Difference in Differences with Latent Group Structures*

Young Ahn
Department of Economics
University of Pennsylvania
youngahn@sas.upenn.edu

Hiroyuki Kasahara
Vancouver School of Economics
University of British Columbia
hkasahar@mail.ubc.ca

September 25, 2023

Abstract

This paper investigates the estimation of average treatment effects on the treated (ATT) from the panel data in the presence of latent group structures, where potential outcome distributions depend on latent types. We examine a scenario where the parallel trends assumption holds when conditioned on latent types, but may not be maintained in aggregate, resulting in an inconsistent standard difference-in-difference estimator. We demonstrate that the latent group-specific ATT (LGATT) can be identified if parallel trend assumptions and other regularity conditions are met for latent types. We present the conditions under which latent group structures are identified from the pre-treatment period data. We propose an estimator for the LGATT that minimizes a weighted least squares criterion function, using weights derived from the estimated posterior probabilities of each latent type using pre-treatment data.

1 Introduction

The Difference-in-Differences (DiD) method is among the most widely used techniques for evaluating the causal effect of policy changes in non-experimental settings within empirical economics (Currie et al., 2020). As reviewed by Roth et al. (2023), numerous recent methodological papers have been written on DiD methods, relaxing some of the assumptions inherent in the canonical DiD model with two time periods (e.g., Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021).

A key requirement for the DiD method is the parallel trends assumption. If this assumption is violated, the estimates derived from the DiD method may be biased, leading to incorrect inferences about the causal effect of the treatment. Although the parallel trends assumption cannot be directly tested—it involves the counterfactual outcome of the treatment group in the absence of treatment—its validity is often assessed indirectly through pre-treatment trends.

*We thank Wenhui Bao and Jiayun Xu for excellent research assistance. All mistakes are our own.

A prevalent issue with current methods for conducting pre-trend analysis is the lack of guidance on how to proceed in the presence of a significant pre-trend. Even when statistical tests on pre-trends suggest that parallel trends may not hold, researchers may still be interested in understanding the treatment effect, especially when the deviation from parallel trends is minimal. However, there is currently no consensus on how to proceed with this type of analysis, and the conventional approach offers little guidance in such circumstances.

This paper proposes a new DiD method with latent group structures, which is applicable in situations where the parallel trends assumption is violated in aggregate. Our proposed method weakens the conventional parallel trends assumption by classifying units into a set of latent groups—such that within a group, the first differenced outcomes for the treated and control units follow the same distribution in the pre-treatment period. This ensures that the pre-trend holds conditional on a latent type.

By considering the multi-period setting with staggered treatment adoption (e.g., Callaway and Sant'Anna, 2021), the causal parameter of our interest in this paper is the **Latent group average treatment effects on the treated (LGATT)** for latent type j and the treatment timing cohort g defined by

$$\mu_{g,t}^{j} = \mathbb{E}\left[Y_{it}(g) - Y_{it}(0) \,\middle|\, \underbrace{G_i = g}_{\text{treated at } g} \;,\; \underbrace{Z_i = j}_{\text{latent type}} \right] \quad \text{for } t = g, g+1, ..., T,$$

where $\{Y_{it}(g) : g \in \mathcal{G}\}$ is a potential outcome across different treatment timings, where $g = 0$ indicates that the unit is "never treated," $D_{it} \in \{0, 1\}$ is the binary treatment variable, $G_i \in \{0, \underline{g}, \underline{g}+1, ..., \bar{g}\}$ is the treatment timing, and $Z_i \in \{1, 2, ..., J\}$ is latent type. Once the LGATT is estimated, we may aggregate them across latent types and treatment timing cohorts using user-specified weights to estimate a target parameter of economic interest.

We derive the conditions under which latent structures and the LGATT are identified from the short panel data. The key identification condition for latent structures is that the observed outcome follows a Markov process, where we analyze a sufficient condition for the Markov assumption in the potential outcome framework. Based on our identification analysis, we propose an estimator that minimizes a weighted least squares criterion function as

$$(\hat{\boldsymbol{\gamma}}^{j}(g), \Delta\hat{\boldsymbol{\delta}}^{j}) = \arg \min_{\Delta\boldsymbol{\delta}^{j}, \boldsymbol{\gamma}^{j}(g)} \sum_{i \in \mathcal{I}_g} \sum_{t=g}^{T} (\Delta Y_{it} - \Delta\delta_t^{j} - \gamma_{g,t}^{j} D_{it})^2 \hat{\tau}_i^{j}, \tag{1}$$

where $\mathcal{I}_g := \{i : G_i \in \{0, g\}\}$ is a set of indices for the units with $G_i = 0$ or g in the sample, $\hat{\tau}_i^{j}$ is an estimated posterior type probability of being the j-th latent type obtained from the pre-treatment

period's data. Then, the LGATT for latent type j and timing cohort g is estimated as

$$\hat{\mu}^j_{g,t} = \sum_{s=g}^{t} \hat{\gamma}^j_{g,s} \quad \text{for } t \geq g.$$

The estimator $\hat{\boldsymbol{\gamma}}^j(g)$ in equation (1) is a version of two-way fixed effects estimator based on the first-differenced transformation to eliminate individual unit's fixed effect while using the latent type-specific posterior-probabilities as weights. We establish consistency and asymptotic normality of our estimator when the data in the pre-treatment periods is generated follows a Gaussian finite mixture model.

A standard two-way fixed effects estimator for the DiD is formulated as

$$(\hat{\boldsymbol{\mu}}^{\text{did}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}_t) = \underset{\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\delta}}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \delta_t - \mu_t D_{it})^2.$$

When the number of latent type is one (i.e., $J = 1$), our estimator in (1) implements a version of two-way fixed effects estimator for the DiD based on the first differenced transformation by solving $\underset{\gamma,\boldsymbol{\delta}}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} (\Delta Y_{it} - \Delta\delta_t - \gamma_t D_{it})^2$.

Our proposed estimator is also closely related to but distinct from the Synthetic DiD by Arkhangelsky et al. (2021):

$$(\hat{\boldsymbol{\mu}}^{\text{sdid}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}_t) = \underset{\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\delta}}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \delta_t - \mu_t D_{it})^2 \hat{\omega}^{\text{sdid}}_i \hat{\lambda}^{\text{sdid}}_t.$$

Here, $\hat{\omega}^{\text{sdid}}_i$ is chosen such that the average outcome of control units aligns with those of the treated units in the pre-treatment periods, i.e., $\sum_{i=1}^{N_{co}} \omega_{it} Y_{it} \approx \sum_{i=N_{co}+1}^{N} Y_{it}$ for the pre-treatment period, where N_{co} is the number of control units. This is achieved by assigning higher weights to control units that exhibit similar pre-trends to treated units on average.[1] Instead of constructing synthetic control units that average similar pre-trends to the treated units for the entire sample, our method categorizes treated and control units into a finite number of latent classes. Within a latent class, both treated and control units share similar pre-trends.

Our method offers advantages over the Synthetic DiD. Firstly, the consistency and asymptotic normality of the Synthetic DiD require long panel data, where the length of pre-treatment periods needs to grow to infinity for the asymptotic approximation to be valid. This is a significant disadvantage of the Synthetic DiD in practice, as the length of pre-treatment periods is often short in many empirical applications. In contrast, we demonstrate that, under regularity conditions, the LGATT is identified from short panel data—as short as six pre-treatment periods—and our inference procedure is valid even when the length of panel data is short. Secondly, our method allows

[1] In the Synthetic DiD, $\hat{\lambda}^{\text{sdid}}_t$ balances pre-treatment time periods with post-treatment periods. See Section 2 of Arkhangelsky et al. (2021).

us to estimate the average treatment effects on the treated separately for each latent group, aiding our understanding of the extent to which the treatment effects are heterogeneous across different units. However, our method also has disadvantages over the Synthetic DiD method in that we require the first-differenced observed outcome to follow a Markov process, which in turn implies that potential outcomes need to have limited dependence over time. Furthermore, for practical implementation, we estimate the posterior probabilities of being each of latent types based on parametric models.

While our DiD estimator relaxes the parallel trends assumption by assuming that it holds conditional on an *unobserved* latent group, some papers propose DiD estimators that are valid when the parallel trends assumption holds conditional on *observed* covariates (Heckman et al., 1997, 1998; Abadie, 2005; Sant'Anna and Zhao, 2020). These methods are extended to staggered DiD setups by Callaway and Sant'Anna (2021) and Wooldridge (2021).

Several recent studies have explored alternative approaches for situations where the parallel trends assumption might be violated. Rambachan and Roth (2023) discuss the extent to which outcomes are affected by the violation of parallel trends and suggest a sensitivity analysis todetermine the magnitude of post-treatment parallel trends violation that would nullify a specific conclusion. Roth and Sant'Anna (2023) investigate the conditions under which parallel trends can be maintained for all monotonic transformations of the outcome. Roth (2022) has also developed tools for conducting power analyses and estimating potential distortions from pre-testing under hypothesized violation of parallel trends.

Bonhomme and Manresa (2015) introduced a flexible yet parsimonious approach for modeling unobserved heterogeneity in linear panel data models, known as Grouped Fixed-Effects (GFE). In their conclusion, they suggested potential applications of GFE in difference-in-difference designs as a means to relax parallel trend assumptions in the conclusion section. However, to the best of our knowledge, no existing studies have formally applied the GFE approach in a difference-in-difference context. Expanding on this idea, our study addresses the identification issue that arises when the time dimension of panel data is limited and proposes an estimator based on soft classifications, with posterior probabilities of each latent type constructed from pre-treatment data. It is important to note that the GFE method necessitates conditions where both the cross-sectional and time dimensions of the panel tend towards infinity. In contrast, our study considers a setup where the time-dimension is fixed.

To illustrate our method in an empirical context, we revisit Biasi and Moser (2021) that studies the impact of copyrights on science by exploiting an exogenous change toward weaker copyrights during the World War II. We also reevaluate the impact of the Chinese Wikipedia blockade in mainland China on the contributions of Chinese-speaking users in regions including Taiwan, Hong Kong, Singapore, and beyond, as examined by Zhang and Zhu (2011).

We have developed an R package for LGATT estimation, which is available at the following GitHub repository: https://github.com/bayesiahn/groupdid/. For replication purposes, the codes for both the numerical simulations and the empirical example presented in this paper are provided

as R markdown files.[2]

The remainder of this paper is structured as follows: Section 2 introduces a potential outcome model with latent group structures. Section 3 provides identification analysis. In Section 4, we develop an estimator. Section 5 presents simulation results, and Section 6 follows with empirical applications.

Example

Suppose that an econometrician would like to estimate the impact of a public policy. For illustration, let us consider a three-period model such that outcomes of individual i, Y_{it}, represent pre-treatment observations in the first two periods $t = 1, 2$ and post-treatment observation for $t = 3$. With the potential outcome notation, let $Y_{it}(1)$ denote the potential outcome of unit i in t period when the unit is exposed to the treatment in period 3. Likewise, let $Y_{it}(0)$ denote the potential outcome when unit i remains untreated for all periods.

Consider a public policy that is introduced just after the $t = 2$ period to unit i, leading to treatment effects $Y_{i3}(1) - Y_{i3}(0)$. Suppose that the outcome Y_{it} in period t is determined by the following model:

$$Y_{it}(0) - Y_{it-1}(0) = D_{it} \underbrace{[Y_{it}(1) - Y_{it}(0)]}_{\text{treatment effects}} + \eta_{it},$$

where D_{it} is one if ith unit is treated in period t and zero otherwise, and η_{it} is an idiosyncratic shock. We assume that η_{it} is independent of D_{it} and Y_{it-1} for all t and i. For the rest of the example, let D_i denote the treatment status of the ith unit, i.e., $D_i = D_{i3}$.

Throughout the example, we assume that there are two types of individuals indexed by $Z_i \in \{1, 2\}$ such that $\mathbb{E}[\eta_{it} \mid Z_i = j] = \Delta\delta_t^j$ all $j = 1, 2$. The type Z_i here captures differences in the trends in outcomes, with $\Delta\delta_t^1 < \Delta\delta_t^2$ for all t. We also allow for heterogenous treatment effects across groups, with $\mu^j = \mathbb{E}[Y_{i3}(1) - Y_{i3}(0) \mid Z_i = j, D_i = 1]$. Note that the aggregate treatment effects on the treated (ATT) μ can be written as:

$$\begin{aligned}
\mu &= \mathbb{E}[Y_{i3}(1) - Y_{i3}(0) \mid D_i = 1] \\
&= \sum_{j=1}^{2} \mathbb{E}[Y_{i3}(1) - Y_{i3}(0) \mid D_i = 1, Z_i = j] \Pr(Z_i = j \mid D_i = 1) \\
&= \sum_{j=1}^{2} \mu^j \Pr(Z_i = j \mid D_i = 1),
\end{aligned}$$

which is the weighted average of the treatment effects across groups, where the weights are the probabilities of belonging to each group as treated units.

[2]These can be found in the 'examples/simulation' and 'examples/wikipedia' directories, respectively.

Selection into Treatment. We revisit classic critique of difference-in-differences estimators due to selection bias (LaLonde, 1986). Suppose that the government introduces a job-training program for low-income households to enhance their job skills to improve earnings. Let Y_{it} denote the income level of individual i in period t. The program is targeted to households with income in period 2, Y_{i2}, below a certain threshold, $B > 0$. The DiD estimate of the ATT then identifies:

$$\begin{aligned}
\mu_{DiD} &= \mathbb{E}[Y_{i3} - Y_{i2} \mid D_i = 1] - \mathbb{E}[Y_{i3} - Y_{i2} \mid D_i = 0] \\
&= \mathbb{E}[Y_{it}(1) - Y_{it}(0) + \eta_{i3} \mid D_i = 1] - \mathbb{E}[\eta_{i3} \mid D_i = 0] \\
&= (\mu^1 + \Delta\delta_3^1) \Pr(Z_i = 1 \mid D_i = 1) + (\mu^2 + \Delta\delta_3^2) \Pr(Z_i = 2 \mid D_i = 1) \\
&\quad - \Delta\delta_3^1 \Pr(Z_i = 1 \mid D_i = 0) - \Delta\delta_3^2 \Pr(Z_i = 2 \mid D_i = 0) \\
&= \underbrace{\mu}_{\text{ATT}} + \underbrace{(\Delta\delta_3^2 - \Delta\delta_3^1) \left[\Pr(Z_i = 1 \mid D_i = 0) - \Pr(Z_i = 1 \mid D_i = 1)\right]}_{\text{selection bias}}
\end{aligned}$$

The later term can add negative bias to the DiD estimate if selection into treatment is present. Since the program is targeted to households with lower income, unit i is more likely to belong to the first type $Z_i = 1$ if she was given treatment, i.e., $\Pr(Z_i = 1 \mid D_i = 1) > \Pr(Z_i = 2 \mid D_i = 1)$, if income levels are positively correlated with their changes. Likewise, we have $\Pr(Z_i = 1 \mid D_i = 0) < \Pr(Z_i = 2 \mid D_i = 0)$, which implies that $\Pr(Z_i = 1 \mid D_i = 0) < \Pr(Z_i = 1 \mid D_i = 1)$, i.e., the latter term is negative. Hence, the DiD estimate of the ATT is biased downward in this case.

In fact, in this example, the parallel trends assumption does not hold. In our context, the parallel trends assumption can be stated as $\mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid D_i = 0\right]$, i.e., the mean trends in untreated potentials are identical between treated and control groups. The difference in the mean trends can be written as

$$\begin{aligned}
&\mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid D_i = 0\right] \\
&= \mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid \Delta Y_{i2} \leq B\right] - \mathbb{E}\left[Y_{i3}(0) - Y_{i2}(0) \mid \Delta Y_{i2} > B\right] \\
&= \sum_{j=1}^{2} \Delta\delta_2^j \left[\Pr(Z_i = j \mid D_i = 1) - \Pr(Z_i = j \mid D_i = 0)\right] \\
&= (\Delta\delta_2^2 - \Delta\delta_2^1) \left[\Pr(Z_i = 1 \mid D_i = 0) - \Pr(Z_i = 1 \mid D_i = 1)\right],
\end{aligned}$$

which is strictly negative if $\Pr(Z_i = 1 \mid D_i = 0) < \Pr(Z_i = 1 \mid D_i = 1)$ as above.

2 The Model

The setup expands on that of Callaway and Sant'Anna (2021). Consider a model with T periods, where $t \in \mathcal{T} := \{1, 2, ..., T\}$ represents a specific time period. Unit i belongs to one of J latent groups, where the number of latent types J is assumed to be known by the econometrician. Let $Z_i \in \mathcal{J} := \{1, ..., J\}$ indicate the latent group to which unit i belongs, referred to as the *latent*

type. The population probability of being the j-th latent type is denoted by $\pi^j := \Pr(Z_i = j)$ for $j = 1, 2, ..., J$.

Each unit may receive a binary treatment at different periods or may never be treated within the T periods. Let D_{it} be a binary indicator for the i-th unit's treatment status at t and let G_i be the period in which unit i receives her treatment for the first time. We assume that treatment is an absorbing state, i.e., once a unit is treated, then it remains treated for the rest of periods:

$$D_{it} = \begin{cases} 0 & \text{if } t < G_i \text{ or } G_i = 0, \\ 1 & \text{if } t \geq G_i. \end{cases}$$

If unit i is never treated within the sample period, we denote the never treated unit with $G_i = 0$. The econometrician observes an outcome Y_{it} and treatment timing G_i but does not observe unit i's latent type Z_i. We call a set of units that are first treated in time g as *cohort* g. G_i takes a value on $\mathcal{G} = \{\underline{g}, ..., \bar{g}, 0\}$, where $\underline{g}$ and $\bar{g}$ denote the earliest and latest first-time treatment periods across all units within the sample period, where $1 \leq \underline{g} \leq \bar{g} \leq T$. When $\underline{g} = \bar{g}$, all treated units receive their first-time treatment in the same period.

We adopt the potential outcomes framework of Robins (1986), where the potential outcome may depend on the entire sequence of treatment assignments over the T periods. Denote unit i's potential outcome in period t if she is treated for the first time at time g by $Y_{it}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$, where $\mathbf{0}_{g-1}$ and $\mathbf{1}_{T-g+1}$ denote a vector of zeros and ones with length $g-1$ and $T-g+1$, respectively. Because treatment is an absorbing state, the sequence of treatment assignments is solely determined by the first treatment period. Therefore, we simplify notation by denoting $Y_{it}(g) := Y_{it}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$ and $Y_{it}(0) := Y_{it}(\mathbf{0}_T)$. With this notation, the observed outcome Y_{it} is related to potential outcomes as:

$$Y_{it} = \sum_{g \in \mathcal{G}} \mathbb{I}\{G_i = g\} Y_{it}(g). \tag{2}$$

The joint distribution of potential outcomes, $\{(Y_{it}(\underline{g}), Y_{it}(\underline{g}+1), ..., Y_{it}(\bar{g}), Y_{it}(0)) : t = 1, ..., T\}$, is assumed to be the same within the identical latent group but may be different between latent groups. The causal estimand of our primary interest is the **latent group average treatment effects on treated (LGATT)** for latent group j and cohort g:

$$\mu^j_{g,t} = \mathbb{E}\left[Y_{it}(g) - Y_{it}(0) \mid G_i = g, Z_i = j\right] \quad \text{for } t = g, g+1, ..., T, \tag{3}$$

where $g \in \mathcal{G} \backslash \{0\}$, and $j \in \mathcal{J}$. We are also interested in estimating the average treatment effects on treated (ATT) for cohort g:

$$\mathbb{E}\left[Y_{it}(1) - Y_{it}(0) \mid G_i = g\right] = \sum_{j=1}^{J} \Pr(Z_i = j | G_i = g) \mu^j_{g,t}.$$

There are two difficulties in identifying the treatment parameter $\mu^j_{g,t}$ from the observed data.

First, we are not able to observed the untreated potential outcome $Y_{it}(0)$ for the unit treated after time g. Second, we are not able to observe the latent type Z_i.

If we know the latent type, then we may identify the treatment parameter $\mu^j_{g,t}$ by imposing a set of assumptions similar to those adopted in the recent difference-in-differences literature (e.g., Callaway and Sant'Anna, 2021; Roth et al., 2023). In particular, we assume that the parallel trends holds within each of latent groups, i.e., untreated potential outcomes are parallel to the treated potential outcomes within a latent group.

Assumption 1. *(Latent-type specific parallel trends) For all $t, t' \in \mathcal{T}$, all $j \in \mathcal{J}$, and all $g \in \mathcal{G}$,*

$$\mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid G_i = g, Z_i = j\right] = \mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid G_i = 0, Z_i = j\right]. \tag{4}$$

This assumption corresponds to Assumption 4 of Callaway and Sant'Anna (2021) but we require that the parallel trends hold when conditioned on latent types. Crucially, the parallel trends are presumed not only for post-treatment but also for pre-treatment periods. This is because, in our identification analysis, latent structures are determined to ensure that parallel trends during pre-treatment periods hold within each latent group.

