Fused Extended Two-Way Fixed Effects for Difference-in-Differences With Staggered Adoptions

Gregory Faletto*

University of Southern California

April 26, 2024

Abstract

To address the bias of the canonical two-way fixed effects estimator for difference-in-differences under staggered adoptions, Wooldridge (2021) proposed the *extended two-way fixed effects* estimator, which adds many parameters. However, this reduces efficiency. Restricting some of these parameters to be equal (for example, subsequent treatment effects within a cohort) helps, but *ad hoc* restrictions may reintroduce bias. We propose a machine learning estimator with a single tuning parameter, *fused extended two-way fixed effects* (FETWFE), that enables automatic data-driven selection of these restrictions. We prove that under an appropriate sparsity assumption FETWFE identifies the correct restrictions with probability tending to one, which improves efficiency. We also prove the consistency, oracle property, and asymptotic normality of FETWFE for several classes of heterogeneous marginal treatment effect estimators under either conditional or marginal parallel trends, and we prove the same results for conditional average treatment effects under conditional parallel trends. We demonstrate FETWFE in simulation studies and an empirical application.

Keywords: Difference-in-differences, staggered adoptions, treatment effect heterogeneity, bridge regression, fused lasso.

1 Introduction

We investigate difference-in-differences estimation under staggered adoptions. Suppose N units are observed at $T \geq 2$ times, where our theoretical results will assume N grows asymptotically while T is fixed. For convenience we assume a balanced panel where all units are observed at all times, so we have NT observations.

Suppose $R \leq T - 1$ cohorts start receiving a treatment at times $r \in \mathcal{R} = \{r_1, r_2, \ldots, r_R\} \subseteq \{2, \ldots, T\}$. For example, we might want to investigate the effect of a law or policy that is implemented state-by-state across time (Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Borusyak et al., 2024), or the effect of a product or service that is released across different regions over time (De Chaisemartin and d'Haultfoeuille, 2020). We consider settings where the treatment is an absorbing state, so that once a unit starts being treated, they continue receiving treatment thereafter, but this can be relaxed with little complication. Since no cohort enters at time $t = 1$ and we assume the probability of never selecting into treatment is strictly positive for all units (see Assumption F2 in Section 2), for sufficiently large N we have untreated units available as a baseline comparison for all treated cohorts at all times.

For $i \in \{1, \ldots, N\} =: [N]$ and $t \in [T]$, we encode the treatment status for each unit in a random variable $W_i \in \{0\} \cup \mathcal{R}$, where $W_i = r$ if unit i is in cohort $r \in \mathcal{R}$ or $W_i = 0$ if unit i is never

---

*gregory.faletto@marshall.usc.edu

treated. Let $\tilde{y}_{(it)}(r)$ be the potential outcome for unit i at time t if they were in cohort r. We will be interested in estimating treatment effects of the form

$$\tau_{\text{ATT}}(r,t) := \mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(it)}(0) \mid W_i = r\right], \qquad r \in \mathcal{R}, t \geq r, \tag{1.1}$$

as well as conditional treatment effects that depend on the value of unit covariates; see (1.6), to come. It is well-known that in the $T = 2$, $\mathcal{R} = \{2\}$ case, under no anticipation and parallel trends assumptions

$$\tau_{\text{ATT}}(2,2) = \mathbb{E}\left[\tilde{y}_{i2}(2) - \tilde{y}_{i1}(2) \mid W_i = 2\right] - \mathbb{E}\left[\tilde{y}_{i2}(0) - \tilde{y}_{i1}(0) \mid W_i = 0\right], \tag{1.2}$$

and this treatment effect can be estimated consistently by the canonical two-way fixed effects regression

$$\tilde{y}_{it} = \alpha_i^* + \gamma_t^* + \tau^* \cdot \sum_{r \in \mathcal{R}} \mathbb{1}\{W_i = r\}\mathbb{1}\{t \geq r\} + \epsilon_{it}, \tag{1.3}$$

where α_i^* are fixed effects corresponding to units, γ_t^* are fixed effects corresponding to time, $\mathbb{1}\{\cdot\}$ is an indicator variable that equals 1 on the event $\cdot$ and 0 otherwise, ϵ_{it} is a noise term, and for conciseness we denote $\tilde{y}_{it} := \tilde{y}_{it}(W_i)$, the observed response for unit i at time t.

In the more general $T \geq 2$ and arbitrary $\mathcal{R}$ setting, one might hope that regression (1.3) would aggregate the treatment effects in a sensible way to estimate some form of an average treatment effect, but it is now well-known that this isn't the case, and many alternative estimators have been proposed (Borusyak and Jaravel, 2018; De Chaisemartin and d'Haultfoeuille, 2020; Sun and Abraham, 2021; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021).

1.1 The Extended Two-Way Fixed Effects Estimator

Most of the alternative estimators proposed in these and other works depart from ordinary least squares (OLS) estimation, but Wooldridge (2021)[1] shows that linear regression can still be unbiased provided that enough parameters are estimated. He first shows that (1.2) is also the estimand for τ in the regression

$$\tilde{y}_{it} = \eta^* + \nu_2^* \cdot \mathbb{1}\{W_i = 2\} + \gamma_2^* + \tau^* \cdot \mathbb{1}\{W_i = r\}\mathbb{1}\{t = 2\} + \epsilon_{it}, \tag{1.4}$$

where η can be interpreted as the expected outcome for the never-treated units at $t = 1$, ν_2 is the selection bias for the treated units at $t = 1$, and γ_2 is the trend in the untreated potential outcomes. Wooldridge proposes generalizing this to the arbitrary T case and allowing for each parameter to be linear in covariates rather than constant. This leads to a linear model on cohort dummy variables, time dummies, covariates, treatment indicators, and interactions of those terms: for all $i \in [N]$, $t \in [T]$, and $r \in \mathcal{R}$,

$$\tilde{y}_{it} = \eta^* + \gamma_t^* + \boldsymbol{X}_i^\top(\boldsymbol{\kappa}^* + \boldsymbol{\xi}_t^*) + \sum_{r \in \mathcal{R}} \mathbb{1}\{W_i = r\}\left(\nu_r^* + \boldsymbol{X}_i^\top \boldsymbol{\zeta}_r^* + \mathbb{1}\{t \geq r\}(\tau_{rt}^* + \dot{\boldsymbol{X}}_{(ir)}^\top \boldsymbol{\rho}_{rt}^*)\right) + \epsilon_{it}. \tag{1.5}$$

Here $\boldsymbol{X}_i \in \mathbb{R}^d$ is a set of time-invariant (pre-treatment) covariates, $\eta^* + \boldsymbol{X}_i^\top \boldsymbol{\kappa}^*$ is the conditionally expected outcome of the never-treated units at $t = 1$, the $\gamma_t^* + \boldsymbol{X}_i^\top \boldsymbol{\xi}_t^*$ terms are the conditional trends in the untreated potential outcomes, $\nu_r^* + \boldsymbol{X}_i^\top \boldsymbol{\zeta}_r^*$ is the conditional selection bias for cohort r, and $\tau_{rt}^* + \dot{\boldsymbol{X}}_{(ir)}^\top \boldsymbol{\rho}_{rt}^*$ is the conditional treatment effect for cohort r at time t. (See Assumption (LINS) in Section 3.2 for more details.)

