Finitely Heterogeneous Treatment Effect in Event-study[*]

Myungkou Shin[†]

December 6, 2022

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 across different treatment timings. In this paper, I relax the parallel trend assumption by including a latent type variable and develop a *conditional* two-way fixed-effects model. With a finite support assumption on the latent type variable, I show that an extremum classifier consistently estimates the type assignment. Then I solve the endogeneity problem of the selection into treatment by conditioning on the latent type, through which the treatment timing is correlated with the outcome. I also allow treatment to affect units of different types differently and thus directly model and estimate type-level heterogeneity in treatment effect.

Keywords: event-study, difference-in-differences, panel data, heterogeneity,

classification, K-means clustering, unsupervised learning

JEL classification codes: C13, C14, C23

[†]Kenneth C. Griffin Department of Economics, University of Chicago. email: myungkoushin@uchicago.edu

1 Introduction

The event-study 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 empirical microeconomics. An example of empirical contexts where the event-study design is most frequently used is policy reforms. A policy reform is often gradually expanded within a country, instead of being adopted across the whole country at once. Thus, economists use the variation in the timing of policy adoption to estimate the effect of the policy reform; for example, Meghir and Palme (2005) use variation in the timing of education reform across municipalities and Havnes and Mogstad (2011) use variation in the timing of welfare benefit expansion across municipalities. Natural disasters and health shocks are also examples of empirical contexts where event-study design is often used. For example, Gallagher and Hartley (2017) use the timing of hurricane Katrina to study its effect on financial market outcomes of flooded households. Fadlon and Nielsen (2019) use the timing of heart attacks to study their effect on family members' health behaviors.

The most canonical estimation strategy in the event-study design is a difference-in-difference (diff-in-diff) estimator, used when there are two units and two time periods. When the dataset has more than two time periods, the diff-in-diff estimator is extended to the two-way fixed-effects (TWFE) regression specification. In this paper, I modify the TWFE regression specification to a *conditional* TWFE regression model with a latent unit-level type variable. In the model, outcome is modelled with unit fixed-effects, type-specific time fixed-effects, type-specific dynamic treatment effects, and observable control covariates with linear coefficients. The latent type variable denotes unit heterogeneity in treatment timing. Thus, the distribution of treatment timing may vary across different types. With the latent type variable, I make following assumptions. Firstly, I assume that conditioning on the type variable, treatment timing is independent of potential outcomes. Thus, the usual event-study estimation approaches for treatment effect, such as the TWFE regression specification, do not have the selection bias problem[1] when applied to units of the same type. Secondly, I assume that the type variable has a finite support and the types are well-separated in terms of pretreatment outcomes, motivated by Bonhomme and Manresa (2015). This allows us to find units of the same type, by looking at

---

[1] The term 'selection bias' can be used in different contexts but in this paper, what I refer to as the selection bias is the selection into treatment.

the time series of pretreatment outcomes.

The *conditional* TWFE model provides an econometrics framework that can be used when a researcher suspects that the treatment timing is not completely exogenous in the unconditional TWFE specification. Suppose that a researcher uses the unconditional TWFE specification to estimate treatment effect, even though the true model for the dataset is the *conditional* TWFE model. Then, the treatment effect estimator will be biased since the type-specific time fixed-effects, which are not controlled in the TWFE specification, have nonzero correlation with the treatment timing; hence the selection bias. Under the assumption that the latent type variable is recovered from the pretreatment outcomes, I solve the selection bias problem by comparing units with the same pretreatment outcomes. In this sense, the *conditional* TWFE model allows us to use the event-study type estimation approach even when we suspect that the treatment timing is not completely random, as long as we believe that the heterogeneity in treatment timing is recovered from the pretreatment outcomes. In addition, the *conditional* TWFE model allows us to explore unobserved heterogeneity in treatment effect. By assuming that the latent type variable is recovered from the pretreatment outcomes, the *conditional* TWFE model connects the unobserved heterogeneity in the pretreatment outcomes to the unobserved heterogeneity in treatment effect.

In estimation, I propose a least-square estimator to estimate the *conditional* TWFE model. Though the least-square estimator is not analytically solvable, I propose an iterative algorithm that finds a local minimum with little computational burden. The least-square estimator is consistent and asyptotically normal, as the number of units goes to infinity, under some regular assumptions. The key assumption is that (a polynomial function of) the number of pretreatment time periods goes to infinity faster than the number of units. I use this assumption to show that the probability of the least-square estimator misallocating the types is negligible in the asymptotic distribution.

To provide an empirical illustration of my method, I 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 TWFE specification. I apply the *conditional* TWFE estimator and find interesting patterns between the pretreatment trend in school segregation index and the treatment effect of terminating school desegregation plans. Specifically, I find strong segregation effect from terminating school desegregation plans in school districts where segregation index was worsening even before the termination, whereas I do not find significant segregation effect in school districts where segregation index was relatively stable over the time.

In Section 2, I formally discuss the *conditional* TWFE model. In Section 3, I propose a least-square estimator and an iterative algorithm to solve the optimization problem. In Section 4, I discuss asymptotic results on the estimator. In Section 5, I provide the empirical illustration by revisiting Lutz (2011).

1.1 Related Literature

In this paper, I make contribution to the group fixed-effects model literature. (Bonhomme and Manresa, 2015; Su et al., 2016; Wang and Su, 2021) The *conditional* TWFE model of this paper can be understood as a group fixed-effects model in the sense that the latent type variable induces a grouping structure and treatment effects and time fixed-effects of the model are assumed to be type-specific. This paper makes contribution to the group fixed-effects literature by developing a variant of group fixed-effects model for the event-study design and providing asymptotic theory for the model. The nontrivial part of the adaptation comes from the fact that one of the regressors in the model has a specific structure, i.e., the staggered adoption of the treatment.

Secondly, I make contribution to the event-study literature, by proposing a new econometric framework that relaxes the parallel trend assumption. The synthetic control method (Abadie et al., 2010, 2015; Arkhangelsky et al., 2021) relaxes the parallel trend assumption by using interactive fixed-effects modelled with a factor model, instead of using unit fixed-effects and time fixed-effects. Between the factor model and the *conditional* TWFE model, there is no clear order in terms of generality. At each time period, the interactive fixed-effects in the factor model are not restricted across units while the type-specific time fixed-effects in this paper are restricted in the sense that there can be only finite types. However, within each unit, across time periods, the interactive fixed-effects have linear structure while the type-specific time fixed-effects do not have any restriction. The same applies to the literature that directly estimates the factor model to estimate treatment effect: Xu (2017); Moon and Weidner (2015). In a slight different path, Rambachan and Roth (2022) considers another relaxation of the parallel trend assumption and develops a partial identification result.

Lastly, this paper contributes to the rapidly growing literature on heterogeneous treatment effect. The growing literature highlights the negative weighting problem of the TWFE specification under treatment effect heterogeneity (De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Borusyak et al., 2021). This literature mostly focuses on how to estimate the average treatment effect when there is unobserved heterogeneity in treatment effect while this paper aims to explicitly model the unobserved heterogeneity and estimates the heterogeneity in treatment effect. In Section 4, I discuss an extended *conditional* TWFE model where the treatment effect is allowed to be heterogeneous within each type and apply the solutions of the heterogeneous treatment effect literature to estimate the type-specific treatment effect, free of the negative weighting problem.

2 Model

Let us consider a potential outcome model with staggered adoption for a panel data with N units and $T + 1 = T_0 + T_1 + 1$ time periods: for $i = 1, \cdots, N$ and $t = -T_0 - 1, \cdots, 0, \cdots, T_1 - 1$,

$$Y_{it} = \sum_{r=-\infty}^{\infty} Y_{it}(r) \mathbf{1}_{\{t=E_i+r\}},$$

$$= Y_{it}(-1) + \sum_{r \neq -1, r=-\infty,}^{\infty} \left(Y_{it}(r) - Y_{it}(-1) \right) \mathbf{1}_{\{t=E_i+r\}}. \tag{1}$$

E_i denotes the treatment timing of unit i. $T_0 + 1$ is the number of periods where no unit is treated and T_1 the number of periods where some units are treated; at the aggregate level, $t < 0$ means pretreatment and $t \geq 0$ means treatment. The outcome Y_{it} is constructed with $Y_{it}(r)$, the potential outcome for unit i at time t when unit t is treated at $t - r$. We can think of positive r as lags and negative r as leads; at the individual level, $r < 0$ means pretreatment and $r \geq 0$ means treatment. For example, $Y_{it}(0)$ is the potential outcome for unit i at time t when unit i is treated at time t; unit i is treated at time t. $Y_{it}(-1)$ is the potential outcome for unit i at time t whe unit i is treated at time $t + 1$; unit i is untreated at time t. In this model, I assume that treatment timing E_i is observed for every unit i. The model can be easily extended to setup were some units are never treated by letting $E_i = T_1$ for never-treated units and assuming $Y_{it}(r) = Y_{it}(-1)$ for all $r < -1$.

In addition to Y_{it}, a control covariate X_{it} and the treatment timing E_i is observed; a researcher observes $\{Y_{it}, X_{it}, E_i\}$ for $i = 1, \cdots, N$ and $t = -T_0 - 1, \cdots, T_1 - 1$. Also, there exists a unit-level latent type variable. Conditional upon the latent type and the observable covariate, the treatment timing is independent of the potential outcomes, at every time t.

Assumption 1. (UNCONFOUNDEDNESS WITH THE LATENT TYPE) *There exists a latent type variable k_i such that for each t*

$$\{Y_{it}(r)\}_r \perp\!\!\!\perp E_i \mid k_i, \{X_{is}\}_{s=-T_0-1}^{t}$$

with some observable control covariate $X_{is} \in \mathbb{R}^p$.

Assumption 1 is the regular unconfoundedness assumption, but with a latent variable. There exists a notion of sequential exogeneity in Assumption 1; the potential outcomes at time t, $\{Y_{it}(r)\}_r$, are independent of the treatment timing E_i conditional upon the latent type and the information available at time t. However, the potential outcomes for time $t' > t$ can still be correlated with the treatment timing E_i if we are only conditioning on the information available at time t.

Proposition 1. *Under Assumption 1,*

$$\mathbf{E}\left[Y_{it}|E_i, k_i, \{X_{is}\}_{s=-T_0-1}^t\right] = \sum_e \mathbf{E}\left[Y_{it}(t-e)|k_i, \{X_{is}\}_{s=-T_0-1}^t\right] \cdot \mathbf{1}_{\{E_i=e\}}.$$

Proof. This is direct from Assumption 1. □

Proposition 1 is important since it allows us to compare oranges to oranges. To illustrate this, let us consider a simpler cross-sectional potential outcome model:

$$Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i),$$

$$\mathbf{E}[Y_i|D_i, X_i] = \mathbf{E}[Y_i(0)|D_i = 0, X_i] + \underbrace{\left(\mathbf{E}[Y_i(1)|D_i = 1, X_i] - \mathbf{E}[Y_i(0)|D_i = 0, X_i]\right)}_{=\beta(X_i)} \cdot D_i.$$

The intercept $\beta(X_i)$ has no causal interpretation; it is neither treatment effect on treated nor treatment effect on untreated. However, once we assume the unconfoundedness as in Assumption 1, we get $\beta(X_i) = \mathbf{E}[Y_i(1) - Y_i(0)|X_i]$. Thus, Assumption 1 guarantees that we have a causal interpretation.

Also, note that the residual defined as

$$U_{it} = Y_{it} - \sum_e \mathbf{E}\left[Y_{it}(t-e)|k_i, \{X_{is}\}_{s=-T_0-1}^t\right] \cdot \mathbf{1}_{\{E_i=e\}}$$

is mean independent of the treatment timing E_i, after conditioning on the latent type and all the available information at time t:

$$E\left[U_{it}|E_i, k_i, \{X_{is}\}_{s=-T_0-1}^t\right].$$

For the rest of the paper, let us impose more structures on the conditional expectation function.

Assumption 2. (Linear Conditional Expectation of Potential Outcome)

$$\mathbf{E}\left[Y_{it}(r)|k_i, \{X_{is}\}_{s=-T_0-1}^t\right] = \alpha_i + \delta_t(k_i) + \beta_r(k_i)\mathbf{1}_{\{r\neq-1\}} + X_{it}^\mathsf{T}\theta.$$

Assumption 2 imposes linearity and additive separability; roughly put, the expectation is a sum of treatment effect $\beta_r(k_i)\mathbf{1}_{\{r\neq-1\}}$ and untreated potential outcome $\alpha_i + \delta_t(k_i) + X_{it}^\mathsf{T}\theta$. The coefficient on the control covariate, θ, is assumed to be time-invariant and homogeneous across units. While both fixed-effect $\delta_t(k_i)$ and treatment effect $\beta_r(k_i)$ depend on the type k_i, $\delta_t(k_i)$ depends on time t and $\beta_r(k_i)$ depends on relative treatment timing r. In this sense, I call $\delta_t(k_i)$ *time* fixed-effect and $\beta_r(k_i)$ *dynamic* treatment effect. Note that by dropping t and introducing r for β, treatment effect is assumed to be time-invariant and dynamic and that $\beta_{-1}(k_i)$ is dropped to use $Y_{it}(-1)$ as a reference point as in Equation (1). Lastly α_i is unit fixed-effect.

Assumption 1 and 2 imply the following linear outcome model:

$$Y_{it} = \alpha_i + \delta_t(k_i) + \sum_{r \neq -1, r = -\infty}^{\infty} \beta_r(k_i) \mathbf{1}_{\{t = E_i + r\}} + X_{it}^\intercal \theta + U_{it}, \tag{2}$$

$$0 = \mathbf{E}\left[U_{it} | E_i, k_i, \{X_{is}\}_{s=-T_0-1}^{t}\right]. \tag{3}$$

The main goal of this paper is to develop an alternative econometric framework to be used when a researcher suspects treatment endogeneity in the unconditional TWFE specification. Thus, I build upon the unconditional TWFE specification and impose the linear structure as in the unconditional TWFE specification with Assumption 2. However, the idea of using the pretreatment outcomes to control for unit heterogeneity in treatment timing is not necessarily confined to the linear model. Though it will not be discussed in this paper, one can develop a model where the unit heterogeneity in treatment timing can be recovered from the pretreatment outcomes, without the linear structure.[2]

Lastly, let us adopt two more assumptions for tractability.

