Finitely Heterogeneous Treatment Effect in Event-study*

Myungkou Shin†

December 1, 2023

Abstract

Treatment effect estimation strategies in the event-study setup, namely panel data with variation in treatment timing, often use the parallel trend assumption that assumes mean independence of potential outcomes across different treatment timings. In this paper, we relax the parallel trend assumption by assuming a latent type variable and develop a *type-specific* parallel trend assumption. With a finite support assumption on the latent type variable, we show that an extremum classifier consistently estimates the type assignment. Based on the clasfification result, we propose a type-specific diff-in-diff estimator for the type-specific CATT. By estimating the CATT with regard to the latent type, we study heterogeneity in treatment effect, in addition to heterogeneity in baseline outcomes.

Keywords: event-study, difference-in-differences, panel data, heterogeneity, classification, K-means clustering

JEL classification codes: C13, C14, C23

*I am deeply grateful to Stéphane Bonhomme, Christian Hansen and Azeem Shaikh, who have provided me invaluable support and insight. I would also like to thank Manasi Deshpande, Ali Hortaçsu, Guillaume Pouliot, Max Tabord-Meehan, Alex Torgovitsky, Martin Weidner and the participants of the metrics advising group at the University of Chicago and the Nuffield Econometrics Seminar at the University of Oxford for their comments and inputs. I acknowledge the support from the European Research Council through the grant ERC-2018-CoG-819086-PANEDA. Any and all errors are my own.

†Department of Economics, University of Oxford. email: myungkou.shin@economics.ox.ac.uk

1 Introduction

The event-study design is an empirical methodology whose popularity among empirical researchers has risen tremendously over the time. The seminal works by Ball and Brown (1968) and Fama et al. (1969) started a huge literature in financial economics that utilizes the random timing of shocks in capital markets to build empirical evidence of asset pricing theory. The increase in the use of the event-study design was not confined to the field of financial economics. Empirical economists in fields ranging from labor economics to education and environmental economics soon realized the benefit of utilizing variations in treatment timing and the event-study design has become one of the most widely used tools for causal analysis in applied microeconomics.

Unlike financial economics where the event-study research design often does not rely on never-treated units, applied microeconomists usually use the event-study research design when there is a well-defined comparison group of never-treated units. Using the never-treated units as control units, applied microeconomists extend the difference-in-difference (diff-in-diff) approach and allow for dynamic treatment effects in the event-study design. The key identifying assumption of the diff-in-diff style event-study design is the parallel trend assumption: temporal differences of untreated potential outcomes are mean independent of treatment status/treatment timing. The parallel trend assumption is a concise and powerful assumption that identifies treatment effects on treated units, while allowing for unobserved unit-level heterogeneity in outcome level. However, when the unit-level heterogeneity goes beyond heterogeneity in outcome level, an estimator using the parallel trend assumption may be biased. To address this issue, numerous alternatives to the parallel trend assumption have been proposed: parametric model for unit-specific trend, cluster-specific time trend when observable clustering structure is given, interactive fixed-effect models, synthetic control, etc (see Allegretto et al. (2017) for the first two alternatives).

Our framework contributes to the literature of relaxing the parallel trend assumption and incorporating additional unit-level heterogeneity than that in level. However, instead of

allowing for more generalized and flexible heterogeneity as they do in interactive fixed-effect models and the synthetic control literature, we only focus on subclass of models where units are finitely heterogeneous. At the cost of imposing restriction on across-the-unit variation, we allow for more flexible patterns of across-the-time variation, compared to parametrized unit-specific trends. Also, by modelling the unit-level heterogeneity to vary discretely, we motivate construction and estimation of subpopulation treatment effect and explore treatment effect heterogeneity, which is not very straightforward in models with continuously heterogeneous units. Though restrictive, the finite heterogeneity assumption is not new to the economics literature, especially in the context of finite mixture models (see Bonhomme et al. (2019); Kasahara et al. (2015) among others).

Let us begin by introducing what the finite heterogeneity actually means in our framework. In the model, we assume a unit-level latent type variable to model unit-level heterogeneity that goes beyond heterogeneity in level. Given the latent type variable, we assume that the usual parallel trend assumption holds, but only for units with the same type: 'type-specific parallel trend.' In a simple two periods case, the type-specific parallel trend assumption can be written as follows: with some latent type variable k_i,

$$\mathbf{E}[Y_{i2}(\infty) - Y_{i1}(\infty)|k_i, D_{i2} = 1] = \mathbf{E}[Y_{i2}(\infty) - Y_{i1}(\infty)|k_i, D_{i2} = 0]. \tag{1}$$

$Y_{it}(\infty)$ denotes untreated potential outcome of unit i at time t and D_{i2} denotes if unit i is treated at time $t = 2$; all units are untreated at time $t = 1$.

The type-specific parallel trend assumption can be understood as an extension of a conditional parallel trend assumption, by replacing the observable pretreatment covariate X_i in the conditioning set of the conditional parallel trend assumption with the latent type variable k_i (see Abadie (2005); Sant'Anna and Zhao (2020); Callaway and Sant'Anna (2021) among others). The conditional parallel trend assumption is used when the observable covariates associated with the dynamics of the untreated potential outcomes are not balanced

across treatment timings. Following the same spirit, we use the latent type variable to model the unit-level unobserved heterogeneity that affects the dynamics of the untreated potential outcomes and is potentially not balanced across treatment timing.

The difference between the conditional parallel trend setup and our setup is that the type-specific parallel trend assumption in this paper uses a latent variable not observed by the econometrician; the types need to be estimated. For that end, we assume two additional assumptions. Firstly, we assume that the latent type variable k_i has a finite support; hence, the unit-level heterogeneity varies only finitely. Using the finite variation in the baseline parallel trend, we construct type-specific treatment effect parameters and explore treatment effect heterogeneity by treating the type structure as a strata. Secondly, we assume that the types are well separated in the domain of the pretreatment outcomes. Note that the separation assumption by itself does not guarantee that the units can be correctly classified into their type when given a finite number of pretreatment periods. However, when the number of pretreatment periods grows to infinity, the decomposition of the pretreatment outcomes into noise and signal, i.e. the type, becomes more accurate, giving us a consistent classification result.[1]

Given the type-specific parallel trend assumption with well-separated and finite types, we propose a two-step estimation procedure. In the first step, we use the pretreatment outcomes to classify units into the finite number of types. For classification, we use K-means clustering algorithm. In the main model of the paper, we assume (a dynamic version of) (1) and use the canonical K-means clustering. In Section 5, an extended model where the trend also depends on the observable information X_{it} is discussed. In that section, we assume X_{it} and k_i are linearly separable in the pretreatment outcome model and use a modified version of the K-means clustering algorithm as in Bonhomme and Manresa (2015). Given the classification result, the second step of the estimation procedure is to estimate the

[1]The asymptotic classification with large T and finite types is mostly closely related to the group fixed-effect model of Bonhomme and Manresa (2015) (also, see (Su et al., 2016; Wang and Su, 2021; Janys and Siflinger, 2024; Ando and Bai, 2016; Mugnier, 2022) among others).

conditional treatment effect on treated units, using the estimated types as given.[2] We call our estimator 'type-specific diff-in-diff' estimator.

To discuss asymptotic properties of the treatment effect estimators, we first show that the probability of first-step missclassification goes to zero when the number of pretreatment time periods grows at a polynomial rate of the number of units. Given that the number of pretreatment time periods grows sufficiently faster compared to the number of units, the type-specific diff-in-diff estimators are consistent and asymptotically normal under some regular assumptions. These asymptotic results are supported by Monte Carlo simulations.

To provide an empirical illustration of our method, we revisit Lutz (2011) that studies the effect of terminating school desegregation plans on racial segregation index at the school district level. Lutz (2011) uses the variation in the timing of the district court ruling that terminates court-mandated school desegregation plans and uses the first-differenced outcomes with time fixed-effects. To relax the universal parallel trend assumption used in Lutz (2011), we apply the type-specific parallel trend assumption and find interesting patterns between the pretreatment trend in school segregation index and the treatment effect of terminating school desegregation plans. Specifically, we find strong segregation effect from terminating school desegregation plans in school districts where segregation index was worsening even before the termination, whereas we do not find significant segregation effect in school districts where segregation index was rising slower.

This paper contributes to the large literature of panel data models where interactive fixed-effects models are used to control for unit heterogeneity across treatment timings: see Abadie et al. (2010); Arkhangelsky et al. (2021); Athey et al. (2021); Hsiao et al. (2012); Freyaldenhoven et al. (2019); Xu (2017); Chernozhukov et al. (2019); Callaway and Karami (2023); Janys and Siflinger (2024) among others. In most cases, the interactive fixed-effect

[2]This two-step property of the estimation procedure closely relates to the stratification exercise used in estimating subpopulation treatment effect (see Abadie et al. (2018)). The goal of the stratification (i.e. classification in this paper's terminology) is to find groups of units whose (estimated) counterfactual untreated outcomes are similar; in this paper, we use the finite type assumption with long pretreatment periods to justify the stratification/classification step.

models assume that the error term is mean zero conditioning on the unit-level factor and therefore nest the type-specific parallel trend assumption by treating the unit-level factor as the type variable. Our framework can be thought of as a special case of the interactive fixed-effect model with a finite support on the factor; Janys and Siflinger (2024) takes the same approach. Also, as discussed in Athey et al. (2021), one way to compare various estimation procedures suggested in the literature is to compare weights on untreated outcomes that the estimators use in constructing a counterfactual untreated outcome. The type-specific diff-in-diff estimator in this paper applies uniform weights to the untreated observations within the given type to construct a counterfactual outcome. In that sense, the set of weights we consider in this paper is larger than that of the canonical diff-in-diff, but smaller than that of, e.g., synthetic diff-in-diff from Arkhangelsky et al. (2021).

As with the type-specific diff-in-diff estimator, most of the papers in this event-study and interactive fixed-effect model literature rely on large pretreatment periods as well. Notable exceptions are Callaway and Karami (2023); Freyaldenhoven et al. (2019). Callaway and Karami (2023) do not use a long pretreatment periods by using control variables with time-invariant coefficients in the outcome model as instruments. Freyaldenhoven et al. (2019) also do not require a long pretreatment by using external variables to control for the unit-by-time unobserved heterogeneity. The need for this extra information is the cost of using small pretreatment periods.

Outside of the literature that uses the interactive fixed-effect model, Rambachan and Roth (2022) suggests an alternative framework to relax the parallel trend assumption and have partial identification result.