Wooldridge calls (1.5) the *extended two-way fixed effects* (ETWFE) estimator (Wooldridge, 2021, Equation 6.33) and shows that in this regression τ_{rt}^* is unbiased for $\tau_{\text{ATT}}(r,t)$ under conditional no-anticipation and parallel trends assumptions; see Assumptions (CNAS) and (CCTS) in Section 3. The ETWFE estimator can also be used to estimate conditional average treatment effects on the treated units

$$\tau_{\text{CATT}}(r,t,\boldsymbol{x}) := \mathbb{E}[\tilde{y}_{(it)}(r) - \tilde{y}_{(it)}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}], \qquad i \in [N], r \in \mathcal{R}, t \in \{r, \dots, T\}. \tag{1.6}$$

---

[1]Along with Sun and Abraham (2021).

1.2 Restrictions in the ETWFE Model

As Wooldridge (2021, Section 6.5) points out, the ETWFE estimator (1.5) may contain a large number of parameters to estimate if R, T, or d is moderately large relative to N. In this case, ordinary least squares estimation of (1.5) may lead to very imprecise estimates of at least some of the parameters, trading the bias problem that model (1.3) has for a low-efficiency (or high-variance) problem.

Wooldridge (2021) suggests some possible ways to mitigate this problem by imposing restrictions in order to reduce the number of parameters to be estimated. For example, one might assume that the treatment effects only vary across cohorts and not with time, so that $\tau_{rt} = \tau_r$ for all $r \in \mathcal{R}$, $t \geq r$. (Other works that consider similar restrictions include Borusyak and Jaravel (2018) and Borusyak et al. (2024).) If many of these restrictions hold, then imposing them results in more efficient estimation of the treatment effects than regression (1.5). However, imposing these restrictions in an *ad hoc* way could reintroduce bias if the wrong restrictions are imposed, or leave unnecessary inefficiency if not enough restrictions are imposed.

The first goal of this paper is to propose a novel estimator that imposes these restrictions in an automatic, data-driven way. We propose the *fused extended two-way fixed effects* (FETWFE) estimator, a novel machine learning method for difference-in-differences estimation. Before we describe FETWFE, we discuss sparse fusion penalties in linear regression, which play a key role in our estimator.

1.3 Sparse Fusion Penalties

Given a $N \times p$ real-valued design matrix $\boldsymbol{X}$ and a response $\boldsymbol{y} \in \mathbb{R}^N$, the well-known lasso (Tibshirani, 1996) estimator is defined by the optimization problem

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}(\lambda_N) \in \ &\operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i \right)^2 + \lambda_N \sum_{j=1}^{p} |\beta_j| \right\} \\
= \ &\operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_N \|\boldsymbol{\beta}\|_1 \right\},
\end{aligned}
$$

where λ_N is a tuning parameter that can be chosen from a set of candidate values by cross-validation or a criterion like BIC, and recall that the ℓ_q norm[2] of an n-vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top$ is defined as $\|\boldsymbol{v}\|_q := \left(\sum_{i=1}^{n} |v_i|^q \right)^{1/q}$. The ℓ_1 penalty term leads to *sparse* estimated coefficients $\hat{\boldsymbol{\beta}}(\lambda_N)$ with entries set exactly equal to 0.

The *fused lasso* (Tibshirani et al., 2005) adds a second tuning parameter and ℓ_1 penalty term:

$$
\hat{\boldsymbol{\beta}}(\lambda_N^{(1)}, \lambda_N^{(2)}) \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i \right)^2 + \lambda_N^{(1)} \sum_{j=1}^{p} |\beta_j| + \lambda_N^{(2)} \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right\}.
$$

The fused lasso tends to yield solutions that both set some individual coefficients exactly equal to 0 (due to the first ℓ_1 penalty term) and also set the differences between some adjacent pairs of coefficients exactly equal to 0 (due to the second ℓ_1 penalty term). That is, some adjacent pairs of coefficients are set exactly equal to each other, or *fused* together.

FETWFE uses ℓ_q fusion penalties for $q \in (0, 2]$ (*bridge* regularization) to shrink certain sets of parameters in (1.5) towards each other in a way that aligns with intuition and previously proposed restrictions. See Figure 1 for a visual depiction of which of the marginal average treatment effects $\hat{\tau}_{rt}$ we propose penalizing towards each other, and see further details about the FETWFE penalty term (5.1) in Section 5. A lasso estimator is within our framework if one chooses $q = 1$. Choices

---

[2]It is a norm for finite $q \geq 1$, though we also consider $q \in (0, 1)$, in which case $\|\cdot\|_q$ does not satisfy the triangle inequality.

of $q \in (0, 1]$ will result in sparse solutions, and choices of $q \in (0, 1)$ will allow for convergence at $1/\sqrt{N}$ rates and asymptotic normality of the estimated coefficients. These fusion penalties allow the restrictions to be selected in a hands-off, automatic way that comes with theoretical guarantees.
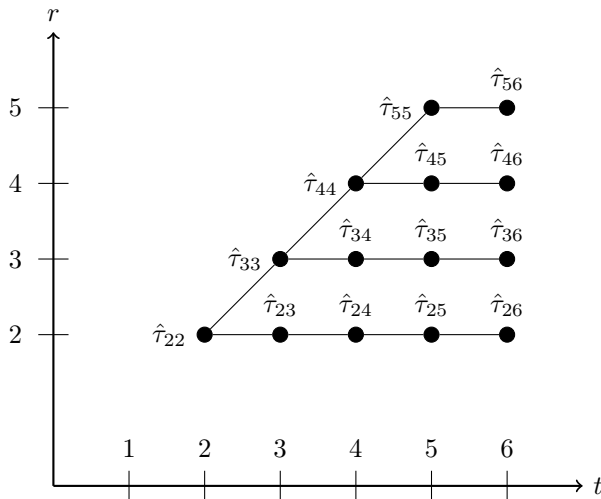

Figure 1: *Visualization of which of the estimated marginal average treatment effect terms $\hat{\tau}_{rt}$ from regression (1.5) (which estimate the average treatment effects $\tau_{ATT}(r, t)$ from Equation 1.1) we penalize towards each other in the FETWFE penalty (5.1). In this setting, $T = 6$ and $\mathcal{R} = \{2, \ldots, 5\}$. The horizontal axis depicts time and the vertical axis depicts cohorts. FETWFE works well under an assumption that the linked treatment effects tend to be close together, and at least some of them are exactly equal. See further details in Section 5.*

1.4 Demonstrating the Efficacy of FETWFE

The second goal of this paper is to provide theoretical guarantees for the performance of FETWFE. We prove that under a suitable sparsity assumption on the true coefficients and regularity conditions, FETWFE learns the correct restrictions with probability tending to one (Theorem 6.2). FETWFE then estimates only the needed parameters, improving efficiency. We also prove that for $q \in (0, 2]$ our proposed estimators of two classes of heterogeneous marginal average treatment effects are consistent under either conditional or marginal parallel trends, and two classes of conditional average treatment effect estimators are consistent under conditional parallel trends (Theorem 6.1). To the best of our knowledge, this is the first work establishing the consistency of a difference-in-differences estimator for marginal average treatment effects under either form of parallel trends. For estimating propensity-weighted conditional average treatment effects, the generalized propensity scores can be estimated by an arbitrary nonparametric estimator with theoretical convergence guarantees, so our theory applies to propensity score estimators as simple as multinomial logit or as flexible as deep neural networks (Farrell et al., 2021) or random forests (Gao et al., 2022).