Assumption 3. (FINITE SUPPORT)

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

The finiteness of the type k_i allows us to use the readily available literature of unsupervised partitioning methods to estimate the type.

Assumption 4. (NO ANTICIPATION)

$$\beta_r^0(k) = 0 \quad \forall k \ and \ r < 0.$$

Assumption 4 is the standard parallel trend assumption in the event-study setup.

Here I would like to make two observations. Firstly, Assumptions 1-2 and 4 together complete the *conditional* TWFE regression model with a latent conditioning variable. Note that the standard TWFE regression model is nested in Equations (2) and (3) by imposing $\delta_t(k) = \delta_t(k')$ and $\beta_r(k) = \beta_r(k')$ for every k, k'. Secondly, Assumptions 2-3 reduce the heterogeneity across untreated potential outcomes to two channels: firstly, the unit fixed-effect α_i and secondly, the type-specific time fixed-effect $\delta_t(k_i)$. With the unit fixed-effects, unit heterogeneity is flexible across units but fixed across time periods. With the type-specific time fixed-effects, unit heterogeneity is restricted at each time periods with the finite type variable k_i, but flexible across time periods.

---

[2]One noteworthy remark to be made here is that the linear structure helps us in estimating the latent type variable and the model parameters simultaneously. For models with more complex structure, a two-step estimation method can be more suitable. In the first step, the type assignment will be separately estimated with the pretreatment outcomes and in the second step, the model parameters are estimated with the first step type assignment estimates as given.

3 Estimation

To proceed in more details, let us adopt following notations:

$$\gamma := \begin{pmatrix} k_1 & \cdots & k_N \end{pmatrix}^{\mathsf{T}} \in \Gamma,$$

$$\Gamma := \{1, \cdots, K\}^N,$$

$$\beta = \begin{pmatrix} \beta_{-T}(1) & \cdots & \beta_{T_1-1}(K) \end{pmatrix}^{\mathsf{T}},$$

$$\alpha = \begin{pmatrix} \alpha_1 & \cdots & \alpha_N \end{pmatrix}^{\mathsf{T}},$$

$$\delta = \begin{pmatrix} \delta_{-T_0-1}(1) & \cdots & \delta_{T_1-1}(K) \end{pmatrix}^{\mathsf{T}},$$

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. β, α and δ are vectors of dynamic treatment effects, unit fixed-effects and time fixed-effects. With these, we can construct an objective function to minimize:

$$\widehat{Q}^c(\theta, \beta, \alpha, \delta, \gamma)$$
$$= \frac{1}{N(T+1)} \sum_{i=1}^N \sum_{t=-T_0-1}^{T_1-1} \left(Y_{it} - \alpha_i - \delta_t(k_i) - X_{it}^{\mathsf{T}}\theta - \sum_{r \neq -1; r=-T}^{T_1-1} \beta_r(k_i)\mathbf{1}_{\{t=E_i+r\}} \right)^2.$$

With parameter spaces Θ, $\mathcal{B}$, $\mathcal{A}$, $\mathcal{D}$, Γ, a candidate estimator is

$$\left(\hat{\theta}, \hat{\beta}, \hat{\alpha}, \hat{\delta}, \hat{\gamma} \right) = \arg\min_{(\theta, \beta, \alpha, \delta, \gamma) \in \Theta \times \mathcal{B} \times \mathcal{A} \times \mathcal{D} \times \Gamma} \widehat{Q}^c(\theta, \beta, \alpha, \delta, \gamma).$$

However, using both α and δ poses the multicollinearity problem. Thus, I will modify the objective function $\widehat{Q}^c$ so that all the variables in it are first-differenced, relieving us of the burden of estimating α.[3] The dot notation below is used to indicate that the variables are first-differenced. Also, motivated by A4, a proportion of treatment effects for pretreatment leads is suppressed to be zero, for tractability. Note that treatment effects are included for all treatment lags but only for l pretreatment leads. The objective function is

$$\widehat{Q}(\theta, \dot{\beta}, \dot{\delta}, \gamma) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=-T_0}^{T_1-1} \left(\dot{Y}_{it} - \dot{\delta}_t(k_i) - \dot{X}_{it}^{\mathsf{T}}\theta - \sum_{r \neq -1; r=-l}^{T_1-1} \dot{\beta}_r(k_i)\mathbf{1}_{\{t=E_i+r\}} \right)^2$$

---

[3]An alternative of mean-differencing is discussed in Section 4.

with $\dot{Y}_{it} = Y_{it} - Y_{i,t-1}$ and $\dot{X}_{it} = X_{it} - X_{i,t-1}$ and the estimator is

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

Note that the type-specific time fixed-effects $\delta_t(k)$ as well as the type-specific dynamic treatment effects $\beta_t(k)$ are first-differenced. Thanks to the staggered adoption structure, the first-differenced treatment effects are uniform across units with different treatment timing E_i. For every t, k and r,

$$\dot{\delta}_t(k) = \delta_t(k) - \delta_{t-1}(k),$$

$$\dot{\beta}_r(k) = \beta_r(k) - \beta_{r-1}(k).$$

3.1 Algorithm

The algorithm that I use to solve the minimization problem (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. First-difference Y_{it} and X_{it}. Denote the first-differenced variables with a dot: $\dot{Y}_{it}$ and $\dot{X}_{it}$.

2. **(update θ, β)** Given the type assignment $\gamma^{(s)}$ from the s-th iteration, construct indicator variables for each relative treatment timing r and the assigned type k: $\mathbf{1}_{\{t=E_i+r, k_i^{(s)}=k\}}$ for $r = -l, \cdots, -2, 0, \cdots, T_1 - 1$ and $k = 1, \cdots, K$. Be aware that some of the indicators need to be dropped based on the type assignment since some types may not get assigned units for every treatment timing. Also, construct indicator variables for each time s and the assigned type k: $\mathbf{1}_{\{t=s, k_i^{(s)}=k\}}$ for $s = -T_0, \cdots, T_1 - 1$ and $k = 1, \cdots, K$. By running OLS regression of $\dot{Y}_{it}$ on $\dot{X}_{it}$ and the indicators, we get $\hat{\delta}_t^{(s)}(k)$, $\hat{\beta}_r^{(s)}(k)$ and $\hat{\theta}^{(s)}$.

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

$$\min_{k\in\{1,\cdots,K\}} \sum_{t=-T_0}^{T_1-1} \left(\dot{Y}_{it} - \dot{X}_{it}^{\intercal}\hat{\theta}^{(s)} - \sum_{r\neq-1; r=-l}^{T_1-1} \hat{\beta}_r^{(s)}(k)\mathbf{1}_{\{t=E_i+r\}} - \hat{\delta}_t^{(s)}(k)\right)^2.$$

If some of $\hat{\beta}^{(s)}$ are not estimated in the previous step, use the values from $\hat{\beta}^{(s-1)}$.

4. Repeat Step 2-3 until Step 3 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

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

The suggested algorithm quickly attains a local minimum of the minimization problem (4), which is computationally difficult, with the space for the type assignment having the cardinality of N^K. In the application I used in Section 5, the algorithm mostly converged within 20 iterations. The iterative algorithm proposed here has two stages. In the first stage, the algorithm estimates δ, β and θ by running a OLS regression on the first-differenced variables. In the second stage, the algorithm reassigns a type for each unit, by finding the type that minimizes the squared sum of residuals evaluated with the type-specific time fixed-effects and the type-specific dynamic treatment effects.

Since the iterative algorithm does not conduct an exhaustive search, there is a possibility that it might not converge to a global minimum. 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. There is growing literature on estimating the number of types using an information criterion, though the rigorous application of those to the model in this paper is yet to be investigated.

4 Asymptotic Results

In this section, I discuss the asymptotic theory on the estimator suggested in Section 3. The true parameters are denoted with superscript 0: the true DGP is

$$Y_{it} = \alpha_i^0 + \delta_t^0(k_i^0) + \sum_{r=0}^\infty \beta_r^0(k_i^0)\mathbf{1}_{\{t=E_i+r\}} + X_{it}^\mathsf{T}\theta^0 + U_{it}, \tag{5}$$

$$0 = \mathbf{E}\left[U_{it}|E_i, k_i^0, \{X_{is}\}_{s=-T_0}^t\right].$$

After first-differencing,

$$\dot{Y}_{it} = \dot{\delta}_t^0(k_i^0) + \sum_{r=0}^\infty \dot{\beta}_r^0(k_i^0)\mathbf{1}_{\{t=E_i+r\}} + \dot{X}_{it}^\mathsf{T}\theta^0 + \dot{U}_{it}$$

with $\dot{U}_{it} = U_{it} - U_{i,t-1}$.

To construct consistency results of the estimator defined in (4), let us adopt following assumptions.

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

 a. $\Theta \subset [-M, M]^p$, $\mathcal{B} \subset [-M, M]^{T+1}$ *and* $\mathcal{D} \subset [-M, M]^{T+1}$.

10

b. *Independence and identical distribution across units:* $\{U_{it}, X_{it}, E_i, k_i\}_t \stackrel{iid}{\sim} F.$

c. *(Finite moments) For any t, s and $q_1, q_2 \in \mathbb{N} \cup \{0\}$ such that $q_1 + q_2 \leq 4$,*

$$\mathbf{E}\left[U_{it}{}^{q_1} U_{is}{}^{q_2}\right], \mathbf{E}\left[||X_{it}||^2\right], \mathbf{E}\left[U_{it}{}^2 | \{X_{is}\}_{s=-T_0-1}^t\right], \mathbf{E}\left[U_{i,t-1}{}^2 | \{X_{is}\}_{s=-T_0-1}^t\right] \leq M.$$

d. *(Sequential exogeneity)*

$$\mathbf{E}\left[U_{it} \big| E_i, k_i, \{X_{is}\}_{s=-T_0-1}^t, \{U_{is}\}_{s=-T_0-1}^{t-1}\right] = 0$$

$$\mathbf{E}\left[X_{it} - X_{i,t-1} \big| E_i, k_i, \{X_{is}\}_{s=-T_0-1}^{t-1}, \{U_{is}\}_{s=-T_0-1}^{t-1}\right] = 0$$

e. *(No multicollinearity) Given an arbitrary type assignment $\gamma = \begin{pmatrix} k_1 & \cdots & k_N \end{pmatrix}$, let $\bar{\dot{X}}_{k \wedge \tilde{k} \wedge e, t}$ denote the mean of $\dot{X}_{jt}$ within units such that $k_j^0 = k$, $k_j = \tilde{k}$, and $E_i = e$, and let $\rho_{N,T}(\gamma)$ denote the minimum eigenvalue of the following matrix:*

$$\frac{1}{NT} \sum_i \sum_t \left(\dot{X}_{it} - \bar{\dot{X}}_{k_i^0 \wedge k_i \wedge E_i, t}\right) \left(\dot{X}_{it} - \bar{\dot{X}}_{k_i^0 \wedge k_i \wedge E_i, t}\right)^{\mathsf{T}}.$$

Then, $\min_{\gamma \in \Gamma} \rho_{N,T}(\gamma) \stackrel{p}{\to} \rho > 0$ as $N, T \to \infty$.

Note that Assumption 5.d is stricter than what I derived from Assumptions 1-3. Assumption 5.e means that for some (k, e, t), the subpopulation of units who share the same type k and share the treatment timing e should show sufficient variation in $\dot{X}_{it}$ at time t. If $\dot{X}_{it}$ is discrete, a criterion for the 'sufficient' variation is that the support of $\dot{X}_{it}$ within the subpopulation should contain at least $K + 1$ points. Note that this does not have to hold true for every (k, e, t), but for a set of (k, e, t)s whose measure is positive.

Theorem 1. *Assume model (5). Let Assumption 3-5 hold. Then, as N and T go to infinity,*

$$\hat{\theta} \stackrel{p}{\to} \theta^0.$$

Proof. See Appendix. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Theorem 1 is a consistency result for $\hat{\theta}$. For a consistency result on the type assignment estimator and the treatment effects estimators, I will argue that the suggested estimator is equivalent with the OLS estimator with the true type assignment as given, in probability.

Assumption 6.

    a. $\frac{T_0}{T_1} \to \tau \in (0,1)$ *as* $T \to \infty$.

    b. *For all* $k \in \{1, \cdots, K\}$, $\mu(k) = \Pr\left\{k_i^0 = k\right\} > 0$

    c. $\Pr\left\{\max_i E_i \geq \tilde{E}\right\} = o(1)$ *uniformly for all* N, *as* $\tilde{E} \to \infty$.

    d. *Let* $0 \leq l \leq T_0$ *increase with* T, *satisfying* $\frac{T_0 - l}{T} \to \tau_l \in (0,1)$ *as* $T \to \infty$. *Then, for all* $k, \tilde{k} \in \{1, \cdots, K\}$ *such that* $k \neq \tilde{k}$,

$$\left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k})\right)^2\right)^{\frac{1}{2}} \xrightarrow{p} c(k, \tilde{k}) > 0.$$

    e. *There exist* $d, \varepsilon^* > 0$ *such that for any* $k \neq \tilde{k}$, $e \in \mathbb{N}$, *and* $\varepsilon > 0$,

$$\Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \sum_{r=0}^{\infty}\left(\dot{\beta}_r^0(k) - \dot{\beta}_r^0(\tilde{k})\right)\mathbf{1}_{\{t=e+r\}}\right)^2 \leq \varepsilon^*\right\} \leq \exp(-dT),$$

$$\Pr\left\{\left|\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \sum_{r=0}^{\infty}\left(\dot{\beta}_r^0(k) - \dot{\beta}_r^0(\tilde{k})\right)\mathbf{1}_{\{t=e+r\}}\right)\dot{U}_{it}\right| > \varepsilon\right\} \leq \exp(-dT)$$

    when T *is large.*

    f. *There exists* $M^* > 0$ *such that for any* $\varepsilon > 0$ *and* $\nu > 0$,

$$\Pr\left\{\left|\frac{1}{T}\sum_t \dot{U}_{it}\right| > \varepsilon\right\} = o(T^{-\nu})$$

$$\Pr\left\{\frac{1}{T}\sum_t \|\dot{X}_{it}\| > M^*\right\} = o(T^{-\nu})$$

    as N, T *go to* ∞.

