression is

$$
\begin{aligned}
(II) &= \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i\in C} \sum_{j\in E} (F\mu_i + \epsilon_i)' X_i \left(X_i' X_i \right)^{-1} X_i' \epsilon_j \\
&= \frac{1}{N_C \sqrt{T}} \sum_{i\in C} \mu_i' \left(\frac{F' X_i}{T^r} \right) \left(\frac{X_i' X_i}{T} \right)^{-1} \left(\sum_{j\in E} \frac{X_i' \epsilon_j}{\sqrt{N_E T}} \right) \\
&\quad + \frac{1}{N_C T^r} \sum_{i\in C} \frac{\epsilon_i' X_i}{\sqrt{T}} \left(\frac{X_i' X_i}{T} \right)^{-1} \left(\sum_{j\in E} \frac{X_i' \epsilon_j}{\sqrt{N_E T}} \right),
\end{aligned}
$$

which is $O_p\left(\frac{1}{\sqrt{T}}\right)$ by Assumptions MX(ii), (iii) and (vii). The third expression is

$$
\begin{aligned}
(III) &= \frac{1}{N_C \sqrt{N_E} T^r} \sum_{i\in C} \sum_{j\in E} \begin{pmatrix} \mu_i' F' 1_{post,j} + \epsilon_i' 1_{post,j} & \mu_i' F' X_j + \epsilon_i' X_j \end{pmatrix} \\
&\quad \times \begin{pmatrix} T_{1j} & 1_{post,j}' X_j \\ X_j' 1_{post,j} & X_j' X_j \end{pmatrix}^{-1} \begin{pmatrix} 1_{post,j}' \epsilon_j \\ X_j' \epsilon_j \end{pmatrix} \\
&= \frac{1}{\sqrt{N_E}} \sum_{j\in E} \begin{pmatrix} \bar{\mu}_C' \frac{F' 1_{post,j}}{T^{r+0.5}} + \frac{\sum_{i\in C} \epsilon_i' 1_{post,j}}{N_C T^{r+0.5}} & \bar{\mu}_C' \frac{F' X_j}{T^{r+0.5}} + \frac{\sum_{i\in C} \epsilon_i' X_j}{N_C T^{r+0.5}} \end{pmatrix} \\
&\quad \times \begin{pmatrix} \frac{T_{1j}}{T} & \frac{1_{post,j}' X_j}{T} \\ \frac{X_j' 1_{post,j}}{T} & \frac{X_j' X_j}{T} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1_{post,j}' \epsilon_j}{\sqrt{T}} \\ \frac{X_j' \epsilon_j}{\sqrt{T}} \end{pmatrix} \\
&= \mu_0' \frac{1}{\sqrt{N_E} T^r} \sum_{j\in E} F' P_{[1_{post,j}, X_j]} \epsilon_j + O_p\left(\frac{1}{T^r \sqrt{N_C}} \right).
\end{aligned}
$$

The last line follows from (25) and Assumptions E(iii), IE(iv) and MX(vii). The leading term on the last line is $O_p(1)$ by the assumed hypothesis and Assumptions MX(i)-(iv). The fourth expression (IV) is $O_p\left(\frac{1}{\sqrt{T}}\right) = o_p(1)$ by a similar reasoning. The result follows by combining the above expressions.

9.8 Proof of Lemma A8

For simplicity, we give the proof for the case with a single factor.

First, note that $S_{1\epsilon} := \frac{1}{\sqrt{T}} 1_{post,i}' \epsilon_i$ and $S_{X\epsilon} := \frac{1}{\sqrt{T}} X_i' \epsilon_i$ are asymptotically normal by Assumption D_i. To show the asymptotic normality of $S_{F\epsilon} := \frac{1}{T^r} \sum_{t=1}^T f_t \epsilon_{it}$, we consider different specifications of f_t. The I(0) factor case is immediate from Assumption D_i by setting $f_t = u_t$.

Suppose f_t is an I(1) process generated by u_t. More precisely, f_t is defined by the recursion $f_t = f_{t-1} + u_t$. Define the function

$$U_T(a) = \frac{1}{\sqrt{\omega_{uu}T}} \sum_{s=1}^{[Ta]} u_s, \quad a \in [0,1],$$

where $\omega_{uu} := \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T} u_s)$ is finite and strictly positive (Assumption D_i(iii)). Similarly, define

$$W_T(a) = \frac{1}{\sqrt{\omega_{\epsilon\epsilon}T}} \sum_{s=1}^{[Ta]} \epsilon_{is}, \quad a \in [0,1],$$

where $\omega_{\epsilon\epsilon} := \lim_{T\to\infty} Var(T^{-1/2}\sum_{s=1}^{T} \epsilon_{is})$ is finite and strictly positive (Assumption D_i(iii)). Introduce the process $S_T(\cdot)$ defined by

$$S_T(a) = \sum_{t=0}^{[Ta]-1} U_T\left(\frac{t}{T}\right)\left[W_T\left(\frac{t+1}{T}\right) - W_T\left(\frac{t}{T}\right)\right], \quad a \in [0,1].$$

By FCLT for weakly dependent processes (DeJong and Davidson (2000)), we have

$$(U_T, W_T, S_T) \Longrightarrow (\mathcal{U}, \mathcal{W}, \mathcal{S} + \lambda_{u\epsilon}), \tag{26}$$

where $\mathcal{U}$, $\mathcal{W}$ are Wiener processes, $\mathcal{S} := \int \mathcal{U}d\mathcal{W}$, and $\lambda_{u\epsilon} := \lim_{T\to\infty} T^{-1}\sum\sum_{s<t} E(u_s\epsilon_{it})$.

Now let us simplify the sample mean $S_{F\epsilon}$ as follows:

$$\begin{aligned}
S_{F\epsilon} &= \frac{1}{T}\sum_{t=1}^{T} f_t\epsilon_{it} = \frac{1}{T}\sum_{t=1}^{T} f_{t-1}\epsilon_{it} + \frac{1}{T}\sum_{t=1}^{T} u_t\epsilon_{it} \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{t-1} u_s\epsilon_{it} + \frac{1}{T}\sum_{t=1}^{T} u_t\epsilon_{it} \\
&= \sqrt{\omega_{uu}\omega_{\epsilon\epsilon}} S_T(1) + \frac{1}{T}\sum_{t=1}^{T} u_t\epsilon_{it},
\end{aligned}$$

where $S_T(1) = \frac{1}{\sqrt{\omega_{uu}\omega_{\epsilon\epsilon}T}}\sum_{t=0}^{T-1}\sum_{s=0}^{t} u_s\epsilon_{i,t+1}$. Since (u_t, ϵ_{it}) satisfies Assumption D_i, the probability limit $\mu_{u\epsilon} := \mathrm{plim}_{T\to\infty} T^{-1}\sum_{t=1}^{T} u_t\epsilon_t$ exists. By (26), we have

$$S_{F\epsilon} \Longrightarrow \int_0^1 \mathcal{U}d\mathcal{W} + \lambda_{u\epsilon} + \mu_{u\epsilon}.$$

Strong exogeneity of f_t with respect to ϵ_{it} implies that $E(u_s\epsilon_{it}) = 0$ for all s, t, so that both $\lambda_{u\epsilon}$ and $\mu_{u\epsilon}$ are zero. Since $\mathcal{W}$ is a Gaussian process, we obtain asymptotic normality after conditioning on $F = \{f_t\}_t$. More precisely, we have

$$S_{F\epsilon}|_F \xrightarrow{d} N(0, \Omega_{F\epsilon}), \tag{27}$$

where $\Omega_{F\epsilon} := \lim_{T\to\infty} Var(S_{F\epsilon}|F)$. In the case of $I(1)$ factor, we have $\Omega_{F\epsilon} = \omega_{uu}\omega_{\epsilon\epsilon}\int_0^1 \mathcal{U}^2(s)ds$.

The above argument can be easily extended to the case with an $I(r)$ factor for positive integers r. Let $\mathcal{U}_1, \ldots, \mathcal{U}_r$ be r independent Wiener processes on $[0,1]$. It follows that (27) holds for the sample mean $S_{F\epsilon} = T^{-r}\sum_{t=1}^{T} f_t\epsilon_{it}$ with $\Omega_{F\epsilon} = \omega_{u_1u_1}\cdots\omega_{u_ru_r}\omega_{\epsilon\epsilon}\int_0^1\int_0^{s_r}\cdots\int_0^{s_2}\mathcal{U}_1^2(s_1)ds_1\cdots ds_r$.

Suppose f_t is a deterministic trend with maximal polynomial order p. The sample mean becomes

$$S_{F\epsilon} = \frac{1}{T^{p+1/2}}\sum_{t=1}^{T} f_t\epsilon_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\frac{f_t}{T^p}\right)\epsilon_t.$$

It is easy to check that $f_t\epsilon_{it}/T^p$ satisfies the conditions in Assumption D_i: (i) is immediate as f_t is deterministic, while (ii) and the finiteness of $\Omega_{F\epsilon}$ in (iii) follow because f_t/T^p is uniformly bounded: there exists $M < \infty$ such that $|f_t/T^p| < M$ for all t and T. Applying CLT to $S_{F\epsilon}$ conditional on F, we see that (27) remains valid, with $\Omega_{F\epsilon} > 0$ as guaranteed by the stated assumption.

9.9 Proof of Theorem 1

Fix $i \in E$ throughout the proof.
(a) The estimator is

$$\hat{\delta}_i = (1'_{post,i} M_{[\hat{F},X_i]} 1_{post,i})^{-1} 1'_{post,i} M_{[\hat{F},X_i]} y_i.$$

From the DGP, we obtain:

$$\begin{aligned}\hat{\delta}_i - \bar{\Delta}_i &= \frac{1'_{post,i} M_{[\hat{F},X_i]} (F\mu_i + X_i\beta_i + \epsilon_i)/T}{1'_{post,i} M_{[\hat{F},X_i]} 1_{post,i}/T} \\ &= \frac{1'_{post,i} M_{[F,X_i]} \epsilon_i/T}{1'_{post,i} M_{[F,X_i]} 1_{post,i}/T} + o_p(1),\end{aligned}$$

where the last step is by lemmas A5(a) and (b). The probability limit of the denominator is strictly positive *a.s.* by Assumption AI_i(i). Let us turn to the numerator. Denote η_{it} the residuals from the time series regression of $1_{\{t>T_{0i})}$ on f_t and x_{it}. Since $1_{\{t>T_{0i})}$, f_t, x_{it} and ϵ_{it} as stochastic processes satisfy the mixing condition in Assumption D_i, so do η_{it} and hence $\eta_{it}\epsilon_{it}$. It follows from Assumption D that the numerator converges in probability as follows: as $T \to \infty$,

$$\frac{1}{T} 1'_{post,i} M_{[F,X_i]} \epsilon_i = \frac{1}{T} \sum_{t=1}^{T} \eta_{it}\epsilon_{it} \xrightarrow{p} E(\eta_{it}\epsilon_{it}),$$

but then the limit is zero because $E(\eta_{it}\epsilon_{it}) = E[\eta_{it} E(\epsilon_{it}|f_t, x_{it})] = 0$ by Assumption E(i). The result follows.