Even when the parallel trends assumption holds within each latent group, the parallel trends assumption does not generally hold in aggregate as shown in the following example.

Example 1. *Suppose that there are two latent types with $\mathcal{J} = \{1, 2\}$ and the timing of receiving treatment is the same across all units with $\mathcal{G} = \{0, g\}$ for some $g \in \mathcal{T}$. If Assumption 1 holds, $\Pr(Z_i = 1|G_i = g) \neq \Pr(Z_i = 1|G_i = 0)$, and $\mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid G_i = 0, Z_i = 1\right] \neq \mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid G_i = 0, Z_i = 2\right]$, then*

$$\mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid G_i = g\right] \neq \mathbb{E}\left[Y_{it}(0) - Y_{it'}(0) \mid G_i = 0\right].$$

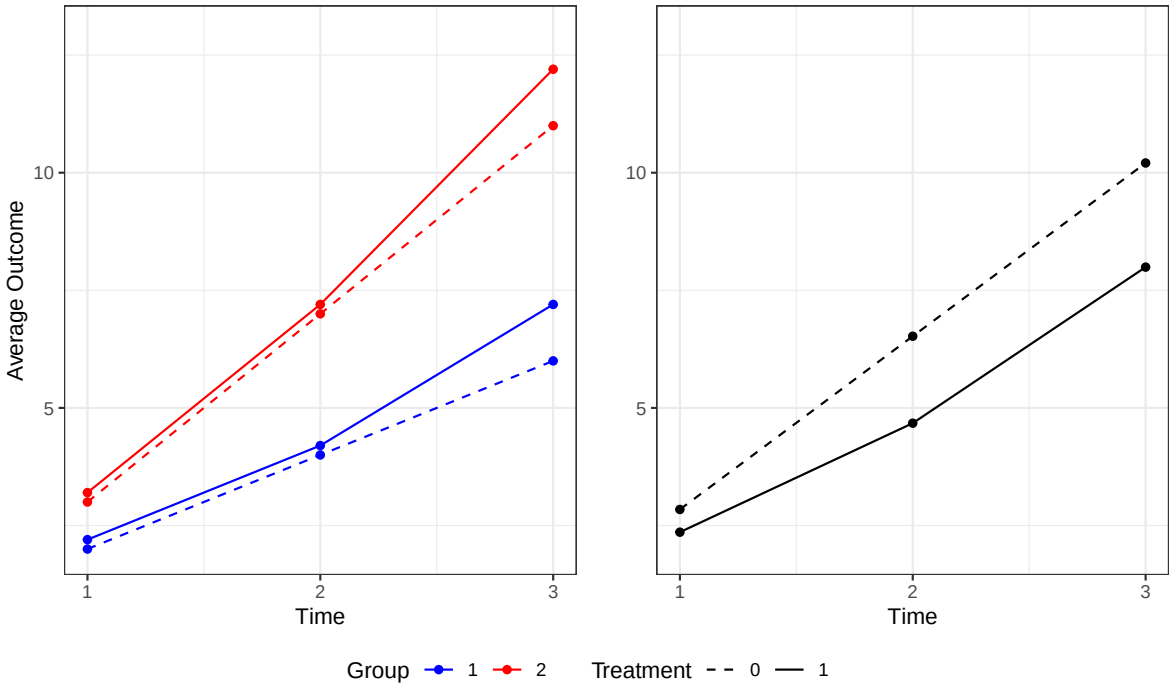
Figure 2 provides a graphical representation of Example 1 using a potential outcome model (12)-(13) with two latent types and three periods. In the left figure, the parallel trends assumption holds within each latent group, as evidenced by the blue and red lines signifying the first and second latent types, respectively, and the solid and dotted lines denoting treated and untreated units, respectively. Conversely, the right figure demonstrates the violation of the parallel trends assumption at the aggregate level when latent types are not taken into account.

This example demonstrates the usefulness of our proposed DiD method within latent groups. By identifying the latent structure that satisfies Assumption 1, we can estimate the causal effect by applying the DiD method within each latent group, even when the parallel trends assumption is not met at the aggregate level.

Additionally, we assume that the treatment status has no influence on observations prior to its implementation, i.e., there is no anticipation of treatment in treated observations.

Assumption 2. *(No anticipatory effects) For all $g \in \mathcal{G}$ and $t \in \{1, ..., g-1\}$, $Y_{it}(g) = Y_{it}(0)$.*

Figure 1: Parallel trends assumption holds within each latent group but not in aggregate.



Notes: The figure is generated using a potential outcome model (12)-(13) with two latent types and three periods, where the treatment occurs at $t = 3$. The parameters employed are: $\pi^1 = \pi^2 = 0.5$, $\delta_t^j = 2 \times j \times t$ for $j = 1, 2$, $\mathbb{E}[\epsilon_{it}^2] = 1$, and $\mu_t^j = t - 1$ if $t = 3$. Treatment is assigned for individuals with $\Delta Y_{i2} < B$ where $B = 3$. In the left figure, the parallel trends assumption is satisfied within each latent group, with blue and red lines representing the first and second latent types, respectively, and solid and dotted lines indicating treated and untreated units, respectively. The right figure illustrates the violation of the parallel trends assumption at the aggregate level when latent types are not conditioned on.

To guarantee the identification of latent type-specific treatment parameters, we assume that the conditional probability of receiving the first treatment at g or never receiving the treatment, given the latent type, is bounded away from both zero and one.

Assumption 3. *(Overlap) There exists a positive constant $\epsilon > 0$ such that $\epsilon < \Pr(G_i = g|Z_i = j) < 1 - \epsilon$ for all $g = \mathcal{G}$ and $j \in \mathcal{J}$.*

Under Assumptions 1-3, we may show that, for $g = \underline{g}, ..., \bar{g}$ and $t \geq g$, the LGATT is written as:[3]

$$\begin{aligned}\mu_{g,t}^{j} &= \mathbb{E}\left[Y_{it}(g) - Y_{it}(0) \mid G_i = g, Z_i = j\right] \\ &= \underbrace{\mathbb{E}\left[Y_{it} - Y_{i,g-1} \mid G_i = g, Z_i = j\right]}_{\text{the } j\text{-th group change for } G_i=g} - \underbrace{\mathbb{E}\left[Y_{it} - Y_{i,g-1} \mid G_i = 0, Z_i = j\right]}_{\text{the } j\text{-th group change for } G_i=0}. \end{aligned} \tag{5}$$

Consequently, the LGATT can be identified using the average difference-in-differences in the observed outcomes between the treatment group ($G_i = g$) and the control group ($G_i = 0$), conditional on the latent variable Z_i. Since Z_i is not directly observable, it is necessary to identify the underlying latent group structure to estimate the LGATT.

We write $\mu_{g,t}^{j}$ in terms of telescoping sums as

$$\mu_{g,t}^{j} = \sum_{s=g}^{t} \gamma_{g,s}^{j},$$

where

$$\gamma_{g,s}^{j} := \mathbb{E}\left[\Delta Y_{is} \mid G_i = g, Z_i = j\right] - \mathbb{E}\left[\Delta Y_{is} \mid G_i = 0, Z_i = j\right] \tag{6}$$

with $\Delta Y_{is} := Y_{is} - Y_{i,s-1}$.

The identification of $\gamma_{g,s}^{j}$ depends on the first-differenced outcomes ΔY_{it} in (6). Thus, we focus on these first-differenced outcomes, rather than the outcomes themselves, as the variables in our dataset. Utilizing the parallel trends within each latent group, we identify latent structures where different latent groups follow distinct trends.

In our analysis, we consider the scenario where $\Delta \boldsymbol{Y}_i := (\Delta Y_{i2}, ..., \Delta Y_{iT})^\top$ is a random vector with its support $\Delta\boldsymbol{\mathcal{Y}} = (\Delta\mathcal{Y})^{T-1}$. Here, $\boldsymbol{W}_i := (\Delta \boldsymbol{Y}_i^\top, G_i)^\top \in \Delta\boldsymbol{\mathcal{Y}} \times \mathcal{G}$ denotes the vector of

[3]Note that

$$\begin{aligned}\mathbb{E}\left[Y_{it}(0) \mid G_i = g, Z_i = j\right] &= \mathbb{E}\left[Y_{i,g-1}(0) \mid G_i = g, Z_i = j\right] + \mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \mid G_i = g, Z_i = j\right] \\ &= \mathbb{E}\left[Y_{i,g-1}(0) \mid G_i = g, Z_i = j\right] + \mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \mid G_i = 0, Z_i = j\right] \\ &= \mathbb{E}\left[Y_{i,g-1}(g) \mid G_i = g, Z_i = j\right] + \mathbb{E}\left[Y_{it}(0) - Y_{i,g-1}(0) \mid G_i = 0, Z_i = j\right] \\ &= \mathbb{E}\left[Y_{i,g-1} \mid G_i = g, Z_i = j\right] + \mathbb{E}\left[Y_{it} - Y_{i,g-1} \mid G_i = 0, Z_i = j\right],\end{aligned}$$

where the second and third equalities follow from the parallel trends and no anticipatory effects assumptions, respectively. Hence, we have $\mu_{g,t}^{j} = \mathbb{E}\left[Y_{it}(g) \mid G_i = g, Z_i = j\right] - \mathbb{E}\left[Y_{it}(0) \mid G_i = g, Z_i = j\right] = \mathbb{E}\left[Y_{it} - Y_{i,g-1} \mid G_i = g, Z_i = j\right] - \mathbb{E}\left[Y_{it} - Y_{i,g-1} \mid G_i = 0, Z_i = j\right]$.

first-differenced outcomes and treatment timing for unit i. Our model assumes that the data are randomly sampled from a finite mixture model.

Assumption 4. *(Random Sampling from a Finite Mixture Distribution) (a) We observe a sample of n i.i.d. draws $\{\boldsymbol{W}_i\}_{i=1}^n$, where $\boldsymbol{W}_i \overset{i.i.d.}{\sim} F_{\boldsymbol{W}}(\boldsymbol{w})$, (b) The cumulative distribution function $F_{\boldsymbol{W}}(\boldsymbol{w})$ follows a finite mixture representation $F_{\boldsymbol{W}}(\boldsymbol{w}) = \sum_{j=1}^J \pi^j F^j_{\boldsymbol{W}}(\boldsymbol{w})$, where $\pi^j := \Pr(Z_i = j)$ and*

$$F^j_{\boldsymbol{W}}(\boldsymbol{w}) := \Pr(\boldsymbol{W} \leq \boldsymbol{w} | Z_i = j), \tag{7}$$

(c) $F^j_{\boldsymbol{W}}(\boldsymbol{w})$ satisfies Assumptions 1-3 with the relationship between observed outcome and potential outcomes given in (2). (d) The true number of components defined as the smallest integer J such that the data distribution function admits the representation (7) is known.

In Assumption 4(d), we assume that the true number of components is known. In our analysis, J denotes the true number of components.

3 Identification

For $\boldsymbol{w} = (\Delta \boldsymbol{y}^\top, g)^\top \in \Delta\mathcal{Y} \times \mathcal{G}$, denote the conditional density function of $\Delta \boldsymbol{Y}$ given $G_i = g$ and $Z_i = j$ by $f^j_{\Delta \boldsymbol{Y}|G}(\Delta \boldsymbol{y}|g)$ and let $p^j_g := \Pr(G_i = g | Z_i = j)$. Let

$$f_{\boldsymbol{W}}(\boldsymbol{w}) := \sum_{j=1}^J \pi^j f^j_{\boldsymbol{W}}(\boldsymbol{w}) \quad \text{with} \quad f^j_{\boldsymbol{W}}(\boldsymbol{w}) := f^j_{\Delta \boldsymbol{Y}|G}(\Delta \boldsymbol{y}|g) p^j_g \tag{8}$$

so that $F_{\boldsymbol{W}}(\boldsymbol{w}) = \sum_{j=1}^J \pi^j \sum_{g' \leq g} \int_{\Delta \boldsymbol{y}' \leq \Delta \boldsymbol{y}} f_{\Delta \boldsymbol{Y}|G}(\Delta \boldsymbol{y}'|g') p^j(g') d\Delta \boldsymbol{y}'$.

Let $\boldsymbol{W}^s_{2,i} := (\Delta Y_{i2}, ..., \Delta Y_{is}, G_i)^\top \in (\Delta\mathcal{Y})^{s-1} \times \mathcal{G}$ for $s \leq T$, and define

$$f^j_{\boldsymbol{W}^s_2}(\boldsymbol{w}^s_2) := \int ... \int f^j_{\boldsymbol{W}}(\boldsymbol{w}^s_2, w_{s+1}, ..., w_T) dw_{s+1} ... dw_T$$

for $j = 1, ..., J$.

Given the realized value $\boldsymbol{W}^s_{2,i} = \boldsymbol{w}^s_{2,i}$, we may express the posterior probability of being type j as

$$\tau^j(\boldsymbol{w}^s_2) := \Pr(Z = j | \boldsymbol{W}^s_2 = \boldsymbol{w}^s_2) = \frac{\pi^j f^j_{\boldsymbol{W}^s_2}(\boldsymbol{w}^s_2)}{\sum_{k=1}^J \pi^k f^k_{\boldsymbol{W}^s_2}(\boldsymbol{w}^s_2)}, \tag{9}$$

We characterize the LGATT in terms of the posterior probability of being type j given the pre-treatment period's observations. The following proposition suggests that if we are able to identify the posterior type probability, $\tau^j(\boldsymbol{w}^s_2)$, and the conditional probability of being type j given the treatment timing G_i, we can then identify $\gamma^j_{g,t}$, and consequently, $\mu^j_{g,t}$, from the observed data.

Assumption 5. *ΔY_{it} is independent of $\{\Delta Y_{it-s} : \ s \geq 2\}$.*

Proposition 1. *Under Assumptions 1-5, for $g = \underline{g}, ..., \bar{g}$, and for $t = g, ..., T$,*

$$\gamma_{g,t}^{j} = \frac{\mathbb{E}\left[\tau^{j}(\boldsymbol{W}_{2}^{g-2})\Delta Y_{it} \mid G_i = g\right]}{\Pr(Z_i = j \mid G_i = g)} - \frac{\mathbb{E}\left[\tau^{j}(\boldsymbol{W}_{2}^{g-2})\Delta Y_{it} \mid G_i = 0\right]}{\Pr(Z_i = j \mid G_i = 0)}. \tag{10}$$

The characterization of the LGATT in (10) relies on the identification of latent types from the pre-treatment periods data.

Given the characterization in Propositions 1, we may identify $\gamma_{g,t}^{j}$ if we identify $\tau^{j}(\cdot)$ and $\Pr(Z_i = j \mid G_i = g)$'s.

In order to identify $\tau^{j}(\cdot)$ and $\Pr(Z_i = j \mid G_i = g)$, we impose the following Markov assumption. Let $f^{j}_{\Delta Y_{it}|\Delta Y_{i,t-1},...,\Delta Y_{i,t-s},G}(\Delta y_{it}|\Delta y_{i,t-1}, ..., \Delta y_{i,t-s}, g)$ denote the conditional probability density function of ΔY_{it} given $Z_i = j$ and $(\Delta Y_{i,t-1}, ..., \Delta Y_{i,t-s}, G_i) = (\Delta y_{i,t-1}, ..., \Delta y_{i,t-s}, g)$ for $0 \leq s \leq t-1$.

Assumption 6. *(Markov) For all $j \in \mathcal{J}$, conditional on $Z_i = j$, $\{\Delta Y_{it} : t = 2, ..., T\}$ follows a (non-stationary) first-order Markov process conditional on $G_i = g$, i.e., for $t = 2, ..., T$ and all $g \in \mathcal{G}$,*

$$f^{j}_{\Delta Y_{it}|\Delta Y_{i,t-1},...,\Delta Y_{i,t-s},G}(\Delta y_{it}|\Delta y_{i,t-1}, ..., \Delta y_{i2}, g) = f^{j}_{\Delta Y_{it}|\Delta Y_{i,t-1},G}(\Delta y_{it}|\Delta y_{i,t-1}, g).$$

Under Assumption 6, for $\boldsymbol{w} = (\Delta y_{i2}, ..., \Delta y_{iT}, g)$, the mixture model (8) is written as

$$f_{\boldsymbol{W}}(\boldsymbol{w}) := \sum_{j=1}^{J} \pi^{j} p_{g}^{j} f^{j}_{\Delta Y_{i2}|G}(\Delta y_{i2}|g) \prod_{t=1}^{T} f^{j}_{\Delta Y_{it}|\Delta Y_{i,t-1},G}(\Delta y_{it}|\Delta y_{i,t-1}, g), \tag{11}$$

where $p_{g}^{j} := \Pr(G = g|Z = j)$.

By extending the argument in Kasahara and Shimotsu (2009), Carroll et al. (2010), and Hu and Shum (2012), we may establish the nonparametric identification of the mixture model (8) as the following proposition states.

Proposition 2. *Suppose that Assumptions 4, 6, and 11 holds. Then, we may uniquely identify $\{\pi^{j}, f^{j}_{\boldsymbol{W}}(\boldsymbol{w}) : j \in \mathcal{J}\}$ from $f(\boldsymbol{w})$.*

Because $\tau^{j}(\cdot)$ and $\Pr(Z_i = j|G = g)$ can be identified from $\{\pi^{j}, f^{j}_{\boldsymbol{W}}(\boldsymbol{w}) : j \in \mathcal{J}\}$, in view of (5), (10) and Proposition 3, we may identify $\gamma_{g,t}^{j}$ and $\mu_{g,t}^{j}$ from $f_{\boldsymbol{W}}(\boldsymbol{w})$.

Proposition 3. *Under Assumptions 1-6 and 11, we may uniquely identify $\gamma_{g,t}^{j}$ and $\mu_{g,t}^{j}$ from $f_{\boldsymbol{W}}(\boldsymbol{w})$ for all $g \in \mathcal{G}$, $t \geq g$, and $j \in \mathcal{J}$.*

4 Continuous Outcome

The conjunction of the Markov assumption, as detailed in Assumption 6, along with the parallel trends and no anticipation assumptions, as outlined in Assumptions 1-2, forms the central identify-

ing assumption of this study. Under these assumptions, we now develop a regression specification based on the potential outcomes framework and propose an estimator for the LGATT.

4.1 Regression specification and potential outcome framework

For the j-th latent type with $Z_i = j$, decompose the potential outcome of untreated as

$$Y_{it}(0) = \delta_t^j + \alpha_i + \epsilon_{it}(0), \tag{12}$$

where δ_t^j is period-specific intercept for latent type j defined by $\delta_t^j := E[Y_{it}(0)|Z_i = j]$, $\alpha_i := \frac{1}{T}\sum_{t=1}^{T}(Y_{it}(0) - \delta_t^j)$ for $Z_i = j$ is a random variable that is mean zero across units but time-invariant within each unit, and $\epsilon_{it}(0) := Y_{it}(0) - \delta_t^j - \alpha_i$ is a time-varying random variable. Note that $\mathbb{E}[\alpha_i|Z_i = j] = E[\epsilon_{it}(0)|Z_i^* = j] = 0$ by construction.

Given δ_t^j and α_i in (12), we express the treated potential outcomes as

$$Y_{it}(g) = \delta_t^j + \mu_{g,t}^j D_{it} + \alpha_i + \epsilon_{it}(g), \quad \text{for } t \geq g \text{ and } g = \underline{g}, ..., \bar{g}, \tag{13}$$

where $\mu_{g,t}^j$ is defined by (5) while $\epsilon_{it}(g)$ is *defined* as

$$\epsilon_{it}(g) := Y_{it}(g) - \delta_t^j - \mu_{g,t}^j D_{it} - \alpha_i.$$

In the following, for notational brevity, we focus on the subsample of the data set for $G_i = g$ and $G_i = 0$. When conditioned on $G_i = g$ or $G_i = 0$, we have $Y_{it} = \mathbb{I}\{G_i = g\}Y_{it}(g) + \mathbb{I}\{G_i = 0\}Y_{it}(0)$. Consequently, under Assumption 2, ΔY_{it} is written as

$$\begin{aligned}\Delta Y_{it} &= \mathbb{I}\{G_i = g\}\Delta Y_{it}(g) + \mathbb{I}\{G_i = 0\}\Delta Y_{it}(0) \\ &= \begin{cases} \Delta Y_{it}(0) & \text{if } t \leq g-1, \\ D_{ig}(Y_{ig}(g) - Y_{i,g-1}(0)) + (1 - D_{ig})\Delta Y_{ig}(0) & \text{if } t = g, \\ D_{ig}\Delta Y_{it}(g) + (1 - D_{ig})\Delta Y_{it}(0) & \text{if } t > g. \end{cases}\end{aligned}$$

For $t \leq g - 1$, using the notations in (12)-(13), ΔY_{it} is expressed as

$$\Delta Y_{it} = \Delta\delta_t^j + \underbrace{\Delta\epsilon_{it}(0)}_{=\eta_{it}}. \tag{14}$$

For $t = g$,

$$\Delta Y_{ig} = \Delta\delta_g^j + \underbrace{\mu_g^j(g)}_{=\gamma_g^j(g)} D_{ig} + \underbrace{D_{ig}(\epsilon_{ig}(g) - \epsilon_{i,g-1}(0)) + (1 - D_{ig})\Delta\epsilon_{ig}(0)}_{:=\eta_{ig}}. \tag{15}$$

For $t > g$, noting that $D_{it} = D_{i,t-1}$ when $G_i = 0$ or g,

$$\Delta Y_{it} = \Delta\delta_t^j + \underbrace{\Delta\mu_{g,t}^j}_{=\gamma_{g,t}^j} D_{it} + \underbrace{D_{it}\Delta\epsilon_{it}(g) + (1 - D_{it})\Delta\epsilon_{it}(0)}_{:=\eta_{it}}. \tag{16}$$

Therefore, from (14)-(16), we have

$$\Delta Y_{it} = \Delta\delta_t^j + \gamma_{g,t}^j D_{it} + \eta_{it} \quad \text{for } Z_i = j. \tag{17}$$

The following proposition states that η_{it} is mean-independent of D_{it} conditional on $Z_i = j$ under the parallel trends and no-anticipation assumptions.