    Assumptions 6.a-b guarantees that each type is large and there are enough population-level pretreatment periods for each type. Thus, the types can be recovered from the pretreatment outcomes under Assumptions 6.a-b. Assumption 6.d makes it possible for us to distinguish each type by looking only at those pretreatment periods. Assumption 6.c allows us to only consider finitely many treatment scenarios. When the distribution of the treatment timing E_i does not depend on N, the independence assumption from A5.b combined with A6.c implies that there are only finite treatment timings. When the distribution of E_i depends on N, more flexible treatment timing distribution is allowed.

    Assumptions 6.e-f impose weak dependency on random processes in the model. Assumption 6.e firstly assumes two different types are distinct from each other in terms of their time fixed-effects in probability.

Also, Assumption 6.e assumes that the time fixed-effect difference is orthogonal to individual error $\dot{U}_{it}$ in probability. In both cases, the rate of convergence is exponential. A sufficient condition for this is that the type-specific time fixed-effects and the type-specific dynamic treatment effects are all degenerate random variables and the individual errors $\{U_{it}\}_t$ are iid. Lastly, Assumption 6.f assumes that the tail probability of $U_{it} - U_{it-1}$ and $X_{it} - X_{it-1}$ go to zero fast.

Theorem 2. *Assume model* (5). *Let Assumptions 3-6 hold. Let* $\left(\hat{\theta}^{ols\intercal}, \hat{\beta}^{ols}, \hat{\delta}^{ols}\right)^{\intercal}$ *denote the OLS estimator of* (5) *when the true type assignment* γ^0 *is given. Then, as* N, T *go to infinity,*

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

for any $\nu > 0$. *Moreover, given two treatment timings for type* k,

$$\hat{\delta}_t(k) = \hat{\delta}_t^{ols}(k) + o_p(T^{-\nu})$$

$$\hat{\beta}_r(k) = \hat{\beta}_r^{ols}(k) + o_p(T^{-\nu})$$

for any $\nu > 0$.

Proof. See Appendix. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Theorem 2 shows that the probability of wrongly estimating the type assignment goes to zero when N/T^ν goes to zero for some $\nu > 0$. To finalize the consistency result and derive asymptotic distribution of the dynamic treatment effect estimator $\hat{\beta}_r(k)$, let us discuss the asymptotic behavior of the OLS estimators when the true type assignment is given. Let

$$\tilde{X}_{it} = \dot{X}_{it} - \overline{\dot{X}}_{\cdot t}(k_i^0) - \overline{\dot{X}}_{t-E_i \cdot}(k_i^0)$$

where

$$\overline{\dot{X}}_{\cdot t}(k) = \frac{1}{\sum_{i=1}^N \mathbf{1}_{\{k_i^0=k\}}} \sum_{i=1}^N \dot{X}_{it} \mathbf{1}_{\{k_i^0=k\}},$$

$$\overline{\dot{X}}_{r\cdot}(k) = \begin{cases} \frac{1}{\sum_{i=1}^N \sum_{t=-T_0}^{T_1-1} \mathbf{1}_{\{k_i^0=k,t-E_i=r\}}} \sum_{i=1}^N \sum_{t=-T_0}^{T_1-1} \left(\dot{X}_{it} - \overline{\dot{X}}_{\cdot t}(k)\right) \mathbf{1}_{\{k_i^0=k,t-E_i=r\}}, & \text{if } r \notin R_0, \\ 0, & \text{if } r \in R_0, \end{cases}$$

with $R_0 = \{-l, -l-1 \cdots\} \cup \{-1\}$. $\overline{\dot{X}}_{\cdot t}(k)$ is the mean of $\dot{X}_{it}$ for time t and type k and $\overline{\dot{X}}_{r\cdot}(k)$ is the mean of $\left(\dot{X}_{it} - \overline{\dot{X}}_{\cdot t}(k)\right)$ for relative treatment timing r and type k. $\tilde{X}_{it}$ is the residual of $\dot{X}_{it}$ from projection onto

13

the indicators for the time fixed-effects and dynamic treatment effects. Note that $\overline{X}_{r.}(k)$ is zero for relative treatment timings that are not used in the estimation.

Assumption 7.

a. *With l satisfying A6.d, there exist positive definite matrices Σ_θ and Ω_θ such that*

$$\Sigma_\theta = \plim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=T_0}^{T_1-1} \tilde{X}_{it}\tilde{X}_{it}^{\intercal}$$

$$\Omega_\theta = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=-T_0}^{T_1-1} \sum_{s=-T_0}^{T_1-1} \mathbf{E}\left[\tilde{X}_{it}\tilde{X}_{js}^{\intercal}\dot{U}_{it}\dot{U}_{js}\right]$$

b. *As N, T go to infinity,*

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=T_0}^{T_1-1} \tilde{X}_{it}\dot{U}_{it} \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\Omega_\theta\right).$$

c. *There are only finite treatment timings: $\Pr\{E_i \leq E^*\} = 1$.*

Corollary 1. *Assume model (5). Let Assumption 3-7 hold and suppose that there exist at least two treatment timings for type k. As $N, T, T^{\nu^*}/N$ go to infinity for some $\nu^* > 0$,*

$$\hat{\theta} \xrightarrow{p} \theta^0,$$

$$\sum_{r'=0}^{r} \hat{\beta}_{r'}(k) \xrightarrow{p} \beta_r^0(k),$$

and

$$\sqrt{NT}\left(\hat{\theta} - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1}\right),$$

$$\sqrt{N}\left(\sum_{r'=0}^{r} \hat{\beta}_{r'}(k) - \beta_r^0(k)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{1}^{\intercal}\Sigma_r^{-1}\Omega_r\Sigma_r^{-1}\mathbf{1}\right),$$

with some consistently estimable positive definite matrices Σ_r, Ω_r as defined in Appendix.

Proof. See Appendix. □

Note that the estimators $\hat{\beta}_{r'}(k)$ are summed from $r' = 0$ to $r' = r$ to construct an estimator for $\beta_r(k)$ due to first-differencing. For any type that there are at least two treatment timings with positive probabilities, every type-specific dynamic treatment effect is consistently estimated. When every individual in a given type is treated at once, treatment effects and time fixed-effects cannot be disentangled.

4.1 Extension

4.1.1 Additional heterogeneity: treatment effect

In Theorem 2, the type assignment estimation is driven by the pretreatment type-specific time fixed-effects: $\{\delta_t(k)\}_{t<0,k}$. Thus, as long as the type-specific time fixed-effects satisfy Assumption 6, the type assignment can be consistently estimated at the rate of $NT^{-\nu}$ with some $\nu > 0$, by using only the pretreatment periods. Motivated by this observation and the heterogeneous treatment effect literature, let us further relax the model:

$$Y_{it} = \alpha_i + \delta_t(k_i) + \sum_{r=0}^{\infty} \beta_{ir} \mathbf{1}_{\{t=E_i+r\}} + X_{it}^{\mathsf{T}}\theta + U_{it}, \tag{6}$$

$$0 = \mathbf{E}\left[U_{it}|E_i, k_i^0, \{X_{is}\}_{s=-T_0}^{t}\right].$$

Note that the dynamic treatment effects are not anymore functions of the type, but subscripted with unit indices.

A recent development in the event-study literature has shown that the presence of such heterogeneity in treatment effect creates a negative weighting problem for the TWFE specification (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Borusyak et al., 2021). Under (6), the type-specific treatment effect estimator from this paper faces the same problem since it converges to the TWFE estimator under the true type assignment, as discussed in Theorem 2. Fortunately, since the type assignment estimation is driven by the pretreatment time fixed-effects, the available solutions in the literature are directly applicable once the type assignment is estimated using only the pretreatment observations. These solutions solve the problem by using some researcher-chosen weighting rather than the one induced by the TWFE specification. For example, De Chaisemartin and d'Haultfoeuille (2020) and Sun and Abraham (2021) uses an uniform weighting while Callaway and Sant'Anna (2021) uses an inverse of propensity score. By applying the suggestion from Sun and Abraham (2021) after estimating the type assignment,

$$\hat{\beta}_0^{het}(k) = \frac{1}{\sum_{i=1}^{N} \mathbf{1}_{\{\hat{k}_i=k, E_i=0\}}} \sum_{i=1}^{N} (Y_{i1} - Y_{i0}) \mathbf{1}_{\{\hat{k}_i=k, E_i=0\}}$$

$$- \frac{1}{\sum_{i=1}^{N} \mathbf{1}_{\{\hat{k}_i=k, E_i>0\}}} \sum_{i=1}^{N} (Y_{i1} - Y_{i0}) \mathbf{1}_{\{\hat{k}_i=k, E_i>0\}}$$

is an interpretable estimator for the dynamic treatment effect at relative treatment timing $r = 0$ for type k. As N/T^{ν^*} goes to zero with some $\nu^* > 0$, a similar argument as that for Corollary 1 holds and an asymptotic distribution for $\hat{\beta}_0^{het}(k)$ is obtained.
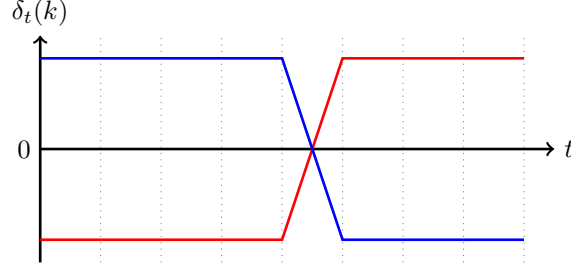
Figure 1: constant time fixed-effects with a structural change

4.1.2 Additional heterogeneity: slope on control covariate

As we assume that there exists heterogeneity across units in terms of time fixed-effects and treatment effects, a natural next direction of introducing more heterogeneity is in the slope coefficient for control covariate θ. Let us consider the following relaxed model:

$$Y_{it} = \alpha_i + \delta_t(k_i) + \sum_{r=1}^{\infty} \beta_r(k_i)\mathbf{1}_{\{t=E_i+r\}} + X_{it}^{\mathsf{T}}\theta(k_i) + U_{it}. \tag{7}$$

Note that the slope coefficient on the control covariate is now a function of the type. Since $\theta(k)$ is time-invariant, the existing asymptotic discussion of Bonhomme and Manresa (2015) is directly applicable to (7) by only using the pretreatment periods and $\theta(k)$ is consistently estimatable for each k.

4.1.3 Strict exogeneity and mean-differencing

In the main specification, the outcome variable Y_{it} and the control covariate X_{it} are first-differenced to remove the unit fixed-effects α_i. For the type assignment estimation, it is assumed that the first-differenced time fixed-effects and dynamic treatment effects show enough variation across types in Assumption 6.d-e. However, it is possible that the time fixed-effects vary across types only in level, but not in first differences. A straightforward example is constant time fixed-effects with a structural change presented as in Figure 1. In this case, the distance between the two types evaluated with first-differencing used in Assumption 6.d will be much smaller than that with mean-differencing:

$$0 \approx \frac{1}{T}\sum_t \Big(\delta_t(k) - \delta_{t-1}(k) - (\delta_t(k') - \delta_{t-1}(k'))\Big)^2 \ll \frac{1}{T}\sum_t \Big(\delta_t(k) - \bar{\delta}(k) - (\delta_t(k') - \bar{\delta}(k'))\Big)^2.$$

This calls for mean-differencing.

Fortunately, though not free of cost, mean-differencing is also feasible. By adopting a stronger assumption of strict exogeneity and using mean-differenced variables, accordingly modified versions of Theorem 1, 2 and

Corollary 1 hold.

Assumption 5.*d'* (Strict exogeneity)

$$\mathbf{E}\left[U_{it}\big|E_i, k_i, \{X_{is}\}_{s=-T_0-1}^{T_1-1}, \{U_{is}\}_{s=-T_0-1}^{T_1-1}\right] = 0$$

Under Assumption 5.d', Step 1 of the proof for Theorem 1 is proven similarly. Then, the rest of the proofs for Theorem 2 and Corollary 1 follow by modifying Assumption 5-6 accordingly to accommodate for mean-differenced variables and mean-differenced treatment effects and time fixed-effects, though a rigorous exposition of such accommodations is not discussed in this paper. The estimation results of the following application under the mean-differencing are presented in Appendix.

5 Application

To see how the estimation method suggested in this paper fares with a real dataset, I 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 should be free of the court-ordered plans once it eradicated the effects of previous 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, I 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 : # of black students in school j, $\quad$ w_j : # of white students in school j

J_i : the set of school in school district i,

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

The segregation index ranges from 0 to 100, with 100 being perfectly segregated schools and 0 being perfectly representative schools.[4]

I followed the data cleaning process in the paper and chose the timespan of 1989-2007 to form a balanced panel of school districts that were under a court-ordered desegregation plan in 1991, which gave me 102 school districts. In estimation, I included four indicators for pretreatment leads ($r = -5, \cdots, -2$) and nine indicators for treatment lags ($r = 0, \cdots, 7$ and all treatment lags beyond the eighth lag). Also, I set the dynamic treatment effects for pretreatment leads to be shared across types, for bigger power:

$$\dot{Y}_{it} = \alpha_i + \delta_t(k_i) + \sum_{r=-5}^{-2} \beta_r \mathbf{1}_{\{t=E_i+r\}} + \sum_{r=0}^{7} \beta_r(k_i) \mathbf{1}_{\{t=E_i+r\}} + \beta_8(k_i) \mathbf{1}_{\{t \geq E_i+8\}} + \dot{X}_{it}^{\mathsf{T}}\theta + U_{it}.$$

For the purpose of comparison, here I present the main empirical specification of Lutz (2011):

$$\dot{Y}_{it} = \delta_{jt} + \sum_{r=-l}^{T_1-1} \beta_r \cdot \sum_{r'=-l}^{r} \mathbf{1}_{\{t=E_i+r'\}} + X_i^{\mathsf{T}}\theta_t + U_{it}$$

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. In my main specification, I use time-varying control covariates X_{it}, with time-invariant coefficient θ. This deviation is made to resemble the canonical TWFE regression specification. Secondly, Lutz (2011) uses a type-specific 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 and imposed that $\beta_r(k) = \beta_r(k')$ for any k and k' whereas I used the data to estimate the type assignment.