(b) Define $S_{Ti} := \frac{1}{\sqrt{T}} 1'_{post,i} M_{[F,X_i]} \epsilon_i$ and $R_{Ti} := \frac{1}{T} 1'_{post,i} M_{[F,X_i]} 1_{post,i}$. By lemma A5(a) and (b), we have the following approximation

$$\sqrt{T}(\hat{\delta}_i - \bar{\Delta}_i) = S_{Ti}/R_{Ti} + o_p(1) \tag{28}$$

as $N_C, T \to \infty$ and $\sqrt{N_C}/T \to 0$.

Let us first study the denominator R_{Ti}. Assumption AI_i(i) ensures that $R_{Ti} \xrightarrow{p} \rho_i$ and that R_{Ti} is strictly positive *a.s.* for all sufficiently large T.

Now let us turn to the numerator S_{Ti}. By Assumption D_i, the random terms $S_{1\epsilon} := \frac{1}{\sqrt{T}} 1'_{post,i}\epsilon_i$ and $S_{X\epsilon} := \frac{1}{\sqrt{T}} X_i'\epsilon_i$ have a limiting normal distribution. Furthermore, conditional on $\tilde{G}_i$, we deduce from (27) that $S_{F\epsilon} := \Upsilon^{-1} F'\epsilon_i$ is asymptotically normal. This implies that, conditional on $\tilde{G}_i$, S_{Ti}, being a linear combination of $S_{1\epsilon}$, $S_{X\epsilon}$ and $S_{F\epsilon}$ with $O_p(1)$ coefficients, has a normal distribution in the limit (by lemma A8). More precisely, we have

$$S_{Ti}|_{\tilde{G}_i} \xrightarrow{d} N(0, \xi_i^2) \tag{29}$$

as $T \to \infty$. The limiting conditional distribution has mean zero (by Assumption ES) and variance $\xi_i^2 := \mathrm{plim}_{T\to\infty} \xi_{Ti}^2$, which is strictly positive *a.s.* (by Assumption AI_i(ii)).

Now, for a given T, define the conditional variance $\xi_{Ti}^2 := Var(S_{Ti}|\tilde{G}_i)$. Assumption AI_i(ii) ensures that $\xi_i^2 := \mathrm{plim}_{T\to\infty} \xi_{Ti}^2 > 0$ *a.s.*, which implies that there exists $c > 0$ uniformly over T such that $\xi_{Ti}^2 \geq c$ *w.p.a.*1 as $T \to \infty$. Normalizing S_{Ti} by the conditional standard deviation, we obtain $\xi_{Ti}^{-1} S_{Ti}|_{\tilde{G}_i} \xrightarrow{d} N(0,1)$ as $T \to \infty$. As the limiting distribution is independent of $\tilde{G}_i$, the conditioning set may be dropped, yielding

$$\xi_{Ti}^{-1} S_{Ti} \xrightarrow{d} N(0,1) \tag{30}$$

as $T \to \infty$. The result immediately follows from (28)-(30) and by noting that $\sigma_{Ti}^2 = \xi_{Ti}^2/R_{Ti}^2$.

9.10 Proof of Theorem 2

(a)(i) Assume homogeneous ITET. From the DGP, we obtain:

$$\begin{aligned}
\left|\hat{\delta}^{mg} - \bar{\bar{\Delta}}\right| &= \left| \frac{1}{N_E T} \sum_{i \in E} \frac{1'_{post,i} M_{[\hat{F}, X_i]} (F\mu_i + X_i \beta_i + \epsilon_i)}{1'_{post,i} M_{[\hat{F}, X_i]} 1_{post,i}/T} \right| \\
&= \left| \frac{1}{N_E T} \sum_{i \in E} \frac{1'_{post,i} M_{[\hat{F}, X_i]} (F\mu_i + X_i \beta_i + \epsilon_i)}{1'_{post,i} M_{[F, X_i]} 1_{post,i}/T} \right| + o_p(1) \\
&\leq \frac{1}{\sqrt{N_E T}} \frac{\left| \frac{1}{\sqrt{N_E T}} \sum_{i \in E} 1'_{post,i} M_{[F, X_i]} \epsilon_i \right|}{\inf_i 1'_{post,i} M_{[F, X_i]} 1_{post,i}/T} + o_p(1) \\
&= O_p\left(\frac{1}{\sqrt{N_E T}}\right),
\end{aligned}$$

where the second step follows from lemma A5(a), the third step from lemma A6, and the last step from Assumptions AI_{mg}(i) and (ii). The remainder term has mean zero in the limit by Assumption E(i).

(a)(ii) After normalization, the simple mean-group estimator is decomposed as follows:

$$\sqrt{N_E T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) = \frac{1}{\sqrt{N_E}} \sum_{i \in E} \frac{S_{Ti}}{R_{Ti}} + o_p(1) \tag{31}$$

as $N_C, T \to \infty$ and $T/N_C \to \infty$. Denote $\tilde{G} := \{\tilde{G}_i\}_{i \in E}$. The key step is to show that

$$\left. \frac{1}{\sqrt{N_E} \bar{\varsigma}_{N_E,T}} \sum_{i \in E} \frac{S_{Ti}}{R_{Ti}} \right|_{\tilde{G}} \xrightarrow{d} N(0,1) \tag{32}$$

as $N_E, T \to \infty$ jointly, where $\bar{\varsigma}^2_{N_E,T} := Var[\sqrt{N_E T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})|\tilde{G}]$. To this end we check the conditions behind Liapounov's theorem (Theorem 5.10 of White (2001)):

- For each $i \in E$, $E(S_{Ti}/R_{Ti}|\tilde{G}) = 0$ by Assumption ESS.
- For some $p > 2$, there exists $0 < C < \infty$ (uniformly over $i \in E$ and T) such that $E(|S_{Ti}/R_{Ti}|^p|\tilde{G}) \leq C \quad w.p.a.1$ as $T \to \infty$. To show this, we first note from the proof of Theorem 2(a) that $E(|S_{Ti}|^p|\tilde{G}) \leq \tilde{C} < \infty \quad w.p.a.1$ as $T \to \infty$. By Assumption AI_{mg}(i), we note that $R_{Ti} > 0$ uniformly over $i \in E$ as $T \to \infty$. More precisely, for any fixed $i \in E$ and given $\tilde{G}$, there exists $\tilde{c} > 0$ such that $R_{Ti} \geq \tilde{c}$ for all sufficiently large T. It follows that $E(|S_{Ti}/R_{Ti}|^p|\tilde{G}) \leq \tilde{C}/\tilde{c}^p < \infty$ $w.p.a.1$ as $T \to \infty$. The result is proved by setting $C = \tilde{C}/\tilde{c}^p$, which is independent of i and T.
- There exists $c > 0$ uniformly over N_E and T such that $\bar{\varsigma}^2_{N_E,T} \geq c \quad w.p.a.1$ as $N_E, T \to \infty$. To show this, we first compute $\bar{\varsigma}^2_{N_E,T}$ as

  $$\bar{\varsigma}^2_{N_E,T} = \frac{1}{N_E} \sum_{i \in E} \frac{\bar{\zeta}^2_{Ti}}{R^2_{Ti}} + o_p(1).$$

  Since $R_{Ti} \leq 1$ unformly over i, we have

  $$\frac{1}{N_E} \sum_{i \in E} \frac{\bar{\zeta}^2_{Ti}}{R^2_{Ti}} \geq \frac{1}{N_E} \sum_{i \in E} \bar{\zeta}^2_{Ti} = \bar{\zeta}^2_{N_E,T} \geq c \quad w.p.a.1 \tag{33}$$

  as $N_E, T \to \infty$. Assumption AI_{mg}(ii) guarantees that such a lower bound $c > 0$ exists and is uniform over N_E and T.

By Liapounov's theorem, the Lindeberg condition is satisfied, i.e., for all $\varepsilon > 0$, we have

$$\lim_{N_E,T\to\infty} \frac{1}{N_E\bar{\varsigma}^2_{N_E,T}} \sum_{i\in E} E\left[\left(\frac{S_{Ti}}{R_{Ti}}\right)^2 1_{\{|S_{Ti}/R_{Ti}|>\varepsilon\sqrt{N_E}\bar{\varsigma}_{N_E,T}\}} \middle| \tilde{G}\right] = 0 \ a.s..$$

Applying the joint CLT result (Theorem 2 of Phillips and Moon (1999)) yields (32).

Since $\bar{\varsigma}^2_{N_E,T}$ is strictly positive as observed above, we can thus normalize (31) by its conditional standard deviation, meanwhile using (32), and obtain

$$\sqrt{N_E T}\bar{\varsigma}^{-1}_{N_E,T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})\Big|_{\tilde{G}} \xrightarrow{d} N(0,1)$$

as $N_E, N_C, T \to \infty$ jointly and $T/N_C \to \infty$. The conditioning set $\tilde{G}$ may be dropped as the limiting distribution is independent of $\tilde{G}$.

(b)(i) Assume heterogeneous ITET. From the DGP, we obtain:

$$\begin{aligned}
\hat{\delta}^{mg} - \bar{\bar{\Delta}} &= \frac{1}{N_E}\sum_{i\in E} \frac{1'_{post,i}M_{[\hat{F},X_i]}(1_{post,i}\upsilon_i + F\mu_i + X_i\beta_i + \epsilon_i)}{1'_{post,i}M_{[\hat{F},X_i]}1_{post,i}} \\
&= \frac{1}{N_E}\sum_{i\in E}\upsilon_i + \frac{1}{N_E T}\sum_{i\in E} \frac{1'_{post,i}M_{[\hat{F},X_i]}(F\mu_i + X_i\beta_i + \epsilon_i)}{1'_{post,i}M_{[\hat{F},X_i]}1_{post,i}/T} \\
&= E(\upsilon_i|i\in E) + O_p\left(\frac{1}{\sqrt{N_E T}}\right),
\end{aligned}$$

where the last step follows from part (a)(i) and Assumption RT$_{mg}$. Note that $E(\upsilon_i|i \in E) = 0$ by construction.