This paper also closely relates to the rapidly growing literature on heterogeneous treatment effect: see De Chaisemartin and d'Haultfoeuille (2020); Sun and Abraham (2021); Callaway and Sant (2021); Goodman-Bacon (2021); Borusyak et al. (2021); Baker et al. (2022) among others. Callaway and Sant'Anna (2021) is particularly close to this paper in the sense that they also consider a conditional parallel trend assumption. This literature discusses the negative

weighting problem that arises in the standard TWFE specification when there is treatment effect heterogeneity across units and provides treatment effect estimators that are robust to this problem. We build upon this literature and construct the type-specific diff-in-diff estimator to be robust to the treatment effect heterogeneity. While doing so, we introduce a new element to the literature: documenting the unobserved treatment effect heterogeneity. Most of the existing literature assume a parallel trend type assumption with observable information. Thus, though their models allow for arbitrarily heteregeneous treatment effect, the treatment effect parameters discussed in the literature are functions of treatment timing or other observable covariates. In this paper, we document treatment effect heterogeneity in a way that goes further than the observable information, using the latent type variable.

The rest if the paper is organized as follows. In Section 2, we formally discuss the type-specific parallel trend assumption. In Section 3, we propose the two-step estimation procedure for treatment effect estimation. In Section 4-5, we discuss the asymptotic results on the estimator. In Section 6, we give some simulation results on the finite-sample performance of the estimator. In Section 7, we provide some empirical illustration of the type-specific diff-in-diff estimator by revisiting Lutz (2011).

2 Model

In the main model of the paper, we consider a setup where an econometrician observes a panel data with binary treatment: $\left\{\{Y_{it}, D_{it}\}_{t=-T_0-1}^{T_1-1}\right\}_{i=1}^{n}$. Y_{it} is the outcome variable for unit i at time t and $D_{it} \in \{0,1\}$ is the binary treatment variable for unit i at time t. D_{it} follows the staggered adoption scheme; $D_{it} \leq D_{it+1}$. $E_i = \min\{t : D_{it} = 1\}$ denotes the treatment timing of unit i. There are $n = N_0 + N_1$ units and $T+1 = T_0 + T_1 + 1$ time periods, with the unit index ranging $i = 1, \cdots, n$ and the time index ranging $t = -T_0 - 1, \cdots, T_1 - 1$. N_0 denotes the number of units that are never treated and N_1 denotes the number of units that are treated at some time $0 \leq t \leq T_1 - 1$. For never-treated units, let $E_i = \infty$.

WLOG let $\{1, \cdots, N_0\}$ be the set of the never-treated units. $T_0 + 1$ denotes the number of population pretreatment periods and T_1 denotes the number of population treatment periods; $\sum_{i=1}^{n} D_{it} = 0$ for all $t < 0$. $t < 0$ denotes pretreatment periods at the population level and $t \geq 0$ denotes population treatment periods at the population level. T_1 is fixed. Throughout the paper, we use the potential outcome setup to discuss treatment effect:

$$Y_{it} = Y_{it}(E_i).$$

$Y_{it}(e)$ is the potential outcome of unit i at time t when their treatment timing is e. Thus, for some $Y_{it}(e)$, $t < e$ means untreated potential outcome and $t \geq e$ means treated potential outcome.

The key assumption of this paper is that there exists a unit-level latent type variable. Conditional upon the latent type, the parallel trend assumption and the no anticipation assumption hold.

Assumption 1. (Type-Specific Parallel Trend) *There exists a latent type variable k_i such that for any t, s*

$$\mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, E_i\right] = \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i\right]$$

Assumption 2. (No Anticipation) *for any $t < e$*

$$\mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty) | k_i, E_i\right] = 0$$

Assumption 1-2 allow us to have causal interpretation. Fix two time periods (s, t) and a treatment timing e such that $s < e \leq t$. The conditional average treatment effect on treated units (CATT) for time t, type k and treatment timing e can be rewritten as follows:

$$\begin{aligned} CATT_t(k, e) &= \mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty) | k_i = k, E_i = e\right] \\ &= \mathbf{E}\left[Y_{it}(e) - Y_{is}(\infty) | k_i = k, E_i = e\right] - \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i = k, E_i = e\right] \\ &= \mathbf{E}\left[Y_{it}(e) - Y_{is}(e) | k_i = k, E_i = e\right] - \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i = k\right]. \end{aligned} \tag{2}$$

Suppose $\{k_i\}_{i=1}^n$ is observable to the econometrician. When $\Pr\{E_i > t | k_i = k\} > 0$, the second term $\mathbf{E}[Y_{it}(\infty) - Y_{is}(\infty) | k_i = k]$ is identified. When $\Pr\{E_i = e | k_i = k\} > 0$, the first term $\mathbf{E}[Y_{it}(e) - Y_{is}(e) | k_i = k, E_i = e]$ is identified. Thus, when $\{k_i\}_{i=1}^n$ is known and $\Pr\{E_i = e | k_i = k\} \cdot \Pr\{E_i > t | k_i = k\} > 0$, $CATT_t(k, e)$ is identified.

Note that the CATT parameter in (2) takes treatment timing E_i as a conditioning variable and focuses on a specific time period t. The full-fledgedness of $CATT_t(k, e)$ is useful when the researcher is interested in treatment effect heterogeneity across both time periods and types. Though both dimensions of the treatment effect heterogeneity may be of interest depending on contexts, we focus on an aggregated CATT parameter in this paper, to highlight the treatment effect heterogeneity across types. To construct the (aggregated) dynamic CATT parameter, we take the average of (2) across (t, e) while maintaining the relative treatment timing $t - e$ fixed: for some $r \geq 0$,

$$\beta_r(k) := \mathbf{E}\left[\mathbf{E}\left[Y_{i,e+r}(e) - Y_{i,e+r}(\infty) | k_i = k, E_i\right] | k_i = k, E_i \leq T_1 - r\right].$$

$\beta_r(k)$ is the r-times-lagged conditional average treatment effect on treated units. Note that $\beta_r(k)$ is dynamic and type-specific. A sufficient condition for identification of $\beta_r(k)$ when $\{k_i\}_{i=1}^n$ is observable to the econometrician is

$$\Pr\{E_i \leq T_1 - r | k_i = k\} \cdot \Pr\{E_i = \infty | k_i = k\} > 0.$$

Now that we have established identification results for CATT when the types are known, let us adopt two additional assumptions for classification.

Assumption 3. (Finite Support)

$$k_i \in \{1, \cdots, K\}.$$

The finiteness of the type k_i from Assumption 3 allows us to use the readily available literature of unsupervised partitioning methods to estimate the type. In particular, we use the

K-means minimization problem, which will be discussed in detail in Section 3. Then, we use the conventional K-means clustering algorithm to solve the K-means minimization problem and take the classification result as our 'estimated' types.

For the classification result to be consistent, we assume the following separation assumption.

Assumption 4. (Well-Separated Types) *whenever* $k \neq k'$,

$$\frac{1}{T_0} \sum_{t=-T_0}^{-1} \Big(\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k\right] - \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k'\right] \Big)^2 \to c(k,k') > 0$$

as $T_0 \to \infty$.

To discuss separation of types, Assumption 4 uses

$$\mathbf{E}[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k],$$

the conditional mean of the first-differenced never-treated potential outcomes. For any two different types, the l_2 norm of the difference between their conditional means is strictly nonzero. Thanks to the strong separation of types from Assumption 4, the K-means minimization consistently estimates the type when T_0 grows at a polynomial rate of n. Note that the separation assumption is in relation to time trends of the never-treated potential outcomes. From Assumptions 1-2, we have

$$\mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k\right] = \mathbf{E}\left[Y_{it}(e) - Y_{it-1}(e) | k_k = k, E_i = e\right]$$

whenever $t < e$. Thus, Assumption 4 can be applied not only to the never-treated units, but also to the pretreatment outcomes of the treated units.

The classification of n units into K types is a crucial part of the estimation procedure that the performance of the treatment effect estimators hugely depends on. Consider a very

simple case where $K = 2$ and model the untreated potential outcomes as follows: for $t \leq 0$,

$$Y_{it}(\infty) = \delta(k_i) + U_{it}, \qquad U_{it} \overset{iid}{\sim} \mathcal{N}(0,1).$$

WLOG let $\delta(1) < \delta(2)$. Find that $\bar{Y}_i(\infty) = \frac{1}{T_0+1}\sum_{t=-T_0-1}^{-1} Y_{it}(\infty) \sim \mathcal{N}\left(\delta(k_i), \frac{1}{T_0+1}\right)$. It is easy to see that for any fixed T_0,

$$\begin{aligned}\Pr\left\{\bar{Y}_i(\infty) \geq \bar{Y}_j(\infty) | k_i = 1, k_j = 2\right\} &= \Pr\left\{\bar{U}_i - \bar{U}_j \geq \delta(2) - \delta(1) | k_i = 1, k_j = 2\right\} \\ &= \Phi\left(\sqrt{\frac{T_0+1}{2}}\Big(\delta(2) - \delta(1)\Big)\right)\end{aligned}$$

is nonzero. When T_0 is fixed, the probability of imperfect classification is nonzero. Thus, we need large pretreatment periods ($\Leftrightarrow T_0 \gg 0$), in addition to the strong separation ($\Leftrightarrow \delta(2) - \delta(1) > 0$). When we do not have both conditions satisfied and thus units are potentially misclassified, the treatment effect estimator suffer from bias. For example, when U_{i0} is serially correlated with $\{U_{it}\}_{t<0}$, the bias may not even be proportional to the misclassification probability:

$$\mathbf{E}\left[|U_{i0} - U_{j0}| \Big| \bar{U}_i - \bar{U}_j \geq \delta(2) - \delta(1), k_i = 1, k_j = 2\right] \gg \Phi\left(\sqrt{\frac{T_0+1}{2}}\Big(\delta(2) - \delta(1)\Big)\right).$$

Armstrong et al. (2022) discuss a similar type of bias in the context of weak factors in an interactive fixed-effect model and suggest an estimation method robust to the weak factors. we believe applying Armstrong et al. (2022) to the finite type structure of this paper and controlling the misclassification bias would be an interesting avenue for future research. On the other side, we may give up on the perfect classification and adopt a different set of assumptions to identify the type-specific CATT. Ahn and Kasahara (2023) uses the same type-specific parallel trend assumption and additionally assume that the first differences of the potential outcomes satisfy the Markov property. Using the Markov property, Ahn and Kasahara (2023) identify the type-specific CATT without relying on large T_0.