We also show that our treatment effect estimators are asymptotically normal for $q \in (0, 1)$ and we provide feasible, provably consistent finite-sample variance estimators (Theorems 6.4 and D.3). For probability-weighted treatment effect estimates, we propose both an asymptotically normal test statistic if the generalized propensity scores are estimated on a separate data set from FETWFE (Theorem 6.4b) and an asymptotically subgaussian test statistic if both models are estimated on the same data set (Theorem 6.4c). The data-driven restrictions FETWFE learns improve parameter estimation while preserving asymptotic unbiasedness and consistency, allowing for the construction of asymptotically valid confidence intervals.

Under the assumptions of Theorem 6.3 we show that FETWFE has an oracle property (Fan and Li, 2001; Fan and Peng, 2004): the asymptotic variance of its estimates depend only on the parameters of the smaller restricted model, and it converges at the same $1/\sqrt{N}$ rate as an OLS-estimated model on the true restricted model even if the number of covariates grows to infinity with

N. Our theoretical guarantees leverage extensions of the bridge regression theory of Kock (2013) (Theorems F.4 and F.5 in the appendix) that may be of independent interest.

Finally, a third goal of this paper is to show that FETWFE is useful in practice through simulation studies that we conduct in Section 7 comparing the FETWFE estimator to competitor estimators. We also illustrate FETWFE in an empirical application in Section 8.

1.5 Related Literature

There has been an abundance of recent research on difference-in-differences and it is not possible to cite all of the work in this stream; see De Chaisemartin and d'Haultfoeuille (2023); Sun and Shapiro (2022); Roth et al. (2023), and Callaway (2023) for recent reviews. Lechner et al. (2011) reviews earlier literature. Chiu et al. (2023) review the literature from a political science perspective. de Chaisemartin and D'Haultfœuille (2023) and Arkhangelsky and Imbens (2023) review causal inference in panel data more broadly. Beyond those reviews and the papers we have already cited, throughout the paper we cite papers in the difference-in-differences literature and relate them to FETWFE; see Sections 3 and 4 and Appendix A, in particular.

Besides Wooldridge (2021), there are many other estimators that target similar estimands to ours, and we discuss several of these in Section 4, though none of these are machine learning estimators. It is also natural to compare FETWFE to other machine learning and non- and semi-parametric difference-in-differences estimators. Generally these methods are more flexible in the kinds of models than they can estimate than FETWFE (which is a regularized linear model, though our theoretical results allow for nonparametric estimation of the generalized propensity scores), but much less flexible in the estimands they target. Further, these methods tend to require assumptions that an estimator is available that achieves consistency at particular rates, so applying these results would involve further checking the required assumptions for a particular estimator. In contrast, we provide "end-to-end" assumptions for the consistency of three classes of estimators (5.6), (5.9), and (5.10) and the oracle property and asymptotic normality of (5.9) and (5.10).

The parametric and semi-parametric estimators proposed by Zimmert (2020), Sant'Anna and Zhao (2020), and Chang (2020) yield theoretical results for a wide variety of machine learning estimators with theoretical guarantees under conditional parallel trends assumptions, but only in the $T = 2$ setting and only for a single average treatment effect estimand (specifically, our estimand in Equation 4.11). Blackwell et al. (2024) accomplish similar goals with different estimands under different assumptions, focusing mainly on a setting with $T = 2$ and a single mediator of interest. Nie et al. (2021) propose related non-parametric estimators that achieve $1/\sqrt{N}$ convergence and asymptotic normality in the $T = 2$ repeated cross-sections setting only for an average treatment effect, and consistency for conditional average treatment effects at much slower rates, in contrast to the $1/\sqrt{N}$ rate for conditional average treatment effects estimated by FETWFE (see Theorem D.3). Hatamyar et al. (2023) provide similar semi-parametric estimators for both marginal and conditional average treatment effects in the staggered adoptions setting with arbitrary T, but they do not provide any theoretical guarantees.

Gavrilova et al. (2023) propose an approach using random forests that is somewhat more flexible in its estimands. Their approach allows for arbitrary T, though they focus on the case with only one treated cohort that starts an absorbing treatment at time $t = 2$, providing a brief sketch as to how their method might be extended to allow for staggered adoptions. Further, their method requires splitting the data and estimating completely separate models to estimate the treatment effect of the treated cohort at each time $2, 3, \ldots, T$, while FETWFE allows for the full sample to be used to jointly estimate all treatment effects simultaneously. Indeed, a central idea of our method is that the fusion penalties we use allow FETWFE to borrow strength across cohorts and time in order to improve estimation efficiency. (None of the above works mentions fusion penalties.)

Our work also contributes to the much broader topic of estimating conditional average treatment effects. In Section 4 we mention some prominent works in this stream of literature.

Outside of causal inference, our approach is clearly closely related to both the fused lasso (Tib-

shirani et al., 2005) and the generalized lasso (Tibshirani and Taylor, 2011); see (5.4). Our work also draws much from the bridge regression theory of Kock (2013), and, in turn, Huang et al. (2008). Kaddoura and Westerlund (2022) propose a linear panel data model with a fused lasso-like penalty, though not for causal inference. Park and Yamauchi (2023) propose a related Bayesian approach. Heiler and Mareckova (2021) also propose a penalized panel data model that combines categories, but not in the causal inference setting. Similarly, Kwon (2021) proposes a shrinkage estimator for linear panel data models (though not for causal inference) and cites more related work in this stream.

FETWFE also relates more specifically to the generalized fused lasso (Höfling et al., 2010; Xin et al., 2014). Like FETWFE, the generalized fused lasso involves applying ℓ_q penalties to the absolute differences between a variety of pairs of coefficients, though the generalized fused lasso only uses the $q = 1$ case. Some works that prove theoretical results for the generalized fused lasso include Viallon et al. (2013) and Viallon et al. (2016). Viallon et al. (2013) prove consistency for the generalized fused lasso under different assumptions from ours, including a restriction to the fixed p case. They prove selection consistency and asymptotic normality only for an *adaptive* (Zou, 2006) version of their method. Viallon et al. (2016) prove similar results (with similar limitations) for a broader class of generalized linear models.