(b)(ii) After normalization, the simple mean-group estimator is decomposed as follows:

$$\sqrt{N_E}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) = \frac{1}{\sqrt{N_E}}\sum_{i\in E}\upsilon_i + o_p(1)$$

as $T, N_C \to \infty$ and $\frac{T}{N_C} \to 0$. By Assumption RT$_{mg}$(iii), the variance $\bar{\bar{\varsigma}}^2_{N_E,T} = Var(N_E^{-1/2}\sum_{i\in E}\upsilon_i) > 0$ for some $c > 0$ for sufficiently large N_E. The conditions in Assumption RT$_{mg}$ enable us to apply CLT for mixing sequences (Theorem 5.20 of White (2001)), yielding

$$\sqrt{N_E}\bar{\bar{\varsigma}}^{-1}_{N_E,T}(\hat{\delta}^{mg} - \bar{\bar{\Delta}}) \xrightarrow{d} N(0,1)$$

as $T, N_E, N_C \to \infty$ and $\frac{T}{N_C} \to 0$.

9.11 Proof of Theorem 3

(a)(i) Assume homogeneous ITET. From the DGP, we obtain:

$$\begin{aligned}
\left|\hat{\delta}^{pool} - \bar{\bar{\Delta}}\right| &\le \frac{1}{\sqrt{N_E T}} \frac{\left|\frac{1}{\sqrt{N_E T}}\sum_{i\in E} 1'_{post,i}M_{[\hat{F},X_i]}(F\mu_i + X_i\beta_i + \epsilon_i)\right|}{\frac{1}{N_E T}\sum_{j\in E} 1'_{post,j}M_{[\hat{F},X_j]}1_{post,j}} \\
&= \frac{1}{\sqrt{N_E T}} \frac{\left|\frac{1}{\sqrt{N_E T}}\sum_{i\in E} 1'_{post,i}M_{[F,X_i]}\epsilon_i\right|}{\frac{1}{N_E T}\sum_{j\in E} 1'_{post,j}M_{[F,X_j]}1_{post,j}} + o_p(1) \\
&= O_p\left(\frac{1}{\sqrt{N_E T}}\right),
\end{aligned}$$

where the second step is by lemmas A5(a) and A6, and the last step follows from Assumptions AI_{pl}(i) and (ii). The remainder term has mean zero in the limit by Assumption E(i).

(a)(ii) Denote $\tilde{G} := \{\tilde{G}_i\}_{i\in E}$. The key step of the proof involves showing

$$\frac{1}{\sqrt{N_E}\bar{\zeta}_{N_E,T}} \sum_{i\in E} S_{Ti}\Bigg|_{\tilde{G}} \xrightarrow{d} N(0,1) \tag{34}$$

as $N_E, T \to \infty$ jointly, where $\bar{\zeta}^2_{N_E,T} := Var(N_E^{-1/2}\sum_{i\in E} S_{Ti}|\tilde{G})$. To this end, we proceed by checking the conditions behind the Liapounov's theorem (Theorem 5.10 of White (2001)):

- For each $i \in E$, $E(S_{Ti}|\tilde{G}) = 0$ by Assumption ESS.
- For some $p > 2$, there exists $0 < C < \infty$ (uniformly over $i \in E$ and T) such that $E(|S_{Ti}|^p|\tilde{G}) \le C$ $w.p.a.1$ as $T \to \infty$. To show this, we first fix N_E and $i \in E$. By a similar argment that leads to (29) in the proof of Theorem 1 (except that the conditioning set is now set to $\tilde{G}$), we have $S_{Ti}|_{\tilde{G}} \xrightarrow{d} N(0, \zeta_i^2)$ as $T \to \infty$, where $\zeta_i^2 \ge 0$ is the asymptotic conditional variance (may be zero, in which case the limiting distribution is degenerate). Denote $\mu_z^{(p)} = E|Z|^p$ where $Z \sim N(0,1)$. Since all moments exist for the standard normal distribution, for any $p > 2$ and $c_1 > 0$, there exists $\tilde{T} = \tilde{T}(c_1, p, i, N_E)$ such that for all $T > \tilde{T}$, we have
$$E(|S_{Ti}|^p|\tilde{G}) \le \zeta_i^p \mu_z^{(p)} + c_1 \quad a.s.. \tag{35}$$
Set $\tilde{\tilde{T}} = \tilde{\tilde{T}}(c_1, p, N_E) = \max_{i\in E} \tilde{T}(c_1, p, i, N_E)$. Then we see that the inequality (35) holds for all $T > \tilde{\tilde{T}}$. By Assumption D(ii), there exists $C_2 > 0$ such that $\max_{i\in E} \zeta_i^2 \le C_2\zeta^2$ for sufficiently large N_E, where $\zeta^2 := \mathrm{plim}_{N_E\to\infty} N_E^{-1}\sum_{i\in E}\zeta_i^2$ is well defined according to Assumption AI_{pl}(ii). It follows from (35) that
$$\max_{i\in E} E(|S_{Ti}|^p|\tilde{G}) \le \max_{i\in E} \zeta_i^p \mu_z^{(p)} + c_1 \le C_2^{p/2}\zeta^p\mu_z^{(p)} + c_1.$$
This holds for all $p > 2$. The upper bound is independent of N_E and T and is finite (as ζ^2 exists by Assumption AI_{pl}(ii)). The claim holds by setting $C = C_2^{p/2}\zeta^p\mu_z^{(p)} + c_1$, which is independent of i and T.
- There exists $c > 0$ uniformly over N_E and T such that $\bar{\zeta}^2_{N_E,T} \ge c$ $w.p.a.1$ as $N_E, T \to \infty$. This follows immediately from Assumption AI_{pl}(ii).

By Liapounov's theorem, the Lindeberg condition is satisfied, i.e., for all $\varepsilon > 0$, we have

$$\lim_{N_E,T\to\infty} \frac{1}{N_E\bar{\zeta}^2_{N_E,T}} \sum_{i\in E} E(S^2_{Ti}1_{\{|S_{Ti}|>\varepsilon\sqrt{N_E}\bar{\zeta}_{N_E,T}\}}|\tilde{G}) = 0 \ a.s..$$

Applying the joint CLT result in Theorem 2 of Phillips and Moon (1999), we obtain (34).

Denote $\bar{R}_{N_E,T} := N_E^{-1}\sum_{i\in E} R_{Ti}$. Next, lemmas A5(a) and A6 imply that, as $N_C, T \to \infty$ and $T/N_C \to 0$,

$$\begin{aligned}\sqrt{N_E T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}}) &= \frac{1}{\sqrt{N_E}} \frac{\sum_{i\in E} S_{Ti}}{\bar{R}_{N_E,T}} + o_p(1) \\ &= \frac{\bar{\zeta}_{N_E,T}}{\bar{R}_{N_E,T}} \frac{1}{\sqrt{N_E}\bar{\zeta}_{N_E,T}} \sum_{i\in E} S_{Ti} + o_p(1). \end{aligned} \tag{36}$$

To normalize (36) we obtain the conditional variance as follows:

$$\bar{\sigma}^2_{N_E,T} = Var\left(\frac{1}{\sqrt{N_E}} \frac{\sum_{i\in E} S_{Ti}}{\bar{R}_{N_E,T}}\Bigg|\, \tilde{G}\right) + o(1) = \frac{\bar{\zeta}^2_{N_E,T}}{\bar{R}^2_{N_E,T}} + o(1).$$

For large N_E and T, we see that $\bar{\sigma}^2_{N_E,T} > 0$ because $\bar{\zeta}^2_{N_E,T} > 0$ (Assumption AI$_{pl}$(ii)) and $\bar{R}_{N_E,T} \leq 1$. After normalization by the conditional standard deviation, it follows from (36) that

$$\sqrt{N_E T}\bar{\sigma}^{-1}_{N_E,T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})\Big|_{\tilde{G}} = \frac{1}{\sqrt{N_E}\bar{\zeta}_{N_E,T}} \sum_{i \in E} S_{Ti} + o_p(1)\Bigg|_{\tilde{G}} \xrightarrow{d} N(0,1)$$

as $N_E, N_C, T \to \infty$ jointly and $T/N_C \to \infty$. We may drop the conditioning set $\tilde{G}$ as the limiting distribution is independent of $\tilde{G}$.

(b)(i) Assume heterogeneous ITET. From the DGP, we obtain:

$$\begin{aligned}
\hat{\delta}^{pool} - \bar{\bar{\Delta}} &= \frac{\sum_{i \in E} 1'_{post,i} M_{[\hat{F},X_i]}(F\mu_i + X_i\beta_i + \upsilon_i 1_{post,i} + \epsilon_i)}{\sum_{i \in E} 1'_{post,i} M_{[\hat{F},X_i]} 1_{post,i}} \\
&= \frac{1}{\sqrt{N_E}} \frac{\frac{1}{\sqrt{N_E T}} \sum_{i \in E} 1'_{post,i} M_{[F,X_i]} 1_{post,i} \upsilon_i}{\frac{1}{N_E T} \sum_{i \in E} 1'_{post,i} M_{[F,X_i]} 1_{post,i}} + \frac{1}{\sqrt{N_E T}} \frac{\frac{1}{\sqrt{N_E T}} \sum_{i \in E} 1'_{post,i} M_{[F,X_i]} \epsilon_i}{\frac{1}{N_E T} \sum_{i \in E} 1'_{post,i} M_{[F,X_i]} 1_{post,i}} + o_p(1) \\
&=: (I) + (II) + o_p(1).
\end{aligned}$$

From part (a)(i), the second term (II) is of order $O_p\left(\frac{1}{\sqrt{N_E T}}\right)$, so it remains to study (I). Let $R_{Ti} := 1'_{post,i} M_{[F,X_i]} 1_{post,i}/T$, so that

$$(I) = \frac{\frac{1}{N_E} \sum_{i \in E} R_{Ti} \upsilon_i}{\frac{1}{N_E} \sum_{i \in E} R_{Ti}}.$$

Assumption AI$_{pl}$(i) implies that there exists some $c_1 > 0$ such that the denominator of (I) is $\frac{1}{N_E} \sum_{i \in E} R_{Ti} \geq c_1 \;\; w.p.a.1$ for all sufficiently large N_E and T. Let us turn to the numerator. The conditional mean $E\left(\frac{1}{\sqrt{N_E}} \sum_{i \in E} R_{Ti} \upsilon_i \Big| MSR_T\right) = 0$ a.s. by Assumption RT$_{pl}$(i), and the conditional variance $Var\left(\frac{1}{\sqrt{N_E}} \sum_{i \in E} R_{Ti} \upsilon_i \Big| MSR_T\right) \leq c_2 \;\; w.p.a.1$ for all large N_E and T by Assumption RT$_{pl}$(iv). This implies that $Var\left(\frac{1}{\sqrt{N_E}} \sum_{i \in E} R_{Ti} \upsilon_i\right) \leq c_2$ for all large T and N_E, and so $\frac{1}{\sqrt{N_E}} \sum_{i \in E} R_{Ti} \upsilon_i$ is $O_p(1)$. It follows that $(I) = O_p\left(\frac{1}{\sqrt{N_E}}\right)$, and hence $\hat{\delta}^{pool} - \bar{\bar{\Delta}} = o_p(1)$.