3 Estimation

The estimation procedure is two-step. The first step is to estimate the type using the K-mean minimization problem. The second step is to take the estimated type as given and estimate CATT. To describe the estimation procedure, let us adopt following notations:

$$\gamma := (k_1, \cdots, k_n) \in \Gamma,$$
$$\Gamma := \{1, \cdots, K\}^n,$$
$$\delta := \{\delta_t(k)\}_{t,k}$$

γ is a $n \times 1$ vector of a type assignment. Γ is a set of all possible type assignments; n units are assigned with K different types. δ is a vector of type-specific time trend given time t and type k:

$$\delta_t(k) = \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty) | k_i = k\right].$$

Note that whenever $t < e$, $\delta_t(k_i) = \mathbf{E}\left[Y_{it} - Y_{it-1} | k_i, E_i = e\right]$ from Assumptions 1-2.

In the first step of estimating types, we only use a subset of the given data; we use population pretreatment periods of all units. With the population pretreatment periods, we construct an objective function with mean squared error:

$$\widehat{Q}(\delta, \gamma) = \frac{1}{nT_0} \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} \left(Y_{it} - Y_{it-1} - \delta_t(k_i)\right)^2 \tag{3}$$

and the resulting first-step classifier is

$$\left(\hat{\delta}, \hat{\gamma}\right) = \arg\min_{(\delta,\gamma)\in\mathcal{D}\times\Gamma} \widehat{Q}\left(\delta, \gamma\right). \tag{4}$$

$\mathcal{D} = [-M, M]^{T_0}$ with some $M > 0$. The minimization problem in (3) is called K-mean minimization problem; the solution to the K-means minimization problem is a grouping structure with K groups, defined with K centeroids. In our minimization problem (3), the

centeroids are denoted with $\{\delta_t(1)\}_{t<0}, \cdots, \{\delta_t(K)\}_{t<0}$ and the grouping structure is denoted with $k_1, \cdots, k_n$.

The algorithm that we use to obtain (4) is a conventional K-means clustering algorithm. Given an initial type assignment $\gamma^{(0)} = \left(k_1^{(0)}, \cdots, k_n^{(0)}\right)$,

1. **(update δ)** Given the type assignment $\gamma^{(s)}$ from the s-th iteration, estimate $\hat{\delta}_t^{(s)}(k)$ by letting
$$\hat{\delta}_t^{(s)}(k) = \frac{\sum_{i=1}^n (Y_{it} - Y_{it-1})\, \mathbf{1}\{k_i^{(s)} = k\}}{\sum_{i=1}^n \mathbf{1}\{k_i^{(s)} = k\}}$$
whenever the denominator is not zero.

2. **(update γ)** Update $k_i^{(s)}$ for each i by letting $k_i^{(s+1)}$ be the solution to the following minimization problem: for $i = 1, \cdots, N$,
$$\min_{k \in \{1, \cdots, K\}} \sum_{t=-T_0}^{-1} \left(Y_{it} - Y_{it-1} - \hat{\delta}_t^{(s)}(k)\right)^2.$$

3. Repeat Step 1-2 until Step 2 does not update $\hat{\gamma}$, or some stopping criterion is met. For stopping criterion, one can set a maximum number of iteration or a minimum update in $\hat{\delta}^{(s)}$: set S and ε such that the iteration stops when
$$s \geq S \quad \text{or} \quad \left\| \hat{\delta}^{(s)} - \hat{\delta}^{(s-1)} \right\|_\infty \leq \varepsilon.$$

The iterative algorithm proposed here has two stages. In the first stage, the algorithm estimates δ by taking sample means. In the second stage, the algorithm reassigns a type for each unit, by finding the type that minimizes the distance between $\{Y_{it} - Y_{it-1}\}_t$ and $\{\delta_t(k)\}_t$. The algorithm quickly attains a local minimum of the minimization problem (3). In the application we used in Section 5, the algorithm mostly converged within 20 iterations.

Since the iterative algorithm does not conduct an exhaustive search, there is a possibility that it might not converge to a global minimum; the computational burden of the

exhaustive search is extremely heavy since the space for the type assignment has cardinality of n^K. Thus, it is recommended that a random initial type assignment be drawn multiple times and the associated local minima be compared. Another concern is the choice of K. The number of the types in the data generating process is treated as known to the econometrician, which is unlikely in the empirical context. As discussed in Bai and Ng (2002); Bonhomme and Manresa (2015); Janys and Siflinger (2024), an information criterion can be used to estimate the number of type K. More discussion on the choice of K is given in the Supplementary Appendix.

Given the first-step classification result, the type-specific diff-in-diff estimator for the full-fledged CATT parameter $CATT_t(k,e)$ can be constructed by taking sample means for each type. In the case of the dynamic CATT parameter $\beta_r(k)$, the type-specific diff-in-diff estimator would be

$$\hat{\beta}_r(k) = \sum_{e \leq T_1 - 1 - r} \frac{\hat{\mu}(k,e)}{\sum_{e' \leq T_1 - 1 - r} \hat{\mu}(k,e')} \cdot \widehat{CATT}_{e+r}(k,e)$$

where

$$\widehat{CATT}_t(k,e) = \sum_{i=1}^{n} (Y_{it} - Y_{i,e-1}) \left(\frac{\mathbf{1}\{\hat{k}_i = k, E_i = e\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = e\}} - \frac{\mathbf{1}\{\hat{k}_i = k, E_i = \infty\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = \infty\}} \right)$$

$$\hat{\mu}(k,e) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k, E_i = e\}.$$

$\widehat{CATT}_t(k,e)$ is the type-specific diff-in-diff estimator for $CATT_t(k,e)$ as defined in (2) and $\hat{\mu}(k,e)$ is the estimator for

$$\mu(k,e) := \Pr\{k_i = k, E_i = e\}.$$

Note that this is merely one way of constructing an estimator for the dynamic CATT parameter $\beta_r(k)$. As discussed in Callaway and Sant'Anna (2021), there is no straightforward

choice in picking which time differences to use in a diff-in-diff type approach. Though the estimator described above takes one period before the treatment timing to construct a time difference, other choices such as two periods before the treatment timing are equally valid as long as the parallel trend assumption holds for every time period.[3] Also, the estimator uses the never-treated units to use as control units. When there is no never treated units, the latest treatment cohort can take up the same role. In that case, the definition of the dynamic CATT $\beta_r(k)$ will be adjusted in a way that it does not include the latest treatment cohort anymore.

Similarly, we can extend $\hat{\beta}_r(k)$ for $r < -1$ and construct estimators for

$$\mathbf{E}\left[\mathbf{E}\left[Y_{i,e+r}(e) - Y_{i,e+r}(\infty)|k_i = k, E_i\right] k_i = k, E_i \leq T_1\right]$$

for some $r < -1$. From Assumption 2, $\mathbf{E}\left[Y_{i,e+r}(e) - Y_{i,e+r}(\infty)|k_i = k, E_i\right] = 0$ whenever $r < -1$. Thus, though Assumption 1 does not have a testable implication, we can use $\hat{\beta}_r(k)$ for $r < -1$ to test Assumption 2. This is equal to the widely used 'no pretreatment test' in the event-study literature.

Note that the classification result from (4) satisfy that

$$\hat{\delta}_t(k) = \frac{\sum_{i=1}^{n} (Y_{it} - Y_{it-1}) \mathbf{1}\{\hat{k}_i = k\}}{\sum_{i=1}^{n} \mathbf{1}\{\hat{k}_i = k\}}.$$

$\hat{\delta}_t(k)$ puts equal weights over $Y_{it} - Y_{it-1}$ for units with the same estimated type k. In light of this, we can compare the type-specific diff-in-diff estimator with existing methods in terms of the weights that it considers. Consider a simple cases where there is only one post-treatment period and only one treated unit: $T_1 = 1$ and $N_1 = 1$. $E_i = \infty$ for every $i \leq N_0$ and $E_n = 0$. We consider an arbitrary treatment effect estimator $\hat{\beta}$ which can be written as a weighted

[3] Roth and Sant'Anna (2023a) discusses efficiency of these diff-in-diff type estimates when the treatment timing is truly random. More discussion on this is given the Appendix.

sum of Y_{it}: $\hat{\beta} = \sum_{i,t} w_{it} Y_{it}$. In a simple diff-in-diff estimator using $t \in \{-1, 0\}$, the weight is

$$w_{it}^{did} = \mathbf{1}\{i = n, t = 0\} - \mathbf{1}\{i = n, t = -1\} - \frac{\mathbf{1}\{i \leq N_0, t = 0\}}{N_0} + \frac{\mathbf{1}\{i \leq N_0, t = -1\}}{N_0}.$$

In the case of the synthetic control (see Abadie et al. (2010, 2015),

$$w_{it}^{sc} = \mathbf{1}\{i = n, i = 0\} - \sum_{j=1}^{N_0} w_j^* \mathbf{1}\{i = j, t = 0\}$$

where $\{w_j^*\}_{j \leq N_0}$ are solution to the following minimization:

$$\min_w \sum_{t=-T_0-1}^{-1} \left(Y_{nt} - \sum_{i=1}^{N_0} w_i Y_{it} \right)^2.$$

subject to $\sum_{i=1}^{N_0} w_i = 1$ and $w_i \geq 0$. In the case of the type-specific diff-in-diff,

$$w_{it}^{tdid} = \mathbf{1}\{i = n, t = 0\} - \mathbf{1}\{i = n, t = -1\} - \sum_{j=1}^{N_0} w_j^{**} \Big(\mathbf{1}\{i = j, t = 0\} - \mathbf{1}\{i = j, t = -1\} \Big)$$

where $\{w_j^{**}\}_{j \leq N_0}$ are (a function of) the solution to the following minimization:

$$\min_w \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} \left((Y_{it} - Y_{it-1}) - \sum_j w_{ij} (Y_{jt} - Y_{jt-1}) \right)^2$$

subject to $w_{ij} = \mathbf{1}\{k_i = k_j\} \big/ \sum_l \mathbf{1}\{k_l = k_i\}$ for some $\{k_i\}_{i=1}^n \in \{1, \cdots, K\}^n$. Based on the optimized w_{ij}, we get $w_j^{**} = w_{nj} \big/ \sum_{j' \neq n} w_{nj'}$.

Compared to the diff-in-diff estimator, the type-specific diff-in-diff estimator admits more flexible cross-sectional weights by possibly using only a subset of never-treated units. Compared to the synthetic control estimator, the type-specific diff-in-diff estimator is less flexible in terms of the cross-sectional weights since it is dichotomous cross-sectionally; a never-treated unit gets a uniform weight if and only if it shares the same type with the treated unit. However, the synthetic control estimator assigns nonzero weights only to contem-

poraneous outcomes while the type-specific diff-in-diff estimator takes temporal difference. Lastly, though the weights are not as straightforward as with other methods discussed here, the synthetic diff-in-diff estimator from Arkhangelsky et al. (2021) also uses a weighted sum type estimator and assigns flexible weights both cross-sectionally and intertemporally.