Proposition 4. *(Mean independence) Suppose that Assumption 1-4 holds. Then, conditional on $Z_i = j$, η_{it} defined by (14)-(16) for $t = 2, ..., T$ is mean independent of D_{it}.*

Given the representation (17), the observed outcome $\{\Delta Y_{it} : t = 2, ..., T\}$ follows a first-order Markov process when the regression residuals are serially independent.

Assumption 7. *$(\epsilon_{it}(0), \epsilon_{it}(g))$ is independent of $(\epsilon_{is}(0), \epsilon_{is}(g))$ for all $t \neq s$ conditional on $G_i = g$ for all $g \in \mathcal{G}$.*

Assumption 7 implies Assumptions 5-6.

Proposition 5. *Suppose that Assumptions 1-4, and 7 hold and the potential outcomes are generated as in (12)-(13). Then, Assumption 5-6 holds.*

Assumption 7 can be relaxed by requiring that $(\epsilon_{it}(0), \epsilon_{it}(g))$ is independent of $(\epsilon_{is}(0), \epsilon_{is}(g))$ for all t and s such that $|t - s| \geq r$ for some $r \geq 2$ when the length panel data is sufficiently large.

Leveraging the result of Propositions 4 and 5, we are now poised to construct an estimator for the LGATT.

4.2 Soft-classification estimator

Define the parameter $\boldsymbol{\theta}_g := ((\boldsymbol{\beta}_g^1)^\top, ..., (\boldsymbol{\beta}_g^J)^\top)^\top \in \Theta_{\boldsymbol{\theta}_g}$ with $\boldsymbol{\beta}_g^j := ((\Delta\delta_g^j, \gamma_{g,g}^j), ..., (\Delta\delta_T^j, \gamma_{g,T}^j))^\top$ for $j = 1, 2, ..., J$ and $g \in \mathcal{G}$. Collect the parameter $\boldsymbol{\theta}_g$ for $g = \underline{g}, ..., \bar{g}$ into a vector as $\boldsymbol{\vartheta} := (\boldsymbol{\theta}_{\underline{g}}^\top, ..., \boldsymbol{\theta}_{\bar{g}}^\top)^\top \in \Theta_{\boldsymbol{\vartheta}} := \prod_{g=\underline{g}}^{\bar{g}} \Theta_{\boldsymbol{\theta}_g}$. Denote the true value of $\boldsymbol{\vartheta}$ by $\boldsymbol{\vartheta}^0 = ((\boldsymbol{\theta}_{\underline{g}}^0)^\top, ..., (\boldsymbol{\theta}_{\bar{g}}^0)^\top)^\top$.

In view of (17) and Proposition 4, we may identify the true value of $\boldsymbol{\theta}_g$ denoted by $\boldsymbol{\theta}_g^0$ as

$$\boldsymbol{\theta}_g^0 = \arg\min_{\boldsymbol{\theta}_g \in \Theta_{\boldsymbol{\theta}_g}} \mathbb{E}\left[\sum_{j=1}^{J} \mathbb{I}\{Z_i = j\} \sum_{t=g}^{T} \left(\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it}\right)^2 \middle| G_i \in \{0, g\}\right]. \tag{18}$$

By the law of iterated expectations, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{J}\mathbb{I}\{Z_i=j\}\sum_{t=g}^{T}\left(\Delta Y_{it}-\Delta\delta_t^j-\gamma_{g,t}^j D_{it}\right)^2\middle| G_i\in\{0,g\}\right]\\
&=\mathbb{E}\left[\sum_{j=1}^{J}\mathbb{E}\left[\mathbb{I}\{Z_i=j\}\sum_{t=g}^{T}\left(\Delta Y_{it}-\Delta\delta_t^j-\gamma_{g,t}^j D_{it}\right)^2\middle| \boldsymbol{W}_{2,i}^{g-2}\right]\middle| G_i\in\{0,g\}\right]\\
&=\mathbb{E}\left[\sum_{j=1}^{J}\tau^j(\boldsymbol{W}_{2,i}^{g-2})\sum_{t=g}^{T}\left(\Delta Y_{it}-\Delta\delta_t^j-\gamma_{g,t}^j D_{it}\right)^2\middle| G_i\in\{0,g\}\right], \qquad (19)
\end{aligned}$$

where $\boldsymbol{W}_{2,i}^{g-2}=(\boldsymbol{W}_{i2},....\boldsymbol{W}_{i,g-2})^\top$, and the third equality follows from the independence between $\boldsymbol{W}_{2,i}^{g-2}$ and $\{\boldsymbol{W}_{g,i}^T, D_{ig},...,D_{iT}\}$. In view of this equation, the first order condition for the minimization problem in (18) leads to the equation (10).

This suggests that we may estimate $\boldsymbol{\theta}_g$ by minimizing the following sample analogue criterion function of (18):

$$\hat{\boldsymbol{\theta}}_g=\arg\min_{\boldsymbol{\theta}_g\in\Theta_{\boldsymbol{\vartheta}_g}}\sum_{i\in\mathcal{I}_g}\sum_{j=1}^{J}\hat{\tau}_g^j(\boldsymbol{W}_{2,i}^{g-2})\sum_{t=g}^{T}(\Delta Y_{it}-\Delta\delta_t^j-\gamma_{g,t}^j D_{it})^2 \quad \text{for } g=\underline{g},...,\bar{g}, \qquad (20)$$

where $\mathcal{I}_g:=\{i: G_i\in\{0,g\}\}$ is a set of indices for the units with $G_i=0$ or g in the sample and $\hat{\tau}_g^j(\boldsymbol{W}_{2,i}^{g-2})$ is a consistent estimator for $\tau^j(\boldsymbol{W}_{2,i}^{g-2})$.

Let $\eta_{it}:=\epsilon_{it}(0)-\epsilon_{it-1}(0)$ for $t\le g-2$. For estimation, we assume that η_{it} follows a Gaussian distribution conditional on η_{it-1} and Z_i:

Assumption 8. *For $t=1,2,...,\bar{g}-2$, $\eta_{i2}|Z_i=j\overset{iid}{\sim}N(0,\sigma_\epsilon^2)$ and $\eta_{it}|\eta_{it-1},Z_i=j\overset{iid}{\sim}N(\rho\eta_{t-1},(1-\rho^2)\sigma_\epsilon^2)$.*

Let $\boldsymbol{\psi}:=(\pi^1,...,\pi^{J-1},(\Delta\boldsymbol{\delta}_{\bar{g}}^1)^\top,...,(\Delta\boldsymbol{\delta}_{\bar{g}}^J)^\top,\rho,\sigma_\epsilon^2)^\top\in\Theta_{\boldsymbol{\psi}}$ with $\Delta\boldsymbol{\delta}_{\bar{g}}^j:=(\Delta\delta_2^j,...,\Delta\delta_{\bar{g}-2}^j)^\top$ and $\pi^J=1-\sum_{j=1}^{J-1}\pi^j$. Denote the true value of $\boldsymbol{\psi}$ by $\boldsymbol{\psi}^0$. Assumption 8 with $\epsilon_{it}(0)\overset{iid}{\sim}N(0,\sigma^2)$ implies that $\eta_{it}|\eta_{it-1}\sim N(\rho\eta_{it-1},(1-\rho^2)\sigma_\epsilon^2)$ with $\rho=1/2$. We estimate $\boldsymbol{\psi}$ using pre-treatment observations by the maximum likelihood estimator as

$$\hat{\boldsymbol{\psi}}=\arg\max_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}}\sum_{i=1}^{n}\log L(\boldsymbol{W}_{2,i}^{\bar{g}-2},G_i;\boldsymbol{\psi}) \qquad (21)$$

with

$$L(\boldsymbol{W}_{2,i}^{\bar{g}-2},G_i;\boldsymbol{\psi}):=\sum_{j=1}^{J}\pi^j\left\{\mathbb{I}\{G_i=0\}f_{\bar{g}}(\boldsymbol{W}_{2,i}^{\bar{g}-2};\Delta\boldsymbol{\delta}_{\bar{g}}^j,\sigma_\epsilon^2)+\sum_{g=\underline{g}}^{\bar{g}}\mathbb{I}\{G_i=g\}f_g(\boldsymbol{W}_{2,i}^{g-2};\Delta\boldsymbol{\delta}_g^j,\sigma_\epsilon^2)\right\},$$

where $\Delta\boldsymbol{\delta}_g^j := (\Delta\delta_2^j, ..., \Delta\delta_{g-2}^j)^\top$ and

$$f_g(\boldsymbol{W}_{2,i}^{g-2}; \Delta\boldsymbol{\delta}_g^j, \rho, \sigma_\epsilon^2) := \frac{1}{\sigma_\epsilon}\phi\left(\frac{\eta_{i2}(\Delta\delta_t^j)}{\sigma_\epsilon}\right)\prod_{t=3}^{g-2}\frac{1}{\sqrt{1-\rho^2}\sigma_\epsilon}\phi\left(\frac{\eta_{it}(\Delta\delta_t^j) - \rho\eta_{it-1}(\Delta\delta_{t-1}^j)}{\sqrt{1-\rho^2}\sigma_\epsilon}\right)$$

with $\eta_{it}(\Delta\delta_t^j) := \Delta Y_{it} - \Delta\delta_t^j$ and $\phi(t) =: \exp(-t^2/2)/\sqrt{2}\pi$.

The posterior probabilities of latent type j given $\boldsymbol{W}_2^{g-2}$ and $G \in \{0, g\}$ depend on $\boldsymbol{\psi}$ as

$$\tau_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) := \frac{\pi^j f_g(\boldsymbol{W}_2^{g-2}; \Delta\boldsymbol{\delta}_g^j, \rho, \sigma_\epsilon^2)}{\sum_{k=1}^J \pi^k f_g(\boldsymbol{W}_2^{g-2}; \Delta\boldsymbol{\delta}_g^k, \rho, \sigma_\epsilon^2)}, \tag{22}$$

and we set $\hat{\tau}_g^j(\boldsymbol{W}_2^{g-2}) := \tau_g^j(\boldsymbol{W}_2^{g-2}; \hat{\boldsymbol{\psi}})$ in (20).

Let

$$m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g, \boldsymbol{\psi}) := \mathbb{I}\{G \in \{0, g\}\}\nabla_{\boldsymbol{\theta}_g}\sum_{j=1}^J \tau_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi})\sum_{t=g}^T(\Delta Y_t - \Delta\delta_t^j - \gamma_{g,t}^j D_t)^2,$$
$$s(\boldsymbol{W}_2^{\bar{g}-2}, G; \boldsymbol{\psi}) := \nabla_{\boldsymbol{\psi}^\top}\log L(\boldsymbol{W}_2^{\bar{g}-2}, G; \boldsymbol{\psi}),$$

and define

$$\boldsymbol{M}_{\boldsymbol{\theta},g} := \mathbb{E}[\nabla_{\boldsymbol{\theta}^\top} m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)],\ \boldsymbol{M}_{\boldsymbol{\psi},g} := \mathbb{E}[\nabla_{\boldsymbol{\psi}^\top} m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)],\ \boldsymbol{S} := \mathbb{E}[\nabla_{\boldsymbol{\psi}^\top} s(\boldsymbol{W}_2^{\bar{g}-2}, G; \boldsymbol{\psi}^0)],$$
$$\boldsymbol{\Omega}_{g,g'} = \mathbb{E}\left[\left\{m_g^0(\boldsymbol{W}_g^T, G) + \boldsymbol{M}_{\boldsymbol{\psi},g}\boldsymbol{S}^{-1}s^0(\boldsymbol{W}_2^{\bar{g}-2}, G)\right\}\left\{m_{g'}^0(\boldsymbol{W}_{g'}^T, G) + \boldsymbol{M}_{\boldsymbol{\psi},g'}\boldsymbol{S}^{-1}s^0(\boldsymbol{W}_2^{\bar{g}-2}, G)\right\}^\top\right],$$

where $m_g^0(\boldsymbol{W}_g^T, G) := m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)$ and $s^0(\boldsymbol{W}_2^{\bar{g}-2}, G) := s(\boldsymbol{W}_2^{\bar{g}-2}, G; \boldsymbol{\psi}^0)$.

Recall that the number of components, J, is the smallest integer such that the representation of the data density for $\boldsymbol{W}_{2,i}^{g-2}$ given by $\sum_{j=1}^J \pi^j f(\boldsymbol{W}_{2,i}^{g-2}; \Delta\boldsymbol{\delta}^j, \sigma_\epsilon^2)$ admits the true density function.

Assumption 9. *(a) $(\boldsymbol{\vartheta}^0, \boldsymbol{\psi}^0) \in int(\Theta_{\boldsymbol{\vartheta}} \times \Theta_{\boldsymbol{\psi}})$, where $\Theta_{\boldsymbol{\vartheta}} \times \Theta_{\boldsymbol{\psi}}$ is compact. (b) $\Delta\boldsymbol{\delta}^{j,0} \neq \Delta\boldsymbol{\delta}^{k,0}$ for $j \neq k$ and $j, k = 1, ..., J$. (c) $\boldsymbol{\beta}_g^{j,0} \neq \boldsymbol{\beta}_g^{k,0}$ for $j \neq k$ and $j, k = 1, ..., J$ and for $g = \underline{g}, ..., \bar{g}$. (d) J is known. (e) $\mathbb{E}\Delta Y^{2+\delta} < \infty$ for some $\delta > 0$.*

Proposition 6. *Assumptions 1-4 and 7-9 hold. Then, (a) $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0$, (b) $\hat{\boldsymbol{\vartheta}} \xrightarrow{p} \boldsymbol{\vartheta}^0$, and (c) $\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where*

$$\boldsymbol{\Sigma} := \begin{pmatrix} \boldsymbol{\Sigma}_{\underline{g},\underline{g}} & \boldsymbol{\Sigma}_{\underline{g},\underline{g}+1} & \cdots & \boldsymbol{\Sigma}_{\underline{g},\bar{g}} \\ \boldsymbol{\Sigma}_{\underline{g}+1,\underline{g}} & \boldsymbol{\Sigma}_{\underline{g}+1,\underline{g}+1} & \cdots & \boldsymbol{\Sigma}_{\underline{g}+1,\bar{g}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{\bar{g},\underline{g}} & \boldsymbol{\Sigma}_{\bar{g},\underline{g}+1} & \cdots & \boldsymbol{\Sigma}_{\bar{g},\bar{g}} \end{pmatrix}.$$

with $\boldsymbol{\Sigma}_{g,g'} = \boldsymbol{M}_{\boldsymbol{\theta},g}^{-1}\boldsymbol{\Omega}_{g,g'}(\boldsymbol{M}_{\boldsymbol{\theta},g'}^{-1})^\top$.

When Assumption 8 does not hold because the Gaussian parametric model is a mis-specified model, the estimator $\hat{\boldsymbol{\psi}}$ is no longer consistent. Consequently, $\hat{\boldsymbol{\vartheta}}$ becomes inconsistent and is

subject to asymptotic bias. Nonetheless, we may analyze an asymptotic distribution of $\hat{\boldsymbol{\vartheta}}$ under mis-specification as in Gallant and White (1988).

To analyze a misspecified case, define the pseudo true value of $\boldsymbol{\vartheta}$ by

$$\boldsymbol{\vartheta}_* = \arg\max_{\boldsymbol{\vartheta}\in\Theta_{\boldsymbol{\vartheta}}} \mathbb{E}\left[\sum_{j=1}^{J} \tau_g^j(\boldsymbol{W}_{2,i}^{g-2};\boldsymbol{\psi}_*)\sum_{t=g}^{T}\left(\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it}\right)^2 \middle| G_i \in \{0,g\}\right],$$

where

$$\boldsymbol{\psi}_* = \arg\max_{\boldsymbol{\psi}\in\Theta_{\boldsymbol{\psi}}} \mathbb{E}\left[\log\left(\sum_{j=1}^{J}\pi^j\left\{\mathbb{I}\{G_i=0\}f_{\bar{g}}(\boldsymbol{W}_{2,i}^{\bar{g}-2};\Delta\boldsymbol{\delta}_{\bar{g}}^j,\sigma_\epsilon^2) + \sum_{g=\underline{g}}^{\bar{g}}\mathbb{I}\{G_i=g\}f_g(\boldsymbol{W}_{2,i}^{g-2};\Delta\boldsymbol{\delta}_g^j,\sigma_\epsilon^2)\right\}\right)\right].$$

When Assumption 8 does not hold, the value of $\boldsymbol{\vartheta}^0$ defined in (18), $\boldsymbol{\vartheta}_* \neq \boldsymbol{\vartheta}^0$ because $\tau^j(\boldsymbol{W}_{2,i}^{g-2};\boldsymbol{\psi}_*) \neq \mathbb{E}[Z_i = j|\boldsymbol{W}_{2,i}^{g-2}]$. The following proposition shows that the asymptotic distribution of $\hat{\boldsymbol{\vartheta}}$ under a misspecification.

Assumption 10. *(a) $(\boldsymbol{\vartheta}_*,\boldsymbol{\psi}_*) \in int(\Theta_{\boldsymbol{\vartheta}}\times\Theta_{\boldsymbol{\psi}})$, where $\Theta_{\boldsymbol{\vartheta}}\times\Theta_{\boldsymbol{\psi}}$ is compact. (b) $\Delta\boldsymbol{\delta}_*^j \neq \Delta\boldsymbol{\delta}_*^k$ for $j\neq k$ and $j,k=1,...,J$. (c) $\boldsymbol{\beta}_{*g}^j \neq \boldsymbol{\beta}_{*g}^k$ for $j\neq k$ and $j,k=1,...,J$ and for $g=\underline{g},...,\bar{g}$. (d) J is known. (e) ΔY_{it} has a finite variance.*

Proposition 7. *Assumptions 1-4 and 7, 10 hold. Then, (a) $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}_*$, (b) $\hat{\boldsymbol{\vartheta}} \xrightarrow{p} \boldsymbol{\vartheta}_*$, and (c) $\sqrt{n}(\hat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_*) \xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma}_*)$, where $\boldsymbol{\Sigma}_*$ is defined similarly to $\boldsymbol{\Sigma}$ in Proposition 6 but replacing $\boldsymbol{\psi}^0$ and $\boldsymbol{\vartheta}^0$ with $\boldsymbol{\psi}_*$ and $\boldsymbol{\vartheta}_*$, respectively.*

4.3 Hard-classification estimator

We also consider an estimator based on "hard classification" of each unit. Namely, given the posterior probability estimate $\hat{\tau}_g^j(\boldsymbol{W}_{2,i}^{g-2})$ in (22), define an estimator for the value of Z_i as

$$\hat{Z}_i = j \quad \text{if} \quad \hat{\tau}_g^j(\boldsymbol{W}_{2,i}^{g-2}) > \hat{\tau}_g^k(\boldsymbol{W}_{2,i}^{g-2}) \text{ for all } k\neq j.$$

Then, we propose the following estimator based on hard-classification:

$$\tilde{\boldsymbol{\theta}}_g = \arg\min_{\boldsymbol{\theta}_g\in\Theta_{\boldsymbol{\theta}_g}} \sum_{i\in\mathcal{I}_g}\sum_{j=1}^{J}\mathbb{I}\{\hat{Z}_i=j\}\sum_{t=g}^{T}(\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it})^2. \tag{23}$$

When the time length T is fixed and short, the hard-classification estimator $\tilde{\boldsymbol{\theta}}_g$ is an inconsistent estimator of $\boldsymbol{\theta}_g^0$ because the misclassification of $\hat{Z}_i$ does not approach zero even when $n\to\infty$.