1.6 Organization of the Paper

After we introduce some notation, we establish our setup and basic assumptions in Section 2 and present our difference-in-differences assumptions in Section 3. In particular, we provide a testable sufficient condition (CIUN) for several kinds of estimators, including FETWFE, to be consistent under either marginal or conditional parallel trends. In Section 4 we introduce causal estimands that FETWFE can be used to estimate, and in Section 5 we propose the fused extended two-way fixed effects model, as well as our estimators for our proposed causal estimands. We present our theoretical guarantees for FETWFE in Section 6. In Section 7 we demonstrate the superior performance of FETWFE in our setting against extended two-way fixed effects (ETWFE), bridge-penalized ETWFE, and a slightly more flexible model than (1.3) in simulation studies. We illustrate the usefulness of FETWFE in an empirical application using data from Stevenson and Wolfers (2006) in Section 8. Finally, Section 9 concludes.

1.7 Notation

Throughout we will use subscripts like $\tilde{y}_{(it)}$ to refer to the ith unit at the tth time, $\tilde{\boldsymbol{y}}_{(i\cdot)} \in \mathbb{R}^T$ to refer to all response observations corresponding to unit i, and so on. For matrices, $\boldsymbol{Z}_{(it)\cdot}$ refers to the row corresponding to unit i at time t, $\boldsymbol{Z}_{(i\cdot)j}$ refers to the T observations of the jth feature in unit i, and so on. We will assume that unit-specific covariates are time invariant, and it will be more convenient to refer to the covariates for unit i with the notation $\boldsymbol{X}_i$.

Recall that for a sequence of random variables $\{X_N\}$ and a sequence of positive real numbers $\{a_N\}$, we say that $X_N = \mathcal{O}_{\mathbb{P}}(a_N)$ if for every $\epsilon > 0$ there exists a finite $M_\epsilon > 0$ and $N_\epsilon > 0$ such that for all $N > N_\epsilon$,

$$\mathbb{P}\left(\left| \frac{X_N}{a_N} \right| > M_\epsilon \right) < \epsilon.$$

If $a_N \to 0$, $X_N = \mathcal{O}_{\mathbb{P}}(a_N)$ is sufficient for $X_N \xrightarrow{p} 0$, and a_N can be interpreted as the rate of convergence (DasGupta, 2008, Chapter 1).

2 Basic Setup and Assumptions

We will expand on the setup from the introduction. For each $i \in [N]$, assume

$$\tilde{\boldsymbol{y}}_{(i\cdot)} = \mathbb{E}[\tilde{\boldsymbol{y}}_{(i\cdot)} \mid W_i, \boldsymbol{X}_i] + c_i \mathbf{1}_T + \boldsymbol{u}_{(i\cdot)},$$

where $c_i \in \mathbb{R}$ and $\boldsymbol{u}_{(i \cdot)} \in \mathbb{R}^T$ are both random variables and $\mathbf{1}_T$ is a T-vector with 1 in every entry.

Assumption (F1): $\mathbb{E}[c_i \mid W_i, \boldsymbol{X}_i] = 0$, $\mathbb{E}[\boldsymbol{u}_{(i \cdot)} \mid c_i, W_i, \boldsymbol{X}_i] = \mathbf{0}$, $\mathrm{Var}(c_i \mid W_i, \boldsymbol{X}_i) = \sigma_c^2$, and $\mathrm{Var}(\boldsymbol{u}_{(i \cdot)} \mid c_i, W_i, \boldsymbol{X}_i) = \sigma^2 \boldsymbol{I}_T$ almost surely, where σ_c^2 and σ^2 are assumed to be known and finite and $\boldsymbol{I}_T \in \mathbb{R}^{T \times T}$ is the identity matrix.

With Assumption (F1) we can use a generalized least squares transform to create a data set with conditionally uncorrelated rows. Let $\boldsymbol{\epsilon}_{(i \cdot)} := c_i \mathbf{1}_T + \boldsymbol{u}_{(i \cdot)}$, and observe that for every $i \in [N]$ almost surely $\mathrm{Var}(\boldsymbol{\epsilon}_{(i \cdot)} \mid W_i, \boldsymbol{X}_i) = \sigma^2 \boldsymbol{I}_T + \sigma_c^2 \mathbf{1}_T \mathbf{1}_T^\top =: \boldsymbol{\Omega} \in \mathbb{R}^{T \times T}$. Under (F1), for every $i \in [N]$ almost surely $\mathbb{E}[\sigma \boldsymbol{\Omega}^{-1/2} \boldsymbol{\epsilon}_{(i \cdot)} \mid W_i, \boldsymbol{X}_i] = \mathbf{0}$ and

$$\mathrm{Var}(\sigma \boldsymbol{\Omega}^{-1/2} \boldsymbol{\epsilon}_{(i \cdot)} \mid W_i, \boldsymbol{X}_i) = \sigma^2 \boldsymbol{\Omega}^{-1/2} \mathrm{Var}(\boldsymbol{\epsilon}_{(i \cdot)} \mid W_i, \boldsymbol{X}_i) \boldsymbol{\Omega}^{-1/2} = \sigma^2 \boldsymbol{I}_T. \tag{2.1}$$

We can carry out this transformation for any known finite and symmetric positive definite $\boldsymbol{\Omega} = \mathrm{Var}(\boldsymbol{\epsilon}_{(i \cdot)} \mid W_i, \boldsymbol{X}_i)$. All of our results can be easily adapted to apply for any such $\boldsymbol{\Omega}$, though we focus on the specified setting for concreteness.

Next we make an assumption on the joint distribution of the potential outcomes, treatment, and covariates. Let

$$\pi_r(\boldsymbol{x}) := \mathbb{P}(W_i = r \mid \boldsymbol{X}_i = \boldsymbol{x}), \qquad r \in \{0\} \cup \mathcal{R} \tag{2.2}$$

be the conditional probability of a unit's treatment status, a propensity score.

Assumption (F2): The N joint random variables $(W_i, \boldsymbol{X}_i, c_i, \boldsymbol{u}_{(i \cdot)})_{i=1}^N$ are independent and identically distributed (iid). Further, assume that for all $r \in \{0\} \cup \mathcal{R}$, almost surely $0 < \pi_r(\boldsymbol{X}_i) < 1$.

Notice that for each $\boldsymbol{X}_i$ all of the potential outcomes $\tilde{y}_{it}(r)$ are well-defined under Assumption (F2) since treatment probabilities are always strictly between 0 and 1. So far we have made no assumptions on the relationship between the treatment assignments and the potential outcomes for a given unit, but we do so in the next section.

3 Difference-in-Differences Assumptions

We will use the generalizations of the no anticipation and common trends assumptions described in the introduction—Assumptions (CNAS) and (CCTS), respectively—that were proposed by Wooldridge (2021), in addition to some variations and other assumptions.

Assumption (CNAS) ("conditional no-anticipation with staggered interventions"): For treatment cohorts $r \in \mathcal{R}$, almost surely

$$\mathbb{E}[\tilde{y}_{(it)}(r) - \tilde{y}_{(it)}(0) \mid W_i = r, \boldsymbol{X}_i] = 0 \qquad \forall i \in [N], t < r.$$

This requires that before treatment, in conditional expectation each individual's expected outcome is identical to what it would have been if the unit were never treated. Some works (two recent examples are Caetano et al. (2022) and Ghanem et al. 2024) define the potential outcomes of the units at each time in terms of their treatment status at that time rather than in terms of their cohort membership. In our notation, this amounts to assuming that $\tilde{y}_{(it)}(r) = \tilde{y}_{(it)}(0)$ almost surely for $t < r$. As Wooldridge (2021, Section 6.3) points out, this stronger assumption is sufficient for (CNAS). Such a framework may not be troubling, but we only need the weaker assumption (CNAS). See further discussion in Wooldridge (2021, Sections 5.3, 6.1, 6.3).