4 Asymptotic Results

In this section, we discuss the asymptotic theory on the estimator proposed in Section 3. Firstly, to derive the classification result for the type estimator defined in (4), let us adopt following assumptions.

Assumption 5. *With some $M > 0$,*

a. *(iid across units)* $\left(\{Y_{it}(e)\}_{e,t}, E_i, k_i\right) \overset{iid}{\sim} F$.

b. *(finite moments) For every e, t and k,* $\mathbf{E}\left[Y_{it}(e)^4 \middle| k_i = k\right] \leq M$.

c. *(long pretreatment)* $T_0 \to \infty$ *as* $n \to \infty$.

d. *(no measure zero types) For all* $k \in \{1, \cdots, K\}$, $\Pr\{k_i = k\} > 0$

e. *(weakly dependent, thin-tailed errors) With some positive constant d_1 and a,*

$$\left\{Y_{it}(e) - Y_{it-1}(e) - \mathbf{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|k_i\right]\right\}_{t=-T_0}^{-1}$$

is strongly mixing with mixing coefficient $\alpha[t]$ such that $\alpha[t] \leq \exp(-at^{d_1})$ uniformly over e. Also, with some positive constant d_2 and b, $Y_{it}(e)$ satisfies the following tail probability: for any $y > 0$,

$$\Pr\left\{\left|Y_{it}(e) - \mathbf{E}\left[Y_{it}(\infty)|k_i\right]\right| \geq y\right\} \leq \exp\left(1 - (y/b)^{d_2}\right)$$

uniformly over e and $t < 0$.

Assumption 5-c assumes that the number of population pretreatment periods T_0 grows to infinity as n goes to infinity. Assumption 5-d assumes that each type realizes with positive probability. Assumption 5-e assumes that for $t < 0$, tail probability of $Y_{it}(e) - \mathbf{E}[Y_{it}(\infty)|k_i]$ goes to zero exponentially and the first difference of $Y_{it}(e) - \mathbf{E}[Y_{it}(\infty)|k_i]$ is weakly dependent in the sense that it is strongly mixing with mixing coefficient decreasing exponentially in t.

Theorem 1. *Let Assumptions 1-5 hold. Then, up to some permutation on* $\{1, \cdots, K\}$,

$$\Pr\left\{\sup_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} > 0\right\} = o(n{T_0}^{-\nu}) + o(1) \quad \forall \nu > 0$$

as $n \to \infty$.

Proof. Theorem 1 is nested in Theorem 2 by connecting Assumption 5 to Assumption 7. Assumption 5-b induces parts of Assumption 7-b,d concerning U_{it} and $\delta_t(k)$, by letting $U_{it} = Y_{it}(E_i) - \mathbf{E}[Y_{it}(\infty)|k_i]$. Assumption 5-e provides the weak dependence conditions for which the proof for Theorem 2 uses Assumption 7-g. □

Theorem 1 puts a bound on the misclassification probability; the rate on the bound is shared by other papers in the group fixed-effect literature (Bonhomme and Manresa, 2015; Mugnier, 2022). When the bound goes to zero as $n \to \infty$, an asymptotic distribution can be derived for the type-specific diff-in-diff estimator $\hat{\beta}_r(k)$.

For asymptotic results on the type-specific diff-in-diff estimator $\hat{\beta}_r(k)$, let us adopt the additional assumption below. Recall that $\mu(k, e) = \Pr\{k_i = k, E_i = e\}$.

Assumption 6. *For each* $k = 1, \cdots, K$, *there exists some* $\bar{r}_k \geq 0$ *such that*

$$\bar{r}_k = \max\left\{r \geq 0 : \mu(k, T_1 - 1 - \bar{r}_k) \cdot \mu(k, \infty) > 0\right\}.$$

For any t, s *and* e, $\mathrm{Var}\left(Y_{it}(e) - Y_{is}(e)|k_i, E_i\right) > 0$.

Assumption 6 assumes that each type has nonzero measure of never-treated units and finds

an upper bound $\bar{r}_k$ on how far the dynamic treatment effects can be estimated. Note that

$$\bar{r}_k \geq r \quad \Rightarrow \quad \mu(k,e) > 0 \text{ for some } e \text{ such that } e + r \leq T_1 - 1.$$

For every $0 \leq r \leq \bar{r}_k$, r-times-lagged treatment effect can be estimated for type k.

Corollary 1. *Let Assumptions 1-6 hold. There exists some $\nu^* > 0$ such that $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. Then, for any k and $r \leq \bar{r}_k$ with some permutation on $\{1, \cdots, K\}$,*

$$\sqrt{n}\left(\hat{\beta}_r(k) - \beta_r(k)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

with some $\sigma^2 > 0$, as $n \to \infty$.

Proof. Corollary 1 is nested in Corollary 3. □

Remark 1. The asymptotic variance has a consistent estimator, whose expression is given in the Supplementary Appendix, along with the proof of Corollary 1.

Remark 2. In formulating the dynamic CATT parameter $\beta_r(k)$, treatment timing distribution is used as weights. Similar asymptotic results as in Corollary 1 hold for many other choices of weights: e.g. uniform weights across treatment timing.

5 Extension

In this section, we extend our main model by adding observed covariates $X_{it} \in \mathbb{R}^p$ to the model. The control covariates X_{it} gives us an extra source of heterogeneity in outcomes across different units and different times. For the classification to be successful, we need to decompose the variation of the outcome variable into the variation from the control covariates X_{it} and the variation from the latent type k_i. For that end, assume the following linear model

for untreated potential outcome: for $t < 0$,

$$Y_{it} - Y_{it-1} = \delta_t(k_i) + X_{it}^{\intercal}\theta + U_{it}. \tag{5}$$

Note that the interpretation of $\delta_t(k)$ is changed. Within the linear model, $\delta_t(k)$ in (5) is not the conditional mean of first-differenced potential outcome anymore since there exists $X_{it}^{\intercal}\theta$. Thus, we call $\delta_t(k)$ the type-specific time fixed-effects.

Given the model (5), we can construct a similar objective function from before and solve the K-means minimization problem for classification:

$$(\theta, \delta, \gamma) = \arg\min_{\theta,\delta,\gamma} \frac{1}{nT_0} \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} \left(Y_{it} - Y_{it-1} - \delta_t(k_i) - X_{it}^{\intercal}\theta\right)^2$$

The K-means algorithm needs to be slightly adjusted to account for X_{it}: given an initial type assignment $\gamma^{(0)} = \left(k_1^{(0)}, \cdots, k_N^{(0)}\right)$,

1. **(update θ and δ)** Given the type assignment $\gamma^{(s)}$ from the s-th iteration, construct indicator variables for each time s and the assigned type k: $\mathbf{1}\{t = s, k_i^{(s)} = k\}$ for $s = -T_0 \cdots, -1$ and $k = 1, \cdots, K$. By running OLS regression of $Y_{it} - Y_{it-1}$ on X_{it} and the indicators, we update $\hat{\delta}_t^{(s)}(k)$ and $\hat{\theta}^{(s)}$.

2. **(update γ)** Update $k_i^{(s)}$ for each i by letting $k_i^{(s+1)}$ be the solution to the following minimization problem: for $i = 1, \cdots, N$,

$$\min_{k \in \{1,\cdots,K\}} \sum_{t=-T_0}^{-1} \left(Y_{it} - Y_{it-1} - \hat{\delta}_t^{(s)}(k) - X_{it}^{\intercal}\hat{\theta}^{(s)}\right)^2.$$

3. Repeat Step 1-2 until Step 2 does not update $\hat{\gamma}$, or some stopping criterion is met. For stopping criterion, one can set a maximum number of iteration or a minimum update

in $\hat{\beta}^{(s)}$ and $\hat{\delta}^{(s)}$: set S and ε such that the iteration stops when

$$s \geq S \quad \text{or} \quad \max\left\{ \left\| \hat{\theta}^{(s)} - \hat{\theta}^{(s-1)} \right\|_\infty, \left\| \hat{\delta}^{(s)} - \hat{\delta}^{(s-1)} \right\|_\infty \right\} \leq \varepsilon.$$

We use Assumption 7 to derive classification results on the type estimation.

Assumption 7. *With some $M, \tilde{M} > 0$,*

a. *(iid across units) $\left(\{X_{it}, U_{it}\}_{t<0}, E_i, k_i\right) \overset{iid}{\sim} F$.*

b. *(compact parameter space) For every t and k, $|\delta_t(k)| \leq M$. $\|\theta\|_2 \leq M$.*

c. *(well-separated types) Whenever $k \neq k'$,*

$$\frac{1}{T_0} \sum_{t=-T_0}^{-1} \left(\delta_t(k) - \delta_t(k')\right)^2 \to c(k, k') > 0$$

as $n \to \infty$.

d. *(strict exogeneity and finite moments)*

For every $t < 0$, $\mathbf{E}\left[U_{it} | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\right] = 0$ and $\mathbf{E}\left[{U_{it}}^4 | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\right] \leq M$.

For every $t, s < 0$, $\mathbf{E}\left[{X_{it}}^\intercal X_{is}\right] \leq M$. For any $\nu > 0$,

$$\Pr\left\{ \frac{1}{T_0} \sum_{t=-T_0}^{-1} \|X_{it}\|_2 \geq \tilde{M} \right\} = o\left({T_0}^{-\nu}\right)$$

as $n \to \infty$.

e. *(long pretreatment) $T_0 \to \infty$ as $n \to \infty$.*

f. *(no measure zero types) For all $k \in \{1, \cdots, K\}$, $\Pr\{k_i = k\} > 0$*

g. *(weakly dependent, thin-tailed errors) With some positive constant d_1 and a, $\{U_{it}\}_{t=-T_0}^{-1}$ is strongly mixing with mixing coefficient $\alpha[t]$ such that $\alpha[t] \leq \exp(-at^{d_1})$. Also, with*

some positive constant d_2 and b, U_{it} satisfies the following tail probability: for any $u > 0$,

$$\Pr\{|U_{it}| \geq u\} \leq \exp\left(1 - (u/b)^{d_2}\right)$$

uniformly over i and $t < 0$.

h. *(no multicollinearity) Given an arbitrary type assignment $\tilde{\gamma} = \left(\tilde{k}_1, \cdots, \tilde{k}_n\right)$, let $\bar{X}_{k \wedge \tilde{k}, t}$ denote the mean of X_{it} among units such that $k_i = k$ and $\tilde{k}_i = \tilde{k}$. Let $\rho_n(\tilde{\gamma})$ denote the minimum eigenvalue of the following matrix:*

$$\frac{1}{nT_0} \sum_{i=1}^{n} \sum_{t=-T_0}^{-1} \left(X_{it} - \bar{X}_{k_i \wedge \tilde{k}_i, t}\right) \left(X_{it} - \bar{X}_{k_i \wedge \tilde{k}_i, t}\right)^{\intercal}.$$

Then, $\min_{\tilde{\gamma} \in \Gamma} \rho_n(\tilde{\gamma}) \xrightarrow{p} \rho$ as $n \to \infty$.