By Proposition 6(a), the classifier $\hat{Z}_i$ is asymptotically equivalent to the classifier of latent type defined by

$$Z_i^* = j \quad \text{if} \quad \tau^{j,0}(\boldsymbol{W}_{2,i}^{g-2}) > \tau^{k,0}(\boldsymbol{W}_{2,i}^{g-2}) \text{ for all } k\neq j,$$

where

$$\tau^{j,0}(\boldsymbol{W}_{2,i}^{g-2}) := \frac{\pi^{j,0} f(\boldsymbol{W}_{2,i}^{g-2}; \Delta\boldsymbol{\delta}^{j,0}, (\sigma_\epsilon^0)^2)}{\sum_{k=1}^{J} \hat{\pi}^{k,0} f(\boldsymbol{W}_{2,i}^{g-2}; \Delta\boldsymbol{\delta}^{k,0}, (\sigma_\epsilon^0)^2)}.$$

The classifier Z_i^* misclassifies its latent type with strictly positive probabilities, i.e., $\Pr(Z_i^* \neq Z_i) > 0$. Consequently, $\tilde{\boldsymbol{\theta}}_g$ is asymptotically biased.

To define the probability limit of $\tilde{\boldsymbol{\theta}}_g$, consider a population analogue of (23) and define $\boldsymbol{\theta}_g^*$ by

$$\boldsymbol{\theta}_g^* := \arg\min_{\boldsymbol{\theta}_g \in \Theta_{\boldsymbol{\theta}_g}} \mathbb{E}\left[\sum_{j=1}^{J} \sum_{t=g}^{T} \mathbb{I}\{Z_i^* = j\} (\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it})^2 \middle| G_i \in \{0, g\} \right] \quad \text{for } g = \underline{g}, ..., \bar{g}. \tag{24}$$

In general, $\boldsymbol{\theta}_g^* \neq \boldsymbol{\theta}_g^0$ because $\Pr(Z_i^* \neq j | Z_i = j) > 0$.

Define a binary random variable

$$I_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) := \mathbb{I}\{\tau_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) > \tau_g^k(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) \text{ for all } k \neq j\}.$$

Note that $\mathbb{I}\{\hat{Z}_i = j\} = I_g^j(\boldsymbol{W}_{i,2}^{g-2}; \hat{\boldsymbol{\psi}})$ and $\mathbb{I}\{Z_i^* = j\} = I_g^j(\boldsymbol{W}_{i,2}^{g-2}; \boldsymbol{\psi}^0)$.

Let

$$\tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g, \boldsymbol{\psi}) := \mathbb{I}\{G \in \{0, g\}\} \nabla_{\boldsymbol{\theta}_g} \sum_{j=1}^{J} I_g^j(\boldsymbol{W}_{i,2}^{g-2}; \boldsymbol{\psi}) \sum_{t=g}^{T} (\Delta Y_t - \Delta\delta_t^j - \gamma_{g,t}^j D_t)^2,$$

and define

$$\tilde{\boldsymbol{M}}_{\boldsymbol{\theta},g} := \mathbb{E}[\nabla_{\boldsymbol{\theta}^\top} \tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)], \ \tilde{\boldsymbol{M}}_{\boldsymbol{\psi},g} := \nabla_{\boldsymbol{\psi}^\top} \mathbb{E}[\tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)],$$
$$\tilde{\boldsymbol{\Omega}}_{g,g'} = \mathbb{E}\left[\left\{ \tilde{m}_g^0(\boldsymbol{W}_g^T, G) + \tilde{\boldsymbol{M}}_{\boldsymbol{\psi},g} \boldsymbol{S}^{-1} s^0(\boldsymbol{W}_2^{\bar{g}-2}, G) \right\} \left\{ \tilde{m}_{g'}^0(\boldsymbol{W}_{g'}^T, G) + \tilde{\boldsymbol{M}}_{\boldsymbol{\psi},g'} \boldsymbol{S}^{-1} s^0(\boldsymbol{W}_2^{\bar{g}-2}, G) \right\}^\top \right],$$

where $\tilde{m}_g^0(\boldsymbol{W}_g^T, G) := \tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)$.

The following proposition derives the asymptotic distribution of $\tilde{\boldsymbol{\theta}}_g$.

Proposition 8. *Assumptions 1-4 and 7-9 hold. Then, (a)* $\tilde{\boldsymbol{\theta}}_g \xrightarrow{p} \boldsymbol{\theta}_g^*$ *and (b)* $\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*) \xrightarrow{d} N(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}})$, *where*

$$\tilde{\boldsymbol{\Sigma}} := \begin{pmatrix} \tilde{\boldsymbol{\Sigma}}_{\underline{g},\underline{g}} & \tilde{\boldsymbol{\Sigma}}_{\underline{g},\underline{g}+1} & \cdots & \tilde{\boldsymbol{\Sigma}}_{\underline{g},\bar{g}} \\ \tilde{\boldsymbol{\Sigma}}_{\underline{g}+1,\underline{g}} & \tilde{\boldsymbol{\Sigma}}_{\underline{g}+1,\underline{g}+1} & \cdots & \tilde{\boldsymbol{\Sigma}}_{\underline{g}+1,\bar{g}} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\boldsymbol{\Sigma}}_{\bar{g},\underline{g}} & \tilde{\boldsymbol{\Sigma}}_{\bar{g},\underline{g}+1} & \cdots & \tilde{\boldsymbol{\Sigma}}_{\bar{g},\bar{g}} \end{pmatrix}.$$

with $\tilde{\boldsymbol{\Sigma}}_{g,g'} = \tilde{\boldsymbol{M}}_{\boldsymbol{\theta},g}^{-1} \tilde{\boldsymbol{\Omega}}_{g,g'} (\tilde{\boldsymbol{M}}_{\boldsymbol{\theta},g'}^{-1})^\top$.

4.4 Multiplier bootstrap

We implement the following multiplier bootstrap to compute the confidence intervals.

1. We generate $\xi_i^b \overset{i.i.d.}{\sim} N(0,1)$ for $i = 1, ..., n$ and $b = 1, 2, ..., B$ and compute

$$\boldsymbol{\mathcal{T}}_{n,g}^b = \hat{\boldsymbol{M}}_{\boldsymbol{\theta}_g}^{-1} \times \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_g(\boldsymbol{W}_{g,i}^T, G_i; \hat{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) + \hat{\boldsymbol{M}}_{\boldsymbol{\psi},g} \hat{\boldsymbol{S}}^{-1} s(\boldsymbol{W}_{2,i}^{\bar{g}-2}, G_i; \hat{\boldsymbol{\psi}}) \right) \xi_i^b$$

for $g = \underline{g}, ..., \bar{g}$ and $b = 1, 2, ..., B$, where $\hat{\boldsymbol{M}}_{\boldsymbol{\theta}_g}$, $\hat{\boldsymbol{M}}_{\boldsymbol{\psi}}$, and $\hat{\boldsymbol{S}}_g$ are sample analogue estimators of $\boldsymbol{M}_{\boldsymbol{\theta}_g}$, $\boldsymbol{M}_{\boldsymbol{\psi}}$, and $\boldsymbol{S}_g$. Let $\{\{\boldsymbol{\mathcal{T}}_{n,g}^b\}_{g=\underline{g}}^{\bar{g}}\}_{b=1}^B$ be the compute values of $\boldsymbol{\mathcal{T}}_{n,g}^b$.

2. Let $\boldsymbol{\theta}_g^*[k]$ and $\boldsymbol{\mathcal{T}}_{n,g}^b[k]$ denote the kth element of $\boldsymbol{\theta}_g^*$ and $\boldsymbol{\mathcal{T}}_{n,g}^b$, respectively. Given $\alpha \in (0, 1/2)$, e.g., $\alpha = 0.05$, compute the $\alpha/2$ and $(1-\alpha/2)$ quantile of $\{\boldsymbol{\mathcal{T}}_{n,g}^b[k]\}_{b=1}^B$ denoted by $q_{n,g,\alpha/2}^B[k]$ and $q_{n,g,1-\alpha/2}^B[k]$, respectively. Then, we construct the $100(1-\alpha)$ percentile confidence interval for $\boldsymbol{\theta}_g^*[k]$ as

$$CI_{n,g}^B[k] = [\hat{\boldsymbol{\theta}}_g[k] - q_{n,g,1-\alpha/2}^B[k]/\sqrt{n},\ \hat{\boldsymbol{\theta}}_g[k] - q_{n,g,\alpha/2}^B[k]/\sqrt{n}].$$

Then,

$$\Pr(\boldsymbol{\theta}_g^*[k] \in CI_{n,g}^B[k]) \to 1 - \alpha, \quad \text{as } n, B \to \infty.$$

4.5 Implementation

We describe implementation of the proposed estimators. Our proposed estimation procedure consists of the following two steps: first, we run the EM algorithm on the pre-treatment observations to estimate the posterior probabilities $\hat{\tau}^j(\boldsymbol{W}_{2,i}^{g-2})$ for each i, j, and g. Second, we solve the weighted least squares problem in (20) using the estimated posterior probabilities from the first step as weights. The entire procedure can be summarized as follows:

Algorithm 1. *(Two-step estimation procedure)*

0. Initiate the process with a given starting value, $\boldsymbol{\psi}_{(0)} \in \Theta_{\boldsymbol{\psi}}$.

1. Estimate ψ by employing the following EM algorithm, starting with $s = 0$:

(a) (E-step) Compute the posterior probabilities from $\boldsymbol{\psi}_{(s)}$ for each g, i, and j:

$$\hat{\tau}_{(s)}^j(\boldsymbol{W}_{2,i}^{g-2}) = \frac{\pi_{(s)}^j f\left(\boldsymbol{W}_{2,i}^{g-2}; \boldsymbol{\psi}_{(s)}^j, \sigma_{\epsilon(s)}^2\right)}{\sum_{k=1}^J \pi_{(s)}^k f\left(\boldsymbol{W}_{2,i}^{g-2}; \boldsymbol{\psi}_{(s)}^k, \sigma_{\epsilon(s)}^2\right)}$$

(b) (M-step) Update the parameters by maximum likelihood estimation:

$$\boldsymbol{\psi}_{(s+1)} = \arg\max_{\boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{J} \hat{\tau}_{(s)}^j(\boldsymbol{W}_{2,i}^{g-2}) f(\boldsymbol{W}_{2,i}^{g-2}; \Delta\boldsymbol{\delta}^j, \sigma_\epsilon^2) \right),$$

(c) Set $s = s + 1$ and repeat until convergence.

2. *Using the weights* $\hat{\tau}^j(\boldsymbol{W}_{2,i}^{g-2})$ *from the first step, solve the following weighted least squares problem:*

$$\hat{\boldsymbol{\theta}}_g = \arg\min_{\boldsymbol{\theta}_g \in \Theta_{\boldsymbol{\theta}_g}} \sum_{i \in \mathcal{I}_g} \sum_{j=1}^{J} \hat{\tau}^j(\boldsymbol{W}_{2,i}^{g-2}) \sum_{t=g}^{T} (\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it})^2$$

It is worth noting the EM algorithm is executed only on the pre-treatment observations and we do not require any iterative repetition on post-treatment observations for estimating the parameters of interest, γ_t^j. This is because we identify latent group structures based solely on pre-treatment trends.

The algorithm above is described for the soft-classification case. In the case of hard classification, the posterior probabilities $\hat{\tau}^j(\boldsymbol{W}_{2,i}^{g-2})$ can be replaced with the hard-classification $\mathbb{I}\{\hat{Z}_i = j\}$ in the second step of the algorithm, by defining $\hat{Z}_i := \arg\max_{j=1,\ldots,J} \hat{\tau}^j(\boldsymbol{W}_{2,i}^{g-2})$.

5 Test for the parallel trends in pre-treatment periods

When we have a sufficiently long length of the pre-treatment periods, we may identify the latent types using the data from $t = 2$ to, say, $t = g - 3$ and test if the parallel trends assumption holds at $t = g - 1$. Specifically, we estimate

$$\tilde{\boldsymbol{\theta}}_g = \arg\min_{\boldsymbol{\theta}_g \in \Theta_{\boldsymbol{\theta}}} \sum_{i \in \mathcal{I}_g} \sum_{j=1}^{J} \hat{\tau}^j(\boldsymbol{W}_{2,i}^{g-3}) \sum_{t=g-1}^{T} (\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it})^2, \tag{25}$$

Then, we test $H_0 : \gamma^j(g, g-1) = 0$. The rejection of this null hypothesis indicates evidence that the parallel trends do not hold at $t = g - 1$. The asymptotic distribution of $\hat{\gamma}^j(g, g-1)$ is obtained analogously to that in Proposition 6 and so testing $H_0 : \gamma^j(g, g-1) = 0$ is straightforward.

6 Simulation

In this section, we investigate the finite-sample performance of our proposed method under an event study setup with $\mathcal{G} = \{T^*, 0\}$, where T^* represents the time of the event. We consider the following data generating process with two components $j = 1, 2$ as:

$$Y_{it} = \alpha_i + \delta_t^j + \mu_t^j D_{it} + \varepsilon_{it}^j \tag{26}$$

where ϵ_{it}^j is assumed to be independently and identically distributed following a normal distribution with mean 0 and standard deviation 1. We employ a total time period of $T = 12$ and a treatment assignment phase of $T^* = 8$. The individual effect α_i is drawn from a normal distribution with mean 0 and standard deviation 0.5.

We assign identical population latent group probabilities as $\pi^1 = \pi^2 = 0.5$, yet with different trends across groups as δ_t^j as $2 \times (j-2) \times t$. We use the threshold rule for treatment assignment so

that ith unit is treated at T^* if $\Delta Y_{i,T^*-1} < B$ and receives no treatment otherwise with $B = 3$. We also set $\mu_t^j = 3 \cdot (2-j) \cdot (t-T^*)$ for $t \geq T^*$ and zero for pre-treatment periods for $j = 1, 2$. Note that this data generating process features heterogenous treatment effects across the groups with $\gamma_t^1 = 3$ and $\gamma_t^2 = 0$ for all post-treatment periods t. Also, we let the parallel trends assumption hold within each latent group, but not in an aggregate level as illustrated in Figure 2, as we have different conditional probabilities of being treated across the groups due to heterogeneity in pretrends.

The simulation is run with a sample size of 400 units, supported by 500 replications. Our simulation results include the estimated trends γ_t^j in Table 1 and LGATTs μ_t^j in Table 2, along with their aggregate treatment effects on treated μ_t in Table 3. The mean estimates are presented, supplemented with their respective standard deviations, enclosed in parentheses. Despite having a small sample size, our LGATT and ATT estimates typically align closely with the population parameters.

It is noteworthy that, in the absence of any latent group structure ($J = 1$), the LGATT estimates align with the ATT estimates derived from the standard DID estimator, which suffer from negative bias. In fact, the sign of the ATT estimate is opposite of the true ATT when latent group structures are not taken into account. This highlights the importance of incorporating a latent group structure when identifying dynamic treatment effects, particularly when parallel trends assumptions do not hold at an aggregate level.

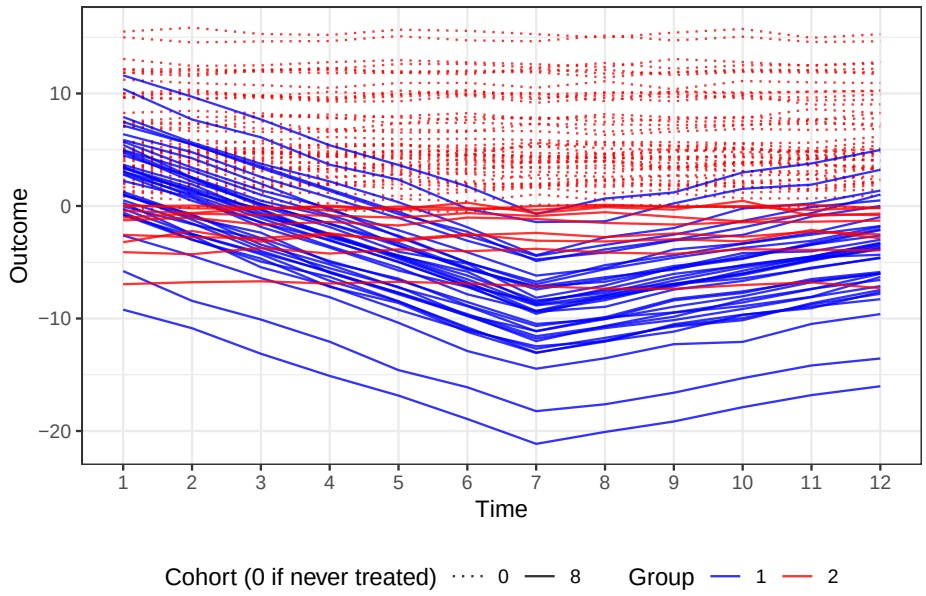
Table 1: Numerical Simulations, Trend Estimates.

	$J=1$	$J=2$	
Parameters	$j=1$	$j=1$	$j=2$
$\gamma_{T^*}^j$	0.95	3.03	0.02
	(0.05)	(0.13)	(0.06)
$\gamma_{T^*+1}^j$	0.94	3.00	-0.01
	(0.05)	(0.13)	(0.06)
$\gamma_{T^*+2}^j$	0.94	3.00	0.00
	(0.05)	(0.12)	(0.06)
$\gamma_{T^*+3}^j$	0.94	3.00	0.00
	(0.05)	(0.13)	(0.06)
$\gamma_{T^*+4}^j$	0.94	3.00	0.00
	(0.05)	(0.13)	(0.06)

7 Empirical Applications

In this section, we illustrate our method in empirical context by revisiting several economic papers that were published in journals. We first describe the data and the empirical strategy in each example, and then present the results. We also provide a brief discussion on the interpretation of the results and the limitations of our method in each example.

Figure 2: Numerical Simulations, Samples.



Notes: The figure is generated from 200 units from the data generating process described in section 6 with identical LGATTs but different treatment assignment probabilities across groups. Note that parallel trends assumptions hold within each latent group (distinguished by different colors) but not in an aggregate level.

Table 2: Numerical Simulations, LGATT Estimates.

	$J=1$	$J=2$	
Parameters	$j=1$	$j=1$	$j=2$
$\mu^j_{T^*}$	0.95	3.03	0.02
	(0.05)	(0.13)	(0.06)
$\mu^j_{T^*+1}$	1.89	6.03	0.02
	(0.08)	(0.12)	(0.06)
$\mu^j_{T^*+2}$	2.83	9.03	0.02
	(0.11)	(0.13)	(0.06)
$\mu^j_{T^*+3}$	3.77	12.03	0.02
	(0.14)	(0.13)	(0.06)
$\mu^j_{T^*+4}$	4.71	15.03	0.02
	(0.18)	(0.12)	(0.06)

Table 3: Numerical Simulations, ATT Estimates

Parameters	$J=1$	$J=2$
μ_{T^*}	0.95	2.60
	(0.05)	(0.13)
μ_{T^*+1}	1.89	5.18
	(0.08)	(0.17)
μ_{T^*+2}	2.83	7.75
	(0.11)	(0.23)
μ_{T^*+3}	3.77	10.33
	(0.14)	(0.28)
μ_{T^*+4}	4.71	12.90
	(0.18)	(0.35)

7.1 The effects of copyrights on science

To illustrate our method in an empirical context, we revisit Biasi and Moser (2021) that studies the impact of copyrights on science by exploiting an exogenous change toward weaker copyrights as a result of World War II. In 1942, the United States has introduced the Book Republication Program (BRP hereafter), which allowed publishers to reprint copies of science books that were owned by the enemy states at the time as a part of the broader war effort to increase the production of scientific knowledge in the United States.

Biasi and Moser (2021) compared the variation in citations received by books subjected to the Book Republication Program (BRP) with those from Switzerland to isolate the causal effect of the policy. This comparison was conducted with the assumption that authors writing in English could serve as an approximation for U.S. authors. Although certain Swiss books were published in German, they were not impacted by the BRP due to Switzerland's neutrality during the war.

This condition facilitates an objective measure for the comparison of new cumulative knowledge derived from the books, thus aiding in analyzing the effects of the policy on American science. This analysis operates on the assumption that both the BRP and Swiss books exhibited parallel trends in citations prior to the implementation of the BRP.

To address possible fundamental differences between the two types of the books, Biasi and Moser (2021) use Mahalanobis propensity score matching methods to create a comparable Swiss book sample for each BRP book by matching based on the research field and pre-BRP non-English-language citations. Their OLS estimates show that citations to BRP books in English compared to Swiss books increased significantly by additional 0.386 citations per year with the complete matched sample, suggesting that weaker copyrights helped cumulating further scientific knowledge in the United States. We revisit the sample used by Biasi and Moser (2021) and apply our method to estimate the LGATTs and the ATTs. Throughout this example, we use the following specification that extends the OLS regression setup Biasi and Moser (2021) used:

$$cites_{it} = \mu_t^j \text{BRP}_{it} + book_i + \delta_t^j + \varepsilon_{it}, \tag{27}$$

where the dependent variable $cites_{it}$ is citations to book i in English in year t. The variable D_{it} is a treatment variable variable that is equal to one if the BRP is applied on book i in year t. The parameter of interest is μ_t^j, which captures the changes in citations to BRP books in English compared to Swiss books after the BRP implementation. As in the original specification, we include both individual book fixed effects $book_i$ and time fixed effects δ_t^j with one caveat: the time fixed effect terms are grouped by the latent component j to capture grouped heterogeneity in trends for the observations in pretreatment periods. We use 6 years of data before the BRP and 20 years of data after the BRP to estimate the LGATTs. Of the 253 books included in the matched book sample, 152 books have complete citation history during the period.