3.1 Parallel Trends Assumptions

We will discuss and use a variety of parallel or common trends assumptions.

Assumption (CCTS) ("conditional common trends with staggered interventions"): For all $r \in \mathcal{R}$, $i \in [N], t \in \{2, \dots, T\}$, almost surely

$$\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i]. \tag{3.1}$$

Assumption (CCTS) requires that each unit's untreated potential outcomes $\tilde{y}_{(it)}(0)$ follow the same trend over time in conditional expectation regardless of when or if they are treated. Assumption (CCTS) is used by Callaway (2023, Assumption 5), and also by Gavrilova et al. (2023, Assumption 1) in the $\mathcal{R} = \{2\}$ case. Assumption (CCTS) has become widely adopted in the $T = 2$ setting; early examples include Heckman et al. (1997, Assumption A-5), Heckman et al. (1998, Equation 12), and Abadie (2005, Assumption 3.1), and some more recent examples include Sant'Anna and Zhao (2020, Assumption 2), Zimmert (2020, Assumption 6(iv)), Chang (2020, Assumption 2.1), Roth et al. (2023, Assumption 6), and Ghanem et al. (2024, Assumption PT-X).

We also consider versions of (CCTS) that only apply before or after treatment begins for each cohort.

Assumption (CCTSB) ("conditional common trends before treatment"): (3.1) holds almost surely for all $r \in \mathcal{R} \setminus \{2\}$, $i \in [N]$, and $t \in \{2, \ldots, r-1\}$.

Assumption (CCTSA) ("conditional common trends after treatment"): (3.1) holds almost surely for all $r \in \mathcal{R}$, $i \in [N]$, and $t \in \{r, \ldots, T\}$.

Assumption (CCTSB) requires that (CCTS) holds only until the period before each unit is treated, while (CCTSA) requires (CCTS) to hold after treatment begins. Many previous works have distinguished between the parallel trends before treatment versus after treatment, noting that (CCTSB) can be directly tested under assumption (CNAS) while (CCTSA) cannot, since untreated potential outcomes are not observed for units after treatment, and noting that parallel trends holding before treatment is no guarantee of holding after treatment. See Bilinski and Hatfield (2018); Kahn-Lang and Lang (2020); Dette (2020); Sun and Abraham (2021); Ban and Kedagni (2022); Callaway and Sant'Anna (2022); Roth (2022), Henderson and Sperlich (2022, Equation 22), Borusyak et al. (2024); Rambachan and Roth (2023), and Callaway (2023, Section 3.3). See Wooldridge (2021, Section 7.1, Equation 7.4) for an example of how (CCTSB) can be tested under (CNAS).

We can contrast (CCTS) with an unconditional or marginal version, which Wooldridge (2021) calls Assumption (CTS).

Assumption (CTS) ("common trends with staggered interventions"): Almost surely for all $r \in \mathcal{R}$

$$\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0], \qquad i \in [N], t \in \{2, \ldots, T\}. \quad (3.2)$$

We will also refer to a version of (CTS) that holds only after treatment:

Assumption (CTSA) ("common trends with staggered interventions after treatment"): (3.2) holds almost surely for all $i \in [N]$, $r \in \mathcal{R}$, and $t \in \{r, \ldots, T\}$.

Because (CTS) does not account for covariates, (CCTS) is often thought of as more plausible than (CTS). We generally agree, though in Appendix A.2 we add some nuance by pointing out that (CTS) can hold when (CCTS) does not hold (Theorem A.2). In such a setting it is not possible to estimate conditional average treatment effects using regression (1.5); we show this formally in Theorem A.5(a) in the appendix. Further, it turns out that even marginal average treatment effect estimates from models like FETWFE that rely on (CCTS) will in general be inconsistent.

However, (CTS) holding is enough to estimate marginal average treatment effects consistently using many estimators, so it is reasonable to wonder whether regression (1.5) might be able to estimate treatment effects consistently under (CTS) and some additional assumptions. The following assumption will turn out to allow FETWFE to consistently estimate marginal average treatment effects if (CTS) holds, even if (CCTS) does not.

Assumption (CIUN) ("conditional independence of trends of the untreated potential outcomes for never-treated units"): Almost surely for all $i \in [N]$ and any $t \in \{2, \ldots, T\}$,

$$\mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0\right] = \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i\right].$$

Assumption (CIUN) requires that the trends in the untreated potential outcomes are mean-independent of the covariates conditional on treatment status for units that are untreated. Since (CIUN) depends only on the observed potential outcomes, it can be tested, as we discuss in the next section.

3.2 Linearity Assumption (Assumption LINS)

We present an assumption that is slightly different than one introduced by Wooldridge (2021) on the functional forms of the conditional distributions of the untreated potential outcomes, the trends $\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0)$, and the difference-in-differences estimands.

Assumption (LINS): For each $N \in \mathbb{N}$, assume we observe d_N covariates, where d_N may be fixed or increasing in N. There exist fixed parameters $\eta^* \in \mathbb{R}$, $\{\nu_r^*\}_{r \in \mathcal{R}} \in \mathbb{R}$, $\{\gamma_t^*\}_{t \in \{2,\dots,T\}} \in \mathbb{R}$, and $\{\tau_{rt}^*\}_{r \in \mathcal{R}, t \geq r} \in \mathbb{R}$, along with fixed sequences of parameters $\{\boldsymbol{\kappa}_N^*\}_{N \in \mathbb{N}} \in \mathbb{R}^{d_N}$, $\{\boldsymbol{\zeta}_{rN}^*\}_{r \in \mathcal{R}, N \in \mathbb{N}} \in \mathbb{R}^{d_N}$, $\{\boldsymbol{\xi}_{tN}^*\}_{t \in \{2,\dots,T\}, N \in \mathbb{N}} \in \mathbb{R}^{d_N}$, and $\{\boldsymbol{\rho}_{rtN}^*\}_{r \in \mathcal{R}, t \geq r, N \in \mathbb{N}} \in \mathbb{R}^{d_N}$ such that for each $N \in \mathbb{N}$, for all $i \in [N]$ and any fixed $\boldsymbol{x} \in \mathbb{R}^{d_N}$ in the support of the random vector $\boldsymbol{X}_i$ it holds that

$$\mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] = \eta^* + \boldsymbol{x}^\top \boldsymbol{\kappa}_N^*, \tag{3.3}$$

$$\mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$
$$-\mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] = \nu_r^* + \boldsymbol{x}^\top \boldsymbol{\zeta}_{rN}^* \qquad \forall r \in \mathcal{R}, \tag{3.4}$$

$$\mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] = \gamma_t^* + \boldsymbol{x}^\top \boldsymbol{\xi}_{tN}^*, \qquad t \in \{2,\dots,T\}, \qquad \text{and} \tag{3.5}$$

$$\mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(i1)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$
$$-\mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] = \tau_{rt}^* + \dot{\boldsymbol{x}}_r^\top \boldsymbol{\rho}_{rtN}^*, \qquad r \in \mathcal{R}, t \geq r, \tag{3.6}$$

where

$$\dot{\boldsymbol{x}}_r := \boldsymbol{x} - \mathbb{E}\left[\boldsymbol{X}_i \mid W_i = r\right]$$

are covariate vectors that are centered with respect to their cohort means. (For conciseness, hereafter we omit N from the subscripts of the parameters defined here that may vary with N.)