(b)(ii) After normalization, the pooled estimator is decomposed as follows:

$$\sqrt{N_E}(\hat{\delta}^{pl} - \bar{\bar{\Delta}}) = \frac{\frac{1}{\sqrt{N_E}} \sum_{i \in E} R_{Ti} \upsilon_i}{\frac{1}{N_E} \sum_{i \in E} R_{Ti}} + o_p(1). \tag{37}$$

Denote $\bar{\bar{\zeta}}^2_{N_E,T} := Var(N_E^{-1/2} \sum_{i \in E} R_{Ti} \upsilon_i | MSR_T)$. As a core step, we show that the numerator in the above decomposition is asymptotically normal. To this end we check the following conditions:

- Assumption RT$_{pl}$(i) implies that $E(R_{Ti} \upsilon_i) = 0$ for all $i \in E$ and all T.
- Assumption RT$_{pl}$(ii) implies that, for some $p \geq 2$, there exists $0 < C < \infty$ (uniformly over all $i \in E$ and T) such that $E(|R_{Ti} \upsilon_i|^p \, |MSR_T) \leq E(|\upsilon_i|^p \, |MSR_T) \leq C \;\; w.p.a.1$ as $T \to \infty$. The first inequality follows from $R_{Ti} \leq 1$ uniformly over all T and i.
- Assumption RT$_{pl}$(iii) implies that, conditional on MSR_T, the sequence $\{R_{Ti} \upsilon_i : i \in E\}$ (T fixed) is mixing with properties as stated in the assumption.
- Assumption RT$_{pl}$(iv) entails that there exists $c > 0$ (uniform over all sufficiently large N_E and T) such that
$$\bar{\bar{\zeta}}^2_{N_E,T} \geq c \;\; w.p.a.1 \tag{38}$$
as $N_E, T \to \infty$.

By CLT for mixing sequences (Theorem 5.20 of White (2001)), we obtain

$$\frac{1}{\sqrt{N_E}\bar{\bar{\zeta}}_{N_E,T}} \sum_{i\in E} R_{Ti}\upsilon_i \Big|_{MSR_T} \xrightarrow{d} N(0,1) \tag{39}$$

as $N_E \to \infty$.

Next, in order to normalize (37) we compute its conditional variance:

$$\bar{\bar{\sigma}}^2_{N_E,T} := Var[\sqrt{N_E}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|MSR_T] = \bar{\bar{\zeta}}^2_{N_E,T}/\bar{R}^2_{N_E,T} + o_p(1).$$

We have $0 < \bar{\bar{\sigma}}^2_{N_E,T} < \infty \;\; w.p.a.1$ as $N_E, T \to \infty$, which follows from:

- $\bar{\bar{\sigma}}^2_{N_E,T} > 0 \;\; w.p.a.1$ as $N_E, T \to \infty$, by (38) and that $R_{Ti} \le 1$ for all T and i.
- $\bar{\bar{\sigma}}^2_{N_E,T} < \infty \;\; w.p.a.1$ as $N_E, T \to \infty$, by the fact that $\bar{R}_{N_E,T} > 0 \;\; w.p.a.1$ (Assumption AI$_{pl}$(i)), and that

$$\begin{aligned}\bar{\bar{\zeta}}^2_{N_E,T} &\le \frac{1}{N_E}\sum_{i,j\in E} |E(R_{Ti}R_{Tj}\upsilon_i\upsilon_j|MSR_T)| \\ &\le \frac{1}{N_E}\sum_{i,j\in E} |E(\upsilon_i\upsilon_j|MSR_T)| \le C < \infty \;\; w.p.a.1..\end{aligned}$$

  The second inequality follows from $R_{Ti} \le 1$, and the third one from Assumption RT$_{pl}$(iii), which implies $\frac{1}{N_E}\sum_{i,j\in E} |Cov(\upsilon_i, \upsilon_j|MSR_T)| \le C$ for some $C < \infty$ uniformly over N_E and T.

We can therefore normalize (37), meanwhile applying (39), and obtain a non-degenerate limiting distribution:

$$\sqrt{N_E}\bar{\bar{\sigma}}^{-1}_{N_E,T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})\Big|_{MSR_T} \xrightarrow{d} N(0,1)$$

as $N_E, N_C, T \to \infty$ jointly and $T/N_C \to \infty$. The conditioning set can be dropped as the limiting distribution is independent of MSR_T.

9.12 Proof of Theorem 4

Using the notations in Theorems 2 and 3, we have the relationships:

$$\sigma^2 = \operatorname*{plim}_{N_E,T\to\infty} \bar{\sigma}^2_{N_E,T}, \quad \tilde{\sigma}^2 = \operatorname*{plim}_{N_E,T\to\infty} \bar{\bar{\sigma}}^2_{N_E,T},$$
$$\varsigma^2 = \operatorname*{plim}_{N_E,T\to\infty} \bar{\varsigma}^2_{N_E,T}, \quad \tilde{\varsigma}^2 = \operatorname*{plim}_{N_E,T\to\infty} \bar{\bar{\varsigma}}^2_{N_E,T}.$$

(a) Rewrite the weighted mean-group estimator as

$$\hat{\delta}^{mg,(\omega)} - \bar{\bar{\Delta}} = \sum_{i\in E} \omega_i \frac{1'_{post,i}M_{G_i}\epsilon_i}{1'_{post,i}M_{G_i}1_{post,i}} + o_p(1).$$

Since the errors ϵ_{it} are $iid(0, \sigma^2_\epsilon)$ by the stated assumption, the conditional variance of $\hat{\delta}^{mg,(\omega)}$ is:

$$V := Var[\sqrt{N_E T}(\hat{\delta}^{mg,(\omega)} - \bar{\bar{\Delta}})|\tilde{G}] = N_E T \sum_{i\in E} \omega_i^2 \frac{\sigma^2_\epsilon}{1'_{post,i}M_{G_i}1_{post,i}} + o_p(1).$$

The solution to the constrained minimization problem

$$\min_{\{\omega_i\}} V \text{ subject to } \sum_{i\in E} \omega_i = 1 \tag{40}$$

is

$$\omega_i^* = \frac{1'_{post,i} M_{G_i} 1_{post,i}}{\sum_{i\in E} 1'_{post,i} M_{G_i} 1_{post,i}},$$

but then $\hat{\delta}^{mg,(\omega^*)} = \hat{\delta}^{pl}$. It follows that $V \geq Var[\sqrt{N_E T}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|\tilde{G}] =: \bar{\sigma}^2_{N_E,T}$ for all N_E and T. In particular, when we choose $\omega_i \equiv N_E^{-1}$, $\hat{\delta}^{mg,(\omega)}$ reduces to $\hat{\delta}^{mg}$, and $V = \bar{\varsigma}^2_{N_E,T}$. By optimality, we have $\bar{\varsigma}^2_{N_E,T} \geq \bar{\sigma}^2_{N_E,T}$. Letting $N_E, T \to \infty$ yields $\varsigma^2 \geq \sigma^2$, which becomes an equality iff ρ_i are identical over $i \in E$.

(b) Under the stated assumption on υ_i, the following inequality holds $a.s.$:

$$Var[\sqrt{N_E}(\hat{\delta}^{pl} - \bar{\bar{\Delta}})|\tilde{G}] = \frac{\sigma_\upsilon^2 \left(\frac{1}{N_E}\sum_{i\in E} R_{Ti}^2\right)}{\left(\frac{1}{N_E}\sum_{i\in E} R_{Ti}\right)^2} \geq \sigma_\upsilon^2 = Var[\sqrt{N_E}(\hat{\delta}^{mg} - \bar{\bar{\Delta}})].$$

In the above we applied Jensen's inequality, which achieves equality iff R_{Ti} are identical over $i \in E$. Since the inequality holds for all N_E and T, the conclusion follows by taking probability limit as $N_E, T \to \infty$.

9.13 Proof of Theorem 5

(a) To show consistency, we decompose $\hat{a}^{mg}$ as follows:

$$\begin{aligned}
\hat{a}^{mg} &= \frac{1}{N_E}\sum_{i\in E}\hat{a}_i = \frac{1}{N_E}\sum_{i\in E}\frac{\bar{u}'_C M_{[1_{post,j},X_i]} y_i}{\bar{u}'_C M_{[1_{post,i},X_i]}\bar{u}_C} \\
&= \frac{1}{N_E}\sum_{i\in E}\frac{\bar{u}'_C M_{[1_{post,j},X_i]} F\mu_i}{\bar{u}'_C M_{[1_{post,i},X_i]}\bar{u}_C} + \frac{1}{N_E}\sum_{i\in E}\frac{\bar{u}'_C M_{[1_{post,j},X_i]}\epsilon_i}{\bar{u}'_C M_{[1_{post,i},X_i]}\bar{u}_C} \\
&=: (I) + (II).
\end{aligned} \tag{41}$$