We estimate models up to three components ($J = 1, 2, 3$) and compute the Bayesian information criterion (BIC) for each model for model selection. The estimated LGATTs for each model are presented in Figure 3. Three latent groups are differentiated by color in the plots. Notably, the first group consistently maintains higher citation trends compared to the other latent groups. In contrast, the second and third groups, initially characterized by low citation numbers prior to the BRP, display a consistent ascent in citations following the policy's initiation. The trends and group averages of the additional third group in the $J = 3$ model bear a strong resemblance to those of the second group. Crucially, the BIC criterion favors the $J = 2$ model over the $J = 1$ or $J = 3$ models, indicating that the model with two latent groups sufficiently captures heterogeneity in trends in the pre-treatment periods.

We additionally report the LGATTs and dynamic ATT estimates for $J = 2$ in Figure 4 and Figure 5 respectively with bootstrap confidence intervals. Interestingly, while the dynamic ATT estimates appear similar across all three models, the aggregate ATT estimates produce different values. The $J = 2$ model yields a marginally higher aggregate ATT estimate with an addition

of 0.705 citations per year, compared to the 0.693 additional citations per year yielded by the $J = 1$ model. This indicates the importance of accounting for grouped heterogeneity within the pretreatment trends. Note that our estimate for the $J = 1$ model within the LGATT framework differs from the OLS estimate reported in Biasi and Moser (2021) as we utilize a subset of their dataset that includes complete pretreatment history.

7.2 The effects of the block of Chinese Wikipedia in mainland China on user contribution

The block on Wikipedia, implemented in October 2005, affected Chinese Wikipedia users in mainland China. However, Chinese-speaking users in regions like Taiwan, Hong Kong, Singapore, and others around the world still had access. Zhang and Zhu (2011) mainly focused on the changes in behaviors of the nonblocked contributors who were active before the block and observed the change in their contribution behavior after the block was implemented. In this example, we explore the potency of social effects on contribution by comparing the shifts in contributions after the block, between users who were socially active prior to the block, and those who were not.

We use 8 weeks of data before the block and 8 weeks of data after the block to estimate the LGATTs. Of the 6,062 contributors who joined Wikipedia before the block, 1,408 are classified as nonblocked contributors and have complete contribution history during the period. It is worth noting that in the original study, the authors utilized a distinct sample of 1,707 contributors that included individuals who joined Wikipedia shortly before the block. We use a different sample of contributors because we need to observe the contribution levels of the contributors before the block for multiple periods.

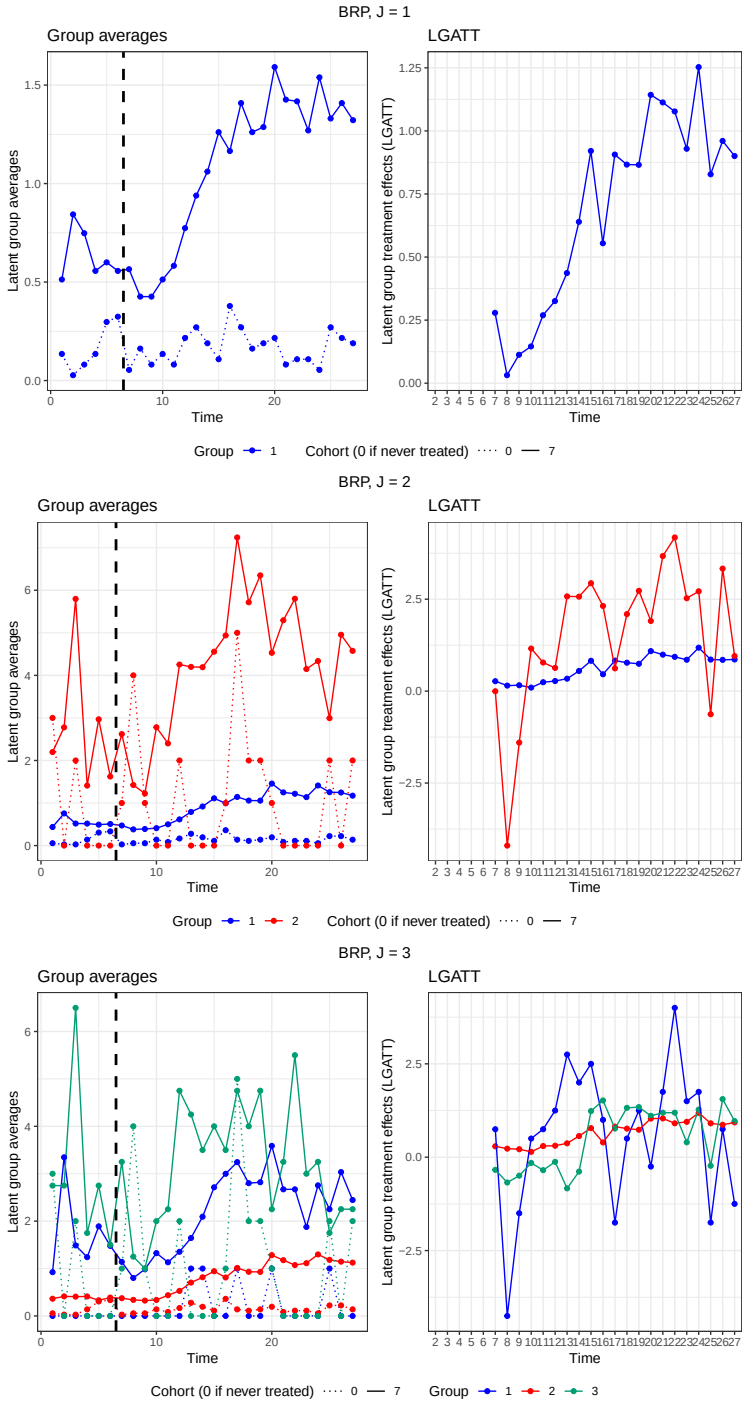
We now examine the social effects of the block by employing the following difference-in-differences specification:

$$contributions_{it} = \beta_0 + \delta_t^j + \mu_t^j D_{it} + \epsilon_{it} \ with \ D_{it} = AfterBlock_t \times SocialInteraction_i \qquad (28)$$

where i indexes the contributors and t indexes the weeks. The dependent variable, $contributions_{it}$, is the logarithm of weekly contributions of each nonblocked contributor to Wikipedia articles, where contribution is measured as the sum of the total characters added and deleted weekly. $AfterBlock_t$ is a dummy that equals one if the time period is after the block, and zero otherwise. To evaluate how the impact was larger for contributors who were actively involved in social interactions with other users, we incorporate two measures for $SocialInteraction_i$: participation in any social activities (social participation) and having a collaborator blocked in mainland China (having any collaborator blocked). 54.4% of the total sample belong to the first group and 85.6% belong to the second group.

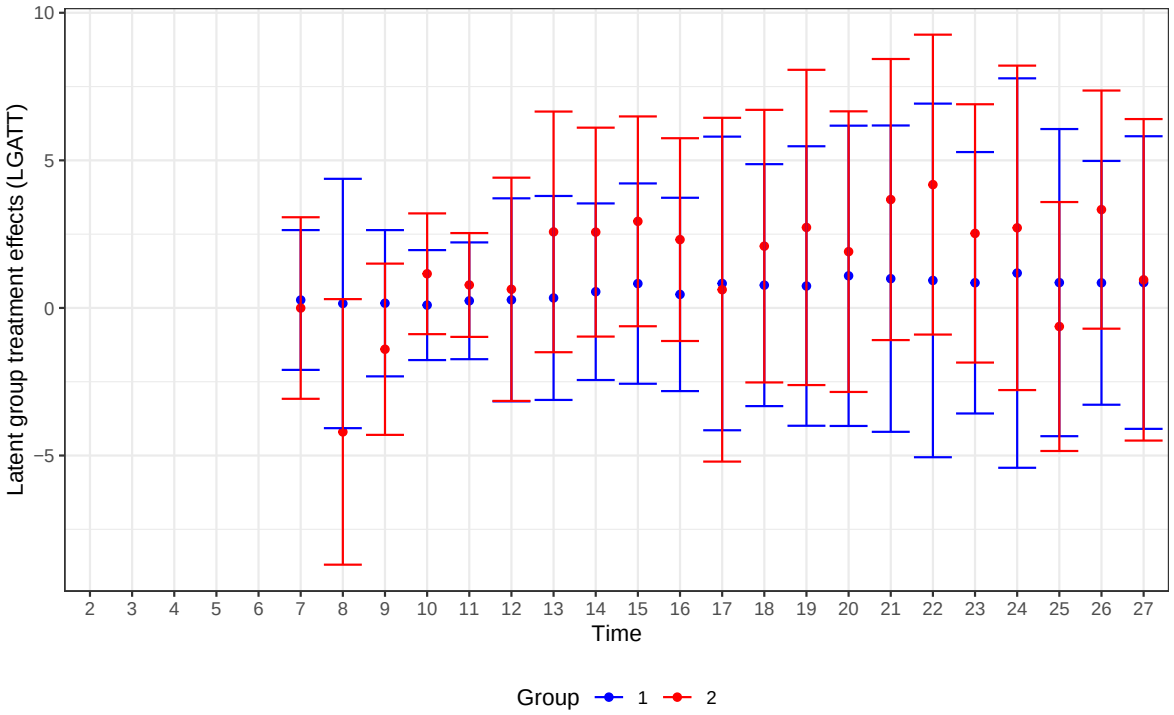
We report LGATTs for two groups in both models in Figure 6 and Figure 7. The LGATTs are estimated using the same procedure as in the simulation study. The two groups are distinguished by color; the first group, which exhibited a relatively lower trend in contribution amounts on average before the block, is represented by red, while the second group, which had a relatively higher

Figure 3: Citations in English on BRP books compared to Swiss books, LGATT.



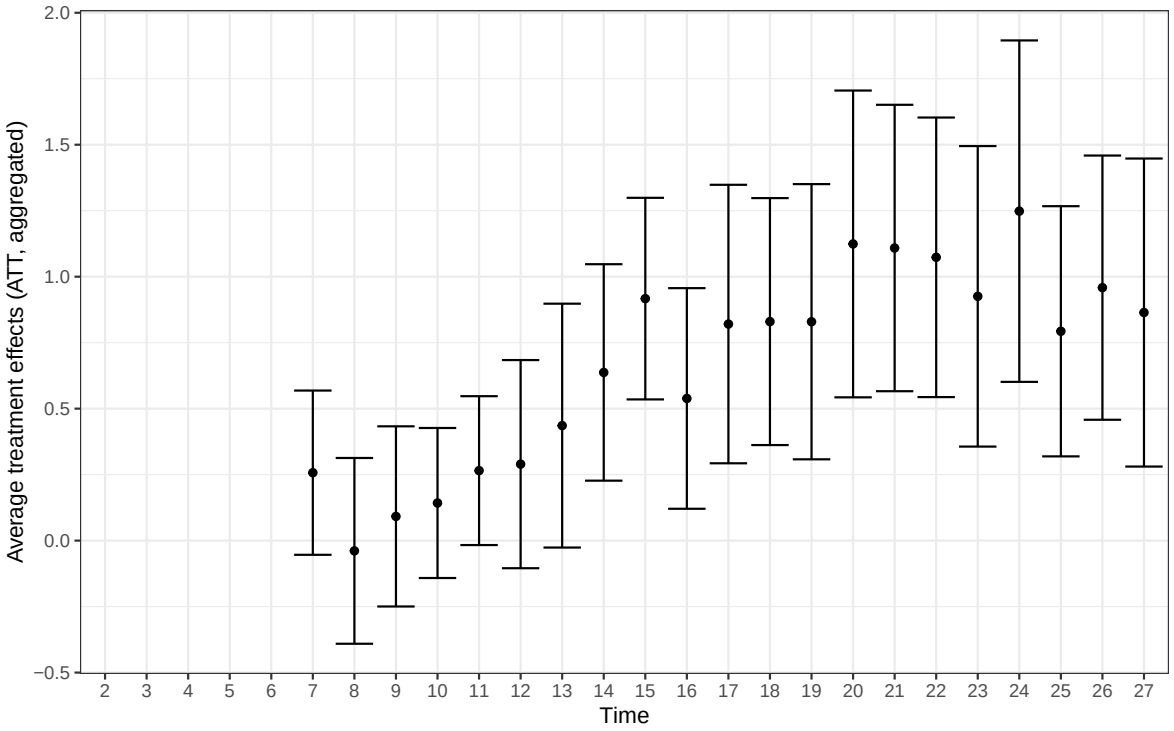
Notes: Plots on the left column present average weekly contribution per cohort and latent group. Latent-group specific average treatment effects on treated (LGATT) are shown on the right. Cohorts and latent groups are represented by line types (dotted lines for books from Switzerland) and colors respectively.

Figure 4: Citations in English on BRP books compared to Swiss books, LGATT for $J = 2$ model.



Notes: Latent-group specific average treatment effects on treated (LGATT) are shown. 95% confidence intervals for all the ATTs are generated from nonparametric bootstraps with 200 draws.

Figure 5: Citations in English on BRP books compared to Swiss books, ATT for $J = 2$ model.



Notes: ATT estimates from the canonical two-way fixed effects estimators are reported in the top figure. The bottom figure presents the ATT estimates computed from the weighted averages of LGATTs with two latent groups. 95% confidence intervals for all the ATTs are generated from nonparametric bootstraps with 200 draws.

trend in contribution levels, is denoted by blue. The prior latent group assignment probability estimates suggest that both groups contain a significant number of contributors. Specifically, the prior probability of belonging to the first group stands at 6.55% and 6.34% for the models with social activity and having any collaborator blocked respectively.

Introducing an additional group with $J = 3$ allows us to identify a new set of units exhibiting lower average treatment effects. In fact, the BIC criterion favors the $J = 3$ model over the other alternative specifications with $J = 1, 2$, or $J = 4$. The third group, which is represented by the green color in both figures, has a prior probability of 9.76% and 9.52% for the models with social activity and having any collaborator blocked respectively. The third group exhibits relatively higher levels of contribution before the block, and the LGATTs for this group are both negative and statistically significant.

On the other hand, their between-group differences in within-group differences between the treated and control groups are not statistically significant, as confirmed in the figures for the ATT in Figure 8 and Figure 9. These results reaffirm that the block negatively impacted contribution levels, irrespective of the contributors' trends of social activity or their interactions with other collaborators prior to the block, supporting the finding from Zhang and Zhu (2011) that social effects significantly motivate contribution.

8 Proofs

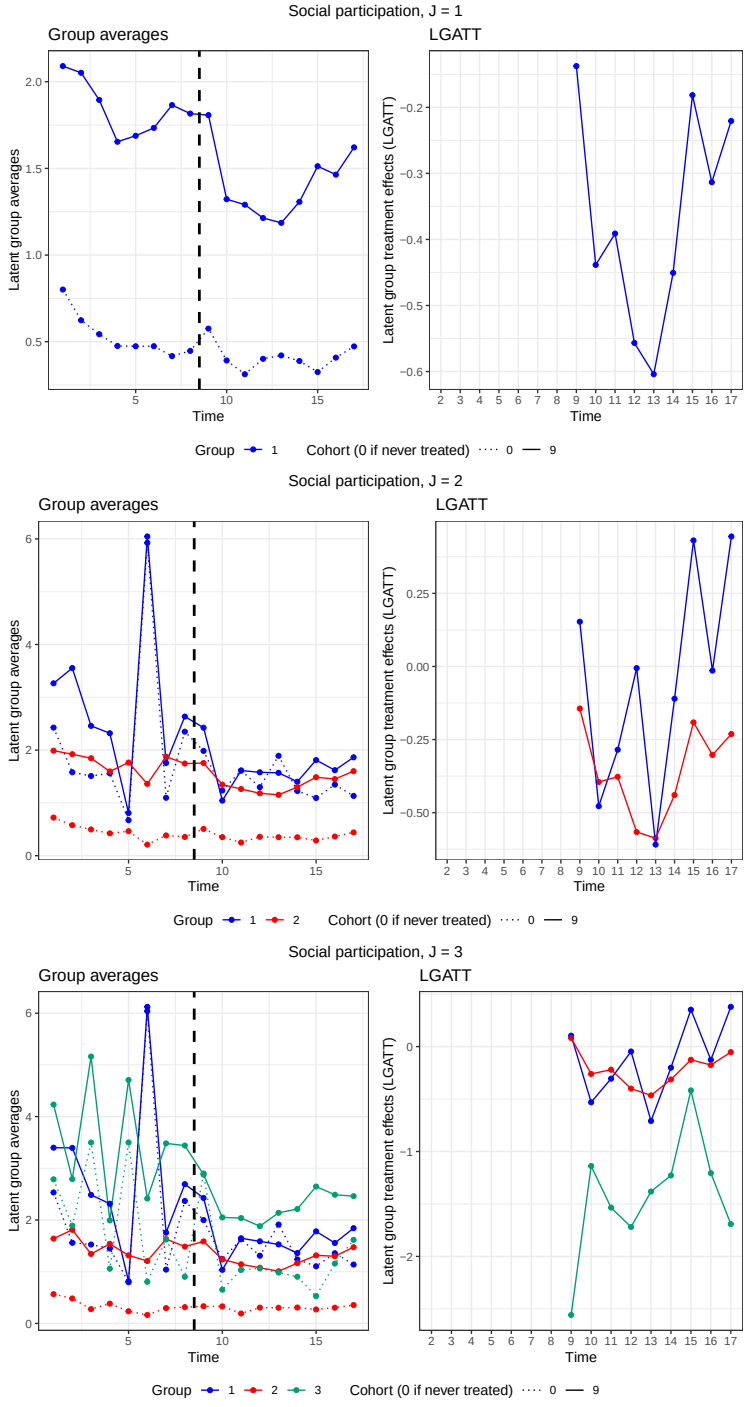
8.1 Proof of Proposition 1

For any $g \in \mathcal{G}$ and $t = 2, 3, ..., T$, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta Y_{it} \mid G_i = g, Z_i = j\right] &= \frac{\Pr(Z_i = j \mid G_i = g)\mathbb{E}\left[1(Z_i = j)\Delta Y_{it} \mid G_i = g, Z_i = j\right]}{\Pr(Z_i = j \mid G_i = g)} \\
&= \frac{\mathbb{E}\left[1(Z_i = j)\Delta Y_{it} \mid G_i = g\right]}{\Pr(Z_i = j \mid G_i = g)} \\
&= \frac{\mathbb{E}\left[\mathbb{E}\left[1(Z_i = j) \mid \boldsymbol{W}_{2,i}^{g-2}\right]\Delta Y_{it} \mid G_i = g\right]}{\Pr(Z_i = j \mid G_i = g)} \\
&= \frac{\mathbb{E}\left[\tau^j(\boldsymbol{W}_{2,i}^{g-2})\Delta Y_{it} \mid G_i = g\right]}{\Pr(Z_i = j \mid G_i = g)},
\end{aligned} \tag{29}$$

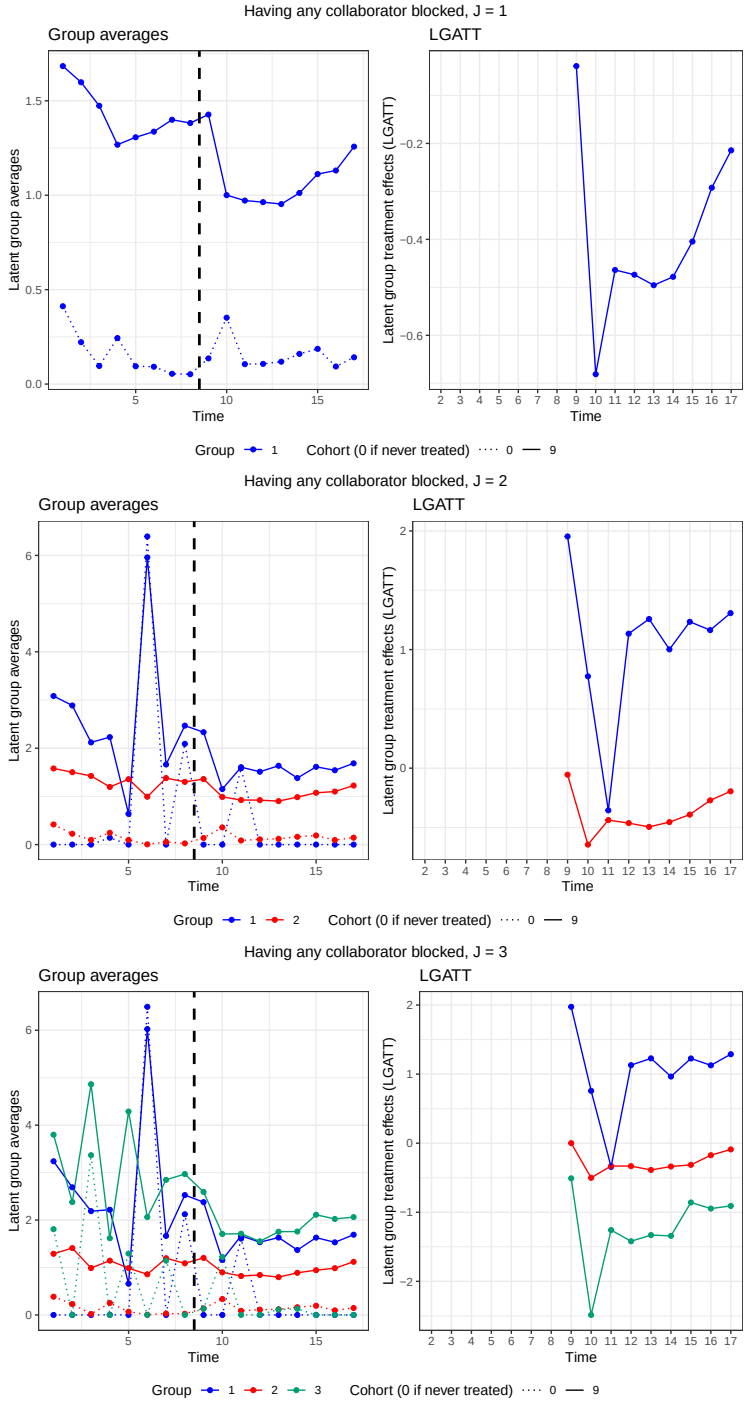
where the second equality follows from the law of iterated expectations and $\mathbb{E}[\Delta Y_{it}|\boldsymbol{W}_{2,i}^{g-2}] = \mathbb{E}[\Delta Y_{it}|G_i]$ because ΔY_{is} is independent of ΔY_{it} for $|t - s| \geq 2$ conditional on G_i. Therefore, the stated result follows from (10) and (29).