Assumption 7-c replaces Assumption 4 in the context of (5). Assumption 7-c assumes that the residual unobserved heterogeneity across units after regressing out X_{it} has finite types and is well-separated in the l_2 norm. Assumption 7-d replaces Assumption 5-b and additionally assumes that for large enough $\tilde{M}$, the probability of $\frac{1}{T_0} \sum_{t=-T_0}^{-1} \|X_{it}\|_2$ being larger than $\tilde{M}$ goes to zero exponentially. Moreover, Assumption 7-d combined with (5) replaces the parallel trend assumption given in Assumptions 1-2, by imposing

$$\mathbf{E}\left[Y_{it} - Y_{it-1} - X_{it}^{\intercal}\theta | k_i, \{X_{is}\}_{s=-T_0-1}^{-1}\right] = \delta_t(k_i).$$

Assumption 7-g replaces Assumption 5-e. Assumption 7-h assumes that there is sufficient variation in X_{it} within each type. For example, for discrete X_{it}, Assumption 7-h is satisfied when $\{X_{it}\}_{t=-T_0}^{-1}$ has at least $K + 1$ values in its support.

Theorem 2. *Let Assumptions 3 and 7 hold. Then, up to some permutation on $\{1, \cdots, K\}$,*

$$\Pr\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} = o\left(nT_0^{-\nu}\right) + o(1) \quad \forall \nu > 0$$

as $n \to \infty$.

Proof. See Supplementary Appendix. □

Remark 3. When X_{it} is time-invariant, i.e. $X_{it} = X_i$, (5) and Assumption 7 can be understood as a linear special case of the conditional parallel trend assumption: for $t < 0$,

$$\mathbf{E}\left[Y_{it}(E_i) - Y_{it-1}(E_i)|k_i, X_i\right] = \delta_t(k_i) + X_i{}^{\intercal}\theta.$$

Remark 4. Instead of assuming a linear structure on the first difference as in (5), we can consider a linear model on the level of the outcome:

$$Y_{it} = \alpha_i + \sum_{s=-T_0}^{t} \delta_s(k_i) + X_{it}{}^{\intercal}\theta + U_{it},$$
$$Y_{it} - Y_{it-1} = \delta_t(k_i) + (X_{it} - X_{it-1})^{\intercal}\theta + U_{it} - U_{it-1}.$$

The same conditions from Assumption 7-d,g and an adjusted version of Assumption 7-h by replacing X_{it} with $X_{it} - X_{it-1}$ give us the same classification result as in Theorem 2.

5.1 Outcome model approach

Once $\{k_i\}_{i=1}^n$ is estimated, we may use the estimated types to estimate various models on post-treatment periods with type-specific parameters. An example of such an outcome model approach is to run a group fixed-effect model as below while plugging in $\hat{k}_i$ for k_i:

$$Y_{it} - Y_{it-1} = \delta_t(k_i) + X_{it}{}^{\intercal}\theta + \sum_{r \geq 0} \Big(\beta_{0r}(k_i) + X_{it}{}^{\intercal}\beta_{1r}(k_i)\Big)\mathbf{1}\{t = E_i + r\} + U_{it}. \tag{6}$$

Directly modelling the outcome model with the observable information X_{it} can be helpful when we are interested in treatment effect heterogeneity and we would like to impose some restrictions on the heterogeneity due to the structure of X_{it}. For example, when X_{it} is continuous and multidimensional, a linearity assumption as in (6) can be helpful in summarizing

how X_{it} interacts with the type k_i, in terms of the treatment effect.

Consider a generalized outcome model for post-treatment outcomes: for $t \geq 0$,

$$Y_{it} - Y_{it-1} = m(X_{it}, k_i; \xi) + U_{it}.^4$$

In the example (6), $\xi = \left(\{\delta_t(k)\}_{t\geq 0,k}, \{\beta_{0r}, \beta_{1r}\}_{r\geq 0,k}, \theta\right)$. Note that the dimension of ξ is fixed; the dimension is $3T_1K + p$ and T_1 and K are fixed. Let $\tilde{\xi}$ be the infeasible least-square estimator for ξ and $\hat{\xi}$ be the plug-in least-square estimator for ξ:

$$\tilde{\xi} = \arg\min_{\xi\in\Xi} \frac{1}{nT_1} \sum_{i=1}^{n} \sum_{t=0}^{T_1-1} \left(Y_{it} - Y_{it-1} - m(X_{it}, k_i; \xi)\right)^2,$$
$$\hat{\xi} = \arg\min_{\xi\in\Xi} \frac{1}{nT_1} \sum_{i=1}^{n} \sum_{t=0}^{T_1-1} \left(Y_{it} - Y_{it-1} - m(X_{it}, \hat{k}_i; \xi)\right)^2.$$

Assumption 8. *Ξ, the parameter space for ξ, is bounded: with some $M > 0$,*

$$\sup_{\xi\in\Xi} \|\xi\|_2 \leq M.$$

The true value ξ lies in the interior of Ξ. Also, the infeasible estimator $\tilde{\xi}$ satisfies that

$$\sqrt{n}\left(\tilde{\xi} - \xi\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$$

with some $\Sigma > 0$ as $n \to \infty$.

Corollary 2. *Let Assumptions 3 and 7-8 hold. There exists some $\nu^* > 0$ such that $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. Then, up to some permutation on $\{1, \cdots, K\}$,*

$$\sqrt{n}\left(\hat{\xi} - \xi\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$$

[4] Though the first-differenced outcome variables are used in the post-treatment outcome model for internal consistency with (5), we can also consider models with outcome variable in level. In that case, one could use unit fixed-effects or treatment-timing-by-type fixed-effects to address unit-level heterogeneity in level.

with $\Sigma > 0$ *from Assumption 8 as* $n \to \infty$.

Proof. The result is direct from finding that

$$\sqrt{n}\left\|\tilde{\xi} - \hat{\xi}\right\|_2 \leq 2\sqrt{n}M\mathbf{1}\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} = o_p(1)$$

since for any $\varepsilon > 0$,

$$\Pr\left\{2\sqrt{n}M\mathbf{1}\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} > \varepsilon\right\} \leq \Pr\left\{\sup_i \mathbf{1}\{\hat{k}_i \neq k_i\} > 0\right\} = o(1)$$

from $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. □

Note that Assumption 8 does not discuss whether the true parameter ξ has sensible causal interpretation as $CATT_t(k, e)$ or $\beta_r(k)$ from Section 3 did. In the example of (6), it is well known that the linear coefficients β_{0r} and β_{1r} may suffer from the bias that comes from the dependence structure in $\mathbf{1}\{t = E_i + r\}$, given treatment effect heterogeneity.[5] Thus, we consider an alternative approach in the next subsection.

5.2 Assignment model approach

Directly modelling the outcome model may be too restrictive in some empirical contexts where the treatment effect depends on the control covariates X_{it} and the type k_i in a more flexible way. A similar concern is addressed in Callaway and Sant'Anna (2021) where the authors consider a conditional parallel trend assumption where the conditioning set is the control covariates X_i. In Callaway and Sant'Anna (2021), authors impose restriction on the assignment model while not imposing any restriction on the treatment effect heterogeneity

[5] It has been discussed that treatment effect estimators from TWFE specification are biased under the parallel trend type assumption (see De Chaisemartin and d'Haultfoeuille (2020); Goodman-Bacon (2021); Borusyak et al. (2021); Sun and Abraham (2021) among others) and potentially distort hypothesis testing (see Baker et al. (2022)). Also, Goldsmith-Pinkham et al. (2022) show that even under a stronger assumption of random treatment, treatment effect estimators still suffer from contamination bias when dynamic treatment effect specification is used.

in terms of the control covariate X_i. In the same spirit, in this subsection, we consider an assignment model approach given a time-invariant control covariate X_i.

With some finite-dimensional parameter ξ, we use a parametric function π_e to model the conditional distribution of the treatment timing E_i given the control covariate X_i and the latent type k_i:[6]

$$\Pr\{E_i = e | k_i, X_i\} = \pi_e(X_i, k_i, \xi).$$

Let $\tilde{\xi}$ be the infeasible maximum likelihood estimator for ξ and $\hat{\xi}$ be the plug-in estimator for ξ:

$$\tilde{\xi} = \arg\min_{\xi \in \Xi} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{e=0}^{T_1-1} \mathbf{1}\{E_i = e\} \log \pi_e(X_i, k_i, \xi) + \mathbf{1}\{E_i = \infty\} \log \pi_\infty(X_i, k_i, \xi) \right),$$
$$\hat{\xi} = \arg\min_{\xi \in \Xi} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{e=0}^{T_1-1} \mathbf{1}\{E_i = e\} \log \pi_e(X_i, \hat{k}_i, \xi) + \mathbf{1}\{E_i = \infty\} \log \pi_\infty(X_i, \hat{k}_i, \xi) \right).$$

An example is an ordered logistic model:

$$\Pr\{E_i \leq e | k_i = k, X_i = x\} = \frac{\sum_{e'=0}^{e} \exp(x^\intercal \theta + \delta_{e'}(k))}{\sum_{e'=0}^{T_1-1} \exp(x^\intercal \theta + \delta_{e'}(k)) + \exp(x^\intercal \theta + \delta_\infty(k))}.$$

In this example, $\xi = (\theta, \delta_e(k))_{k,e}$ and its dimension is fixed: $T_1 K + p$. Using $\hat{\xi}$, the type-specific diff-in-diff estimators can be constructed as follows:

$$\hat{\beta}_r(k) = \sum_{e \leq T_1 - 1 - r} \frac{\hat{\mu}(k, e)}{\sum_{e' \leq T_1 - 1 - r} \hat{\mu}(k, e')} \cdot \widehat{CATT}_{e+r}(k, e)$$

[6] The conditional distribution of E_i given (k_i, X_i) captures how treatment timing depends on the type and the control covariate. However, it does not contain any information on the dependence between k_i and X_i. For that end, we could consider the conditional distribution of k_i given X_i. Given a new draw of X_i, we cannot know k_i; however, we can look at the (estimated) distribution of $k_i | X_i$. Moreover, based on the distribution, a prediction on treatment effect can also be made. A close parallel with the IV literature exists here; we cannot know if a newly drawn unit with covariate X_i is a complier or not, but we can identify the conditional probability of them being a complier given X_i.