The version of this assumption presented by Wooldridge (2021) uses a slightly different expression for (3.6) and requires $d_N = d$ to be fixed, since ETWFE is inconsistent in the high-dimensional setting we consider. Notice the similarity of the quantity on the left side of (3.6) to the canonical two-period difference-in-differences estimand (1.2). Wooldridge (2021) shows that Assumption (LINS), along with Assumptions (CNAS) and (CCTS), make (1.5) correctly specified for the potential outcomes and the the left side of (3.6) equal to the conditional average treatment effects. See also our derivation (A.8) and Theorem C.1(a) in the appendix.

Remark 3.1. *We can gain insight on (CCTS) by re-arranging and using the conditional independence from W_i to arrive at the following equivalent condition:*

$$\mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i = r, \boldsymbol{X}_i] - \mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i = 0, \boldsymbol{X}_i]$$
$$= \mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] - \mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i], \qquad r \in \mathcal{R}, i \in [N], t \in \{2,\dots,T\}. \tag{3.7}$$

(This is similar to a framing used by, for example, Heckman et al. (1998, Equations 10 and 12), Lechner et al. (2011, Equation 3.2), Henderson and Sperlich (2022, Assumption 1), Ban and Kedagni (2022, Section 2.1), and Park and Tchetgen (2023, Section 2.2) in the $T = 2$ setting.) Under (CCTS) and (LINS), (3.4) quantifies the time-invariant difference in conditionally expected potential outcomes between units in each cohort and units in the never-treated cohort that appears on each side of (3.7). The inclusion of the terms in (3.4) provides robustness against time-invariant confounding (see Proposition A.1(a) in the appendix).

On the other hand, under unconfoundedness of the untreated potential outcomes (which we define formally and discuss in more detail in (A.1) in the appendix), (3.4) (and each side of Equation 3.7) is identically 0. Although including these covariates in a regression does not harm unbiasedness under unconfoundedness, it reduces efficiency. Fortunately, we will see later that under the assumptions of Theorem 6.3, if the estimated model contains any coefficients that equal 0 exactly, FETWFE successfully screens out these covariates and achieves the same asymptotic efficiency as a model that omitted them. In this way FETWFE provides robustness without compromising asymptotic efficiency under unconfoundedness, avoiding the kind of trade-off between robustness and asymptotic efficiency that Zimmert (2020) discusses.

Many previous works have also noted that difference-in-differences estimation allows for selection bias and can effectively control for some kinds of unobserved time-invariant confounding. Imai and Kim (2019) explore this idea in depth, and also raise interesting issues surrounding the plausibility of an assumption that time-varying covariates do not affect the observed response and the implications when this assumption is violated. Imai et al. (2023) propose a matching estimator in a setting with time-varying covariates. For more in-depth investigations into the parallel trends assumption, see Ghanem et al. (2024), Roth and Sant'Anna (2023), and Marx et al. (2024). See also our further discussion in Appendix A.

Under Assumption (CIUN) and (LINS), the covariates $\boldsymbol{\xi}_t^*$ all equal $\boldsymbol{0}$. So a test of this null hypothesis (for example, a Wald test) serves as a test for the null hypothesis that both (CIUN) and (LINS) hold.

4 Causal Estimands

Next we will discuss several classes of causal estimands we hope to estimate using FETWFE. The building block of our estimands will be the conditional average treatment effects on the treated units $\tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})$ from (1.6), which matches definition (6.23) in Wooldridge (2021). Estimating $\tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})$ is useful if we are interested in identifying which units are most affected by the treatment at a specific time. Besides economics (Florens et al., 2008; Djebbari and Smith, 2008), domains where conditional average treatment effect estimation have been found useful include personalized medicine (Feinstein et al., 1972; Lesko, 2007; Frankovich et al., 2011; Foster et al., 2011), education (Morgan, 2001; Brand and Xie, 2010; Murphy et al., 2016), targeted marketing (Bottou et al., 2013; Athey et al., 2018; Hitsch and Misra, 2018; Ascarza, 2018; Yang et al., 2020; Ellickson et al., 2022), psychology (Bolger et al., 2019; Winkelbeiner et al., 2019), sociology (Xie et al., 2012; Breen et al., 2015), and political science (Imai and Strauss, 2011; Grimmer et al., 2017).

Many methods for estimating conditional average (or *heterogeneous*[3]) treatment effects have been developed within econometrics over the last decade; see Abrevaya et al. (2015); Athey and Imbens (2016); Wager and Athey (2018); Nie and Wager (2021); Semenova and Chernozhukov (2021), and Fan et al. (2022) for a few examples.

4.1 Conditional Average Treatment Effect Estimands

We can use (1.6) to construct a variety of causal estimands. For any set of finite constants $\{\psi_{rt}\}_{r \in \mathcal{R}, t \in \{r, \dots, T\}}$, consider the estimand

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x}), \tag{4.1}$$

a weighted average of the cohort- and time-specific conditional average treatment effects. This is similar to the framework of estimation targets of Borusyak et al. (2024) and Callaway and Sant'Anna (2021, Equation 3.1), though they study average treatment effects (like our estimand (4.8), to come) rather than conditional average treatment effects. Refer to these papers for many examples of choices of $\{\psi_{rt}\}$. Similar aggregations are also discussed by Brown and Butts (2022, Remark 3.2) and Wooldridge (2021, Section 6.5).

---

[3]Within the differences-in-differences literature, the descriptor "heterogeneous" has often been used in reference to treatment effects that vary across cohorts or across time; e.g., De Chaisemartin and d'Haultfoeuille (2020); Sun and Abraham (2021); Wooldridge (2021). Here we refer to heterogeneity in treatment effects across the distribution of $\boldsymbol{X}_i$, though our model also accommodates heterogeneity in treatment effects across cohorts and time. To avoid ambiguity, we will generally use the term "conditional average treatment effects."

To give one example among many possible estimands of interest using specific $\{\psi_{rt}\}$, we could choose for some $r' \in \mathcal{R}$

$$\psi_{rt} = \begin{cases} \frac{1}{T-r'+1}, & r = r', \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

to define the conditional average treatment effect across time for a unit in cohort r' with covariates $\boldsymbol{x}$:

$$\tau_{\mathrm{CATT}}(r', \boldsymbol{x}) := \frac{1}{T-r'+1} \sum_{t=r'}^{T} \tau_{\mathrm{CATT}}(r', \boldsymbol{x}, t) \tag{4.3}$$

$$= \frac{1}{T-r'+1} \sum_{t=r'}^{T} \mathbb{E}[\tilde{y}_{(it)}(r') - \tilde{y}_{(it)}(0) \mid W_i = r', \boldsymbol{X}_i = \boldsymbol{x}].$$

A version of (4.3) for average (rather than conditional average) treatment effects was considered by Wooldridge (2021, Equation 6.41), and a slightly more general form (also accommodated by our framework) was considered by Goodman-Bacon (2021, Equation 12).