Figure 6: The block of Wikipedia on unblocked users with social participation, LGATT.



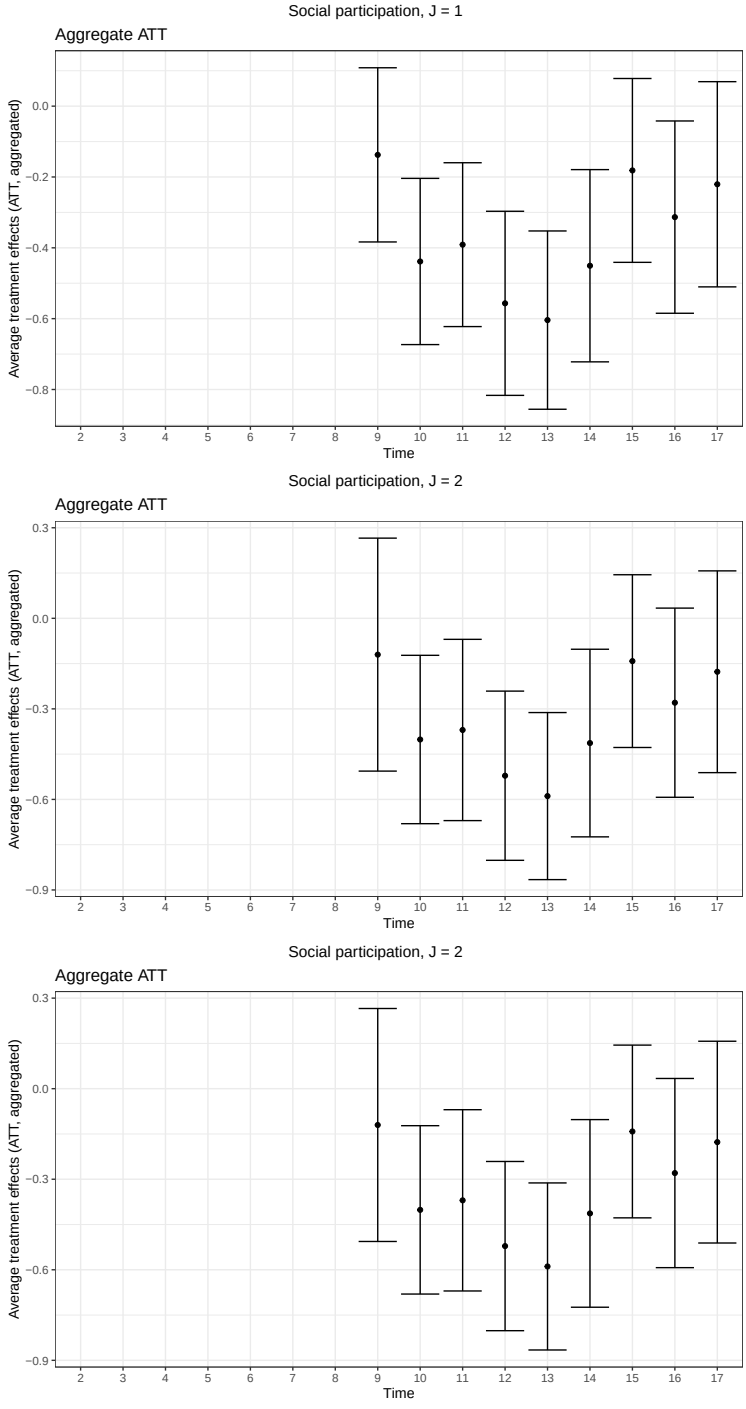
Notes: Plots on the left column present average weekly contribution per cohort and latent group. Latent-group specific average treatment effects on treated (LGATT) are shown on the right. Cohorts and latent groups are represented by line types (dotted lines for contributors with no social activity) and colors respectively. 95% confidence intervals for the LGATT are generated from nonparametric bootstraps with 200 draws.

Figure 7: The block of Wikipedia on unblocked users with a blocked collaborator, LGATT.



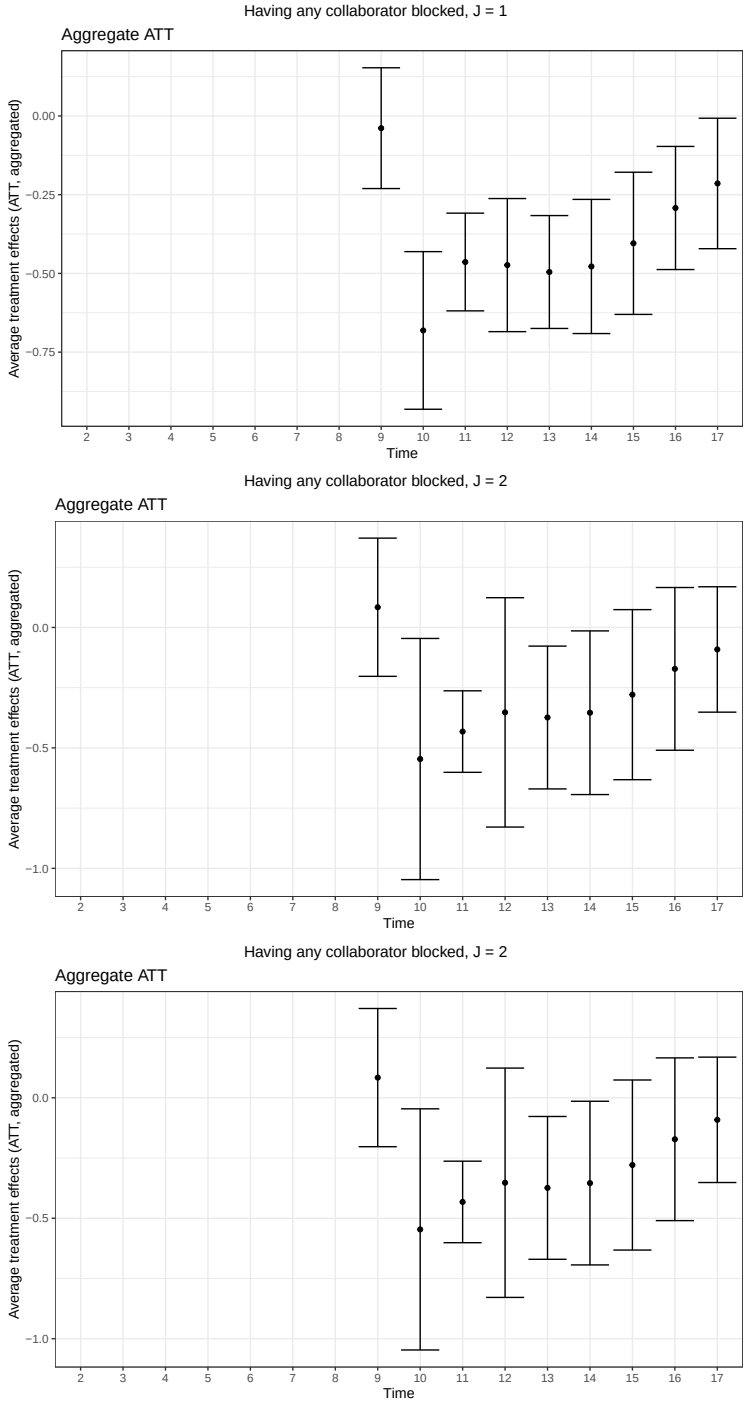
Notes: Plots on the left column present average weekly contribution per cohort and latent group. Latent-group specific average treatment effects on treated (LGATT) are shown on the right. Cohorts and latent groups are represented by line types (dotted lines for contributors with no collaborator being blocked) and colors respectively. 95% confidence intervals for the LGATT are generated from nonparametric bootstraps with 200 draws.

Figure 8: The block of Wikipedia on unblocked users with social participation, ATT.



Notes: ATT estimates from the canonical two-way fixed effects estimators are reported in the top figure. The bottom figure presents the ATT estimates computed from the weighted averages of LGATTs with two latent groups. 95% confidence intervals for all the ATTs are generated from nonparametric bootstraps with 200 draws. The treated group is defined as the contributors who have participated in any social activities.

Figure 9: The block of Wikipedia on unblocked users with a blocked collaborator, ATT.



Notes: ATT estimates from the canonical two-way fixed effects estimators are reported in the top figure. The bottom figure presents the ATT estimates computed from the weighted averages of LGATTs with two latent groups. 95% confidence intervals for all the ATTs are generated from nonparametric bootstraps with 200 draws. The treated group is defined as the contributors who have any collaborator being blocked.

8.2 Proof of Proposition 2

For notational brevity, let $X_{it} := \Delta Y_{it+1}$ and write $f^j_{\Delta Y_{t+1}|\Delta Y_t,G}(\Delta y_{t+1}|\Delta y_t, c)$ as $f^j_{X_t|X_{t-1},G}(x_t|x_{t-1}, c)$ for $c = 0$ and g. Under Assumption 2, $(\Delta Y_2, ..., \Delta Y_{g-1}) = (\Delta Y_2(0), ..., \Delta Y_{g-1}(0))$ regardless of the value of G_i, and therefore $(\Delta Y_2, ..., \Delta Y_{g-1})$ does not depend on G_i. Consequently, we have $f^j_{X_1|G}(x_1|c) = f^j_{X_1}(x_1)$ and $f^j_{X_t|X_{t-1},G}(x_t|x_{t-1}, c)$ as $f^j_{X_t|X_{t-1}}(x_t|x_{t-1})$ for $t = 1, ..., g-2$. Then, (11) with $\boldsymbol{W} = (X_1, X_2, ..., X_{T-1}, G)$ is written as

$$f_{X_1,...,X_{T-1},G}(x_1, ..., x_{T-1}, c) = \sum_{j=1}^{J} \pi^j p^j(c) f^j_{X_1}(x_1) \prod_{t=2}^{g-2} f^j_{X_t|X_{t-1}}(x_t|x_{t-1}) \prod_{t=g-1}^{T-1} f^j_{X_t|X_{t-1},G}(x_t|x_{t-1}, c) \tag{30}$$

We establish the identification when $g \geq 6$ so that the first five periods are pre-treatment periods. We first prove the identification of the latent type probabilities and conditional density functions up to the first five periods. By integrating out $X_{g-1}, ..., X_{T-1}$ in (30), we have

$$f_{X_1,X_2,X_3,X_4,G}(x_1, x_2, x_3, x_4, c) = \sum_{j=1}^{J} \pi^j p^j(c) f^j_{X_1}(x_1) \prod_{t=2}^{4} f^j_{X_t|X_{t-1}}(x_t|x_{t-1}). \tag{31}$$

Define

$$\boldsymbol{L}_{x_3} := \begin{bmatrix} 1 & \cdots & 1 \\ \lambda^1_4(b_1|x_3) & \cdots & \lambda^J_4(b_1|x_3) \\ \vdots & \ddots & \vdots \\ \lambda^1_4(b_{J-1}|x_3) & \cdots & \lambda^J_4(b_{J-1}|x_3) \end{bmatrix} \text{ and } \bar{\boldsymbol{L}}_{x_2,c} := \begin{bmatrix} \bar{\lambda}^1_2(a_1, x_2, c) & \cdots & \bar{\lambda}^J_2(a_1, x_2, c) \\ \vdots & \ddots & \cdots \\ \bar{\lambda}^1_2(a_J, x_2, c) & \cdots & \bar{\lambda}^J_2(a_J, x_2, c) \end{bmatrix}, \tag{32}$$

where $\bar{\lambda}^j_2(x_1, x_2, c) := \pi^j p^j(c) f^j_{X_1}(x_1) f^j_{X_2|X_1}(x_2|x_1)$ with $p^j(c) := \Pr(G_i = c|Z_i = j)$, $\lambda^j_3(x_3|x_2) := f^j_{X_3|X_2}(x_3|x_2)$, and $\lambda^j_4(x_4|x_2) := f^j_{X_4|X_3}(x_4|x_3)$.

Assumption 11. *There exists a value x^*_3 that satisfies the following condition: for every $x_2 \in \Delta\mathcal{Y}_3$, we can find $(\bar{x}_2, \check{x}_2, \bar{x}_3) \in \Delta\mathcal{Y}_3 \times \Delta\mathcal{Y}_3 \times \Delta\mathcal{Y}_4$, $(a_1, ..., a_J) \in (\Delta\mathcal{Y}_2)^J$ and $(b_1, ..., b_{J-1}) \in (\Delta\mathcal{Y}_5)^{J-1}$ such that (a) $\boldsymbol{L}_{x^*_3}$, $\boldsymbol{L}_{x_3}$, $\boldsymbol{L}_{\bar{x}_3}$, $\bar{\boldsymbol{L}}_{\check{x}_2}$, and $\bar{\boldsymbol{L}}_{\bar{x}_2}$ are non-singular, and (b) all the diagonal elements of D_{x_3,x_3} defined in (38) with $x^*_3 = x_3$ take distinct values. Furthermore, (c) for every $(x_3, x_4) \in \Delta\mathcal{Y}_4 \times \Delta\mathcal{Y}_5$, $f^j_{X_4|X_3}(x_4|x_3) > 0$ for $j = 1, ..., J$.*

For each value of $x_2 \in \Delta\mathcal{Y}_3$, choose $(\check{x}_2, \bar{x}_2, \bar{x}_3) \in \Delta\mathcal{Y}_3 \times \Delta\mathcal{Y}_3 \times \Delta\mathcal{Y}_4$, $(a_1, ..., a_J) \in \Delta\mathcal{Y}^J_2$, and $(b_1, ..., b_{J-1}) \in \Delta\mathcal{Y}^{J-1}_5$ that satisfy Assumption 11. Evaluating (30) at $(X_1, X_2, X_3, X_4, G) = (a, x_2, x_3, b, c)$ gives

$$f_{X_1,X_2,X_3,X_4,G}(a, x_2, x_3, b, c) = \sum_{z=1}^{J} \bar{\lambda}^j_2(a, x_2, c) \lambda^j_3(x_3|x_2) \lambda^j_4(b|x_3). \tag{33}$$

Similarly, evaluating $f_{X_1,X_2,X_3,G}(x_1, x_2, x_3, c) = \sum_{j=1}^{J} \pi^j p^j(c) f^j_{X_3|X_2}(x_3|x_2) f^j_{X_2|X_1}(x_2|x_1) f^j_{X_1}(x_1)$ at

$(x_1, x_2, x_3) = (a, x_2, x_3)$ gives

$$f_{X_1,X_2,X_3,G}(a, x_2, x_3, c) = \sum_{j=1}^{J} \bar{\lambda}_2^j(a, x_2, c)\lambda_3^j(x_3|x_2). \tag{34}$$

Denote $q_{x_2,x_3}(a, b, c) := f_{X_1,X_2,X_3,X_4,G}(a, x_2, x_3, b, c)$ and $\bar{q}_{x_2,x_3}(a, c) := f_{X_1,X_2,X_3,G}(a, x_2, x_3, c)$. Evaluating (33) at $a = a_1, ..., a_J$ and $b = b_1, ..., b_{J-1}$ gives $J(J-1)$ equations while evaluating (34) at $a = a_1, ..., a_J$ gives J equations.

Using matrix notation, we collect these $J(J-1) + J = J^2$ equations as

$$\boldsymbol{Q}_{x_2,x_3,c} = \boldsymbol{L}_{x_3}\boldsymbol{D}_{x_3|x_2}\bar{\boldsymbol{L}}_{x_2,c}^{\top}, \tag{35}$$

where $\boldsymbol{L}_{x_3}$ and $\bar{\boldsymbol{L}}_{x_2,c}$ are defined in (32) while

$$\boldsymbol{Q}_{x_2,x_3,c} := \begin{bmatrix} \bar{q}_{x_2,x_3}(a_1, c) & \bar{q}_{x_2,x_3}(a_2, c) & \cdots & \bar{q}_{x_2,x_3}(a_J, c) \\ q_{x_2,x_3}(a_1, b_1, c) & q_{x_2,x_3}(a_2, b_1, c) & \cdots & q_{x_2,x_3}(a_J, b_1, c) \\ \vdots & \vdots & \ddots & \vdots \\ q_{x_2,x_3}(a_1, b_{J-1}, c) & q_{x_2,x_3}(a_2, b_{J-1}, c) & \cdots & q_{x_2,x_3}(a_J, b_{J-1}, c) \end{bmatrix} \tag{36}$$

and $\boldsymbol{D}_{x_3|x_2} := \text{diag}\left(\lambda_3^1(x_3|x_2), ..., \lambda_3^J(x_3|x_2)\right)$. Let x_3^* be the value of x_3 as defined in Assumption Assumption 11. For each x_3, choose $\check{x}_2$, $\bar{x}_2$, and $\bar{x}_3$ that satisfy Assumption 11(a)(b). Evaluating (35) at four different points, $(\check{x}_2, x_3^*)$, $(\bar{x}_2, x_3)$, $(\check{x}_2, \bar{x}_3)$, and $(\bar{x}_2, \bar{x}_3)$ gives

$$\begin{aligned} \boldsymbol{Q}_{\check{x}_2,x_3,c} &= \boldsymbol{L}_{x_3}\boldsymbol{D}_{x_3|\check{x}_2}\bar{\boldsymbol{L}}_{\check{x}_2,c}^{\top}, \quad \boldsymbol{Q}_{\bar{x}_2,\bar{x}_3,c} = \boldsymbol{L}_{\bar{x}_3}\boldsymbol{D}_{\bar{x}_3|\bar{x}_2}\bar{\boldsymbol{L}}_{\bar{x}_2,c}^{\top}, \\ \boldsymbol{Q}_{\check{x}_2,\bar{x}_3,c} &= \boldsymbol{L}_{\bar{x}_3}\boldsymbol{D}_{\bar{x}_3|\check{x}_2}\bar{\boldsymbol{L}}_{\check{x}_2,c}^{\top}, \quad \boldsymbol{Q}_{\bar{x}_2,x_3^*} = \boldsymbol{L}_{x_3^*}\boldsymbol{D}_{x_3^*|\bar{x}_2}\bar{\boldsymbol{L}}_{\bar{x}_2,c}^{\top}. \end{aligned}$$

Then, following the identification argument in Carroll et al. (2010), under Assumption 11(a)(c), we have

$$A_{x_3^*,x_3} := \boldsymbol{Q}_{\check{x}_2,x_3}\boldsymbol{Q}_{\check{x}_2,\bar{x}_3}^{-1}\boldsymbol{Q}_{\bar{x}_2,\bar{x}_3}\boldsymbol{Q}_{\bar{x}_2,x_3^*}^{-1} = \boldsymbol{L}_{x_3}\boldsymbol{D}_{x_3^*,x_3}\boldsymbol{L}_{x_3^*}^{-1}, \tag{37}$$

where

$$\boldsymbol{D}_{x_3^*,x_3} := \boldsymbol{D}_{x_3|\check{x}_2}\boldsymbol{D}_{\bar{x}_3|\check{x}_2}^{-1}\boldsymbol{D}_{\bar{x}_3|\bar{x}_2}\boldsymbol{D}_{x_3^*|\bar{x}_2}^{-1}. \tag{38}$$

We first identify $\boldsymbol{L}_{x_3}$ for all $x_3 \in \Delta\mathcal{Y}_3$ up to an unknown permutation matrix. Evaluating (37) at $x_3^* = x_3$, we have

$$\boldsymbol{A}_{x_3,x_3}\boldsymbol{L}_{x_3} = \boldsymbol{L}_{x_3}\boldsymbol{D}_{x_3,x_3}.$$

Because $\boldsymbol{A}_{x_3,x_3}$ has J distinct eigenvalues under Assumption 11(b), the eignvalues of $\boldsymbol{A}_{x_3,x_3}$ determine the diagonal elements of $\boldsymbol{D}_{x_3,x_3}$ while the right eigenvectors of $\boldsymbol{A}_{x_3,x_3}$ determine the columns of $\boldsymbol{L}_{x_3}$ up to multiplicative constant and the ordering of its columns. Namely, collecting the right

eigenvectors of $\boldsymbol{A}_{x_3,x_3}$ into a matrix in descending order of their eigenvalues, we identify

$$\boldsymbol{B} := \boldsymbol{L}_{x_3}\Delta_{x_3}\boldsymbol{C},$$

where b satisfies $\boldsymbol{A}_{x_3,x_3}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{D}_{x_3,x_3}$, Δ_{x_3} is an unknown permutation matrix, and $\boldsymbol{C}$ is some diagonal matrix with non-zero diagonal elements.