where

$$\widehat{CATT}_t(k,e) = \frac{\sum_{i=1}^n (Y_{it} - Y_{i,e-1})\,\mathbf{1}\{\hat{k}_i = k, E_i = e\}}{\sum_{i=1}^n \mathbf{1}\{\hat{k}_i = k, E_i = e\}} - \frac{\sum_{i=1}^n (Y_{it} - Y_{i,e-1})\,\mathbf{1}\{\hat{k}_i = k, E_i = \infty\}\pi_e(X_i,k,\hat{\xi})/\pi_\infty(X_i,k,\hat{\xi})}{\sum_{i=1}^n \mathbf{1}\{\hat{k}_i = k, E_i = \infty\}\pi_e(X_i,k,\hat{\xi})/\pi_\infty(X_i,k,\hat{\xi})}$$

$$\hat{\mu}(k,e) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\hat{k}_i = k, E_i = e\}.$$

To discuss asymptotic properties of the type-specific diff-in-diff estimator, we adopt the following assumption:

Assumption 9. *With some constant $M > 0$,*

a. *(finite moments) For every e and $t \geq -1$, $\mathbf{E}\left[Y_{it}(e)^4 | k_i, X_i\right] \leq M$.*

b. *(type-specific parallel trend) For every $t, s \geq -1$ and e,*

$$\mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, X_i, E_i\right] = \mathbf{E}\left[Y_{it}(\infty) - Y_{is}(\infty) | k_i, X_i\right]$$

$$\mathbf{E}\left[Y_{it}(e) - Y_{it}(\infty) | k_i, X_i, E_i\right] = 0$$

c. *There exists some $\varepsilon^\pi > 0$ such that $\mu(k,e) > 0 \Rightarrow \Pr\left\{\varepsilon^\pi \leq \inf_{w\in\Xi} \pi_e(X_i,k,w)\right\} = 1$.*

d. *Fix some e and k such that $\mu(k,e) > 0$ and define a function $g : \Xi \to \mathbb{R}$ such that*

$$g(w; X_i) = \frac{\pi_e(X_i,k,w)}{\pi_\infty(X_i,k,w)}.$$

There is a small neighborhood B_ξ around ξ with regard to $\|\cdot\|_2$ such that

i. *g is almost surely twice continuously differentiable on B_ξ;*

ii. *$\frac{\partial}{\partial w} g(w)$ and $\frac{\partial^2}{\partial w \partial w^\intercal} g(w)$ are almost surely bounded by M with regard to $\|\cdot\|_2$ on B_ξ.*

With Assumption 9, we have the following corollary of Theorem 2.

Corollary 3. *Let Assumptions 3 and 6-9 hold by replacing X_{it} with X_i. There exists some $\nu^* > 0$ such that $n/{T_0}^{\nu^*} \to 0$ as $n \to \infty$. Then, up to some permutation on $\{1, \cdots, K\}$,*

$$\sqrt{n}\left(\hat{\xi} - \xi\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\right)$$

with $\Sigma > 0$ from Assumption 8 as $n \to \infty$. In addition, the infeasible estimator $\tilde{\xi}$ admits an asymptotic linear approximation as follows:

$$\sqrt{n}\left(\tilde{\xi} - \xi\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l^{\pi}(X_i, k_i, E_i) + o_p(1)$$

where $\mathbf{E}[l^{\pi}(X_i, k_i, E_i)] = 0$ and $\mathbf{E}\left[l^{\pi}(X_i, k_i, E_i) l^{\pi}(X_i, k_i, E_i)^{\intercal}\right] > 0$. Then, up to some permutation on $\{1, \cdots, K\}$,

$$\sqrt{n}\left(\hat{\beta}_r(k) - \beta_r(k)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

with some $\sigma^2 > 0$, as $n \to \infty$.

Proof. See Supplementary Appendix. □

6 Simulation

In this section, we present simulation results to discuss the finite-sample performance of the type-specific diff-in-diff estimator, compared to other existing estimators in the literature. In a simulated random sample, we set $n = N_0 + 1$ and $T = T_0 + 1$; there are only one treated unit and one post-treatment period.

For the data generating process, we use a simple two-types setup: for $t = -T_0 - 1, \cdots, 0$,

$$\begin{aligned} Y_{it} &= \alpha_i + 2\mathbf{1}\{i = n, t = 0\} - (t-1)\mathbf{1}\{k_i = 1\} + U_{it}, \\ U_{it} &= \rho U_{it-1} + V_{it}. \end{aligned}$$

The unit fixed-effect α_i and the error terms $U_{i,-T_0-1}, V_{it}$ are generated from the following

distribution:

$$\begin{aligned}\left(k_i, \alpha_i, U_{i,-T_0-1}, \{V_{it}\}_t\right) &\sim ind, \\ \alpha_i, U_{i,-T_0-1} \big| k_i &\sim \mathcal{N}(0,1), \\ V_{it} \big| \left(\alpha_i, U_{i,-T_0-1}, k_i\right) &\overset{iid}{\sim} \mathcal{N}(0, 1-\rho^2).\end{aligned}$$

For $i = 1, \cdots, N_0$, $\Pr\{k_i = 1\} = \Pr\{k_i = 2\} = \frac{1}{2}$ and for $i = n$, $\Pr\{k_i = 1\} = 1$; the treated unit is always type 1.[7]The treatment effect is 2 and there are type-specific time fixed-effects only for type 1.

There are a few observations to be made here. Firstly, a simple mean of contemporaneous untreated outcomes, i.e. $1/N_0 \sum_{i \leq N_0} Y_{i0}$, is biased by the same amount of 0.5, for every N_0 and T_0. Secondly, the unit fixed-effects α_i are independent of the type, thus a simple mean of contemporaneous untreated outcomes within type 1, i.e. $\frac{1}{\sum_{i \leq N_0} \mathbf{1}\{k_i=1\}} \sum_{i \leq N_0} Y_{i0} \mathbf{1}\{k_i = 1\}$, is unbiased; taking temporal differences helps only in terms of variance. In this sense, the data generating process is 'fair' to the synthetic control method; the set of weights considered by the synthetic control method includes weights that construct an unbiased couterfactual outcome. Lastly, the types are well separated. For every T_0, the l_2 norm distance between the type-specific time fixed-effects are one.

Table 1 contains the simulated bias and the simulated MSE for the diff-in-diff, the synthetic control, the synthetic diff-in-diff and the type-specific diff-in-diff estimators, when $rho = 0.1$. In addition, it contains some summary statistics for the finite-sample performance of the classification step. As expected, for small T_0, all of the four estimators perform badly since there is little pretreatment information that can be used in constructing the counterfactual outcome. However, for large T_0, both the synthetic diff-in-diff estimator and the type-specific diff-in-diff estimator perform well.

[7]The relative magnitude of signal and noise is taken from the empirical application; when estimated with $K = 10$, the empirical dataset we use in Section 7 give us two stable types separated by $\|\delta(1) - \delta(2)\|_2 = 1.66$. The standard deviation and autocorrelation coefficient of residuals are 1.34 and 0.10.

As for the classification, the type-specific diff-in-diff estimator attains near-perfect classification for relatively small $T_0 = 20$. A rather unintuitive result that the classifier performs better with larger N_0 comes from the fact that the classification has two components: the first is to estimate $\delta_t(k)$ and the second is to classify units based on $\hat{\delta}_t(k)$. Larger N_0 helps in terms of the first component. Thus, there is trade-off in classification accuracy with regard to the number of units n in finite sample.

7 Application

To see how the estimation method suggested in this paper can be applied to a real dataset, we revisit Lutz (2011). Since the Supreme Court ruling on Brown v. Board of Education of Topeka in 1954 that found state laws in US enabling racial segregation in public schools unconstitutional, various efforts have been made to desegregate public schools, including court-ordered desegregation plans. After several decades, another important Supreme Court case was made in 1991; the ruling on Board of Education of Oklahoma City v. Dowell in 1991 stated that school districts could terminate the court-ordered plans once it successfully removed the effects of the segregation. Since the second Supreme Court ruling, school districts started to file for termination of court-ordered desegregation plans, mostly in southern states.

Lutz (2011) used the variation in timing of the district court rulings on the desegregation plan to estimate the effect on racial composition and education outcomes in public schools. The paper uses annual data on mid- and large-sized school districts from 1987 to 2006, obtained from the Common Core of Data (CCD), which contains data on school districts from 1987 to 2006, and the School District Databook (SDDB) of the US census, which contains data on school districts in 1990 and in 2000. To document if a school district was under a court-ordered desegregation plan at the time of the Supreme Court ruling in 1991 and when and if the school district got the desegregation plan dismissed at the district

courts, Lutz (2011) collected data from various published and unpublished sources, including a survey by Rosell and Armor (1996) and the Harvard Civil Rights Project.

Though Lutz (2011) looks at several outcome variable, we focus on one outcome variable, the segregation index: the segregation index for school district i is

$$Y_i = \frac{1}{2} \sum_{j \in J_i} \left| \frac{b_j}{B_i} - \frac{w_j}{W_i} \right| \times 100,$$

$$b_j : \text{ \# of black students in school } j, \quad w_j : \text{ \# of white students in school } j$$

$$J_i : \text{ the set of school in school district } i,$$

$$B_i = \sum_{j \in J_i} b_j, \quad W_i = \sum_{j \in J_i} w_j,$$

The segregation index ranges from 0 to 100, with 100 being perfectly segregated schools and 0 being perfectly representative schools.[8]

We roughly followed the data cleaning process in the paper and chose the timespan of 1988-2007 to form a balanced panel of school districts that were under a court-ordered desegregation plan in 1988-2000, which gave us 64 school districts. The number of students enrolled in 1988 was used as weights across school districts. To estimate types of the school district, we focused on 42 school districts which were under a court-ordered desegregation plan for the entire duration of 1988-2007: never-treated units. By focusing on the never-treated units, we extended the number of time periods used in classification step to be 19: for $t = 1989, \cdots, 2007$,

$$Y_{it} - Y_{it-1} = \delta_t(k_i) + X_{it}{}^{\intercal}\theta + U_{it}.$$

The control covariates X_{it} contain a central city indicator variable, percentage of students who are white, percentage of students who are hispnanic, percentage of students with free/reduced-priced lunch. For the purpose of comparison, here we present the main empir-

[8] In Lutz (2011), the segregation index ranges from zero to one but we rescaled the index for more visibility.

ical specification of Lutz (2011):

$$Y_{it} - Y_{it-1} = \delta_{jt} + \sum_{r=-4}^{10} \beta_r \mathbf{1}\{t = E_i + r\} + X_i^\intercal \theta_t + U_{it} \tag{7}$$

Though two specifications look alike, there are some differences. Firstly, Lutz (2011) uses a time-invariant control covariate X_i, with time-varying coefficient θ_t.[9] In my main specification, we use time-varying control covariates X_{it}, with time-invariant coefficient θ. Adding the time-varying coefficient θ_t is nontrivial extension of the model we consider in this paper since it makes it much more difficult to decompose the variation of Y_{it} into the variation from the type and that from X_{it}. Secondly, Lutz (2011) uses time fixed-effects δ_{jt} based on census region, which assigns every school district into one of the four regions. In the terminology of the model used in this paper, Lutz (2011) took the census region as the true type assignment whereas we used the data to estimate the type assignment.