Figure 2 contains the type-specific dynamic effect estimates where the number of types is set to be 2.[5] From Figure 2, we see that treatment effect is bigger for type 1 and smaller for type 2; the termination of

---

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

[5]In practice, the choice of K should depend on data. In this application, setting $K = 2$ comes at relative low price; increasing K creates a new type to which only a few units are assigned.

court-ordered desegregation plans exacerbated racial segregation more severely for type 1. For reference on the magnitude, the mean of the segregation index was 37 and its standard deviation was around 15 in 1990.
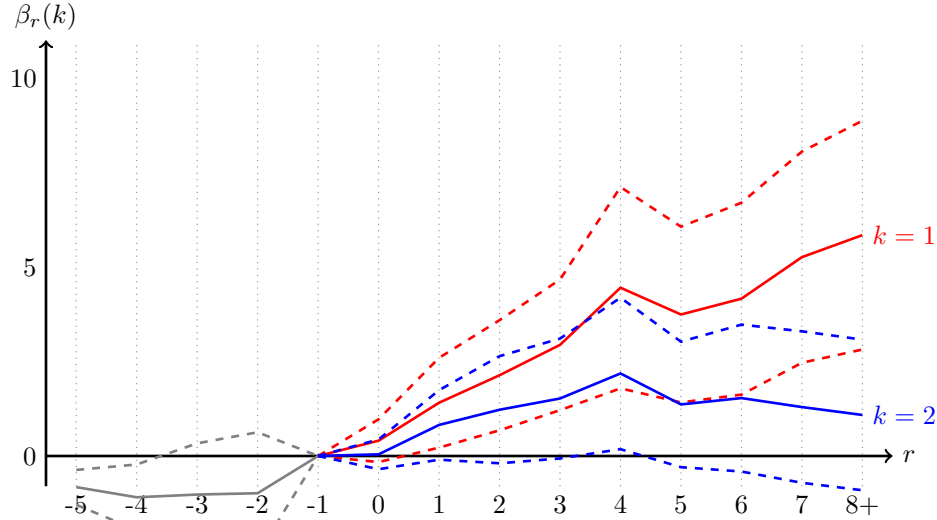

Figure 2: type-specific treatment effect, $K = 2$.

The graph reports the type-specific dynamic effects of terminating court-mandated desegregation plan on the segregation index of a school district.

The segregation index ranges from 0 to 100. In 1990, the average segregation index was 37 and the standard deviation was 15.

$k = 1$ is the first group where the segregation index was improving and $k = 2$ is the second group where the segregation index was mildly worsening.

The confidence intervals are at 0.05 significance level and are computed with asymptotic standard errors clustered at the school district level.

So, estimates on treatment effect suggest that type 1 and type 2 are different; the treatment affects type 1 more. Are these types different in other regards? Table 1 shows us some summary statistics on the outcome variable and other control covariates for type 1 and type 2. Note that we see small mean differences but in most of the cases the difference is not significant individually. However, 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, implying that those estimated types are different from each other.

Lastly, Figure 3 provides us an illustrative evidence that the types are different from each other in another dimension of unobservables as well, other than the treatment effects: type-specific fixed-effects. Figure 3 contains the estimated type-specific time fixed-effects, when $K = 2$. Over the time, type 1 has seen an increase in the segregation index while type 2 has seen a slow decline. This implies that the termination of desegregation plans had a bigger impact on type 1, where the segregation index was already rising. On the other hand, type 2, where racial segregation was being mitigated, the impact was close to zero.

	K = 2		
	(k = 1)	(k = 2)	Diff
Segregation index	29.45	41.12	-11.67
	(16.33)	(17.04)	(3.30)
% (white)	52.70	48.74	3.96
	(20.54)	(22.29)	(4.24)
% (hispanic)	4.31	14.46	-10.16
	(9.88)	(17.95)	(2.85)
enrollment	39695	54647	-14952
	(45920)	(101922)	(15555)
N	50	52	-
joint p-value			0.000

Table 1: Balancedness test

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.
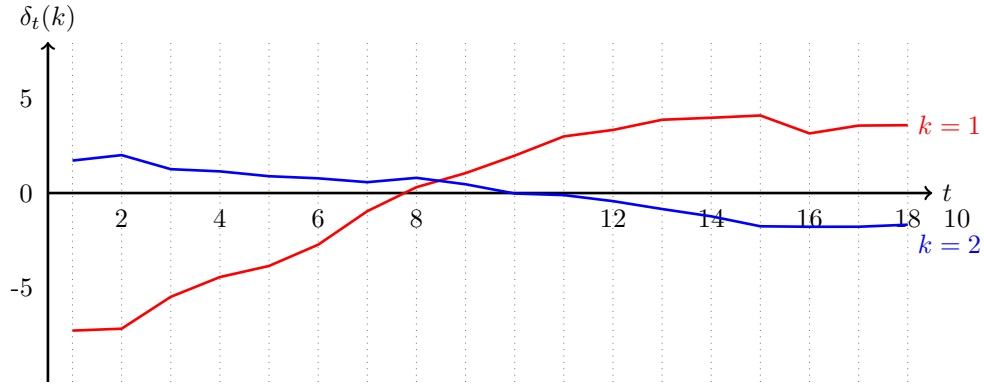

Figure 3: type-specific fixed-effects, $K = 2$

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

The segregation index ranges from 0 to 100. In 1990, the average segregation index was 37 and the standard deviation was 15.

$k = 1$ is the first group where the treatment effects of terminating court-mandated desegregation plan were significantly positive, meaning that the segregation got worse from the treatment, and $k = 2$ is the second group where the the treatment effects were insignificant.

This observation combined with Figure 2 presents numerous 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?

6   Conclusion

In this paper, I motivate an event-study regression model with sequential exogeneity in a panel data setting, from a sequential unconfoundedness assumption with a latent type variable. When the latent type variable has a finite support, the K-means estimator retrieves the true type assignment well. Also, based on the estimated type assignment we can estimate the type-specific treatment effect, given that the type has more than two treatment timings with positive measure. By applying the estimation method to an empirical application, I 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, 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.

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.

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.

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.

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

Fadlon, Itzik and Torben Heien Nielsen, "Family health behaviors," *American Economic Review*, 2019, *109* (9), 3162–91.

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

Gallagher, Justin and Daniel Hartley, "Household finance after a natural disaster: The case of Hurricane Katrina," *American Economic Journal: Economic Policy*, 2017, *9* (3), 199–228.

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

Havnes, Tarjei and Magne Mogstad, "No child left behind: Subsidized child care and children's long-run outcomes," *American Economic Journal: Economic Policy*, 2011, *3* (2), 97–129.

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

Meghir, Costas and Mårten Palme, "Educational reform, ability, and family background," *American Economic Review*, 2005, *95* (1), 414–424.

Moon, Hyungsik Roger and Martin Weidner, "Linear regression for panel with unknown number of factors as interactive fixed effects," *Econometrica*, 2015, *83* (4), 1543–1579.

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

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

In the Appendix, I will abuse the subscript nontation in the following way:

$$\sum_r \hat{\beta}_r(k)\mathbf{1}_{\{t=E_i+r\}} = \dot{\beta}_{t-E_i}(k).$$

Also, even though it is assumed that $\beta_r(k) = 0$ and thus $\dot{\beta}_r(k) = 0$ for all $r \leq -1$, $\dot{\beta}_r(k)$ are still used as placeholders for every $r \geq -l$.

Proof for Theorem 1

Step 1

The first step is to obtain an approximation of the objective function. Note that

$$\widehat{Q}(\theta, \beta, \delta, \gamma) = \frac{1}{NT} \sum_i \sum_t \left(\dot{Y}_{it} - \delta_t(k_i) - \dot{X}_{it}^\intercal \theta - \beta_{t-E_i}(k_i) \right)^2$$

$$= \frac{1}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{X}_{it}^\intercal(\theta^0 - \theta) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) + \dot{U}_{it} \right)^2$$

$$= \frac{1}{NT} \sum_i \sum_t \left\{ \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{X}_{it}^\intercal(\theta^0 - \theta) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) \right)^2 + \dot{U}_{it}^2 \right\}$$

$$+ \frac{2}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{X}_{it}^\intercal(\theta^0 - \theta) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) \right) \dot{U}_{it}.$$

Let

$$\tilde{Q}(\theta, \beta, \delta, \gamma) = \frac{1}{NT} \sum_i \sum_t \left\{ \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{X}_{it}^\intercal(\theta^0 - \theta) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) \right)^2 + \dot{U}_{it}^2 \right\}.$$

Then,

$$\left| \widehat{Q}(\theta, \beta, \delta, \gamma) - \tilde{Q}(\theta, \beta, \delta, \gamma) \right|$$

$$= \left| \frac{2}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{X}_{it}^\intercal(\theta^0 - \theta) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) \right) \dot{U}_{it} \right|$$

$$\leq \left| \frac{2}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) \right) \dot{U}_{it} \right| + \left| \frac{2}{NT} \sum_i \sum_t \dot{X}_{it}^\intercal(\theta^0 - \theta)\dot{U}_{it} \right|. \qquad (8)$$

Firstly, find that

$$
\left| \frac{1}{NT} \sum_i \sum_t \dot{\delta}_t^0(k_i^0) \dot{U}_{it} \right| \leq \sum_{k=1}^K \left| \frac{1}{NT} \sum_i \sum_t \dot{\delta}_t^0(k) \dot{U}_{it} \mathbf{1}_{\{k_i^0=k\}} \right|
$$

$$
\leq \sum_{k=1}^K \left(\frac{1}{T} \sum_t \dot{\delta}_t^0(k)^2 \right)^{\frac{1}{2}} \left(\frac{1}{T} \sum_t \left(\frac{1}{N} \sum_i \dot{U}_{it} \mathbf{1}_{\{k_i^0=k\}} \right)^2 \right)^{\frac{1}{2}}
$$

$$
\leq 4M \sum_{k=1}^K \left(\frac{1}{N^2 T} \sum_i \sum_j \sum_t \dot{U}_{it} \dot{U}_{jt} \mathbf{1}_{\{k_i^0=k\}} \mathbf{1}_{\{k_j^0=k\}} \right)^{\frac{1}{2}}
$$

$$
\leq 4MK \left(\frac{1}{N^2} \sum_i \sum_j \left| \frac{1}{T} \sum_t \dot{U}_{it} \dot{U}_{jt} \right| \right)^{\frac{1}{2}}
$$

$$
\leq 4MK \left(\frac{1}{N^2 T^2} \sum_i \sum_j \sum_t \sum_s \dot{U}_{it} \dot{U}_{jt} \dot{U}_{is} \dot{U}_{js} \right)^{\frac{1}{4}} \xrightarrow{p} 0.
$$

The first two inequalities are from separating the summation into types and applying Cauchy-Schwartz's inequality to over t. The third is from A5.a. The fifth is from applying Jensen's inequality with $x \mapsto x^2$. The convergence in probability comes from A5.b-d. For any N and T,

$$
\mathbf{E}\left[\dot{U}_{it} \dot{U}_{is} \right] = \mathbf{E}\left[(U_{it} - U_{it-1})(U_{is} - U_{it-1}) \right]
$$

$$
\leq \begin{cases} 0, & \text{if } |t-s| > 1 \\ 4M, & \text{if } |t-s| \leq 1 \end{cases},
$$

$$
\mathbf{E}\left[\dot{U}_{it} \dot{U}_{jt} \dot{U}_{is} \dot{U}_{js} \right] \leq \begin{cases} \mathbf{E}\left[\dot{U}_{it} \dot{U}_{is} \right] \cdot \mathbf{E}\left[\dot{U}_{jt} \dot{U}_{js} \right], & \text{if } i \neq j \\ 16M, & \text{if } i = j \end{cases},
$$

$$
\frac{1}{N^2 T^2} \sum_{i,j,t,s} \mathbf{E}\left[\dot{U}_{it} \dot{U}_{jt} \dot{U}_{is} \dot{U}_{js} \right] = \frac{1}{N^2 T^2} \sum_{i \neq j,t,s} \mathbf{E}\left[\dot{U}_{it} \dot{U}_{jt} \dot{U}_{is} \dot{U}_{js} \right] + \frac{1}{N^2 T^2} \sum_{i=j,t,s} \mathbf{E}\left[\dot{U}_{it} \dot{U}_{jt} \dot{U}_{is} \dot{U}_{js} \right]
$$

$$
\leq \frac{16M}{N^2 T^2} \left(3N(N-1)T + NT^2 \right) = \frac{1}{T} \cdot 48M + \frac{1}{N} \cdot 16M.
$$

Since a sequence of random variables whose expectations are uniformly bounded is $O_p(1)$,

$$
\frac{1}{N^2 T^2} \sum_{i,j,t,s} \dot{U}_{it} \dot{U}_{jt} \dot{U}_{is} \dot{U}_{js} = \frac{1}{T} O_p(1) + \frac{1}{N} O_p(1) = o_p(1).
$$

We can repeat this for three other quantities in the first term of (8).