4.1.1 Propensity-Weighted Conditional Average Treatment Effects

A second class of causal estimands can be generated by weighting each $\tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})$ by the probability a unit is in cohort r given that they are treated conditional on $\boldsymbol{x}$,

$$\tilde{\pi}_r(\boldsymbol{x}) := \mathbb{P}(W_i = r \mid W_i \neq 0, \boldsymbol{X}_i = \boldsymbol{x}) = \left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x}) \right)^{-1} \pi_r(\boldsymbol{x}) \qquad \forall r \in \mathcal{R}. \tag{4.4}$$

In particular, for any set of finite constants $\{\psi_{rt}\}$, consider the estimands

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \tilde{\pi}_r(\boldsymbol{x}) \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x}). \tag{4.5}$$

One example of an estimand we can form in this class arises from choosing

$$\psi_{rt} := \begin{cases} 0, & r = 0, \\ \frac{1}{T-r+1}, & \text{otherwise,} \end{cases} \tag{4.6}$$

which yields a conditional average treatment effect by marginalizing (4.3) across treatment status:

$$\tau_{\mathrm{CATT}}(\boldsymbol{x}) := \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \frac{1}{T-r+1} \tilde{\pi}_r(\boldsymbol{x}) \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x}) \tag{4.7}$$

$$= \sum_{r \in \mathcal{R}} \tilde{\pi}_r(\boldsymbol{x}) \tau_{\mathrm{CATT}}(r, \boldsymbol{x})$$

$$= \mathbb{E}\left[\tau_{\mathrm{CATT}}(W_i, \boldsymbol{X}_i) \mid W_i \neq 0, \boldsymbol{X}_i = \boldsymbol{x} \right].$$

This estimand is closely related to many previously considered heterogenous treatment effects; one of many works that considers a similar estimand (outside of the specific context of difference-in-differences) is Wager and Athey (2018, Equation 1).

4.2 Marginal Average Treatment Effects

Outside of conditional average treatment effects, we can also marginalize these conditional treatment effects over the distribution of $\boldsymbol{X}_i$ to get average treatment effects. Since these estimands are

"coarser," it may be possible to get more precise estimates of these estimands given a fixed sample size.

For $\tau_{\text{ATT}}(r, t) = \mathbb{E}\left[\tau_{\text{CATT}}(r, t, \boldsymbol{X}_i)\right]$ from (1.1), consider the classes of estimands

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \tau_{\text{ATT}}(r, t)) \tag{4.8}$$

and

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \tilde{\pi}_r \tau_{\text{ATT}}(r, t), \tag{4.9}$$

where $\tilde{\pi}_r := \mathbb{P}(W_i = r \mid W_i \neq 0) = \mathbb{E}[\tilde{\pi}_r(\boldsymbol{X}_i)]$ is the marginal probability of a unit being in cohort r given that it is treated.

Estimand (1.1) was used by Callaway and Sant'Anna (2021) and Hatamyar et al. (2023, Equation 1), and it matches the estimand from Wooldridge (2021, Equation 6.2) and Brown and Butts (2022, Equation 3). It is also closely related to the *cohort-specific average treatment effect on the treated* (3) from Sun and Abraham (2021) (though we use the acronym CATT to refer to conditional, not cohort, effects) and estimands from Goodman-Bacon (2021).

As we mentioned earlier, the class of estimands (4.8) is the same framework of estimation targets considered by Borusyak et al. (2024) and is similar to aggregations considered by Callaway and Sant'Anna (2021, Equation 3.1), Wooldridge (2021, Section 6.5), Brown and Butts (2022, Remark 3.2), and Hatamyar et al. (2023, Equation 18).

The probability-weighted class of estimands (4.9) bears a resemblance to the probability-weighted estimands proposed by Callaway and Sant'Anna (2021). Theorem D.1 in the appendix extends our consistency result to accommodate classes of estimands with much broader forms of estimated probability weights than (4.5) and (4.9), like estimand (26) of Sun and Abraham (2021) or the broad class of estimands used by Callaway and Sant'Anna (2021). Theorems D.2 and D.3 extend our asymptotic normality results to these more general estimators as well.

As a particular choice of $\{\psi_{rt}\}$, we will again highlight (4.2) for any $r' \in \mathcal{R}$, which for estimand (4.8) marginalizes $\tau_{\text{CATT}}(r', \boldsymbol{X}_i)$ over the distribution of $\boldsymbol{X}_i$:

$$\begin{aligned}
\tau_{\text{ATT}}(r') := \ & \frac{1}{T - r' + 1} \sum_{t=r'}^{T} \tau_{\text{ATT}}(r', t) \tag{4.10} \\
= \ & \frac{1}{T - r' + 1} \sum_{t=r'}^{T} \mathbb{E}\left[\tilde{y}_{(it)}(r') - \tilde{y}_{(it)}(0) \mid W_i = r'\right] \\
= \ & \mathbb{E}\left[\tau_{\text{CATT}}(r', \boldsymbol{X}_i)\right].
\end{aligned}$$

Estimand (4.10) is identical to estimand (3.7) in Callaway and Sant'Anna (2021), and is also the estimand of estimator (6.41) in Wooldridge (2021).

4.2.1 The Average Effect of the Treatment on Treated Units

Lastly, one of the most basic causal parameters of interest is the average effect of treatment on the treated group (ATT). By using the ψ_{rt} from (4.6) with (4.9), we can express this estimand as

$$\tau_{\text{ATT}} := \ \mathbb{E}\left[\tau_{\text{CATT}}(\boldsymbol{X}_i) \mid W_i \neq 0\right] = \mathbb{E}\left[\sum_{r \in \mathcal{R}} \tilde{\pi}_r(\boldsymbol{X}_i) \tau_{\text{CATT}}(r, \boldsymbol{X}_i)\right]. \tag{4.11}$$

Estimand (4.11) is similar to estimand (6.40) proposed by Wooldridge (2021) (a similar estimand is also mentioned by Borusyak et al. 2024), though we weight the cohort average treatment effects $\tau_{\text{CATT}}(r, \boldsymbol{X}_i)$ by the probabilities $\tilde{\pi}_r(\boldsymbol{X}_i)$ rather than taking an average where each average treatment effect in each time period is weighted equally (we could also take an equal average with our

estimand from Equation 4.8). This probability-weighted average is closely related to estimand (26) from Sun and Abraham (2021).

The virtues of estimand τ_{ATT} include summarizing the effect of the treatment in a single number and potentially being a less noisy quantity to estimate. Estimating τ_{ATT} will involve taking a weighted average over many estimated coefficients rather than relying on a single, less precise, estimated coefficient. In contrast, more specific estimands like $\tau_{\text{CATT}}(r, t, \boldsymbol{x})$ will typically be difficult to estimate with precision because their estimators will depend heavily on a small subset of the observed data (though this problem will be mitigated by our fusion penalty, discussed in the next section, which allows us to borrow information from nearby estimators).