We analyze the two terms as follows. The first term on the last line can be rewritten as

$$\begin{aligned}
(I) &= \frac{1}{N_E}\sum_{i\in E}\frac{\bar{u}'_C M_{[1_{post,j},X_i]} F\mu_i/T^{2r}}{\bar{u}'_C M_{[1_{post,i},X_i]}\bar{u}_C/T^{2r}} \\
&= \frac{1}{N_E}\sum_{i\in E}\frac{\mu'_0 F' M_{[1_{post,j},X_i]} F\mu_i/T^{2r}}{\mu'_0 F' M_{[1_{post,i},X_i]} F\mu_0/T^{2r}} + o_p(1).
\end{aligned}$$

as $N_C, T \to \infty$, where the second equality utilizes lemmas A7(a)-(b). The term (II) can be bounded from above as follows:

$$\begin{aligned}
|(II)| &= \left|\frac{1}{N_E}\sum_{i\in E}\frac{\bar{u}'_C M_{[1_{post,j},X_i]}\epsilon_i/T^{2r}}{\bar{u}'_C M_{[1_{post,i},X_i]}\bar{u}_C/T^{2r}}\right| \\
&\leq \frac{\left|\frac{1}{N_E}\sum_{i\in E}\bar{u}'_C M_{[1_{post,j},X_i]}\epsilon_i/T^{2r}\right|}{\inf_i \left|\mu'_0 F' M_{[1_{post,i},X_i]} F\mu_0/T^{2r}\right|} + o_p(1) \\
&\leq \frac{1}{c}\left|\frac{1}{N_E T^{2r}}\sum_{i\in E}\bar{u}'_C M_{[1_{post,j},X_i]}\epsilon_i\right| + o_p(1) \\
&= \frac{1}{c}\frac{1}{\sqrt{N_E}T^r}\left|\frac{1}{\sqrt{N_E}T^r}\sum_{i\in E}\mu'_0 F' M_{[1_{post,j},X_i]}\epsilon_i\right| + o_p(1) \\
&= O_p\left(\frac{1}{\sqrt{N_E}T^r}\right)
\end{aligned}$$

as $N_E, N_C, T \to \infty$ and $T/N_C \to 0$, where the second inequality holds for some $c > 0$ by Assumption AI$_\alpha$(i), and the last equality follows from lemma A7(c). Substituting into (41) and subtracting one from both sides, we obtain

$$\hat{a}^{mg} - 1 = \frac{1}{N_E}\sum_{i\in E}\frac{\mu_0' F' M_{[1_{post,j},X_i]}F(\mu_i-\mu_0)/T^{2r}}{\mu_0' F' M_{[1_{post,i},X_i]}F\mu_0/T^{2r}} + O_p\left(\frac{1}{\sqrt{N_E}T^r}\right).$$

Applying Assumption AI$_\alpha$(i) again, there exists $c > 0$ such that

$$\begin{aligned}|\hat{a}^{mg} - 1| &\le \frac{1}{cN_E T^{2r}}\sum_{i\in E}\left|\mu_0' F' M_{[1_{post,j},X_i]}F(\mu_i-\mu_0)\right| + o_p(1)\\ &\le \frac{1}{c}\|\mu_0\|\left\|\frac{F'F}{T^{2r}}\right\|\cdot\frac{1}{N_E}\sum_{i\in E}\|\mu_i-\mu_0\| + o_p(1),\end{aligned}$$

but then the dominant term on the last line is $o_p(1)$ by the finiteness of μ_0, by Assumption F(i) and by Assumption FLM, which implies $N_E^{-1}\sum_{i\in E}\mu_i \xrightarrow{p} E(\mu_i|i\in E) = \mu_0$.

(b) Under the weak parallel trend hypothesis, we have the decomposition as $N_E, N_C, T \to \infty$ jointly and $T/N_C \to \infty$:

$$\sqrt{N_E}(\hat{a}^{mg} - 1) = \frac{1}{\sqrt{N_E}}\sum_{i\in E}h_{Ti}'(\mu_i - \mu_0) + o_p(1), \tag{42}$$

where $h_{Ti} := \frac{F' M_{[1_{post,i},X_i]}F\mu_0}{\mu_0' F' M_{[1_{post,i},X_i]}F\mu_0}$.

Our goal is to show that

$$\left.\frac{1}{\sqrt{N_E}\bar{\varphi}_{N_E,T}}\sum_{i\in E}h_{Ti}'(\mu_i-\mu_0)\right|_{\tilde{G}} \xrightarrow{d} N(0,1) \tag{43}$$

as $N_E, T \to \infty$ jointly, where $\bar{\varphi}^2_{N_E,T} = Var[N_E^{-1/2}\sum_{i\in E}h_{Ti}'(\mu_i-\mu_0)|\tilde{G}]$. To this end we check the conditions for applying Liapounov's theorem:

- For each $i \in E$, $E[h_{Ti}'(\mu_i-\mu_0)|\tilde{G}] = 0$ $a.s.$ by Assumption FLM2(i) and that h_{Ti} is measurable with respect to $\tilde{G}$.

- For some $p > 2$, there exists $C < \infty$ (uniformly over $i \in E$ and independent of T) such that $E(\|h_{Ti}'(\mu_i-\mu_0)\|^p|\tilde{G}) \le C$ $w.p.a.1$ as $T \to \infty$. To see this, we rewrite $E(\|h_{Ti}'(\mu_i-\mu_0)\|^p|\tilde{G}) = \|h_{Ti}\|^p E(\|\mu_i-\mu_0\|^p|\tilde{G})$. First, we have $E(\|\mu_i-\mu_0\|^p|\tilde{G}) \le C_1 < \infty$ $a.s.$ by Assumption FLM2(ii). On the other hand, by Assumption AI$_\alpha$(i), there exists $c_2 > 0$ uniformly over $i \in E$ such that $\mu_0' F' M_{[1_{post,i},X_i]}F\mu_0/T^{2r} \ge c_2 > 0$ for all sufficiently large T, which implies

  $$\|h_{Ti}\|^p \le c_2^{-p}\left\|F' M_{[1_{post,i},X_i]}F\mu_0/T^{2r}\right\|^p \le c_2^{-p}\|\mu_0\|^p\left\|F'F/T^{2r}\right\|^p.$$

  But then $\|F'F/T^{2r}\| \le \|\Upsilon^{-1}F'F\Upsilon^{-1}\| \le C_3 < \infty$ $w.p.a.1$ as $T \to \infty$ (as $\|\Upsilon^{-1}F'F\Upsilon^{-1}\|$ is bounded in probability by Assumption F(i)), and $\|\mu_0\|$ is bounded (as $\mu_0 = E(\mu_i|i\in E)$ is finite by hypothesis). The claim then follows by picking $C = c_2^{-p}C_3^p C_1 < \infty$. Note that the upper bound C is independent of i and T.

- By Assumption FLM2(iii), we see that, conditional on $\tilde{G}$, $\{h_{Ti}'(\mu_i-\mu_0) : i \in E\}$ is a mixing sequence with mixing coefficient as given in the assumption.

- There exists $c > 0$ such that $\bar{\varphi}^2_{N_E,T} \geq c \;\; w.p.a.1$ as $N_E, T \to \infty$. To show this, we observe that $\mu_0' F' M_{[1_{post,i},X_i]} F \mu_0 / T^{2r} \leq \left\| F'F/T^{2r} \right\| \|\mu_0\|^2 \leq C_1 < \infty$ uniformly over $i \in E \;\; w.p.a.1$ as $N_E, T \to \infty$. On the other hand, Assumption AI_α(ii) implies that

$$E\left(\left\| N_E^{-1/2} T^{-r} \sum_{i \in E} \mu_0' F' M_{[1_{post,i},X_i]} F(\mu_i - \mu_0) \right\|^2 \middle| \tilde{G} \right) \geq c_2$$

$w.p.a.1$ as $N_E, T \to \infty$. The claim holds by setting $c = C_1^{-1} c_2 > 0$.

By Liapounov's theorem, the Lindeberg condition is satisfied, i.e., for all $\varepsilon > 0$, we have

$$\lim_{N_E,T\to\infty} \frac{1}{N_E \bar{\varphi}^2_{N_E,T}} \sum_{i \in E} E\left[\left| h_i'(\mu_i - \mu_0) \right|^2 1_{\{|h_i'(\mu_i-\mu_0)| > \varepsilon \sqrt{N_E} \bar{\varphi}_{N_E,T}\}} \middle| \tilde{G} \right] = 0 \; a.s..$$

Applying the joint CLT result (Theorem 2 of Phillips and Moon (1999)) yields (43).

Finally, applying (43) to the decomposition (42) and normalizing by $\bar{\varphi}_{N_E,T}$, we obtain

$$\sqrt{N_E} \bar{\varphi}^{-1}_{N_E,T} (\hat{a}^{mg} - 1) \Big|_{\tilde{G}} \xrightarrow{d} N(0,1)$$

as $N_E, N_C, T \to \infty$ and $T/N_C \to \infty$. The conditioning set $\tilde{G}$ may be dropped as the limiting distribution is independent of $\tilde{G}$.

9.14 Proof of Proposition S1

Condition (2) implies[9] that 1_{post} is not spanned by the columns of F, i.e.,

$$1_{post} \neq Fv \text{ for all } v \neq 0. \tag{44}$$

Furthermore, condition (2) implies that $G = G_0 + G_1$ is of full column rank (G_1 is the same as G except that the first T_0 rows set to zero), so that $\frac{G'G}{T}$ is invertible and positive definite for any finite T. It follows that there exists $\Psi := \left(\frac{G'G}{T}\right)^{-1/2}$ exists such that $\Psi\Psi = \left(\frac{G'G}{T}\right)^{-1}$.

Now, decompose $1_{post}' P_G 1_{post}/T$ as follows:

$$\frac{1_{post}' P_G 1_{post}}{T} = \left(\frac{1_{post}' G}{T}\right) \left(\frac{G'G}{T}\right)^{-1} \left(\frac{G' 1_{post}}{T}\right). \tag{45}$$

Applying Cauchy-Schwarz inequality, we have

$$\frac{1_{post}' G}{T} < \left(\frac{1_{post}' 1_{post}}{T}\right)^{1/2} \left(\frac{G'G}{T}\right)^{1/2} = \left(\frac{T_1}{T}\right)^{1/2} \Psi^{-1}.$$

The inequality is strict because of (44).[10] Substituting into the decomposition (45), we obtain

$$\frac{1_{post}' P_G 1_{post}}{T} < \left(\frac{T_1}{T}\right)^{1/2} \Psi^{-1}(\Psi\Psi)\Psi^{-1} \left(\frac{T_1}{T}\right)^{1/2} = \frac{T_1}{T},$$

which implies condition (3). The proof is completed.

[9] Suppose 1_{post} is in the range space of F, i.e., there exists a vector v such that $1_{post} = Fv$. Then, there is multicollinearity: $1_{post}' P_F 1_{post} = 1_{post}' 1_{post} = T_1$ and $\hat{\delta}_i^{pcdid}$ is undefined. Importantly, $1_{post} = Fv$ implies $0_{T_0 \times 1} = F_0 v$, which implies that $v' F_0' F_0 v$. This implies that $F_0' F_0$ does not have full rank and the factor loading estimate in GSC estimator is undefined as well.