Secondly, again from applying Cauchy-Schwartz's inequality and Jensen's inequality,

$$
\begin{aligned}
\left| \frac{1}{NT} \sum_i \sum_t \dot{X}_{it}^\intercal (\theta^0 - \theta) \dot{U}_{it} \right| &\leq \frac{1}{N} \sum_i \left| \frac{1}{T} \sum_t \dot{U}_{it} \dot{X}_{it}^\intercal (\theta^0 - \theta) \right| \\
&\leq \frac{1}{N} \sum_i \left\| \frac{1}{T} \sum_t \dot{U}_{it} \dot{X}_{it} \right\| \cdot ||\theta^0 - \theta|| \\
&\leq \frac{2M}{N} \sum_i \left(\frac{1}{T^2} \sum_t \sum_s \dot{U}_{it} \dot{U}_{is} \dot{X}_{it}^\intercal \dot{X}_{is} \right)^{\frac{1}{2}} \\
&\leq \frac{2M}{T^{\frac{1}{2}}} \cdot \frac{1}{N} \sum_i \left(\frac{1}{T} \sum_t \sum_s \dot{U}_{it} \dot{U}_{is} \dot{X}_{it}^\intercal \dot{X}_{is} \right)^{\frac{1}{2}} = \frac{2M}{T^{\frac{1}{2}}} \cdot O_p(1) \xrightarrow{p} 0
\end{aligned}
$$

The convergence in probability is from A5.c-d. For any T,

$$
\sum_s \mathbf{E} \left[\dot{U}_{it} \dot{U}_{is} \dot{X}_{it}^\intercal \dot{X}_{is} \right] = \mathbf{E} \left[\dot{U}_{it}^2 \dot{X}_{it}^\intercal \dot{X}_{it} + \dot{U}_{it} \dot{U}_{i,t+1} \dot{X}_{it}^\intercal \dot{X}_{i,t+1} + \dot{U}_{it} \dot{U}_{i,t-1} \dot{X}_{it}^\intercal \dot{X}_{i,t-1} \right]
$$

$$
\leq (2M + M + M) = 4M
$$

$$
\frac{1}{N} \sum_i \mathbf{E} \left[\left(\frac{1}{T} \sum_t \sum_s \dot{U}_{it} \dot{U}_{is} \dot{X}_{it}^\intercal \dot{X}_{is} \right)^{\frac{1}{2}} \right] \leq 4M + 1.
$$

Thus, $\frac{1}{N} \sum_i \left(\frac{1}{T} \sum_t \sum_s \dot{U}_{it} \dot{U}_{is} \dot{X}_{it}^\intercal \dot{X}_{is} \right)^{\frac{1}{2}} = O_p(1)$ and $\widehat{Q}(\theta, \beta, \delta, \gamma) - \tilde{Q}(\theta, \beta, \delta, \gamma) = o_p(1)$.

Step 2

By plugging in the true parameters,

$$
\tilde{Q}(\theta^0, \beta^0, \delta^0, \gamma^0) = \frac{1}{NT} \sum_{i,t} \dot{U}_{it}^2
$$

$$
\begin{aligned}
\tilde{Q}(\theta, \beta, \delta, \gamma) - \tilde{Q}(\theta^0, \beta^0, \delta^0, \gamma^0) &= \frac{1}{NT} \sum_{i,t} \left(\dot{\delta}_t^0(k_i^0) - \delta_t(k_i) + \dot{X}_{it}^\intercal (\theta^0 - \theta) + \dot{\beta}_{t-E_i}^0(k_i^0) - \beta_{t-E_i}(k_i) \right)^2 \\
&\geq \frac{1}{NT} \sum_{i,t} \left(\dot{X}_{it}^\intercal (\theta^0 - \theta) - \bar{\dot{X}}_{k_i^0 \wedge k_i \wedge E_i, t}^\intercal (\theta^0 - \theta) \right)^2 \\
&= \frac{1}{NT} \sum_{i,t} (\theta^0 - \theta)^\intercal \left(\dot{X}_{it} - \bar{\dot{X}}_{k_i^0 \wedge k_i \wedge E_i, t} \right) \left(\dot{X}_{it} - \bar{\dot{X}}_{k_i^0 \wedge k_i \wedge E_i, t} \right)^\intercal (\theta^0 - \theta) \\
&\geq \min_{\gamma \in \Gamma} \rho_{N,T}(\gamma) \cdot ||\theta^0 - \theta||^2 .
\end{aligned}
$$

Note that the unknowns in $\tilde{Q}(\theta, \beta, \delta, \gamma) - \tilde{Q}(\theta^0, \beta^0, \delta^0, \gamma^0)$ other than $(\theta^0 - \theta)$ are dictated by (t, k_i^0, k_i, E_i). Thus, subtracting the group mean defined with (t, k_i^0, k_i, E_i) from $\dot{X}_{it}^\intercal (\theta^0 - \theta)$ could function as a lower bound for the difference, giving us the first inequality.

Step 3

Since the estimator minimizes the objective function,

$$\tilde{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) = \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) + o_p(1)$$
$$\leq \widehat{Q}(\theta^0, \beta^0, \delta^0, \gamma^0) + o_p(1)$$
$$= \tilde{Q}(\theta^0, \beta^0, \delta^0, \gamma^0) + o_p(1).$$

Therefore from A5.e,

$$\min_{\gamma \in \Gamma} \rho_{N,T}(\gamma) \cdot \left\|\theta^0 - \hat{\theta}\right\|^2 \leq \tilde{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \tilde{Q}(\theta^0, \beta^0, \delta^0, \gamma^0) = o_p(1)$$

$$\left\|\theta^0 - \hat{\theta}\right\|^2 = \frac{1}{\min_{\gamma \in \Gamma} \rho_{N,T}(\gamma)} \cdot \min_{\gamma \in \Gamma} \rho_{N,T}(\gamma) \left\|\theta^0 - \hat{\theta}\right\|^2$$

$$\xrightarrow{p} \frac{1}{\rho} \cdot 0 = 0.$$

Finally,

$$\left|\tilde{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \tilde{Q}(\theta^0, \hat{\beta}, \hat{\delta}, \hat{\gamma})\right|$$

$$= \left| \frac{1}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \dot{X}_{it}^\intercal(\theta^0 - \hat{\theta}) + \dot{\beta}_{t-E_i}^0(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right)^2 \right.$$

$$\left. - \frac{1}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \dot{\beta}_{t-E_i}^0(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right)^2 \right|$$

$$\leq \left| \frac{2}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \dot{\beta}_{t-E_i}^0(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right) \dot{X}_{it}^\intercal(\theta^0 - \hat{\theta}) \right|$$

$$+ \left| \frac{1}{NT} \sum_i \sum_t \left(\dot{X}_{it}^\intercal(\theta^0 - \hat{\theta}) \right)^2 \right|$$

$$\leq \frac{2 \cdot 4M}{NT} \sum_i \sum_t ||\dot{X}_{it}|| \cdot \left\|\theta^0 - \hat{\theta}\right\| + \frac{1}{NT} \sum_i \sum_t ||\dot{X}_{it}||^2 \cdot \left\|\theta^0 - \hat{\theta}\right\|^2 = o_p(1).$$

The second inequality is from A5.a and Cauchy-Schwartz's inequality. Note that for any N and T, both $\frac{1}{NT} \sum_i \sum_t ||\dot{X}_{it}||$ and $\frac{1}{NT} \sum_i \sum_t ||\dot{X}_{it}||^2$ are bounded in expectation by $\max\{2(M+1), 4M\}$ from A5.c, and

thus $O_p(1)$. Since we have shown $\hat{\theta} \xrightarrow{p} \theta^0$, we have the last equality. Then,

$$\frac{1}{NT} \sum_i \sum_t \left\{ \left(\dot{\delta}_t^0(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \dot{\beta}_{t-E_i}^0(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right)^2 + \dot{U}_{it}^2 \right\} = \tilde{Q}(\theta^0, \hat{\beta}, \hat{\delta}, \hat{\gamma})$$

$$= \tilde{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) + o_p(1)$$

$$= \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) + o_p(1)$$

$$\leq \widehat{Q}(\theta^0, \beta^0, \delta^0, \gamma^0) + o_p(1)$$

$$= \frac{1}{NT} \sum_i \sum_t \dot{U}_{it}^2 + o_p(1).$$

□

Proof for Theorem 2

Step 1

Note that $\widehat{Q}(\theta, \beta, \delta, \gamma)$ does not vary for any $(\theta, \tilde{\beta}, \tilde{\delta}, \tilde{\gamma})$ defined with a permutation on $(1, \cdots, K)$: with σ, a permutation on $\{1, \cdots, K\}$, letting $\tilde{k}_i = \sigma(k_i)$, $\tilde{\beta}_r(\sigma(k)) = \beta_r(k)$, and $\tilde{\delta}_t(\sigma(k)) = \delta_t(k)$ gives us $\widehat{Q}(\theta, \beta, \delta, \gamma) = \widehat{Q}(\theta, \tilde{\beta}, \tilde{\delta}, \tilde{\gamma})$. Thus, we need to find a bijection on $\{1, \cdots, K\}$ to match $\hat{k}$ with true k^0. Define a function σ by letting

$$\sigma(k) = \arg\min_{\tilde{k}} \frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\sigma(k)) \right)^2 \tag{9}$$

for each k. First, let us show that σ actually lets the objective go to zero for each k: fix k,

$$\min_{\tilde{k}} \frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\sigma(k)) \right)^2$$

$$\leq \frac{N}{\sum_i \mathbf{1}_{\{k_i^0=k\}}} \frac{T}{T_0 - l} \cdot \min_{\tilde{k}} \frac{1}{NT} \sum_i \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\sigma(k)) \right)^2 \mathbf{1}_{\{k_i^0=k\}}$$

$$\leq \frac{N}{\sum_i \mathbf{1}_{\{k_i^0=k\}}} \frac{T}{T_0 - l} \cdot \frac{1}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \dot{\beta}_{t-E_i}^0(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right)^2 \xrightarrow{p} 0$$

as $N, T \to \infty$. The second inequality is from the fact that $\beta_r(k)$ are forced to be zero for all $r < -l$. With A6.a-b, Theorem 1, and properly chosen l, we have the convergence.

For some $k, \tilde{k}$ such that $k \neq \tilde{k}$,

$$
\left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\hat{\delta}_t(\sigma(k)) - \hat{\delta}_t(\sigma(\tilde{k})) \right)^2 \right)^{\frac{1}{2}}
$$

$$
\geq \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) \right)^2 \right)^{\frac{1}{2}}
$$

$$
- \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\sigma(k)) \right)^2 \right)^{\frac{1}{2}} - \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(\tilde{k}) - \hat{\delta}_t(\sigma(\tilde{k})) \right)^2 \right)^{\frac{1}{2}}
$$

$$
\xrightarrow{p} c(k, \tilde{k}) > 0
$$

from A6.d. Thus, $\Pr \{\sigma \text{ is not bijective}\} \leq \sum_{k \neq \tilde{k}} \Pr \left\{ \sigma(k) = \sigma(\tilde{k}) \right\} \to 0$ as $N, T \to \infty$. Note that σ depends on the dataset.

Step 2

Now, I extend the result of Step 1 to the entire time-series, by showing that for each k and e such that $\mu_e(k) = \Pr \{E_i = e, k_i = k\} > 0$,

$$
\frac{1}{T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\sigma(k)) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(\sigma(k)) \right)^2 \xrightarrow{p} 0 \tag{10}
$$

as $N, T \to \infty$. From A6.b-c, there is at least one such e for each k. Then,

$$
\min_{\tilde{k}} \frac{1}{T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\tilde{k}) + \dot{\beta}_{t-e}^0(k) - \hat{\beta}_{t-e}(\tilde{k}) \right)^2 \tag{11}
$$

$$
\leq \frac{N}{\sum_i \mathbf{1}_{\{k_i^0=k, E_i=e\}}} \cdot \frac{1}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\hat{k}_i) + \dot{\beta}_{t-e}^0(k) - \hat{\beta}_{t-e}(\hat{k}_i) \right)^2 \mathbf{1}_{\{k_i^0=k, E_i=e\}}
$$

$$
\leq \frac{N}{\sum_i \mathbf{1}_{\{k_i^0=k, E_i=e\}}} \cdot \frac{1}{NT} \sum_i \sum_t \left(\dot{\delta}_t^0(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \dot{\beta}_{t-E_i}^0(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right)^2 \xrightarrow{p} 0
$$

as $N, T \to \infty$ from $\mu_e(k) > 0$ and Theorem 1. We would like the solution to (11) to be equal to $\sigma(k)$. Let us denote the solution with $\tilde{k}^*$. Suppose σ is bijective and $\tilde{k}^* \neq \sigma(k) \Leftrightarrow \sigma^{-1}(\tilde{k}^*) \neq k$. Then,

$$\left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(\sigma^{-1}(\tilde{k}^*)) - \dot{\delta}_t^0(k) \right)^2 \right)^{\frac{1}{2}} - c\left(\sigma^{-1}(\tilde{k}^*), k \right)$$

$$- \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(\sigma^{-1}(\tilde{k}^*)) - \hat{\delta}_t(\tilde{k}^*) \right)^2 \right)^{\frac{1}{2}} - \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\hat{\delta}_t(\tilde{k}^*) - \dot{\delta}_t^0(k) \right)^2 \right)^{\frac{1}{2}}$$

$$\leq -c\left(\sigma^{-1}(\tilde{k}^*), k \right)$$

$$\leq - \min_{k' \neq k''} c(k', k'') < 0.$$

The inequality above implies

$$- \sum_{k'} \left| \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k') - \dot{\delta}_t^0(k) \right)^2 \right)^{\frac{1}{2}} - c\left(k', k \right) \right|$$

$$- \sum_{k'} \left| \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\dot{\delta}_t^0(k') - \hat{\delta}_t(\sigma(k')) \right)^2 \right)^{\frac{1}{2}} \right| - \left| \left(\frac{1}{T_0 - l} \sum_{t=-T_0}^{-l-1} \left(\hat{\delta}_t(\tilde{k}^*) - \dot{\delta}_t^0(k) \right)^2 \right)^{\frac{1}{2}} \right|$$

$$\leq - \min_{k' \neq k''} c(k', k'') < 0.$$

Note that the LHS of the inequality is $o_p(1)$. Thus,

$$\Pr \left\{ \tilde{k}^* \neq \sigma(k) \right\} \leq \Pr \left\{ \sigma \text{ is not bijective} \right\} + \Pr \left\{ \sigma \text{ is bijective}, \sigma^{-1}(\tilde{k}^*) \neq k \right\}$$

$$\leq o(1) + \Pr \left\{ o_p(1) \leq - \min_{k' \neq k''} c(k', k'') \right\} = o(1)$$

and consequently

$$\frac{1}{T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\sigma(k)) + \dot{\beta}_{t-e}^0(k) - \hat{\beta}_{t-e}(\sigma(k)) \right)^2 \xrightarrow{p} 0$$

as $N, T \to \infty$. We can repeat this for any e such that $\mu_e(k) > 0$ and the same σ as defined in (9) gives us the convergence in probability to zero.