For the number of types K, we considered $K = 2, \cdots, 10$. The classification result were stable for $K = 8, 9, 10$ in giving us two stables types and outliers and the information criterion minimized at $K = 9$ (for more discussion, see Supplementary Appendix). Thus, we extrapolated the classification results at $K = 9$ to 22 treated units and focused on the two stable types. Table 2 contains within-type balancedenss test using control covariates from $t = 1988$. Within the two stable types, the control covariates are well-balanced across treatment status: treated v. never-treated. Thus, we apply the unweighted type-specific diff-in-diff estimator from Section 3.

Figure 1 contains the type-specific diff-in-diff estimates for the two types of school districts. From Figure 1, we see that treatment effect is bigger for type 1 and smaller for type 2; the termination of court-ordered desegregation plans exacerbated racial segregation more

[9] Instead of using time-varying control covariates as we do, Lutz (2011) used the same covariates but only from the first year they were observed: $X_i = X_{i,-T_0-1}$. Also, Lutz (2011) used four additional variables: number of students, squared number of students, cubed number of students and squared percentage of students with free/reduced-price lunch. These variables were dropped in our analysis due to collinearity issue.

severely for type 1. The pooled regression (7) from Lutz (2011) estimated the dynamic treatment effect to be around 4-5 at $r = 5$, depending on specifications, whereas averaging the type-specific diff-in-diff estimates across type 1 and type 2 from Figure 1 gives us estimate 4.41; the census region fixed-effects control the type fairly well. For reference on the magnitude, the mean of the segregation index was 34 and its standard deviation was around 17 in 1988.

So, estimates on treatment effect suggest that type 1 and type 2 are different; type 1 school districts are more responsive to the treatment. Are these types different in other regards? Table 3 shows us some summary statistics on the outcome variable and other control covariates for type 1 and type 2. The null hypothesis that the entire vector of mean differences between type 1 and type 2 is zero is rejected with a t-test at size 0.05, meaning that those estimated types are significantly different from each other in terms of the observable information X_{it}.

In addition, Figure 2 shows us that the types are different in terms of the type-specific fixed-effects $\delta_t(k)$ as well. Figure 2 contains the estimated type-specific time fixed-effects. Over the time, type 1 has seen a steeper increase in the segregation index while type 2 has seen a slower increase. This implies that the termination of desegregation plans had a bigger impact on type 1, where the segregation index was already rising faster. This observation presents future research questions: for example, why do the school districts that were getting more segregated also get affected more from the dismissal of the desegregation plan?

8 Conclusion

In this paper, we introduce a type-specific parallel trend assumption in a panel data model with a latent type. By assuming the latent type variable has a finite support and is well-separated in a long pretreatment time series, the K-means classifier estimates the true types consistently. Also, based on the estimated types, we estimate the type-specific treat-

ment effect. The type-specific diff-in-diff estimator is useful when we suspect heterogeneity in time trends across units and want to explore the associated treatment effect heterogeneity. By applying the estimation method to an empirical application, we find some interesting empirical results where the estimates on the type-specific treatment effects and those on the type-specific time fixed-effects tell a story: the effect of terminating court-mandated desegregation plans were bigger for school districts where the segregation index was worsening.

References

Abadie, Alberto, "Semiparametric difference-in-differences estimators," *The review of economic studies*, 2005, *72* (1), 1–19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, 2010, *105* (490), 493–505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Comparative politics and the synthetic control method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

Abadie, Alberto, Matthew M Chingos, and Martin R West, "Endogenous stratification in randomized experiments," *Review of Economics and Statistics*, 2018, *100* (4), 567–580.

Ahn, Young and Hiroyuki Kasahara, "Difference in Differences with Latent Group Structures," *working paper*, 2023.

Allegretto, Sylvia, Arindrajit Dube, Michael Reich, and Ben Zipperer, "Credible research designs for minimum wage studies: A response to Neumark, Salas, and Wascher," *ILR Review*, 2017, *70* (3), 559–592.

Ando, Tomohiro and Jushan Bai, "Panel data models with grouped factor structure under unknown group membership," *Journal of Applied Econometrics*, 2016, *31* (1), 163–191.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager, "Synthetic difference-in-differences," *American Economic Review*, 2021, *111* (12), 4088–4118.

Armstrong, Timothy B, Martin Weidner, and Andrei Zeleneev, "Robust estimation and inference in panels with interactive fixed effects," *arXiv preprint arXiv:2210.06639*, 2022.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi, "Matrix completion methods for causal panel data models," *Journal of the American Statistical Association*, 2021, *116* (536), 1716–1730.

Bai, Jushan and Serena Ng, "Determining the number of factors in approximate factor models," *Econometrica*, 2002, *70* (1), 191–221.

Baker, Andrew C, David F Larcker, and Charles CY Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Ball, Ray and Philip Brown, "An empirical evaluation of accounting income numbers," *Journal of accounting research*, 1968, pp. 159–178.

Bonhomme, Stéphane and Elena Manresa, "Grouped patterns of heterogeneity in panel data," *Econometrica*, 2015, *83* (3), 1147–1184.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa, "A distributional framework for matched employer employee data," *Econometrica*, 2019, *87* (3), 699–739.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2021.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Callaway, Brantly and Sonia Karami, "Treatment effects in interactive fixed effects models with a small number of time periods," *Journal of Econometrics*, 2023, *233* (1), 184–208.

Chernozhukov, Victor, Christian Hansen, Yuan Liao, and Yinchu Zhu, "Inference for Heterogeneous Effects using Low-Rank Estimation of Factor Slopes," 2019.

De Chaisemartin, Clément and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–96.

Ding, Peng and Fan Li, "A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment," *Political Analysis*, 2019, *27* (4), 605–615.

Fama, Eugene F, Lawrence Fisher, Michael Jensen, and Richard Roll, "The adjustment of stock prices to new information," *International economic review*, 1969, *10* (1).

Freyaldenhoven, Simon, Christian Hansen, and Jesse M Shapiro, "Pre-event trends in the panel event-study design," *American Economic Review*, 2019, *109* (9), 3307–38.

Ghanem, Dalia, Pedro HC Sant'Anna, and Kaspar Wüthrich, "Selection and parallel trends," *arXiv preprint arXiv:2203.09001*, 2022.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, "Contamination bias in linear regressions," Technical Report, National Bureau of Economic Research 2022.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Hsiao, Cheng, H Steve Ching, and Shui Ki Wan, "A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong Kong with mainland China," *Journal of Applied Econometrics*, 2012, *27* (5), 705–740.

Janys, Lena and Bettina Siflinger, "Mental health and abortions among young women: Time-varying unobserved heterogeneity, health behaviors, and risky decisions," *Journal of Econometrics*, 2024, *238* (1), 105580.

Kasahara, Hiroyuki, Paul Schrimpf, and Michio Suzuki, "Identification and estimation of production function with unobserved heterogeneity," *University of British Columbia mimeo*, 2015.

Lutz, Byron, "The end of court-ordered desegregation," *American Economic Journal: Economic Policy*, 2011, *3* (2), 130–68.

Mugnier, Martin, "Unobserved clusters of time-varying heterogeneity in nonlinear panel data models," Technical Report 2022.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," Technical Report, Working Paper 2022.

Roth, Jonathan and Pedro HC Sant'Anna, "Efficient estimation for staggered rollout designs," *Journal of Political Economy Microeconomics*, 2023.

Roth, Jonathan and Pedro HC Sant'Anna, "When is parallel trends sensitive to functional form?," *Econometrica*, 2023, *91* (2), 737–747.

Sant'Anna, Pedro HC and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 2020, *219* (1), 101–122.

Su, Liangjun, Zhentao Shi, and Peter CB Phillips, "Identifying latent structures in panel data," *Econometrica*, 2016, *84* (6), 2215–2264.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Wang, Wuyi and Liangjun Su, "Identifying latent group structures in nonlinear panels," *Journal of Econometrics*, 2021, *220* (2), 272–295.

Xu, Yiqing, "Generalized synthetic control method: Causal inference with interactive fixed effects models," *Political Analysis*, 2017, *25* (1), 57–76.

APPENDIX

A Parallel trend v. design-based approach

The type-specific parallel trend assumption used in this paper abstract away from an assignment model for the treatment timing and directly impose restrictions on the outcome model. Though the parallel trend type assumptions have their own advantages of being concise and straightforward, the parallel trend assumption hinges on the arbitrary choice of the temporal differences in level. For example, when a researcher is interested in estimating the treatment effect as a percentage change of the outcome variable, they may be motivated use a parallel trend assumption with logged outcome variables:

$$\mathbf{E}\left[\log Y_{it}(\infty) - \log Y_{is}(\infty)|k_i, E_i\right] = \mathbf{E}\left[\log Y_{it}(\infty) - \log Y_{is}(\infty)|k_i\right].$$

On the other hand, a design-based approach such as a unconfoundedness assumption would be free of this commitment to a functional form. When

$$\{Y_{it}(e)\}_{t,e} \perp\!\!\!\perp E_i|k_i, \tag{8}$$

a parallel trend assumption with any functional form would hold.[10] This comes at a cost of assuming conditional (distributional) independence, which is stronger than conditional mean independence as in a parallel trend type assumption. There are some benefits to the design-based approach in addition to being robust to the choice of functional form. Under the parallel trend type assumption, there was no clear choice in which temporal differences to use. However, when we assume random treatment timing or the unconfoundedness assumption, we have some theoretical guidance on this choice. Roth and Sant'Anna (2023a) assume

[10]This statement is only true for the strong unconfoundedness assumption as given in (8). Ding and Li (2019) discuss a simple two period case where no one is treated at $t = 1$ and show that the parallel trend assumption and the unconfoundedness assumption do not nest each other when the unconfoundedness assumption is applied sequentially: $Y_{i1}(\infty) = Y_{i1}(2)$ and $Y_{i2}(\infty) \perp\!\!\!\perp D_{i2}|(k_i, Y_{i1})$

random treatment timing and show that an efficient estimator could be found among diff-in-diff type estimators that uses different weights across different temporal differences. Given the same classification result from Theorem 1, the unconfoundedness assumption (8) can be used to find an efficient type-specific diff-in-diff estimator following the procedure of Roth and Sant'Anna (2023a), using the same argument from the proof for Corollary 2: the classification error is faster than $1/\sqrt{n}$.