[10] Equality is achieved iff 1_{post} is proportional to some linear combinations of the columns of F.

9.15 Proof of Proposition S2

Consider the decomposition

$$\frac{G'G}{T} = \frac{G_0'G_0}{T} + \frac{G_1'G_1}{T} = \frac{T_0}{T}\frac{G_0'G_0}{T_0} + \frac{T_1}{T}\frac{G_1'G_1}{T_1}. \tag{46}$$

Both terms on the right are positive semidefinite in the limit. By (4) and $\kappa < 1$, the first term on the right is positive definite in the limit, and hence $\Sigma := \text{plim}_{T\to\infty} \frac{G'G}{T}$ exists and is invertible.

Next, we study

$$\frac{1_{post}' P_G 1_{post}}{T} = \frac{1_{post}' G}{T}\left(\frac{G'G}{T}\right)^{-1}\frac{G' 1_{post}}{T}.$$

By Cauchy-Schwarz inequality, the term $\frac{G'1_{post}}{T}$ is bounded from above: $\left\|\frac{G'1_{post}}{T}\right\| \leq \left(\frac{T_1}{T}\right)^{1/2}\left\|\frac{G'G}{T}\right\|^{1/2}$. Taking probability limits both sides, we have

$$\underset{T_1,T\to\infty}{\text{plim}}\left\|\frac{G'1_{post}}{T}\right\| \leq \kappa^{1/2}\left\|\Sigma\right\|^{1/2}$$

Note that condition (4) implies that 1_{post} is not in the column span of F in the limit as $T, T_1 \to \infty$. Coupled with the assumption $\kappa > 0$, it follows that the weak inequality is a strict one:

$$\underset{T_1,T\to\infty}{\text{plim}}\left\|\frac{G'1_{post}}{T}\right\| < \kappa^{1/2}\left\|\Sigma\right\|^{1/2}.$$

We thus have

$$\begin{aligned}
\underset{T_1,T\to\infty}{\text{plim}}\frac{1_{post}' P_G 1_{post}}{T} &= \underset{T_1,T\to\infty}{\text{plim}}\left\|\frac{1_{post}'G}{T}\right\| \underset{T\to\infty}{\text{plim}}\left\|\frac{G'G}{T}\right\|^{-1} \underset{T_1,T\to\infty}{\text{plim}}\left\|\frac{G'1_{post}}{T}\right\| \\
&< \kappa^{1/2}\left\|\Sigma\right\|^{1/2}\Sigma^{-1}\left\|\Sigma\right\|^{1/2}\kappa^{1/2} \\
&= \kappa,
\end{aligned}$$

which implies that

$$\underset{T_1,T\to\infty}{\text{plim}}\frac{1_{post}' M_G 1_{post}}{T} = \kappa - \underset{T_1,T\to\infty}{\text{plim}}\frac{1_{post}' P_G 1_{post}}{T} > 0.$$

The proof is completed.

9.16 Proof of Lemma S1

First, let us decompose the estimator $\hat{\delta}_{it}^{gsc}$. For each $i \in E$, we have

$$\begin{aligned}
\hat{\delta}_{it}^{gsc} - \Delta_{it} &= y_{it} - \hat{\beta}_0 - \hat{\lambda}_i'\hat{f}_t - \Delta_{it} \\
&= (\beta_0 - \hat{\beta}_0) + (\lambda_i' f_t - \hat{\lambda}_i'\hat{f}_t) + \tilde{\epsilon}_{it} \\
&= (\beta_0 - \hat{\beta}_0) + (\lambda_i' H'^{-1} H' f_t - \hat{\lambda}_i'\hat{f}_t) + \tilde{\epsilon}_{it} \\
&= (\beta_0 - \hat{\beta}_0) - \hat{\lambda}_i'(\hat{f}_t - H'f_t) - (\hat{\lambda}_i - H^{-1}\lambda_i)' H' f_t + \tilde{\epsilon}_{it}.
\end{aligned}$$

where H is an $\ell \times \ell$ invertible rotation matrix.

Next, let us evaluate the estimation error of $\hat{\delta}_i^{gsc}$. Recall that the target parameter is $\bar{\Delta}_i$. Since $\Delta_{it} = \bar{\Delta}_i + \tilde{\Delta}_{it}$ and $\epsilon_{it} = \tilde{\epsilon}_{it} + \tilde{\Delta}_{it}$, we obtain

$$\begin{aligned}
&\hat{\delta}_i^{gsc} - \bar{\Delta}_i \\
&= \frac{1}{T_1}\sum_{t>T_0}(\hat{\delta}_{it}^{gsc} - \Delta_{it}) + \frac{1}{T_1}\sum_{t>T_0}\tilde{\Delta}_{it} \\
&= (\beta_0 - \hat{\beta}_0) - \hat{\lambda}_i'\left[\frac{1}{T_1}\sum_{t>T_0}(\hat{f}_t - H'f_t)\right] - (\hat{\lambda}_i - H^{-1}\lambda_i)'H'\left(\frac{1}{T_1}\sum_{t>T_0}f_t\right) + \frac{1}{T_1}\sum_{t>T_0}(\tilde{\epsilon}_{it} + \tilde{\Delta}_{it}) \\
&= (\beta_0 - \hat{\beta}_0) - \hat{\lambda}_i'\left[\frac{1}{T_1}\sum_{t>T_0}(\hat{f}_t - H'f_t)\right] - (\hat{\lambda}_i - H^{-1}\lambda_i)'H'\left(\frac{1}{T_1}\sum_{t>T_0}f_t\right) + \frac{1}{T_1}\sum_{t>T_0}\epsilon_{it}. \qquad (47)
\end{aligned}$$

First note that $\beta_0 - \hat{\beta}_0$ is $O_p\left(\frac{1}{\sqrt{T}}\right)$. To proceed, let us evaluate the estimation error of $\hat{\lambda}_i$. From the way the factor loadings are constructed, we see that the factor loading estimator $\hat{\lambda}_i$ is given by

$$\hat{\lambda}_i = \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s y_{is}^0.$$

From the DGP of y_{is}^0, and the observation that $\tilde{\Delta}_{it} = 0$ for $t \leq T_0$, we can decompose it as follows:

$$\begin{aligned}
\hat{\lambda}_i &= \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s f_s'\lambda_i + \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s\epsilon_{is}. \\
&= \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s(f_s'H)(H^{-1}\lambda_i) + \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s\epsilon_{is} \\
&= H^{-1}\lambda_i - \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s(\hat{f}_s' - f_s'H)(H^{-1}\lambda_i) + \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s\epsilon_{is},
\end{aligned}$$

The estimation error of $\hat{\lambda}_i$ is thus given by

$$\hat{\lambda}_i - H^{-1}\lambda_i = -\left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s(\hat{f}_s' - f_s'H)(H^{-1}\lambda_i) + \left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s\epsilon_{is}.$$

We have the uniform bound result due to Bai (2003, Proposition 2):

$$\max_t \left\|\hat{f}_t - H'f_t\right\| = O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\sqrt{\frac{T}{N_C}}\right). \qquad (48)$$

Using (48), we obtain a uniform bound (over i) on the bias of $\hat{\lambda}_i$:

$$\|bias_{\lambda_i}\| = \left\|\left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s(\hat{f}_s' - f_s'H)(H^{-1}\lambda_i)\right\| \leq O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\sqrt{\frac{T}{N_C}}\right).$$

Provided that $T/N_C \to 0$ and $T \to \infty$, we have

$$\hat{\lambda}_i - H^{-1}\lambda_i = \left(\sum_{u=1}^{T_0}\hat{f}_u\hat{f}_u'\right)^{-1}\sum_{s=1}^{T_0}\hat{f}_s\epsilon_{is},$$

and hence

$$\frac{1}{N_E}\sum_{i\in E}(\hat{\lambda}_i - H^{-1}\lambda_i) = \left(\sum_{u=1}^{T_0}\hat{f}_u\hat{f}_u'\right)^{-1}\frac{1}{N_E}\sum_{i\in E}\sum_{s=1}^{T_0}\hat{f}_s'\epsilon_{is}.$$

Substituting back into (47), we have

$$\begin{aligned}
&\hat{\delta}_i^{gsc} - \bar{\Delta}_i \\
=& -\hat{\lambda}_i'\left[\frac{1}{T_1}\sum_{t>T_0}(\hat{f}_t - H'f_t)\right] - \left(\sum_{s=1}^{T_0}\epsilon_{is}\hat{f}_s'\right)\left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}H'\left(\frac{1}{T_1}\sum_{t>T_0}f_t\right) \\
&+\frac{1}{T_1}\sum_{t>T_0}\epsilon_{it} + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\sqrt{\frac{T}{N_C}}\right) \\
=& -\left(\sum_{s=1}^{T_0}\epsilon_{is}\hat{f}_s'\right)\left(\sum_{s=1}^{T_0}\hat{f}_s\hat{f}_s'\right)^{-1}H'\left(\frac{1}{T_1}\sum_{t>T_0}f_t\right) + \frac{1}{T_1}\sum_{t>T_0}\epsilon_{it} + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\sqrt{\frac{T}{N_C}}\right),
\end{aligned}$$

where we use (48) in the last step.

We may express the last line in matrix form:

$$\begin{aligned}
\hat{\delta}_i^{gsc} - \bar{\Delta}_i &= \frac{1}{T_1}1_{post}'\epsilon_i - \frac{1}{T_1}1_{post}'FH(\hat{F}_0'\hat{F}_0)^{-1}\hat{F}_0'\epsilon_i + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\sqrt{\frac{T}{N_C}}\right) \\
&= \frac{1}{T_1}1_{post}'\epsilon_i - \frac{1}{T_1}1_{post}'F(F_0'F_0)^{-1}F_0'\epsilon_i + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\sqrt{\frac{T}{N_C}}\right).
\end{aligned}$$

where we use (48) again in the last step. The result follows immediately.