Before proceeding to the next step, let us drop the σ notation. Based on σ, we can construct a bijection $\tilde{\sigma} : \{1, \cdots, K\} \to \{1, \cdots, K\}$ such that

$$\frac{1}{T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(\tilde{\sigma}(k)) + \dot{\beta}_{t-e}^0(k) - \hat{\beta}_{t-e}(\tilde{\sigma}(k)) \right)^2 \xrightarrow{p} 0$$

as $N, T \to \infty$ for all k, by letting $\tilde{\sigma} = \sigma$ whenever σ is bijective. From now on, I will drop $\tilde{\sigma}$ by always

rearranging $(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma})$ so that $\tilde{\sigma}(k) = k$.

Step 3

Here, we study the probability of the K-means algorithm to assign a wrong type, in terms of one single unit.

$$\Pr\left\{\hat{k}_i \neq k_i^0\right\}$$

$$\leq \sum_{\tilde{k} \neq k_i^0} \Pr\left\{\frac{1}{T}\sum_t \left(\dot{Y}_{it} - \hat{\delta}_t(\tilde{k}) - \dot{X}_{it}^\intercal \hat{\theta} - \hat{\beta}_{t-E_i}(\tilde{k})\right)^2 \leq \frac{1}{T}\sum_t \left(\dot{Y}_{it} - \hat{\delta}_t(k_i^0) - \dot{X}_{it}^\intercal \hat{\theta} - \hat{\beta}_{t-E_i}(k_i^0)\right)^2\right\}$$

$$= \sum_{\tilde{k} \neq k_i^0} \Pr\left\{\frac{2}{T}\sum_t \left(\hat{\delta}_t(k_i^0) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k_i^0) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\right.$$

$$\left. \cdot \left(\dot{Y}_{it} - \frac{\hat{\delta}_t(k_i^0) + \hat{\delta}_t(\tilde{k})}{2} - \dot{X}_{it}^\intercal \hat{\theta} - \frac{\hat{\beta}_{t-E_i}(k_i^0) + \hat{\beta}_{t-E_i}(\tilde{k})}{2}\right) \leq 0\right\}$$

$$= \sum_{\tilde{k} \neq k_i^0} \Pr\left\{\frac{2}{T}\sum_t \left(\hat{\delta}_t(k_i^0) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k_i^0) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\right.$$

$$\left. \cdot \left(\dot{\delta}_t^0(k_i^0) - \frac{\hat{\delta}_t(k_i^0) + \hat{\delta}_t(\tilde{k})}{2} + \dot{X}_{it}^\intercal(\theta^0 - \hat{\theta}) + \dot{\beta}_{t-E_i}^0(k_i^0) - \frac{\hat{\beta}_{t-E_i}(k_i^0) + \hat{\beta}_{t-E_i}(\tilde{k})}{2} + \dot{U}_{it}\right) \leq 0\right\}$$

$$\leq \sum_k \sum_{\tilde{k} \neq k} \Pr\left\{\frac{2}{T}\sum_t \left(\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\right.$$

$$\left. \cdot \left(\dot{\delta}_t^0(k) - \frac{\hat{\delta}_t(k) + \hat{\delta}_t(\tilde{k})}{2} + \dot{X}_{it}^\intercal(\theta^0 - \hat{\theta}) + \dot{\beta}_{t-E_i}^0(k) - \frac{\hat{\beta}_{t-E_i}(k) + \hat{\beta}_{t-E_i}(\tilde{k})}{2} + \dot{U}_{it}\right) \leq 0\right\}.$$

The first inequality is from the Step 2 of the K-means algorithm. Let

$$A_{ik\tilde{k}} = \frac{1}{T}\sum_t \left(\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\dot{U}_{it}$$

$$+ \frac{1}{T}\sum_t \left(\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\dot{X}_{it}^\intercal(\theta^0 - \hat{\theta})$$

$$+ \frac{1}{T}\sum_t \left(\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)$$

$$\cdot \left(\dot{\delta}_t^0(k) - \frac{\hat{\delta}_t(k) + \hat{\delta}_t(\tilde{k})}{2} + \dot{\beta}_{t-E_i}^0(k) - \frac{\hat{\beta}_{t-E_i}(k) + \hat{\beta}_{t-E_i}(\tilde{k})}{2}\right)$$

$$B_{ik\tilde{k}} = \frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)\dot{U}_{it}$$

$$+ \frac{1}{2T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)^2.$$

Note that $A_{ik\tilde{k}}$ depends on the estimator $(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma})$ while $B_{ik\tilde{k}}$ does not. Then,

$$\Pr\left\{\hat{k}_i \neq k_i^0\right\} \leq \sum_k \sum_{\tilde{k} \neq k} \Pr\left\{A_{ik\tilde{k}} \leq 0\right\} \leq \sum_k \sum_{\tilde{k} \neq k} \Pr\left\{B_{ik\tilde{k}} \leq \left|B_{ik\tilde{k}} - A_{ik\tilde{k}}\right|\right\} \tag{12}$$

We will show that $A_{ik\tilde{k}}$ and $B_{ik\tilde{k}}$ are sufficiently close to each other and that $\Pr\left\{B_{ik\tilde{k}} \leq 0\right\}$ converges to zero sufficiently fast.

$$\left|B_{ik\tilde{k}} - A_{ik\tilde{k}}\right|$$

$$= \left|\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(k)\right)\dot{U}_{it}\right| + \left|\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(\tilde{k}) - \hat{\delta}_t(\tilde{k}) + \dot{\beta}_{t-E_i}^0(\tilde{k}) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\dot{U}_{it}\right|$$

$$+ \left|\frac{1}{T}\sum_t \left(\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\dot{X}_{it}^{\intercal}(\theta^0 - \hat{\theta})\right|$$

$$+ \left|\frac{1}{2T}\sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(k)\right)\right.$$

$$\left. \cdot \left(-\dot{\delta}_t^0(k) + \dot{\delta}_t^0(\tilde{k}) - \dot{\beta}_{t-E_i}^0(k) + \dot{\beta}_{t-E_i}^0(\tilde{k}) + \hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\right|$$

$$+ \left|\frac{1}{2T}\sum_t \left(\dot{\delta}_t^0(\tilde{k}) - \hat{\delta}_t(\tilde{k}) + \dot{\beta}_{t-E_i}^0(\tilde{k}) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\right.$$

$$\left. \cdot \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k}) + \hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\right|.$$

We apply Cauchy-Schwartz's inequality to each of the five terms so that we can use the consistency result in (10). For the first term,

$$\left|\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(k)\right)\dot{U}_{it}\right|$$

$$\leq \left(\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(k)\right)^2\right)^{\frac{1}{2}} \left(\frac{1}{T}\sum_t \dot{U}_{it}^2\right)^{\frac{1}{2}} \tag{13}$$

and similarly for the second term. As for the third term, from A5.a,

$$\left|\frac{1}{T}\sum_t \left(\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right)\dot{X}_{it}^{\intercal}(\theta^0 - \hat{\theta})\right|$$

$$\leq \frac{1}{T}\sum_t \left|\hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k})\right| \cdot \|\dot{X}_{it}\| \cdot \left\|\theta^0 - \hat{\theta}\right\|$$

$$\leq 4M\left(\frac{1}{T}\sum_t \|\dot{X}_{it}\|\right) \cdot \left\|\theta^0 - \hat{\theta}\right\| \tag{14}$$

32

Last, for the fourth term, from A5.a,

$$
\left| \frac{1}{2T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(k) \right) \right.
$$

$$
\left. \cdot \left(-\dot{\delta}_t^0(k) + \dot{\delta}_t^0(\tilde{k}) - \dot{\beta}_{t-E_i}^0(k) + \dot{\beta}_{t-E_i}^0(\tilde{k}) + \hat{\delta}_t(k) - \hat{\delta}_t(\tilde{k}) + \hat{\beta}_{t-E_i}(k) - \hat{\beta}_{t-E_i}(\tilde{k}) \right) \right|
$$

$$
\leq 4M \left(\frac{1}{T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-E_i}^0(k) - \hat{\beta}_{t-E_i}(k) \right)^2 \right)^{\frac{1}{2}} \tag{15}
$$

From A5.c, both $\frac{1}{T} \sum_t \dot{U}_{it}^2$ and $\frac{1}{T} \sum_t ||\dot{X}_{it}||$ are bounded in expectation by the same bound for every T and thus $O_p(1)$. To use (10), choose an arbitrary $\eta > 0$ and $\tilde{E} \geq 0$ to focus only on the event of

$$
\left\| \theta^0 - \hat{\theta} \right\|, \left(\frac{1}{T} \sum_t \left(\dot{\delta}_t^0(k) - \hat{\delta}_t(k) + \dot{\beta}_{t-e}^0(k) - \hat{\beta}_{t-e}(k) \right)^2 \right)^{\frac{1}{2}} < \eta \tag{16}
$$

for all k and $e \leq \tilde{E}$ such that $\mu_e(k) > 0$. When (16) is true and $E_i \leq \tilde{E}$, with some constant $C > 0$,

$$
\left| B_{ik\tilde{k}} - A_{ik\tilde{k}} \right| \leq \eta C \left(\left(\frac{1}{T} \sum_t \dot{U}_{it}^2 \right)^{\frac{1}{2}} + \frac{1}{T} \sum_t ||\dot{X}_{it}|| + 1 \right).
$$

Let $D(\eta, \tilde{E})$ be a binary random variable which equals one if (16) holds true for all k and $e \leq \tilde{E}$. From A6.f,

$$
\Pr \left\{ B_{ik\tilde{k}} \leq \left| B_{ik\tilde{k}} - A_{ik\tilde{k}} \right|, D(\eta, \tilde{E}) = 1, E_i \leq \tilde{E} \right\}
$$

$$
\leq \Pr \left\{ B_{ik\tilde{k}} \leq \eta C \left(\left(\frac{1}{T} \sum_t \dot{U}_{it}^2 \right)^{\frac{1}{2}} + \frac{1}{T} \sum_t ||\dot{X}_{it}|| + 1 \right) \right\}
$$

$$
\leq \Pr \left\{ \frac{1}{T} \sum_t \dot{U}_{it}^2 \geq M^{*2} \right\} + \Pr \left\{ \frac{1}{T} \sum_t ||\dot{X}_{it}|| \geq M^* \right\} + \Pr \left\{ B_{ik\tilde{k}} \leq \eta C(2M^* + 1) \right\}. \tag{17}
$$

Note that the first inequality holds for every η and $\tilde{E}$ and C does not depend neither on η nor on $\tilde{E}$.

Now, we let $\Pr \left\{ \frac{1}{T} \sum_i \dot{U}_{it}^2 \geq M^{*2} \right\}, \Pr \left\{ \frac{1}{T} \sum_i ||\dot{X}_{it}|| \geq M^* \right\}, \Pr \left\{ B_{ik\tilde{k}} \leq \eta C(2M^* + 1) \right\} \to 0$. Note that the first two quantities are $o(T^{-\nu})$ for any $\nu > 0$ from A6.f. For the last quantity, let $\eta^* = \frac{\varepsilon^*}{4C(2M^* + 1)}$ with

33

$\varepsilon^* > 0$ from A6.e. Then,

$$\Pr\left\{B_{ik\tilde{k}} \leq \eta^* C(2M^* + 1)\right\}$$

$$\leq \Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)\dot{U}_{it} \leq \eta^* C(2M^* + 1) - \frac{\varepsilon^*}{2}\right\}$$

$$+ \Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)^2 \leq \varepsilon^*\right\}$$

$$\leq \Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)\dot{U}_{it} \leq -\frac{\varepsilon^*}{4}\right\}$$

$$+ \Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)^2 \leq \varepsilon^*\right\}.$$

From A6.e, for any $\nu > 0$,

$$T^\nu \Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)\dot{U}_{it} \leq -\frac{\varepsilon^*}{4}\right\}$$

$$\leq \sum_{e=0}^{T_1} T^\nu \Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-e}^0(k) - \dot{\beta}_{t-e}^0(\tilde{k})\right)\dot{U}_{it} \leq -\frac{\varepsilon^*}{4}\right\}$$

$$\leq T^{1+\nu}\exp(-d_2 T) = o(1)$$

and likewise for $\Pr\left\{\frac{1}{T}\sum_t \left(\dot{\delta}_t^0(k) - \dot{\delta}_t^0(\tilde{k}) + \dot{\beta}_{t-E_i}^0(k) - \dot{\beta}_{t-E_i}^0(\tilde{k})\right)^2 \leq \varepsilon^*\right\}$. Here, all control units, i.e. $E_i = \infty$ and rewritten as $E_i = T_1$. This abuse of notation is harmless thanks to A4.

Finally, going back to (12) and (17), thanks to K being fixed,

$$\Pr\left\{\hat{k}_i \neq k_i^0, D(\eta^*, \tilde{E}) = 1, E_i \leq \tilde{E}\right\} = o(T^{-\nu}) \tag{18}$$

uniformly for $\tilde{E}$.