When a researcher does choose to follow a design-based approach, the question of great interest would be how much more restrictions are imposed when assuming the unconfoundedness, compared to the conditional parallel trend. Roth and Sant'Anna (2023b); Ghanem et al. (2022) provide insights to this question. Roth and Sant'Anna (2023b) show that an equivalent condition for the parallel trend assumption to hold for any monotone transformation of the outcome variable is that the population is divided into two subgroups where the treatment is random for the first subgroup and the untreated potential outcome has time-invariant distribution for the second subgroup. In this sense, the unconfoundedness assumption (8) is indeed strictly stronger than the type-specific parallel trend assumption holding for every monotone transformation of the outcome. Ghanem et al. (2022) provide insight in understanding the cost of assuming an additional parallel trend type assumption incrementally. Given a functional form, Ghanem et al. (2022) provide necessary conditions and sufficient conditions for that specific parallel trend assumption in terms of restrictions on the assignment model.

B Tables and figures

(N_0, T_0)	DiD		SC		synthetic DiD		type-specific DiD			
	bias	MSE	bias	MSE	bias	MSE	bias	MSE	(1)	(2)
(50, 5)	0.56	2.00	0.47	1.93	0.55	1.87	0.57	2.03	0.598	1.000
(50, 10)	0.50	2.03	0.28	1.59	0.25	1.59	0.44	2.06	0.664	0.976
(50, 20)	0.49	1.98	0.19	1.52	0.09	1.43	0.01	1.82	0.940	0.106
(50, 30)	0.45	1.85	0.12	1.31	0.01	1.33	-0.06	1.69	0.980	0.005
(100, 5)	0.47	2.10	0.37	2.01	0.45	1.86	0.47	2.11	0.574	1.000
(100, 10)	0.57	2.13	0.30	1.69	0.28	1.52	0.49	2.09	0.662	0.943
(100, 20)	0.46	2.03	0.18	1.36	0.07	1.39	-0.04	1.82	0.989	0.013
(100, 30)	0.54	1.99	0.21	1.34	0.10	1.35	0.03	1.71	.	0.000

Table 1: simulation results, $\rho = 0.1$

The bias and the MSE are computed for the ATT estimator, with 1,000 randomly generated samples.

$$(1) : \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{k}_i = k_i \middle| \hat{k}_i \neq k_i \text{ for some } i\right]$$

$$(2) : \Pr\left\{\hat{k}_i \neq k_i \text{ for some } i\right\}$$

type 1	treated	untreated	Diff
Segregation index	25.56	28.04	-2.48
	(4.72)	(13.33)	(5.84)
1{central city}	0.80	0.83	-0.03
	(0.45)	(0.41)	(0.26)
% (white)	53.22	63.52	-10.30
	(16.24)	(19.67)	(10.83)
% (hispanic)	2.53	2.03	0.50
	(4.89)	(3.31)	(2.57)
% (free/reduced-price lunch)	40.19	34.05	6.14
	(15.91)	(11.64)	(8.56)
N	5	6	-
joint p-value			0.951

type 2	treated	untreated	Diff
Segregation index	39.53	33.54	5.99
	(13.24)	(20.51)	(5.12)
1{central city}	0.40	0.48	-0.08
	(0.51)	(0.51)	(0.16)
% (white)	49.89	55.08	-5.19
	(23.36)	(22.26)	(7.31)
% (hispanic)	16.03	10.54	5.49
	(17.29)	(15.82)	(5.34)
% (free/reduced-price lunch)	37.74	39.79	-2.05
	(13.61)	(17.29)	(4.76)
N	15	29	-
joint p-value			0.640

Table 2: Within-type balancedness test using $t = 1988$

The table reports the group means of the school district characteristics and their differences. The standard errors are computed at the school district level. The joint p-value is for the null hypothesis that the means of differences across group are all zeros.

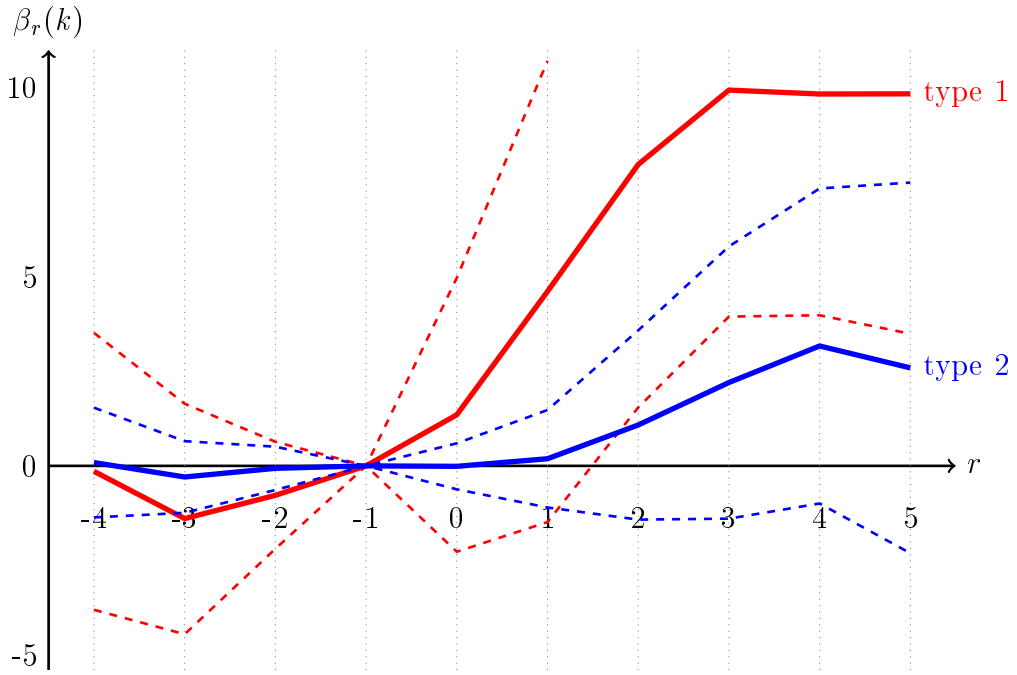

Figure 1: type-specific treatment effect

The graph reports the type-specific diff-in-diff estimates for the effect of terminating court-mandated desegregation plan on the segregation index of a school district.

The segregation index ranges from 0 to 100. In 1988, the average segregation index was 34 and the standard deviation was 17.

Type 1 is the type where the segregation index was rising faster and type 2 is the type where the segregation index was rising slower.

The dashed lines denote the confidence intervals are at 0.05 significance level and are computed with asymptotic standard errors.

	type 1	type 2	Diff
Segregation index	26.91	35.58	-8.67
	(9.97)	(18.42)	(4.09)
1{central city}	0.82	0.45	0.36
	(0.40)	(0.50)	(0.14)
% (white)	58.84	53.31	5.53
	(18.12)	(22.51)	(6.43)
% (hispanic)	2.26	12.41	-10.16
	(3.89)	(16.35)	(2.73)
% (free/reduced-price lunch)	36.84	39.09	-2.25
	(13.39)	(16.00)	(4.70)
N	11	44	-
joint p-value			0.021

Table 3: Across-the-type balancedness test using $t = 1988$

The table reports the group means of the school district characteristics and their differences.

The standard errors are computed at the school district level.

The joint p-value is for the null hypothesis that the means of differences across group are all zeros.

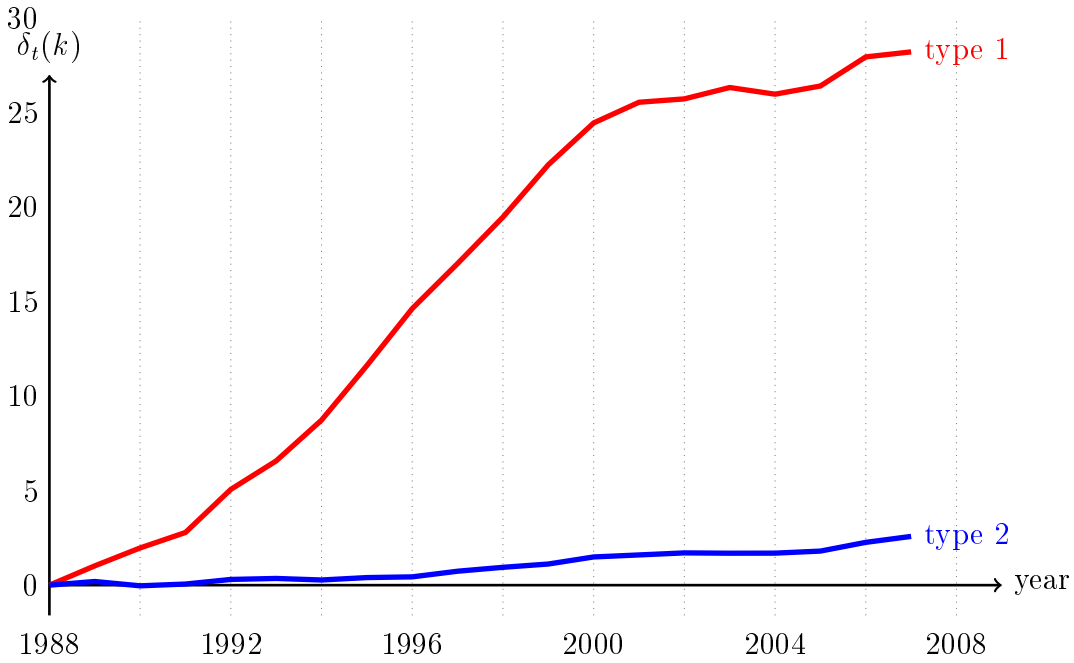

Figure 2: type-specific fixed-effects

The graph reports the type-specific time fixed-effects in the segregation index of a school district.

The trajectory is normalized by letting year 1988 to be 0.

The segregation index ranges from 0 to 100. In 1988, the average segregation index was 34 and the standard deviation was 17.

Type 1 is the type where the treatment effects of terminating court-mandated desegregation plan were significantly positive, meaning that the segregation got worse from the treatment, and type 2 is the type where the the treatment effects were insignificant.