9.17 Proof of Proposition S3

Since $E(f_t) = 0$, it follows that

$$(1-\kappa)b_0 + \kappa b_1 = 0. \tag{49}$$

Moreover, the assumptions imply that

$$\begin{aligned}
\sigma_f^2 := Var(f_t) &= E[Var(f_t|1\{t>T_0\})] + Var[E(f_t|1\{t>T_0\})] \\
&= v_f(1-\kappa) + v_f\kappa + b_0^2(1-\kappa) + b_1^2\kappa - [b_0(1-\kappa)+b_1\kappa]^2 \\
&= v_f + b_0^2(1-\kappa) + b_1^2\kappa \quad \text{(by (49))}.
\end{aligned} \tag{50}$$

Define $m^{(k)} := \frac{1}{T}\sum_t f_t^k$, $m_{pre}^{(k)} := \frac{1}{T}\sum_{t\le T_0} f_t^k$, and $m_{post}^{(k)} := \frac{1}{T}\sum_{t>T_0} f_t^k$. By LLN, we have $m^{(1)} \to 0$, $m_{pre}^{(1)} \to (1-\kappa)b_0$, $m_{post}^{(1)} \to \kappa b_1$, $m^{(2)} \to \sigma_f^2$, $m_{pre}^{(2)} \to (1-\kappa)(v_f+b_0^2)$ and $m_{post}^{(2)} \to \kappa(v_f+b_1^2)$ as $T, T_0, T_1 \to \infty$. Denote $A := \frac{b_1^2}{v_f}$ and $B := \frac{b_1^2}{\sigma_f^2}$.

Let G be a $T\times 2$ matrix with t^{th} row given by $[1, f_t]$, and G_0 be the same as G except that all the entries in the last T_1 rows are set to zero. Now let us compute

$$\begin{aligned}
\frac{1'_{post}G(G'_0G_0)^{-1}G'1_{post}}{T_1} &= \frac{1}{T_1}\frac{T_1^2(\sum_{t\le T_0} f_t^2) - 2T_1(\sum_{t\le T_0} f_t)(\sum_{t>T_0} f_t) + T_0(\sum_{t>T_0} f_t)^2}{T_0(\sum_{t\le T_0} f_t^2) - (\sum_{t\le T_0} f_t)^2} \\
&= \frac{\frac{T_1}{T}m_{pre}^{(2)} - 2m_{pre}^{(1)}\cdot m_{post}^{(1)} + \frac{T_0}{T_1}(m_{post}^{(1)})^2}{\frac{T_0}{T}m_{pre}^{(2)} - (m_{pre}^{(1)})^2} \\
&\to \frac{\kappa(1-\kappa)(v_f+b_0^2) - 2\kappa(1-\kappa)b_0b_1 + \frac{1-\kappa}{\kappa}\kappa^2 b_1^2}{(1-\kappa)^2(v_f+b_0^2) - (1-\kappa)^2 b_0^2} \\
&= \frac{\kappa}{1-\kappa}\frac{v_f + \left(\frac{\kappa}{1-\kappa}\right)^2 b_1^2 + 2\left(\frac{\kappa}{1-\kappa}\right)b_1^2 + b_1^2}{v_f} \\
&= \frac{\kappa}{1-\kappa}\left[1 + \left(\frac{\kappa}{1-\kappa}+1\right)^2 A\right],
\end{aligned}$$

where the second-to-last equality is obtained by using (49). Next we compute

$$\begin{aligned}
\frac{1'_{post}P_G1_{post}}{T_1} &= \frac{1'_{post}G(G'G)^{-1}G'1_{post}}{T_1} \\
&= \frac{1}{T_1}\frac{T_1^2(\sum_t f_t^2) - 2T_1(\sum_t f_t)(\sum_{t>T_0} f_t) + T(\sum_{t>T_0} f_t)^2}{T(\sum_t f_t^2) - (\sum_t f_t)^2} \\
&= \frac{\frac{T_1}{T}m^{(2)} - 2m^{(1)}\cdot m_{post}^{(1)} + \frac{T}{T_1}(m_{post}^{(1)})^2}{m^{(2)} - (m^{(1)})^2} \\
&\to \frac{\kappa\sigma_f^2 + \frac{1}{\kappa}\kappa^2 b_1^2}{\sigma_f^2} \\
&= \kappa\,(1+B) = \kappa\left(1 + \frac{A}{1+\left(\frac{\kappa^2}{1-\kappa}\right)A + \kappa A}\right),
\end{aligned} \tag{51}$$

where the last line is obtained by (52), which is based on (50):

$$B = \frac{A}{1+\left(\frac{\kappa^2}{1-\kappa}\right)A + \kappa A} \Longleftrightarrow A = \frac{B}{1-\left(\frac{\kappa^2}{1-\kappa}\right)B - \kappa B}. \tag{52}$$

Denote $\theta := \frac{\kappa}{1-\kappa}$. The asymptotic variance of GSC estimator is:

$$\begin{aligned}
\operatorname{plim} V^{gsc} &= \operatorname{plim} \sigma_\epsilon^2 \frac{T}{T_1}\left(1 + \frac{1'_{post}G(G'_0G_0)^{-1}G'1_{post}}{T_1}\right) \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\left(1 + \theta\left[1 + (\theta+1)^2 A\right]\right).
\end{aligned} \tag{53}$$

The asymptotic variance of PCDID estimator is:

$$\begin{aligned}
\operatorname{plim} V^{pcdid} &= \operatorname{plim} \frac{T\sigma_\epsilon^2}{1'_{post}M_G1_{post}} = \operatorname{plim} \frac{T}{T_1}\frac{\sigma_\epsilon^2}{1 - 1'_{post}P_G1_{post}/T_1} \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\frac{1}{1-\kappa\left(1+\frac{A}{1+\kappa\theta A + \kappa A}\right)}.
\end{aligned} \tag{54}$$

When $b_1 = b_0 = 0$, we have $A = B = 0$ and therefore $\operatorname{plim} V^{gsc} = \operatorname{plim} V^{pcdid} = \sigma_\epsilon^2 \frac{1}{\kappa(1-\kappa)}$.

When $b_1 \neq 0$ ($b_0 \neq 0$), we have $\text{plim}\, V^{gsc} > \text{plim}\, V^{pcdid}$. To deduce this, observe that $\theta > 0$ (by $0 < \kappa < 1$) and $A > 0$ (by $b_1^2 > 0$ and $v_f > 0$).[11] It follows by simple algebra that

$$\begin{aligned}
\text{plim}\, V^{gsc} - \text{plim}\, V^{pcdid} &= \sigma_\epsilon^2 \frac{1}{\kappa}\left[\left(1+\theta\left[1+(\theta+1)^2 A\right]\right) - \frac{1}{1-\kappa\left(1+\frac{A}{1+\kappa\theta A+\kappa A}\right)}\right] \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\left[1+\theta+\theta(\theta+1)^2 A - \frac{1}{1-\kappa\left(\frac{1+(\kappa\theta+\kappa+1)A}{1+\kappa\theta A+\kappa A}\right)}\right] \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\left[1+\theta+\theta(\theta+1)^2 A - \frac{1+\kappa\theta A+\kappa A}{1-\kappa}\right] \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\frac{[1+\theta+\theta(\theta+1)^2 A](1-\kappa) - (1+\kappa\theta A+\kappa A)}{1-\kappa} \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\frac{\theta(\theta+1)^2 A(1-\kappa) - (\theta+1)\kappa A}{1-\kappa} \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\frac{(1-\kappa)\theta^3 + 2(1-\kappa)\theta^2 - \kappa\theta}{1-\kappa} A \\
&= \sigma_\epsilon^2 \frac{1}{\kappa}\theta^2(\theta+1)A > 0, \qquad (55)
\end{aligned}$$

where the third, the fifth and the sixth steps are obtained from $\theta - \theta\kappa - \kappa = 0$ (as $\theta = \frac{\kappa}{1-\kappa}$).

9.18 Proof of Proposition A

There are $\ell+1$ unknowns $\alpha_j, v_{1j}, v_{2j}, \ldots, v_{\ell j}$ to be solved for. Condition (i) implies $v_{1j} = -\sum_{s=2}^{\ell} v_{2j}$, thus reducing the number of unknowns to ℓ. Now we may express (10) as a system of equations:

$$\begin{pmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{\ell j} \end{pmatrix} = \begin{pmatrix} E(\mu_{1i}|i \in C) & -1 & \cdots & -1 \\ E(\mu_{2i}|i \in C) & 1 & & O \\ \vdots & & \ddots & \\ E(\mu_{\ell i}|i \in C) & O & & 1 \end{pmatrix} \begin{pmatrix} \alpha_j \\ v_{2j} \\ \vdots \\ v_{\ell j} \end{pmatrix}.$$

Let M be the $\ell \times \ell$ coefficient matrix with column vectors denoted by $m_1, \ldots, m_\ell$. It remains to show that restrictions (i) and (ii) hold iff M is invertible.

($\Rightarrow$): Restriction (ii) implies that v_j and $E(\mu_i|i \in C)$ are linearly independent. Restriction (i) implies that v_j is in the linear space spanned by $m_2, \ldots, m_\ell$, which are linearly independent $\ell \times 1$ vectors and have zero column sum. It follows that $m_2, \ldots, m_\ell$ and $E(\mu_i|i \in C)$ are linearly independent, and hence M is invertible.

($\Leftarrow$): Suppose the contrary that (i) holds but (ii) is violated, i.e., there exists $\gamma_j \neq 0$ such that $v_j = \gamma_j E(\mu_i|i \in C)$. It follows that the elements of $E(\mu_i|i \in C) = (1/\gamma_j)v_j$ sum to zero. Since $E(\mu_i|i \in C)$ has zero column sum, it is expressible as a linear combination of $m_2, \ldots, m_\ell$, and hence M is of rank $\ell - 1$ and is singular.