Step 4

In this step let us discuss the probability of assigning a wrong type at least to one unit. As $N, T \to \infty$, for any $\nu > 0$

$$\Pr\left\{\sup_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} > 0\right\}$$

$$\leq \Pr\left\{\sum_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} > 0, D(\eta^*, \tilde{E}) = 1, \max_i E_i \leq \tilde{E}\right\} + \Pr\{D(\eta^*, \tilde{E}) = 0\} + \Pr\left\{\max_i E_i > \tilde{E}\right\}$$

$$\leq N \cdot \Pr\left\{\hat{k}_i \neq k_i^0, D(\eta^*, \tilde{E}) = 1, E_i \leq \tilde{E}\right\} + \Pr\{D(\eta^*, \tilde{E}) = 0\} + \Pr\left\{\max_i E_i > \tilde{E}\right\}$$

$$= o(NT^{-\nu}) + o(1).$$

The last equality holds from A6.c and (18). (18) holds uniformly for any $\tilde{E}$ and thus we can choose $\tilde{E}$ to have $\Pr\{D(\eta^*, \tilde{E}) = 0\} + \Pr\left\{\max_i E_i > \tilde{E}\right\} = o(1)$ as $N, T \to \infty$ without jeopardizing the convergence of the first term.

Step 5

Now, let us consider the OLS estimator given the type assignment is known to econometrician.

$$\left(\hat{\theta}^{ols}, \hat{\delta}^{ols}, \hat{\beta}^{ols}\right) = \arg\min_{\theta, \delta, \beta} \frac{1}{NT} \left(\dot{Y}_{it} - \delta_t(k_i^0) - \dot{X}_{it}^{\mathsf{T}}\theta - \sum_{r \neq -1; r = -l}^{T_1 - 1} \beta_r(k_i^0)\mathbf{1}_{\{t = E_i + r\}}.\right)^2$$

$$= \arg\min_{\theta, \delta, \beta} \widehat{Q}^{ols}(\theta, \beta, \delta).$$

Fix $\varepsilon, \nu > 0$. With any $\tilde{E} > 0$ and $\eta \leq \eta^*$,

$$\Pr\left\{\left|\widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta})\right| > \varepsilon T^{-\nu}\right\}$$

$$\leq \Pr\{D(\eta, \tilde{E}) = 0\} + \Pr\left\{\max_i E_i > \tilde{E}\right\}$$

$$+ \Pr\left\{\left|\widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta})\right| > \varepsilon T^{-\nu}, D(\eta, \tilde{E}) = 1, \max_i E_i \leq \tilde{E}\right\}. \tag{19}$$

To suppress the second probability, apply Cauchy-Schwartz's inequality to get

$$
\left| \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta}) \right| = \left| \frac{2}{NT} \sum_i \sum_t \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \left(\hat{\delta}_t(k_i^0) - \hat{\delta}_t(\hat{k}_i) + \hat{\beta}_{t-E_i}(k_i^0) - \hat{\beta}_{t-E_i}(\hat{k}_i) \right) \right.
$$

$$
\left. \cdot \left(\dot{Y}_{it} - \frac{\hat{\delta}_t(k_i^0) + \hat{\delta}_t(\hat{k}_i)}{2} - \dot{X}_{it}^{\intercal}\hat{\theta} - \frac{\hat{\beta}_{t-E_i}(k_i^0) + \hat{\beta}_{t-E_i}(\hat{k}_i)}{2} \right) \right|
$$

$$
\leq 2 \left(\frac{1}{NT} \sum_i \sum_t \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \right)^{\frac{1}{2}} A_{N,T} = 2 \left(\frac{1}{N} \sum_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \right)^{\frac{1}{2}} A_{N,T}.
$$

Here, $A_{N,T}$ indicates the remainder term from applying Cauchy-Schwartz's inequality. From A5.a and A5.c, $A_{N,T}$ is bounded by a uniform bound for all N, T. With some constant $C > 0$ such that $\Pr\{A_{N,T} > C\} < \tilde{\varepsilon}$ for large N, T,

$$
\Pr \left\{ \left| \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta}) \right| > \varepsilon T^{-\nu}, D(\eta, \tilde{E}) = 1, \max_i E_i \leq \tilde{E} \right\}
$$

$$
\leq \Pr \left\{ \left(\frac{1}{N} \sum_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \right)^{\frac{1}{2}} A_{N,T} > \varepsilon T^{-\nu}, D(\eta, \tilde{E}) = 1, \max_i E_i \leq \tilde{E} \right\}
$$

$$
\leq \Pr \left\{ \left(\frac{1}{N} \sum_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \right) \cdot \frac{C^2}{\varepsilon^2 T^{-2\nu}} > 1, D(\eta, \tilde{E}) = 1, \max_i E_i \leq \tilde{E} \right\} + \tilde{\varepsilon}
$$

$$
\leq \Pr \left\{ \frac{C^2}{\varepsilon^2 T^{-2\nu}} \left(\frac{1}{N} \sum_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \mathbf{1}_{\{E_i \leq \tilde{E}\}} \right) D(\eta, \tilde{E}) > 1 \right\} + \tilde{\varepsilon}
$$

$$
\leq \frac{C^2}{\varepsilon^2 T^{-2\nu}} \mathbf{E} \left[\frac{1}{N} \sum_i \mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \mathbf{1}_{\{E_i \leq \tilde{E}\}} D(\eta, \tilde{E}) \right] + \tilde{\varepsilon} = \frac{C^2}{\varepsilon^2 T^{-2\nu}} \mathbf{E} \left[\mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \mathbf{1}_{\{E_i \leq \tilde{E}\}} D(\eta, \tilde{E}) \right] + \tilde{\varepsilon}
$$

Note that $\mathbf{E} \left[\mathbf{1}_{\{\hat{k}_i \neq k_i^0\}} \mathbf{1}_{\{E_i \leq \tilde{E}\}} D(\eta, \tilde{E}) \right] = \Pr \left\{ \hat{k}_i \neq k_i^0, D(\eta, \tilde{E}) = 1, E_i \leq \tilde{E} \right\} = o(T^{-\nu})$. Thus, by the same logic for Step 4, we have

$$
\left| \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta}) \right| = o_p(T^{-\nu}).
$$

Then, by repeating the same argument for the OLS estimators, not the K-means minimizer, since we did not use any properties of $\left(\hat{\theta}, \hat{\beta}, \hat{\delta} \right)$ which come from its definition,

$$
\left| \widehat{Q}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols}) \right| = o_p(T^{-\nu}).
$$

Lastly, combining these two results,

$$
0 \leq \widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta}) - \widehat{Q}^{ols}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols})
$$

$$
= \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}^{ols}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols}) + o_p(T^{-\nu})
$$

$$
= \widehat{Q}(\hat{\theta}, \hat{\beta}, \hat{\delta}, \hat{\gamma}) - \widehat{Q}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols}, \hat{\gamma}) + o_p(T^{-\nu}) \leq o_p(T^{-\nu}).
$$

The first inequality is from the fact that the OLS estimtor minimizes $\widehat{Q}^{ols}$ and the last inequality is from the fact the K-means estimator minimizes $\widehat{Q}$. Then,

$$
\widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta}) - \widehat{Q}^{ols}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols})
$$

$$
= \frac{1}{NT} \sum_i \sum_t \left(\hat{\delta}_t^{ols}(k_i^0) - \hat{\delta}_t(k_i^0) + \dot{X}_{it}^{\mathsf{T}}(\hat{\theta}^{ols} - \hat{\theta}) + \hat{\beta}_{t-E_i}^{ols}(k_i^0) - \hat{\beta}_{t-E_i}(k_i^0) \right)
$$

$$
\cdot \left(2\dot{Y}_{it} - \hat{\delta}_t^{ols}(k_i^0) - \hat{\delta}_t(k_i^0) - \dot{X}_{it}^{\mathsf{T}}(\hat{\theta}^{ols} + \hat{\theta}) - \hat{\beta}_{t-E_i}^{ols}(k_i^0) - \hat{\beta}_{t-E_i}(k_i^0) \right)
$$

$$
= \frac{1}{NT} \sum_i \sum_t \left(\hat{\delta}_t^{ols}(k_i^0) - \hat{\delta}_t(k_i^0) + \dot{X}_{it}^{\mathsf{T}}(\hat{\theta}^{ols} - \hat{\theta}) + \hat{\beta}_{t-E_i}^{ols}(k_i^0) - \hat{\beta}_{t-E_i}(k_i^0) \right)^2
$$

The last equality is from the first order condition of the OLS estimator. Note that

$$
2\dot{Y}_{it} - \hat{\delta}_t^{ols}(k_i^0) - \hat{\delta}_t(k_i^0) - \dot{X}_{it}^{\mathsf{T}}(\hat{\theta}^{ols} + \hat{\theta}) - \hat{\beta}_{t-E_i}^{ols}(k_i^0) - \hat{\beta}_{t-E_i}(k_i^0)
$$

$$
= \underbrace{\left[2\dot{Y}_{it} - 2\hat{\delta}_t^{ols}(k_i^0) - 2\dot{X}_{it}^{\mathsf{T}}\hat{\theta}^{ols} - 2\hat{\beta}_t^{ols}(k_i^0) \right]}_{=2\hat{U}_{it}^{ols}} + \hat{\delta}_t^{ols}(k_i^0) - \hat{\delta}_t(k_i^0) + \dot{X}_{it}^{\mathsf{T}}(\hat{\theta}^{ols} - \hat{\theta}) + \hat{\beta}_{t-E_i}^{ols}(k_i^0) - \hat{\beta}_{t-E_i}(k_i^0)
$$

and the first order condition gives us

$$
\frac{1}{NT} \sum_i \sum_t Z_{it} \cdot \hat{U}_{it}^{ols} = 0
$$

for any $Z_{it} = \dot{X}_{it}, \mathbf{1}_{\{t=s\}}$ or $\mathbf{1}_{\{t=E_i+r\}}$ with any s, r.

By using the same argument as in Theorem 1 and A6.e, $\left\| \hat{\theta}^{ols} - \hat{\theta} \right\| = o_p(T^{-\nu})$ for any $\nu > 0$. As for fixed-effects, from $\left\| \hat{\theta}^{ols} - \hat{\theta} \right\| = o_p(T^{-\nu})$,

$$
\left(\frac{1}{NT} \sum_i \sum_t \left(\hat{\delta}_t^{ols}(k_i^0) - \hat{\delta}_t(k_i^0) + \hat{\beta}_{t-E_i}^{ols}(k_i^0) - \hat{\beta}_{t-E_i}(k_i^0) \right)^2 \right)^{\frac{1}{2}}
$$

$$
\leq \left(\widehat{Q}^{ols}(\hat{\theta}, \hat{\beta}, \hat{\delta}) - \widehat{Q}^{ols}(\hat{\theta}^{ols}, \hat{\beta}^{ols}, \hat{\delta}^{ols}) \right)^{\frac{1}{2}} + o_p(T^{-\nu})
$$

37

for any $\nu > 0$. Then, for any (e, k) such that $\mu_e(k) > 0$,

$$\frac{1}{T} \sum_t \left(\hat{\delta}_t^{ols}(k) - \hat{\delta}_t(k) + \hat{\beta}_{t-e}^{ols}(k) - \hat{\beta}_{t-e}(k) \right)^2 = o_p(T^{-\nu}),$$

$$\left(\hat{\delta}_t^{ols}(k) - \hat{\delta}_t(k) + \hat{\beta}_{t-e}^{ols}(k) - \hat{\beta}_{t-e}(k) \right)^2 \leq o_p(T^{1-\nu}), \quad \forall t.$$

For any $t < -l$, we have $\left(\hat{\delta}_t^{ols}(k) - \hat{\delta}_t(k) \right)^2 = o_p(T^{1-\nu})$ and thus the K-means estimator and the OLS estimator asymptotically equivalent. Note that the choice of ν was arbitrary so we can swap $1 - \nu$ with $-\nu$. If there exists another treatment timing $e' \neq e$ such that $\mu_{e'}(k) > 0$, we can iteratively recover

$$\hat{\delta}_t(k) = \hat{\delta}_t^{ols}(k) + o_p(T^{-\nu})$$

$$\hat{\beta}_r(k) = \hat{\beta}_r^{ols}(k) + o_p(T^{-\nu})$$

for any $\nu > 0$, for every t and every r estimated for type k. $\square$

Proof for Corollary 1

Consider the OLS regression with the true type assignment known:

$$\dot{Y}_{it} = \hat{\delta}_t^{ols}(k_i^0) + \sum_{r \neq -1; r=-l}^{T_1-1} \hat{\beta}_r^{ols}(k_i^0) \mathbf{1}_{\{t=E_i+r\}} + \dot{X}_{it}^{\mathsf{T}} \hat{\theta}^{ols} + \hat{U}_{it}$$

$$\dot{Y}_{it} - \dot{X}_{it}^{\mathsf{T}} \hat{\theta}^{ols} = \hat{\delta}_t^{ols}(k_i^0) + \sum_{r \neq -1; r=-l}^{T_1-1} \hat{\beta}_r^{ols}(k_i^0) \mathbf{1}_{\{t=E_i+r\}} + \hat{U}_{it}. \tag{20}$$

From A7.a-b,

$$\sqrt{NT} \left(\hat{\theta}^{ols} - \theta^0 \right) = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=T_0}^{T_1-1} \tilde{X}_{it} \tilde{X}_{it}^{\mathsf{T}} \right)^{-1} \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=T_0}^{T_1-1} \tilde{X}_{it} \dot{U}_{it} \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \Sigma_\theta^{-1} \Omega_\theta \Sigma_\theta^{-1} \right)$$

$$= O_p(1).$$

Step 1

Let $N_{ek} = \sum_i \mathbf{1}_{\{k_i^0=k, E_i=e\}}$ and $N_{\cdot k} = \sum_i \mathbf{1}_{\{k_i^0=k\}}$. Fix $\tilde{T}$ and apply the Frisch-Waugh-Lovell to (20) to get

$$\dot{Y}_{it} - \dot{X}_{it}^{\mathsf{T}}\hat{\theta}^{ols} = \sum_{r=0}^{\tilde{T}} \hat{\beta}_r^{ols}(k_i^0)\tilde{W}_{it}^r + \tilde{U}_{it}, \tag{21}$$

$$W_{it}^r = \left(\mathbf{1}_{\{t=E_i+r\}} - \frac{N_{t-r,k_i^0}}{N_{\cdot k_i^0}}\right)$$

$$\tilde{W}_{it}^r = W_{it}^r - \sum_{r' \in R^*} \left(\frac{1}{\sum_{j,s} \mathbf{1}_{\{s=E_j+r', k_j^0=k_i^0\}}} \sum_{j,s} W_{js}^r \mathbf{1}_{\{s=E_j+r', k_j^0=k_i^0\}}\right) \mathbf{1}_{\{t=E_i+r'\}}$$

where $R^* = \{-l, \cdots, -1\} \cup \{\tilde{T}+1, \cdots, T_1-1\}$ and $\tilde{U}_{it}$ and $\tilde{W}_{it}^r$ are orthogonal. W_{it}^r is the residual of $\mathbf{1}_{\{t=E_i+r\}}$ after projecting onto indicators for time fixed-effects. $\tilde{W}_{it}^r$ is the residual of W_{it}^r after projecting onto indicators for dynamic treatment effects that are not used: R^*. Every projection is done within the known type. Note that we need at least two treatment timings to apply the FWL theorem; otherwise the regressors to be projected out, the indicator for dynamic treatment effects, are in the column space of the projection matrix, which is made from the indicators for time fixed-effects.