We can determine the diagonal matrix $\boldsymbol{C}\boldsymbol{D}_{x_3,x_3}$ from the first row of $\boldsymbol{A}_{x_3,x_3}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{D}_{x_3,x_3} = \boldsymbol{L}_{x_3}\Delta_{x_3}\boldsymbol{C}\boldsymbol{D}_{x_3,x_3}$ because the first row of $\boldsymbol{L}_{x_3}\Delta_{x_3}$ is a vector of ones. Then, $\boldsymbol{L}_{x_3}\Delta_{x_3}$ is determined from $\boldsymbol{A}_{x_3,x_3}\boldsymbol{B}$ and $\boldsymbol{C}\boldsymbol{D}_{x_3,x_3}$ as $\boldsymbol{L}_{x_3}\Delta_{x_3} = \boldsymbol{A}_{x_3,x_3}\boldsymbol{B}(\boldsymbol{C}\boldsymbol{D}_{x_3,x_3})^{-1}$ in view of $\boldsymbol{A}_{x_3,x_3}\boldsymbol{B} = \boldsymbol{L}_{x_3}\Delta_{x_3}\boldsymbol{C}\boldsymbol{D}_{x_3,x_3}$. Repeating the above argument for all values of $x_3 \in \Delta\mathcal{Y}_4$, the eigenvalue decomposition algorithm identifies the matrices

$$\tilde{\boldsymbol{L}}_{x_3} := \boldsymbol{L}_{x_3}\Delta_{x_3} \quad \text{for all } x_3 \in \Delta\mathcal{Y}_4, \tag{39}$$

where Δ_{x_3} is an unknown permutation matrix that depends on x_3.

Next, we identify permutation matrices that re-arrange $\boldsymbol{L}_{x_3}\Delta_{x_3}$ in a common order of latent types across different values of x_3 using the identification argument in Higgins and Jochmans (2021). Pre- and post- multiplying (37) by $\tilde{\boldsymbol{L}}_{x_3}^{-1}$ and $\tilde{\boldsymbol{L}}_{x_3^*}$, respectively, we have

$$\tilde{\boldsymbol{D}}_{x_3^*,x_3} := \tilde{\boldsymbol{L}}_{x_3}^{-1}\boldsymbol{A}_{x_3^*,x_3}\tilde{\boldsymbol{L}}_{x_3^*} = \Delta_{x_3}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*} = \Delta_{x_3}^{-1}\Delta_{x_3^*}\left(\Delta_{x_3^*}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*}\right),$$

where the last equality uses the fact that $\Delta_{x_3^*}\Delta_{x_3^*}^{-1}$ is an identity matrix. Because $\Delta_{x_3}^{-1}\Delta_{x_3^*}$ is a permutation matrix, $\tilde{\boldsymbol{D}}_{x_3^*,x_3}$ is a matrix obtained by permutating the rows of the diagonal matrix $\Delta_{x_3^*}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*}$. Therefore, each diagonal element of $\Delta_{x_3^*}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*}$ is identified with the sum of elements in the corresponding column of $\tilde{\boldsymbol{D}}_{x_3^*,x_3}$, and the identification of $\Delta_{x_3^*}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*}$ follows. Then, we may identify $\Delta_{x_3}^{-1}\Delta_{x_3^*}$ as $\Delta_{x_3}^{-1}\Delta_{x_3^*} = \tilde{\boldsymbol{D}}_{x_3^*,x_3}\left(\Delta_{x_3^*}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*}\right)^{-1}$. Therefore, $\boldsymbol{L}_{x_3}$ is identified up to a common permutation matrix $\Delta_{x_3^*}$ that does not depend on x_3 from (39) as

$$\boldsymbol{L}_{x_3}^* := \boldsymbol{L}_{x_3}\Delta_{x_3^*} = \tilde{\boldsymbol{L}}_{x_3}\Delta_{x_3}^{-1}\Delta_{x_3^*} = \tilde{\boldsymbol{L}}_{x_3}\tilde{\boldsymbol{D}}_{x_3^*,x_3}\left(\Delta_{x_3^*}^{-1}\boldsymbol{D}_{x_3^*,x_3}\Delta_{x_3^*}\right)^{-1}. \tag{40}$$

In the next step, we identify $\{\pi^j, p^j(c), f_{X_1}^j(x_1), f_{X_2|X_1}^j(x_2|x_1), f_{X_3|X_2}^j(x_3|x_2), f_{X_4|X_3}^j(x_4|x_3)\}_{j=1}^J$ up to a permutation matrix $\Delta_{x_3^*}$. For this purpose, we evaluate $f_{X_3,X_4|X_2,G}(X_3, X_4|x_2, c) = \sum_{j=1}^J \pi^j p^j(c) f_{X_2,X_3,X_4}^j(x_2, x_3, x_4) / \sum_{k=1}^J \pi^k p^k(c) f_{X_2}^k(x_2)$ at $(x_2, x_3, x_4) = (x_2, x_3, b)$ as

$$\begin{aligned} f_{X_3,X_4|X_2}(x_3, b|x_2, c) &= \frac{\sum_{j=1}^J \pi^j p^j(c) f_{X_2}^j(x_2) f_{X_3|X_2}^j(x_3|x_2) f_{X_4|X_3}^j(b|x_3)}{\sum_{j=1}^J \pi^j p^j(c) f_{X_2}^j(x_2)} \\ &= \sum_{j=1}^J \tilde{\pi}_{x_2}^j(c) f_{X_3|X_2}^j(x_3|x_2) \underbrace{f_{X_4|X_3}^j(b|x_3)}_{=\lambda_4^j(b|x_3)}, \end{aligned} \tag{41}$$

where $\tilde{\pi}^j_{x_2}(c) := \pi^j p^j(c) f^j_{X_2}(x_2) / \sum_{k=1}^J \pi^k p^k(c) f^k_{X_2}(x_2)$. Then, evaluating (41) at $b = b_1, ..., b_{J-1}$ and collecting them into a vector together with $f_{X_3|X_2}(x_3|x_2, c) = \sum_{j=1}^J \tilde{\pi}^j_{x_2}(c) f^j_{X_3|X_2}(x_3|x_2)$ gives

$$\boldsymbol{r}_{x_3|x_2,c} = \boldsymbol{L}_{x_3} \boldsymbol{d}_{x_3|x_2,c} = \boldsymbol{L}^*_{x_3} \Delta^{-1}_{x_3^*} \boldsymbol{d}_{x_3|x_2,c} \tag{42}$$

with

$$\boldsymbol{r}_{x_3|x_2,c} = \begin{pmatrix} f_{X_3|X_2}(x_3|x_2, c) \\ f_{X_3,X_4|X_2}(x_3, b_1|x_2, c) \\ \vdots \\ f_{X_3,X_4|X_2}(x_3, b_{J-1}|x_2, c) \end{pmatrix} \quad \text{and} \quad \boldsymbol{d}_{x_3|x_2,c} = \begin{pmatrix} d^1_{x_3|x_2,c} \\ \vdots \\ d^J_{x_3|x_2,c} \end{pmatrix} := \begin{pmatrix} \tilde{\pi}^1_{x_2}(c) f^1_{X_3|X_2}(x_3|x_2) \\ \vdots \\ \tilde{\pi}^J_{x_2}(c) f^J_{X_3|X_2}(x_3|x_2) \end{pmatrix},$$

where the last equality in (42) follows from (40). Therefore, from (40) and (42), we identify $\tilde{\pi}^j_{x_2}(c) f^j_{X_3|X_2}(x_3|x_2)$ for all values of $(x_2, x_3, c) \in \Delta\mathcal{Y}_3 \times \Delta\mathcal{Y}_4 \times \mathcal{G}$ up to $\Delta_{x_3^*}$ as

$$\Delta^{-1}_{x_3^*} \boldsymbol{d}_{x_3|x_2,c} := \begin{pmatrix} d^{\alpha(1)}_{x_3|x_2,c} \\ \vdots \\ d^{\alpha(J)}_{x_3|x_2,c} \end{pmatrix} = \left(\boldsymbol{L}^*_{x_3}\right)^{-1} \boldsymbol{r}_{x_3|x_2,c}, \tag{43}$$

where

$$\alpha : \{1, 2, ..., J\} \to \{1, 2, ..., J\}$$

is a permutation implied by $\Delta^{-1}_{x_3^*}$. Furthermore, because $\tilde{\pi}^j_{x_2}(c) = \int_{\Delta\mathcal{Y}_4} \tilde{\pi}^j_{x_2} p^j(c) f^j_{X_3|X_2}(x_3|x_2) dx_3$ and $f^j_{X_3|X_2}(x_3|x_2) = [\tilde{\pi}^j_{x_2}(c) f^j_{X_3|X_2}(x_3|x_2)] / \tilde{\pi}^j_{x_2}(c)$, we may identify $f^{\alpha(j)}_{X_3|X_2}(x_3|x_2)$ from $d^{\alpha(j)}_{x_3|x_2,c}$ as

$$f^{\alpha(j)}_{X_3|X_2}(x_3|x_2) := \frac{d^{\alpha(j)}_{x_3|x_2,c}}{\int_{\Delta\mathcal{Y}_4} d^{\alpha(j)}_{x_3'|x_2,c} dx_3'}. \tag{44}$$

Then, we may identify $\boldsymbol{D}_{x_3|x_2}$ up to $\Delta_{x_3^*}$ as

$$\Delta^{-1}_{x_3^*} \boldsymbol{D}_{x_3|x_2} \Delta_{x_3^*} = \text{diag}\left(f^{\alpha(1)}_{X_3|X_2}(x_3|x_2), ..., f^{\alpha(J)}_{X_3|X_2}(x_3|x_2) \right), \tag{45}$$

and $\bar{\boldsymbol{L}}^\top_{x_2,c}$ is identified from (35), (40), and (45) up to $\Delta_{x_3^*}$ as

$$\Delta^{-1}_{x_3^*} \bar{\boldsymbol{L}}^\top_{x_2,c} = \left(\Delta^{-1}_{x_3^*} \boldsymbol{D}_{x_3|x_2} \Delta_{x_3^*}\right)^{-1} \left(\boldsymbol{L}^*_{x_3}\right)^{-1} \boldsymbol{Q}_{x_2,x_3,c}, \tag{46}$$

where the invertibility of $\boldsymbol{D}_{x_3|x_2}$ follows from Assumption 11(c).

Once $\boldsymbol{D}_{x_3|x_2}$ and $\bar{\boldsymbol{L}}_{x_2,c}$ are identified up to $\Delta_{x_3^*}$ as in (45)-(46), we determine $\boldsymbol{\ell}_{x_3}(x_4) := (\lambda^1_4(x_4|x_3), ..., \lambda^J_4(x_4|x_3)) = (f^1_{X_4|X_3}(x_4|x_3), ..., f^J_{X_4|X_3}(x_4|x_3))$ for any $(x_3, x_4) \in \Delta\mathcal{Y}_4 \times \Delta\mathcal{Y}_5$ up

to $\Delta_{x_3^*}$ by constructing

$$\boldsymbol{p}_{x_2,x_3,c}(x_4) := (q_{x_2,x_3,c}(a_1,x_4), q_{x_2,x_3,c}(a_2,x_4), ..., q_{x_2,x_3,c}(a_J,x_4))$$

from the observed data, and using the relationship

$$\boldsymbol{\ell}_{x_3}(x_4)\Delta_{x_3^*} = \left(f^{\alpha(1)}_{X_4|X_3}(x_4|x_3), ..., f^{\alpha(J)}_{X_4|X_3}(x_4|x_3)\right) = \boldsymbol{p}_{x_2,x_3,c}(x_4)(\Delta^{-1}_{x_3^*}\bar{\boldsymbol{L}}^{\top}_{x_2,c})^{-1}(\Delta^{-1}_{x_3^*}\boldsymbol{D}_{x_3|x_2}\Delta_{x_3^*})^{-1} \tag{47}$$

for all values of $(x_3, x_4) \in \Delta\mathcal{Y}_4 \times \Delta\mathcal{Y}_5$. Therefore, $\{f^j_{X_4|X_3}(x_4|x_3)\}_{j=1}^J$ is identified up to $\Delta_{x_3^*}$.

Similarly, we determine $\bar{\boldsymbol{\ell}}_{x_2,c}(x_1) := (\bar{\lambda}^1_2(x_1,x_2,c), ..., \bar{\lambda}^J_2(x_1,x_2,c))^{\top} = (\pi^1 p_1(c) f^1_{X_1}(x_1) f^1_{X_2|X_1}(x_2|a), ..., \pi^J p^j(c) f^J_{X_1}(x_1) f^J_{X_2|X_1}(x_2|a))^{\top}$ up to $\Delta_{x_3^*}$ for any $(x_1,x_2) \in \Delta\mathcal{Y}_2 \times \Delta\mathcal{Y}_3$ and for $c \in \{0, 6, ..., \bar{g}\}$ from (40) and (45) by constructing

$$\bar{\boldsymbol{p}}_{x_2,x_3,c}(x_1) := (\bar{q}_{x_2,x_3,c}(x_1), q_{x_2,x_3,c}(x_1,b_1), q_{x_2,x_3,c}(x_1,b_2), ..., q_{x_2,x_3,c}(x_1,b_{J-1}))$$

and using the relationship

$$\Delta^{-1}_{x_3^*}\bar{\boldsymbol{\ell}}_{x_2,c}(x_1) = \begin{pmatrix} \bar{\lambda}^{\alpha(1)}_2(x_1,x_2,c) \\ \vdots \\ \bar{\lambda}^{\alpha(J)}_2(x_1,x_2,c) \end{pmatrix} = (\Delta^{-1}_{x_3^*}\boldsymbol{D}_{x_3|x_2}\Delta_{x_3^*})^{-1}\left(\boldsymbol{L}^*_{x_3}\right)^{-1}\bar{\boldsymbol{p}}_{x_2,x_3,c}(x_1)^{\top}. \tag{48}$$

Then, $\{\pi^j p^j(c), f^j_{X_1}(x_1), f^j_{X_2|X_1}(x_2|x_1)\}_{j=1}^J$ is identified up to $\Delta_{x_3^*}$ from $\{\bar{\lambda}^{\alpha(j)}_2(x_1,x_2,c)\}_{j=1}^J$ in (48) given $\bar{\lambda}^j_2(x_1,x_2,c) = \pi^j p^j(c) f^j_{X_1}(x_1) f^j_{X_2}(x_2|x_1)$ as

$$\begin{aligned} &\pi^j p^j(c) := \int_{\Delta\mathcal{Y}_1}\int_{\Delta\mathcal{Y}_2} \bar{\lambda}^j_2(x_1,x_2,c)dx_2dx_1, \quad f^j_{X_1}(x_1) := \frac{\int_{\Delta\mathcal{Y}_2}\bar{\lambda}^j_2(x_1,x_2)dx_2}{\pi^j p^j(c)}, \\ &\text{and} \quad f^j_{X_2|X_1}(x_2|x_1) := \frac{\bar{\lambda}^j_2(x_1,x_2)}{\pi^j p^j(c) \times f^j_{X_1}(x_1)} \quad \text{for } j = 1, ..., J. \end{aligned} \tag{49}$$

Repeating the above argument using the subsample of $G_i = 0$ and g, we have $p^j(0) + p^j_g = 1$. Therefore, given the identification of $\pi^j p^j(c)$ for $c \in \{0, g\}$, we may identify π^j as $\pi^j = \pi^j p^j(0) + \pi^j p^j_g$ and the identification of $p^j(c)$ follows as $p^j(c) = (\pi^j p^j(c))/(\pi^j p^j(0) + \pi^j p^j_g)$ for $c \in \{0, g\}$.

Therefore, we identify $\{\pi^j, p^j(c), f^j_{X_1}(x_1), f^j_{X_2|X_1}(x_2|x_1), f^j_{X_3|X_2}(x_3|x_2), f^j_{X_4|X_3}(x_4|x_3)\}_{j=1}^J$ up to a permutation matrix $\Delta_{x_3^*}$. □

8.3 Proof of Proposition 3

In view of Proposition 1, it suffices to show that $\Pr(Z_i = j|G = g)$ and $\tau^j(\boldsymbol{W})$ are identified from $\{\pi^j, f^j_{\boldsymbol{W}}(\boldsymbol{w}) : j \in \mathcal{J}\}$. Using the Bayes' theorem, we may identify $\Pr(Z_i = j|G = g)$ as $\Pr(Z_i = j|G = g) = \frac{\pi^j p^j_g}{\sum_{k=1}^J \pi^k p^k(g)}$, where $p^j_g = \Pr(G_i = g|Z_i = j) = \int f^j_{\boldsymbol{W}}(\boldsymbol{y}', g)d\boldsymbol{y}'$ is identified

from $f^j_{\boldsymbol{W}}(\boldsymbol{w})$. The identification of $\tau^j(\cdot)$ immediately follows from (22) given that $f^j_{\boldsymbol{W}_2^{g-2}}(\boldsymbol{w}_2^{g-2})$ is identified from $f^j_{\boldsymbol{W}}(\boldsymbol{w})$. □

8.4 Proof of Proposition 4

The mean independence of η_{it} for $t \leq g-1$ in (14) follows from $\mathbb{E}[\epsilon_{it}(0)|G_i = g, Z_i = j] = \mathbb{E}[\epsilon_{it}(0)|G_i = 0, Z_i = j] = 0$ by Assumption 2.

To prove the mean independence of η_{it} for $t = g$, because $D_{i,g-1} = 0$ when $G_i = 0$ and $G_i = g$ while $D_{ig} = \mathbb{I}\{G_i = g\}$, it suffices to show that $\mathbb{E}[\eta_{ig}|G_i = 0, Z_i^* = j] = \mathbb{E}[\eta_{ig}|G_i = g, Z_i^* = j] = 0$. Note that

$$\mathbb{E}[\eta_{ig}|G_i = 0, Z_i^* = j] = \mathbb{E}[\Delta\epsilon_{ig}(0)|G_i = 0, Z_i^* = j] = 0$$

because $\mathbb{E}[\epsilon_{ig}(0)|G_i = 0, Z_i^* = j] = \mathbb{E}[\epsilon_{i,g-1}(0)|G_i = 0, Z_i^* = j] = 0$. Second,

$$\begin{aligned}
&\mathbb{E}[\eta_{ig}|G_i = g, Z_i^* = j] \\
&= \mathbb{E}[\epsilon_{ig}(g) - \epsilon_{i,g-1}(0)|G_i = g, Z_i^* = j] \\
&= \mathbb{E}[\epsilon_{ig}(g) - \epsilon_{ig}(0)|G_i = g, Z_i^* = j] + \mathbb{E}[\epsilon_{ig}(0) - \epsilon_{i,g-1}(0)|G_i = g, Z_i^* = j] \\
&= 0,
\end{aligned}$$

where the last equality holds because the definition of $\mu^j_{g,t}$ implies that $\mu^j_{g,t} = \mathbb{E}\left[Y_{it}(g) - Y_{it}(0) \mid G_i = g, Z_i = j\right] = \mathbb{E}\left[\mu^j_{g,t} + \epsilon_{it}(g) - \epsilon_{it}(0) \mid G_i = g, Z_i = j\right]$ for all $t \geq g$ so that $\mathbb{E}[\epsilon_{ig}(g) - \epsilon_{ig}(0)|G_i = g, Z_i^* = j] = 0$ while $\mathbb{E}[\epsilon_{ig}(0) - \epsilon_{i,g-1}(0)|G_i = g, Z_i^* = j] = \mathbb{E}[\epsilon_{ig}(0) - \epsilon_{i,g-1}(0)|G_i = 0, Z_i^* = j] = 0$ by the parallel trend assumptions and $\mathbb{E}[\epsilon_{ig}(0)|Z_i^* = j] = \mathbb{E}[\epsilon_{i,g-1}(0)|Z_i^* = j] = 0$. Therefore, η_{ig} is mean-independent of D_{ig} conditional on latent type.