9.19 Proof of Proposition H

(a) We first note that, under the given DGP, the 2wfe estimator is equivalent to the simple DID estimator, given by

$$\hat{\delta}^{2wfe} := \frac{1}{N_E}\sum_{i \in E} \hat{a}_i - \frac{1}{N_C}\sum_{i \in C} \hat{a}_i,$$

[11] Moreover, $0 < B < A$ because $b_1^2 > 0$ and $\sigma_f^2 > v_f$; by equation (52) and $A, B > 0$, we must have $\left(\frac{\kappa^2}{1-\kappa}\right) B + \kappa B < 1$, which implies $\kappa(1+B) < 1$.

where $\hat{a}_i$ be the OLS estimators of the simple regression $y_i = b_{0i}1 + a_i 1_{post} + e_i$. We may now decompose the $2wfe$ estimator as follows:

$$\begin{aligned}\hat{a}_i &= (1'_{post}M_1 1_{post})^{-1}1'_{post}M_1 y_i \\ &= (1'_{post}M_1 1_{post})^{-1}1'_{post}M_1(\varsigma_i 1 + f + \bar{\bar{\Delta}}1_{post}1_{\{i\in E\}} + \epsilon_i) \\ &= \bar{\bar{\Delta}}1_{\{i\in E\}} + (1'_{post}M_1 1_{post})^{-1}1'_{post}M_1(f + \epsilon_i). \end{aligned} \tag{56}$$

Define $\bar{\epsilon}_E = \frac{1}{N_E}\sum_{i\in E}\epsilon_i$, $\bar{\epsilon}_C = \frac{1}{N_C}\sum_{i\in C}\epsilon_i$. The $2wfe$ estimator takes the following decomposition:

$$\begin{aligned}\hat{\delta}^{2wfe} - \bar{\bar{\Delta}} &= (1'_{post}M_1 1_{post})^{-1}1'_{post}M_1(\bar{\epsilon}_E - \bar{\epsilon}_C) \\ &= (\bar{\epsilon}_{E,post} - \bar{\epsilon}_{C,post}) - (\bar{\epsilon}_{E,pre} - \bar{\epsilon}_{C,pre}).\end{aligned}$$

Since $Var(\epsilon_i) = \sigma_\epsilon^2 I$, where I is the $T\times T$ identity matrix, we have

$$\begin{aligned}Var(\bar{\epsilon}_E - \bar{\epsilon}_C) &= E[(\bar{\epsilon}_E - \bar{\epsilon}_C)(\bar{\epsilon}_E - \bar{\epsilon}_C)'] \\ &= E(\bar{\epsilon}_E\bar{\epsilon}'_E) + E(\bar{\epsilon}_C\bar{\epsilon}'_C) \\ &= \sigma_\epsilon^2\left(\frac{1}{N_E} + \frac{1}{N_C}\right)I.\end{aligned}$$

The variance of $\hat{\delta}^{2wfe}$ is given by

$$\begin{aligned}Var(\hat{\delta}^{2wfe}) &= E[(\hat{\delta}^{2wfe} - \bar{\bar{\Delta}})(\hat{\delta}^{2wfe} - \bar{\bar{\Delta}})'] \\ &= (1'_{post}M_1 1_{post})^{-1}1'_{post}M_1 Var(\bar{\epsilon}_E - \bar{\epsilon}_C)M_1 1_{post}(1'_{post}M_1 1_{post})^{-1} \\ &= \sigma_\epsilon^2\left(\frac{1}{N_E} + \frac{1}{N_C}\right)(1'_{post}M_1 1_{post})^{-1} \\ &= \sigma_\epsilon^2\left(\frac{1}{N_E} + \frac{1}{N_C}\right)\left(\frac{1}{T_0} + \frac{1}{T_1}\right).\end{aligned}$$

(b) The DGP for the control panel in matrix form is given by $y_C = \mu_C f' + \epsilon_C$. Define the control-group averages $\bar{y}_C := y'_C 1/N_C$ (the factor proxy) and $\bar{\mu}_C := \mu'_C 1/N_C$. We rescale the factor loading and factor by setting $\tilde{\mu}_C := \mu_C/\bar{\mu}_C$ and $\tilde{f} = \bar{\mu}_C f$. The DGP now becomes $y_C = \mu_C f' + \epsilon_C = \tilde{\mu}_C\tilde{f}' + \epsilon_C$. Note that $\tilde{\mu}'_C 1/N_C = 1$. It follows that

$$\bar{y}_C = f\bar{\mu}_C + \bar{\epsilon}_C = \tilde{f} + \bar{\epsilon}_C,$$

so that we can solve for the rescaled factor $\tilde{f}$:

$$\tilde{f} = \bar{y}_C - \bar{\epsilon}_C.$$

Using the factor proxy $\bar{y}_C$, the PCDID estimator $\hat{\delta}_i$ is given as follows: for $i \in E$,

$$\begin{aligned}\hat{\delta}_i &= (1'_{post}M_{[1,\bar{y}_C]}1_{post})^{-1}1'_{post}M_{[1,\bar{y}_C]}y_i \\ &= (1'_{post}M_{[1,\bar{y}_C]}1_{post})^{-1}1'_{post}M_{[1,\bar{y}_C]}(\varsigma_i 1 + \tilde{f} + \delta_i 1_{post} + \epsilon_i) \\ &= (1'_{post}M_{[1,\bar{y}_C]}1_{post})^{-1}1'_{post}M_{[1,\bar{y}_C]}(\varsigma_i 1 + \bar{y}_C - \bar{\epsilon}_C + \bar{\bar{\Delta}}1_{post} + \epsilon_i) \\ &= \bar{\bar{\Delta}} + (1'_{post}M_{[1,\bar{y}_C]}1_{post})^{-1}1'_{post}M_{[1,\bar{y}_C]}(\epsilon_i - \bar{\epsilon}_C).\end{aligned}$$

It follows that the simple mean-group estimator has the decomposition:

$$\hat{\delta}^{mg} = \bar{\bar{\Delta}} + (1'_{post}M_{[1,\bar{y}_C]}1_{post})^{-1}1'_{post}M_{[1,\bar{y}_C]}(\bar{\epsilon}_E - \bar{\epsilon}_C).$$

The variance of $\hat{\delta}^{mg}$ is given by

$$\begin{aligned}
Var(\hat{\delta}^{mg}) &= E[(\hat{\delta}^{mg} - \bar{\bar{\Delta}})(\hat{\delta}^{mg} - \bar{\bar{\Delta}})'] \\
&= (1'_{post} M_{[1,\bar{y}_C]} 1_{post})^{-1} 1'_{post} M_{[1,\bar{y}_C]} Var(\bar{\epsilon}_E - \bar{\epsilon}_C) M_{[1,\bar{y}_C]} 1_{post} (1'_{post} M_{[1,\bar{y}_C]} 1_{post})^{-1} \\
&= \sigma_\epsilon^2 \left(\frac{1}{N_E} + \frac{1}{N_C} \right) (1'_{post} M_{[1,\bar{y}_C]} 1_{post})^{-1}.
\end{aligned}$$

Note that $Var(\bar{\epsilon}_C) = Var(\epsilon'_C 1/N_C) = \sigma_\epsilon^2 / N_C$.

(c) It suffices to compute the covariance between $\hat{\delta}^{mg}$ and $\hat{\delta}^{2wfe}$. Since $M_{[1,\bar{y}_C]} M_1 = M_{[1,\bar{y}_C]}$, we obtain

$$\begin{aligned}
Cov(\hat{\delta}^{mg}, \hat{\delta}^{2wfe}) &= E[(\hat{\delta}^{mg} - \bar{\bar{\Delta}})(\hat{\delta}^{2wfe} - \bar{\bar{\Delta}})'] \\
&= (1'_{post} M_{[1,\bar{y}_C]} 1_{post})^{-1} 1'_{post} M_{[1,\bar{y}_C]} Var(\bar{\epsilon}_E - \bar{\epsilon}_C) M_1 1_{post} (1'_{post} M_1 1_{post})^{-1} \\
&= \sigma_\epsilon^2 \left(\frac{1}{N_E} + \frac{1}{N_C} \right) (1'_{post} M_1 1_{post})^{-1} \\
&= Var(\hat{\delta}^{2wfe}).
\end{aligned}$$

As a result,

$$\begin{aligned}
Var(\hat{\delta}^{mg} - \hat{\delta}^{2wfe}) &= Var(\hat{\delta}^{mg}) + Var(\hat{\delta}^{2wfe}) - 2Cov(\hat{\delta}^{mg}, \hat{\delta}^{2wfe}) \\
&= Var(\hat{\delta}^{mg}) - Var(\hat{\delta}^{2wfe}).
\end{aligned}$$

References

[1] Abadie, A., A. Diamond, and J. Hainmueller (2010): Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program, *Journal of American Statistical Association*, 105(490), pp.493–505.

[2] Ahn, S.C., and A.R. Horenstein (2013): Eigenvalue Ratio Test for the Number of Factors, *Econometrica*, 81, pp.1203–1227.

[3] Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018): Matrix Completion Methods for Causal Panel Data Models, Working Paper.

[4] Bai, J. (2003): Inferential Theory for Factor Models of Large Dimensions, *Econometrica*, 71(1), pp.135–171.

[5] Bai, J. (2009): Panel Data Models With Interactive Fixed Effects, *Econometrica*, 77(4), pp.1229–1279.

[6] Bai, J, and S. Ng (2002): Determining the Number of Factors in Approximate Factor Models, *Econometrica*, 70(1), pp.191–221.

[7] Bertrand, M., E. Duflo and S. Mullainathan (2004): How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1) pp.249–275.

[8] Cameron A.C., J.B. Gelbach, and D.L. Miller (2008): Bootstrap-Based Improvements for Inference with Clustered Errors, Review of Economics and Statistics, 90, pp.414–427.

[9] Davidson, J., and R.M. DeJong (2000): The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals II: Fractionally Integrated Processes, *Econometric Theory*, 16, pp.643–666.

[10] Davidson, R., and E. Flachaire (2008): The Wild Bootstrap, Tamed at Last, *Journal of Econometrics*, 146, pp.162–169.

[11] DeJong, R.M., and J. Davidson (2000): The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals I: Weakly Dependent Processes, *Econometric Theory*, 16, pp.621–642.

[12] Gobillon L., and T. Magnac (2016): Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls, *Review of Economics and Statistics*, 98(3), pp.535–551.

[13] Gonçalves, S., and B. Perron (2014): Bootstrapping Factor-Augmented Regression Models, *Journal of Econometrics*, 182, pp.156–173.

[14] Mackinnon J.G., and M.D. Webb (2017): Wild Bootstrap Inference for Wildly Different Cluster Sizes, *Journal of Applied Econometrics*, 32, pp.233–254.

[15] Pesaran, M. H. (2006): Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure, *Econometrica*, 74(4), pp.967–1012.

[16] Phillips, P.C.B., and H.R. Moon (1999): Linear Regression Limit Theorems for Nonstationary Panel Data, *Econometrica*, 67(5), pp.1057–1111.

[17] Politis D.N., and J.P. Romano (1994): The Stationary Bootstrap, *Journal of the American Statistical Association*, 89, pp.1303–1313.

[18] Xu, Y. (2017): Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models, *Political Analysis*, 25, pp.57–76.

[19] Xu, Y. (2017b): Supplementary Materials to Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models.

[20] White, H. (2001): *Asymptotic Theory for Econometricians.* Revised Edition. Academic Press.