From the true model (5),

$$\dot{Y}_{it} - \dot{X}_{it}^{\mathsf{T}}\hat{\theta}^{ols} = \dot{\delta}_t^0(k_i^0) + \sum_{r=0}^{T-1-1} \dot{\beta}_t^0(k_i^0)\mathbf{1}_{\{t=E_i+r\}} + \dot{X}_{it}^{\mathsf{T}}\left(\theta^0 - \hat{\theta}^{ols}\right) + \dot{U}_{it}.$$

With a $(\tilde{T}+1) \times 1$ vector

$$\tilde{W}_{it} = \begin{pmatrix} \tilde{W}_{it}^0 & \cdots & \tilde{W}_{it}^{\tilde{T}} \end{pmatrix}^{\mathsf{T}},$$

the OLS regression of (21) gives us

$$\begin{pmatrix} \hat{\beta}_0^{ols}(k) \\ \vdots \\ \hat{\beta}_{\tilde{T}}^{ols}(k) \end{pmatrix} = \left(\sum_{i,t} \tilde{W}_{it}\tilde{W}_{it}^{\mathsf{T}}\mathbf{1}_{\{k_i^0=k\}}\right)^{-1} \sum_{i,t} \tilde{W}_{it}\left(\dot{Y}_{it} - \dot{X}_{it}^{\mathsf{T}}\hat{\theta}^{ols}\right)\mathbf{1}_{\{k_i^0=k\}}$$

$$= \begin{pmatrix} \beta_0^0(k) \\ \vdots \\ \beta_{\tilde{T}}^0(k) - \beta_{\tilde{T}-1}^0(k) \end{pmatrix} + \left(\sum_{i,t} \tilde{W}_{it}\tilde{W}_{it}^{\mathsf{T}}\mathbf{1}_{\{k_i^0=k\}}\right)^{-1} \sum_{i,t} \tilde{W}_{it}\left(\dot{X}_{it}^{\mathsf{T}}\left(\theta^0 - \hat{\theta}^{ols}\right) + \dot{U}_{it}\right)\mathbf{1}_{\{k_i^0=k\}}.$$

Step 2

Note that $\tilde{W}_{it}$ contains nonzero elements only for finite t from A7.c. Since $0 \le E_i \le E^*$, $\mathbf{1}_{\{t=E_i+r\}}$ is zero for any $t > E^* + r$ and $t < r$. Thus, W_{it}^r is zero for any $t > E^* + r$ and $t < r$. Likewise,

$$\frac{1}{\sum_{j,s} \mathbf{1}_{\{s=E_j+r',k_j^0=k_i^0\}}} \sum_{j,s} W_{js}^r \mathbf{1}_{\{s=E_j+r',k_j^0=k_i^0\}}$$

is zero if $E_j + r' < r$ or $E_j + r' > E^* + r$ ($\Leftarrow r' < r - E^*$ or $r' > E^* + r$). Thus,

$$\sum_{r' \in R^*} \left(\frac{1}{\sum_{j,s} \mathbf{1}_{\{s=E_j+r',k_j^0=k_i^0\}}} \sum_{j,s} W_{js}^r \mathbf{1}_{\{s=E_j+r',k_j^0=k_i^0\}} \right) \mathbf{1}_{\{t=E_i+r'\}}$$

is zero if $t - E_i < r - E^*$ or $t - E_i > E^* + r$ ($\Leftarrow t < r - E^*$ or $t > 2E^* + r$). Ultimately, $\tilde{W}_{it}^r$ is nonzero only if $r - E^* \le t \le r + 2E^*$; we only need to consider $3E^* + 1$ time periods for $\tilde{W}_{it}^r$. Let $\tilde{\mathcal{T}}$ be the set of time periods t where $\tilde{W}_{it}$ is not a zero vector. $\tilde{\mathcal{T}}$ is fixed, after we fix $\tilde{T}$. From A5.b-d, A6.b and A7.c,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \dot{U}_{it} \mathbf{1}_{\{k_i^0=k\}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Omega_{\tilde{T}}\right),$$

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \tilde{W}_{it}^{\mathsf{T}} \mathbf{1}_{\{k_i^0=k\}} \xrightarrow{p} \Sigma_{\tilde{T}}.$$

as $N \to \infty$, with some positive definite matrices $\Omega_{\tilde{T}}$ and $\Sigma_{\tilde{T}}$.

Step 3

Note that as $N, T \to \infty$

$$\sqrt{N}\left(\sum_{r=0}^{\tilde{T}} \hat{\beta}_r^{ols}(k) - \beta_{\tilde{T}}^0(k) \right)$$

$$= \mathbf{1}^{\mathsf{T}} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \tilde{W}_{it}^{\mathsf{T}} \mathbf{1}_{\{k_i^0=k\}} \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \dot{X}_{it}^{\mathsf{T}} \mathbf{1}_{\{k_i^0=k\}} \cdot \sqrt{N}\left(\theta^0 - \hat{\theta}^{ols} \right)$$

$$+ \mathbf{1}^{\mathsf{T}} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \tilde{W}_{it}^{\mathsf{T}} \mathbf{1}_{\{k_i^0=k\}} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \dot{U}_{it} \mathbf{1}_{\{k_i^0=k\}}$$

$$= \mathbf{1}^{\mathsf{T}} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \tilde{W}_{it}^{\mathsf{T}} \mathbf{1}_{\{k_i^0=k\}} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t \in \tilde{\mathcal{T}}} \tilde{W}_{it} \dot{U}_{it} \mathbf{1}_{\{k_i^0=k\}} + o_p(1)$$

since $\sqrt{N}\left(\hat{\theta}^{ols} - \theta^0\right) = \frac{1}{\sqrt{T}}O_p(1) = o_p(1)$ and $\frac{1}{N}\sum_{i=1}^{N}\sum_{t\in\tilde{\mathcal{T}}}\tilde{W}_{it}\dot{X}_{it}^{\intercal}\mathbf{1}_{\{k_i^0=k\}}$ is bounded in expectation from A5.c. As $N, T \to \infty$,

$$\sqrt{N}\left(\sum_{r=0}^{\tilde{T}}\hat{\beta}_r^{ols}(k) - \beta_{\tilde{T}}^0(k)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{1}^{\intercal}\Sigma_{\tilde{T}}^{-1}\Omega_{\tilde{T}}\Sigma_{\tilde{T}}^{-1}\mathbf{1}\right).$$

Lastly, with $\nu^* > 0$ from Corollary 1,

$$\begin{aligned}
\sqrt{NT}\left(\hat{\theta} - \theta^0\right) &= \sqrt{NT}\left(\hat{\theta} - \hat{\theta}^{ols}\right) + \sqrt{NT}\left(\hat{\theta}^{ols} - \theta^0\right) \\
&= \sqrt{\frac{N}{T^{\nu^*}}}\cdot T^{\frac{1+\nu^*}{2}}\left(\hat{\theta} - \hat{\theta}^{ols}\right) + \sqrt{NT}\left(\hat{\theta}^{ols} - \theta^0\right) \\
&= \sqrt{NT}\left(\hat{\theta}^{ols} - \theta^0\right) + o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{N}\left(\sum_{r=0}^{\tilde{T}}\hat{\beta}_r(k) - \beta_{\tilde{T}}^0(k)\right) &= \sqrt{N}\left(\sum_{r=0}^{\tilde{T}}\left(\hat{\beta}_r(k) - \hat{\beta}_r^{ols}(k)\right)\right) + \sqrt{N}\left(\sum_{r=0}^{\tilde{T}}\hat{\beta}_r^{ols}(k) - \beta_{\tilde{T}}^0(k)\right) \\
&= \sqrt{\frac{N}{T^{\nu^*}}}\cdot T^{\frac{\nu^*}{2}}\left(\sum_{r=0}^{\tilde{T}}\left(\hat{\beta}_r(k) - \hat{\beta}_r^{ols}(k)\right)\right) + \sqrt{N}\left(\sum_{r=0}^{\tilde{T}}\hat{\beta}_r^{ols}(k) - \beta_{\tilde{T}}^0(k)\right) \\
&= \sqrt{N}\left(\sum_{r=0}^{\tilde{T}}\hat{\beta}_r^{ols}(k) - \beta_{\tilde{T}}^0(k)\right) + o_p(1).
\end{aligned}$$

$\square$

Extension: mean differencing

As discussed in Section 4.1.3, depending on the variation in the type-specific fixed-effects, mean-differencing may be more suitable than first-differencing, though mean-differencing requires more restrictive assumptions on the DGP. To illustrate the comparison, here I present the estimation results under mean-differencing on the application Lutz (2011).

We observe that the mean-differencing and first-differencing document the heterogeneity roughly in the same direction: type 1 where the segregation index was already rising had a bigger impact from the treatment. Table 3 contains the comparison between the type assignment estimation under mean-differencing and that under first-differencing. 14 units that were assigned type 1 under first-differencing were assigned type 2 under mean-differencing and 3 units that were assigned type 2 under first-differencing were assigned type
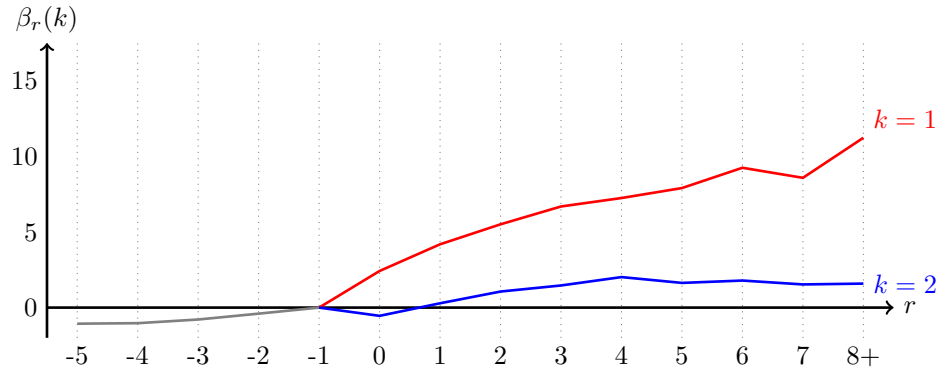
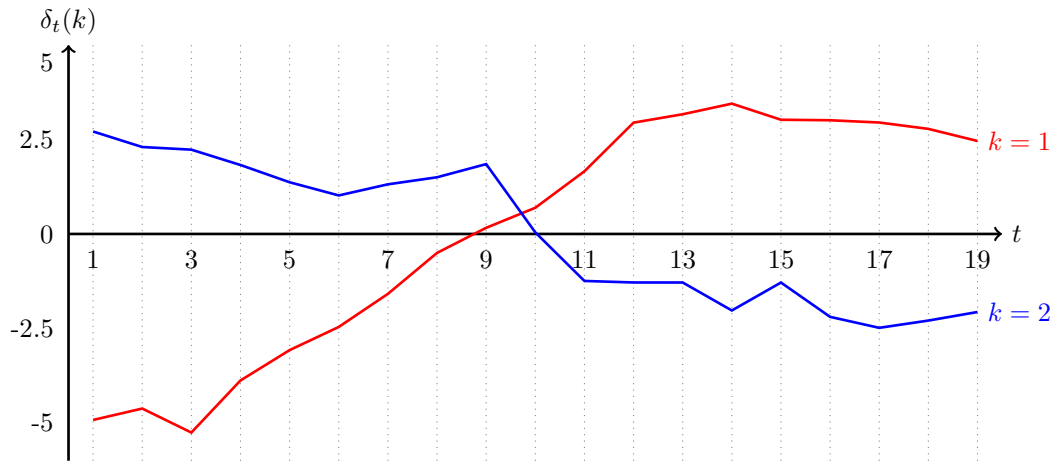Figure 4: type-specific treatment effect, $K = 2$.

Figure 5: type-specific fixed-effects, $K = 2$

	$K = 2$		
	$(k = 1)$	$(k = 2)$	Diff
Segregation index	30.22	38.61	-8.39
	(14.44)	(18.72)	(3.30)
% (white)	51.91	49.92	1.99
	(19.07)	(22.88)	(4.20)
% (hispanic)	4.09	12.82	-8.74
	(8.01)	(17.76)	(2.58)
enrollment	46047	48104	-2057
	(66614)	(87075)	(15301)
N	39	63	-
joint p-value			0.001

Table 2: Balancedness test

1 under mean-differencing; in total 17 out of 102 units were assigned differently across the specifications, implying that the type assignment estimation is fairly robust to the choice of differencing.

	type 1, FD	type 2, FD
type 1, MD	36	3
type 2, MD	14	49

Table 3: Comparison between first-differencing and mean-differencing