For $t > g$, $\mathbb{E}[\eta_{ig}|G_i = 0, Z_i^* = j] = \mathbb{E}[\Delta\epsilon_{it}(0)|G_i = 0, Z_i^* = j] = 0$ while

$$\begin{aligned}
\mathbb{E}[\eta_{ig}|G_i = g, Z_i^* = j] &= \mathbb{E}[\Delta\epsilon_{it}(g)|G_i = g, Z_i^* = j] \\
&= \mathbb{E}[\epsilon_{it}(g) - \epsilon_{it}(0)|G_i = g, Z_i^* = j] - \mathbb{E}[\epsilon_{i,t-1}(g) - \epsilon_{i,t-1}(0)|G_i = g, Z_i^* = j] \\
&\quad + \mathbb{E}[\epsilon_{i,t}(0) - \epsilon_{i,t-1}(0)|G_i = g, Z_i^* = j] \\
&= 0,
\end{aligned}$$

where the last equality follows because $\mathbb{E}[\epsilon_{it}(g) - \epsilon_{it}(0)|G_i = g, Z_i^* = j] = \mathbb{E}[\epsilon_{i,t-1}(g) - \epsilon_{i,t-1}(0)|G_i = g, Z_i^* = j] = 0$ by the definition of $\mu^j_{g,t}$ and $\mu^j(g, t-1)$ while $\mathbb{E}[\epsilon_{i,t}(0) - \epsilon_{i,t-1}(0)|G_i = g, Z_i^* = j] = \mathbb{E}[\epsilon_{i,t}(0) - \epsilon_{i,t-1}(0)|G_i = 0, Z_i^* = j] = 0$ by the parallel trends assumption. Therefore, $\mathbb{E}[\eta_{ig}|G_i = 0, Z_i^* = j] = \mathbb{E}[\eta_{ig}|G_i = g, Z_i^* = j] = 0$. □

8.5 Proof of Proposition 5

Note that $\{\eta_{it}\}_{t=2}^T = \{\Delta\epsilon_{it}(0)\}_{t=2}^T$ when $G_i = 0$ while $\{\eta_{it}\}_{t=2}^T = \{\Delta\epsilon_{i2}(0), ..., \Delta\epsilon_{i,g-1}(0), \epsilon_{ig}(g) - \epsilon_{i,g-1}(0), \Delta\epsilon_{i,g+1}(g), ..., \Delta\epsilon_{iT}\}$ when $G_i = g$ for any $g \in \mathcal{G}\backslash\{0\}$. It follows from Assumption 5 that

η_{it} is independent of $\eta_{i,t-s}$ for any $s \geq 2$ conditional on $G_i = g$ for all $g \in \mathcal{G}$. Therefore, in view of (17), $\{\Delta Y_{it}\}_{t=2}^T$ follows a first-order Markov process conditional on $G_i = g$ for all $g \in \mathcal{G}$. □

8.6 Proof of Proposition 6

The stated result of part (a) follows from Theorem 2.1 of Newey and McFadden (1994), where the theorem's condition (i) follows from Lemma 2.2 of Newey and McFadden (1994) in view of Assumptions 8 and 9 (b)(d); the condition (ii) follows from Assumption 9(a); the condition (iii) follows from Lemma 2.4 of Newey and McFadden (1994) given Assumptions 4, 8, and 9(a). Similarly, part (b) follows from Theorem 2.1 of Newey and McFadden (1994), where the condition (i) holds because $\mathbb{E}\left[\sum_{t=g}^T \left(\Delta Y_{it} - \Delta\delta_t^j - \gamma_{g,t}^j D_{it}\right)^2 \middle| G_i \in \{0, g\}, Z_i = j\right]$ is uniquely minimized at $\boldsymbol{\theta}_g^{j,0}$, the condition (ii) follow from Assumption 9(a), and the condition (iii) follows from Lemma 2.4 of Newey and McFadden (1994).

Part (c) follows from Theorem 6.1 of Newey and McFadden (1994) by verifying conditions (i)-(v) of their Theorem 3.4. Conditions (i) and (ii) hold by Assumption 9(a) and Assumption 8, respectively. For conditions (iii) and (iv), $\mathbb{E}[m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)] = 0$ hold as argued above while $\mathbb{E}[||m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)||]$ and $\mathbb{E}[\sup_{\boldsymbol{\theta}_g} ||\nabla_{\boldsymbol{\theta}_g} m_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g^0, \boldsymbol{\psi}^0)||]$ are finite because both ΔY_{it} and D_{it} have finite variance given Assumption 8 and D_{it} being a binary random variable. Condition (v) follows from Assumption 3. □

8.7 Proof of Proposition 8

Denote $\tilde{m}_{g,i}(\boldsymbol{\theta}_g, \boldsymbol{\psi}) := \tilde{m}(\boldsymbol{W}_{g,i}^T; G_i; \boldsymbol{\theta}_g, \boldsymbol{\psi})$ so that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{m}_{g,i}(\tilde{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) = 0 \tag{50}$$

from (23). Then, the mean value expansion of (50) gives

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}}) - \mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}})]\right) + \sqrt{n}\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}})] + \left(\frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{\theta}_g^\top} \tilde{m}_{g,i}(\bar{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}})\right) \sqrt{n}(\tilde{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*), \tag{51}$$

where $\bar{\boldsymbol{\theta}}_g$ lies between $\tilde{\boldsymbol{\theta}}_g$ and $\boldsymbol{\theta}_g^*$.

Let

$$\nu_n(\boldsymbol{\psi}) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}) - \mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi})]\right). \tag{52}$$

By Lemma 2, $\{\nu_n(\cdot) : n \geq 1\}$ is stochastically equicontinuous and, given $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0$ in the Euclidean norm, we have $\nu_n(\hat{\boldsymbol{\psi}}) - \nu_n(\boldsymbol{\psi}^0) \xrightarrow{p} 0$ (c.f., page 2265 of Andrews, 1994). Therefore, given

$\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)] = 0$, the first term in (51) is written as

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}}) - \mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}})]\right) &= \nu_n(\boldsymbol{\psi}^0) + (\nu_n(\hat{\boldsymbol{\psi}}) - \nu_n(\boldsymbol{\psi}^0)) \\ &= \nu_n(\boldsymbol{\psi}^0) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0) + o_p(1). \end{aligned} \tag{53}$$

For the second term in (51), noting that $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0$ and $\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)] = 0$ from (24), we may apply the mean value theorem to $\sqrt{n}\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}})]$ in conjunction with the continuous mapping theorem to obtain

$$\sqrt{n}\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \hat{\boldsymbol{\psi}})] = (\nabla_{\boldsymbol{\psi}^\top}\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)] + o_p(1))\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^0) \tag{54}$$

where $\mathbb{E}[\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi})]$ is continuously differentiable with respect to $\boldsymbol{\psi}$. It also follows from applying the mean value theorem to $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\boldsymbol{W}_{2,i}^{\bar{g}-2}; \hat{\boldsymbol{\psi}}) = 0$ from (21) with $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}^0$, the law of large numbers, the continuous mapping theorem, and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\boldsymbol{W}_{2,i}^{\bar{g}-2}; \boldsymbol{\psi}^0) = O_p(1)$ that

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^0) = (\boldsymbol{S} + o_p(1))^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\boldsymbol{W}_{2,i}^{\bar{g}-2}; \boldsymbol{\psi}^0) = \boldsymbol{S}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\boldsymbol{W}_{2,i}^{\bar{g}-2}; \boldsymbol{\psi}^0) + o_p(1). \tag{55}$$

Let $\tilde{\boldsymbol{M}}_{g,n}(\boldsymbol{\theta}_g, \boldsymbol{\psi}) := \frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_g^\top}\tilde{m}_{g,i}(\boldsymbol{\theta}_g, \boldsymbol{\psi})$ and $\tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g, \boldsymbol{\psi}) := \mathbb{E}[\nabla_{\boldsymbol{\theta}_g^\top}\tilde{m}_{g,i}(\boldsymbol{\theta}_g, \boldsymbol{\psi})]$. Then, for the third term in (51), we have

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_g^\top}\tilde{m}_{g,i}(\bar{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) = \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0) + (\tilde{\boldsymbol{M}}_{g,n}(\bar{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) - \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)) = \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0) + o_p(1), \tag{56}$$

where the last equality follows from

$$\begin{aligned}||\tilde{\boldsymbol{M}}_{g,n}(\bar{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) - \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)|| &\leq ||\tilde{\boldsymbol{M}}_{g,n}(\bar{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) - \tilde{\boldsymbol{M}}_g(\bar{\boldsymbol{\theta}}_g^*, \hat{\boldsymbol{\psi}})|| + ||\tilde{\boldsymbol{M}}_g(\bar{\boldsymbol{\theta}}_g^*, \hat{\boldsymbol{\psi}}) - \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)|| \\ &\leq \sup_{(\boldsymbol{\theta}_g, \boldsymbol{\psi}) \in \mathcal{N}((\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0), \epsilon)} ||\tilde{\boldsymbol{M}}_{g,n}(\boldsymbol{\theta}_g, \boldsymbol{\psi}) - \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g, \boldsymbol{\psi})|| + o_p(1) \\ &= o_p(1)\end{aligned}$$

where the second inequality follows from $(\bar{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\psi}}) \xrightarrow{p} (\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)$ and the continuous mapping theorem; the last equality follows from Lemma 3, where $\mathcal{N}((\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0), \epsilon)$ is an ϵ-neighborhood of $(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0)$ for some $\epsilon > 0$.

Thus, it follows from (51)-(56) that

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*) = \left(\tilde{\boldsymbol{M}}_{\boldsymbol{\theta}_g} + o_p(1)\right)^{-1} \times \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\tilde{m}_{g,i}(\boldsymbol{\theta}_g^*, \boldsymbol{\psi}^0) + \tilde{\boldsymbol{M}}_{\boldsymbol{\psi},g} \boldsymbol{S}^{-1} s(\boldsymbol{W}_{2,i}^{\bar{g}-2}; \boldsymbol{\psi}^0) \right) + o_p(1) \right\} \quad \text{for } g = \underline{g}, ..., \bar{g},$$

and the stated result follows from the multivariate Lindeberg-Levy Central Limit Theorem. □

8.8 Auxiliary Lemmas

Lemma 1. *The first order condition for the minimization problem in (18) implies the equation (10).*

Lemma 2. $\{\nu_n(\cdot) : n \geq 1\}$ *is stochastically equicontinuous.*

Proof. We verify the conditions A-C of Theorem 1 of Andrews (1994).

Write $\tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g, \boldsymbol{\psi})$ as

$$\tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g, \boldsymbol{\psi}) = \mathbb{I}\{G \in \{0, g\}\} \sum_{j=1}^{J} I_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) \sum_{t=g}^{T} \xi(\Delta Y_t, \boldsymbol{X}_t; \boldsymbol{\beta}_{g,t}^j),$$

where $\xi(\Delta Y_t, \boldsymbol{X}_t; \boldsymbol{\beta}_{g,t}^j) := \boldsymbol{X}_t(\Delta Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}_{g,t}^j)$, $\boldsymbol{X}_t := (1, D_t)^\top$, and $\boldsymbol{\beta}_{g,t}^j := (\Delta\delta_t^j, \gamma_{g,t}^j)^\top$. Notice that $\xi(\Delta Y_t, \boldsymbol{X}_t; \boldsymbol{\beta}_{g,t}^j)$ is Lipschitz in $\boldsymbol{\beta}_{g,t}^j$ because $||\xi(\cdot; \boldsymbol{\beta}_{g,t}^{j,1}) - \xi(\cdot; \boldsymbol{\beta}_{g,t}^{j,2})|| \leq 2||\boldsymbol{\beta}_{g,t}^{j,1} - \boldsymbol{\beta}_{g,t}^{j,2}||$ for all $\boldsymbol{\beta}_{g,t}^{j,1}, \boldsymbol{\beta}_{g,t}^{j,2} \in \Theta_{\boldsymbol{\beta}_{g,t}^j}$ given that $||\boldsymbol{X}_t \boldsymbol{X}_t^\top||_2 \leq 2$. Thus, a class of functions $\{\xi(\cdot; \boldsymbol{\beta}_{g,t}^j) : \boldsymbol{\beta}_{g,t}^j \in \Theta_{\boldsymbol{\beta}_{g,t}^j}\}$ satisfies the Pollard's entropy condition with the envelop given by $2 \vee \sup_{\boldsymbol{\beta}_{g,t}^j \in \Theta_{\boldsymbol{\beta}_{g,t}^j}} |\xi(\cdot; \boldsymbol{\beta}_{g,t}^j)|$ by Theorem 2 of Andrews (1994).

Note also that $I_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) = \prod_{k \neq j} \mathbb{I}\{h_g^{jk}(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) > 0\}$, where $h_g^{jk}(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) := \ln \tau_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) - \ln \tau_g^k(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi})$ for $j \neq k$. Given the definition of $\tau_g^j(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi})$ in (22), $h_g^{jk}(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi})$ is written as

$$h_g^{jk}(\boldsymbol{W}_2^{g-2}; \boldsymbol{\psi}) = a^{jk} + \sum_{t=2}^{g-2} b_t^{jk} \Delta Y_{it},$$

where a^{jk} depends on $\ln \pi^j / \pi^k$, $\{(\Delta\delta_t^j)^2, (\Delta\delta_t^k)^2\}_{t=2}^{g-2}$, and $\{\Delta\delta_t^j \Delta\delta_{t-1}^j\}_{t=3}^{g-2}$ while b_t^{jk} depends on $\{\Delta\delta_s^j, \Delta\delta_s^k\}_{s=t-1}^{t+1}$. Therefore, $\mathcal{H}_g^{jk} := \{h_g^{jk}(\cdot; \boldsymbol{\psi}) : \boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}\}$ is a finite-dimensional vector space of real functions on $\mathcal{W}_2^{g-2}$. It follows from II.Lemma 18 of Pollard (1984) that the class of sets of the form $\{h_g^{jk}(\cdot; \boldsymbol{\psi}) > 0\}$ for $h_g^{jk}(\cdot; \boldsymbol{\psi}) \in \mathcal{H}_g^{jk}$ has polynomial discrimination. Furthermore, by II.Lemma 15 of Pollard (1984), the class of sets of the form $\cap_{k \neq j}\{h_g^{jk}(\cdot; \boldsymbol{\psi}) > 0\}$ also has polynomial discrimination. Therefore, $\{I_g^j(\cdot; \boldsymbol{\psi}) : \boldsymbol{\psi} \in \Theta_{\boldsymbol{\psi}}\}$ is a class of indicator functions of VC sets and satisfies the Pollard's entropy condition with the envelop given by 1.

Then, because $\tilde{m}_g(\boldsymbol{W}_g^T, G; \boldsymbol{\theta}_g, \boldsymbol{\psi})$ is written as the sum and multiplications of classes of functions that satisfy the Pollard's entropy condition, it follows from Theorem 3 of Andrews (1994) that a

class of functions $\{\tilde{m}_g(\cdot,\cdot;\boldsymbol{\theta}_g,\boldsymbol{\psi}) : (\boldsymbol{\theta}_g,\boldsymbol{\psi}) \in \Theta_{\boldsymbol{\theta}_g} \times \Theta_{\boldsymbol{\psi}}\}$ satisfies the Pollard's entropy condition with the envelop $\sum_{j=1}^{J}\sum_{t=g}^{T} 2 \vee \sup_{\boldsymbol{\beta}^j_{g,t}\in\Theta_{\boldsymbol{\beta}^j_{g,t}}} |\xi(\cdot;\boldsymbol{\beta}^j_{g,t})|$. Therefore, the condition A of Theorem 1 of Andrews (1994) holds. Furthermore, the condition B holds because, by Assumption 9(a)(e), $\mathbb{E}\sup_{\boldsymbol{\beta}^j_{g,t}\in\Theta_{\boldsymbol{\beta}^j_{g,t}}} |\xi(\cdot;\boldsymbol{\beta}^j_{g,t})|^{2+\delta} < \infty$. The condition C holds by Assumption 4(a), and the stated result follows from Theorem 1 of Andrews (1994).

□

Lemma 3. *There exists* $\epsilon > 0$ *such that* $\sup_{(\boldsymbol{\theta}_g,\boldsymbol{\psi})\in\mathcal{N}((\boldsymbol{\theta}^*_g,\boldsymbol{\psi}^0),\epsilon)} ||\tilde{\boldsymbol{M}}_{g,n}(\boldsymbol{\theta}_g,\boldsymbol{\psi}) - \tilde{\boldsymbol{M}}_g(\boldsymbol{\theta}_g,\boldsymbol{\psi})|| = o_p(1)$.

Proof. We verify the condition for Lemma 2.4 of Newey and McFadden (1994). By Assumption 4(a), the data are i.i.d. while $\Theta_{\boldsymbol{\theta}_g}$ and $\Theta_{\boldsymbol{\psi}}$ are compact by Assumption 9(a). $\nabla_{\boldsymbol{\theta}_g^\top}\tilde{m}_g(\boldsymbol{W}_2^T, G;\boldsymbol{\theta}_g,\boldsymbol{\psi})$ is continuous at each $(\boldsymbol{\theta}_g,\boldsymbol{\psi})$ with probability one given that ΔY_t is continuously distributed. Finally, given $||I^j_g(\boldsymbol{W}_2^{g-2};\boldsymbol{\psi})|| \leq 1$, $||\nabla_{\boldsymbol{\theta}_g^\top}\tilde{m}_g(\boldsymbol{W}_2^T, G;\boldsymbol{\theta}_g,\boldsymbol{\psi})|| \leq \sum_{j=1}^{J}\sum_{t=g}^{T} ||\boldsymbol{X}_t\boldsymbol{X}_t^\top|| < \infty$ with $\boldsymbol{X}_t := (1, D_t)^\top$. Therefore, the condition for Lemma 2.4 of Newey and McFadden (1994) is satisfied, and the stated result follows from Lemma 2.4 of Newey and McFadden (1994). □

References

Abadie, A. (2005), "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 72, 1–19.

Andrews, D. W. K. (1994), "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, eds. Engle, R. and McFadden, D., Amsterdam: North-Holland, vol. 4, pp. 2247–2794.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021), "Synthetic Difference-in-Differences," *American Economic Review*, 111, 4088–4118.

Biasi, B. and Moser, P. (2021), "Effects of copyrights on science: Evidence from the WWII book republication program," *American Economic Journal: Microeconomics*, 13, 218–260.

Bonhomme, S. and Manresa, E. (2015), "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83, 1147–1184.

Callaway, B. and Sant'Anna, P. H. (2021), "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 225, 200–230, themed Issue: Treatment Effect 1.

Carroll, R. J., Chen, X., and Hu, Y. (2010), "Identification and Estimation of Nonlinear Models Using Two Samples with Nonclassical Measurement Errors," *Journal of Nonparametric Statistics*, 22, 379–399.

Currie, J., Kleven, H., and Zwiers, E. (2020), "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, 110, 42–48.

Gallant, A. R. and White, H. (1988), *A unified theory of estimation and inference for Nonlinear Dynamic models*, Basil Blackwell.

Goodman-Bacon, A. (2021), "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 225, 254–277.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 64, 605–654.

Higgins, A. and Jochmans, K. (2021), "Identification Of Mixtures Of Dynamic Discrete Choices," TSE Working Papers 21-1272, Toulouse School of Economics (TSE).

Hu, Y. and Shum, M. (2012), "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

Kasahara, H. and Shimotsu, K. (2009), "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175.

LaLonde, R. J. (1986), "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, 604–620.

Newey, W. K. and McFadden, D. L. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Amsterdam: North-Holland, vol. 4, pp. 2111–2245.

Pollard, D. (1984), *Convergence of Stochastic processes*, New York: Springer-Verlag.

Rambachan, A. and Roth, J. (2023), "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, rdad018.

Robins, J. M. (1986), "A New Approach To Causal Inference in Mortality Studies With a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 7, 1393–1512.

Roth, J. (2022), "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends," *American Economic Review: Insights*, 4, 305–22.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023), "What's trending in difference-in-differences? A synthesis of the recent econometrics literature," *Journal of Econometrics*.

Roth, J. and Sant'Anna, P. H. C. (2023), "When Is Parallel Trends Sensitive to Functional Form?" *Econometrica*, 91, 737–747.

Sant'Anna, P. H. and Zhao, J. (2020), "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 219, 101–122.

Sun, L. and Abraham, S. (2021), "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 225, 175–199.

Wooldridge, J. M. (2021), "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators," .

Zhang, X. and Zhu, F. (2011), "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia," *American Economic Review*, 101, 1601–1615.