Notice that τ_{ATT} can be obtained not just by marginalizing $\tau_{\text{CATT}}(\boldsymbol{X}_i)$ over the distribution of $\boldsymbol{X}_i$, but also by marginalizing $\tau_{\text{ATT}}(W_i)$ over the distribution of cohort assignments:

$$
\begin{aligned}
\tau_{\text{ATT}} &= \mathbb{E}\left[\tau_{\text{CATT}}(W_i, \boldsymbol{X}_i) \mid W_i \neq 0\right] = \mathbb{E}\left[\mathbb{E}\left[\tau_{\text{CATT}}(W_i, \boldsymbol{X}_i) \mid W_i\right] \mid W_i \neq 0\right] \\
&= \sum_{r \in \mathcal{R}} \tilde{\pi}_r \cdot \mathbb{E}[\tau_{\text{CATT}}(W_i, \boldsymbol{X}_i) \mid W_i = r] = \sum_{r \in \mathcal{R}} \tilde{\pi}_r \tau_{\text{ATT}}(r) \\
&= \sum_{r \in \mathcal{R}} \tilde{\pi}_r \frac{1}{T - r + 1} \sum_{t=r}^{T} \tau_{\text{ATT}}(r, t),
\end{aligned}
$$

which shows that estimand (4.11) is in the class defined by (4.9) and matches estimand (3.11) from Callaway and Sant'Anna (2021). This formulation will also be useful for estimation.

5 Fused Extended Two-Way Fixed Effects

We will carry out regression on the same design matrix $\tilde{\boldsymbol{Z}}$ proposed by Wooldridge (2021) to estimate regression (1.5) by OLS, but we will add fusion penalties to leverage our belief that some of the parameters are equal to each other. This design matrix contains R columns of cohort dummies, $T - 1$ columns of time dummies for times $\{2, \ldots, T\}$, d_N covariates, $\mathfrak{W} := \sum_{r \in \mathcal{R}} (T - r + 1)$ treatment dummies for each possible base treatment effect τ_{rt}, and $d_N(R + T - 1 + \mathfrak{W})$ interactions between covariates and the cohort dummies, time dummies, and treatment dummies. As is typical in penalized regression, we will center $\tilde{\boldsymbol{y}}$ and the columns of $\tilde{\boldsymbol{Z}}$ before estimating our regression. It is not necessary to estimate the intercept term η^* in order to estimate the treatment effects.

5.1 Restrictions in the ETWFE Estimator

Let p_N be the number of columns of $\tilde{\boldsymbol{Z}}$; that is, $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{NT \times p_N}$. We will assume $p_N \leq NT$ (see Assumption R2 later), but as we discussed in the introduction, it is clear that $p_N = R + T - 1 + \mathfrak{W} + d_N(1 + R + T - 1 + \mathfrak{W})$ may be fairly large compared to NT, particularly if d_N is not very small. This could lead to imprecise treatment effect estimates if (1.5) is estimated by OLS. Wooldridge (2021, Section 6.5) suggests alleviating this problem by assuming *restrictions* in the model—assuming some of these parameters equal each other. Some of the possibilities Wooldridge suggests include assuming that the treatment-covariate interactions $\boldsymbol{\rho}_{rt}$ are fixed across time or assuming the within-cohort treatment effects τ_{rt} are fixed across time. Goodman-Bacon (2021) also considers the possibility of treatment effects that are fixed across time. Another weaker assumption Wooldridge (2021) proposes is assuming that different treatment effects do not need to be estimated for every individual time period since treatment starts; instead, these times could be condensed into "early treated" and "late treated" effects, for example.

These suggestions seem sensible, but making these simplifying assumptions without strong justification risks re-introducing the bias that was removed by adding these parameters to begin with. In settings where a practitioner believes some of these restrictions hold but has no clear way to choose which particular parameters may be equal, we instead propose a data-driven approach to estimate which, if any, of these parameters may be equal.

5.2 Fusion Penalization in FETWFE

As described in the introduction, we will assign ℓ_q fusion (Tibshirani et al., 2005) penalties to the differences between estimated parameters which may be equal, and then carry out a penalized linear regression. For example, for the treatment effects, we propose penalizing the absolute differences between treatment effects within the same cohort at adjacent times, $|\tau_{r_k,t} - \tau_{r_k,t-1}|$, and we propose the same penalty structure for interaction effects, $|\rho_{r_k,t,j} - \rho_{r_k,t-1,j}|$ separately for each $j \in [d_N]$. This leverages our belief that some of these effects may be fixed across time while allowing the optimization problem to learn from the data which restrictions to choose.

We also propose penalizing the absolute differences between initial treatment effects for each cohort, $|\tau_{r_k,r_k} - \tau_{r_{k-1},r_{k-1}}|$, and likewise we penalize the absolute differences between adjacent initial interaction effects $|\rho_{r_k,r_k,j} - \rho_{r_{k-1},r_{k-1},j}|$ for each j. This is motivated by the idea that perhaps time since beginning treatment could be the only dimension in which treatment effect changes. It is also natural to penalize the adjacent cohort fixed effects absolute differences $|\nu_{r_k} - \nu_{r_{k-1}}|$ and time fixed effects absolute differences $|\gamma_t - \gamma_{t-1}|$ towards each other, with analogous penalties for the interaction terms $|\zeta_{r_kj} - \zeta_{r_{k-1},j}|$ and $|\xi_{tj} - \xi_{t-1,j}|$.

Finally, we may also want to regularize the coefficients directly, as is common in regressions with large numbers of parameters to estimate relative to the number of observations. Because of the fusion penalties we employ, regularizing only a handful of base terms allows for this regularization. We can also penalize all of the base covariate coefficients directly, allowing us to screen out irrelevant covariates as d_N grows.

In summary, we propose estimating a model like (1.5) but with a penalty term

$$
\begin{aligned}
p(\boldsymbol{\beta}; q) := & \sum_{k=2}^{R} |\nu_{r_k} - \nu_{r_{k-1}}|^q + |\nu_{r_R}|^q + \sum_{t=3}^{T} |\gamma_t - \gamma_{t-1}|^q + |\gamma_T|^q \\
& + \sum_{j=1}^{d_N} \left(|\kappa_j|^q + \sum_{k=2}^{R} |\zeta_{r_k,j} - \zeta_{r_{k-1},j}|^q + |\zeta_{r_R,j}|^q + \sum_{t=3}^{T} |\xi_{tj} - \xi_{t-1,j}|^q + |\xi_{Tj}|^q \right) \\
& + |\tau_{r_1,r_1}|^q + \sum_{k=2}^{R} \left(|\tau_{r_k,r_k} - \tau_{r_{k-1},r_{k-1}}|^q + \sum_{t \geq r_k+1} |\tau_{r_k,t} - \tau_{r_k,t-1}|^q \right) \\
& + \sum_{j=1}^{d_N} \left[|\rho_{r_1,r_1,j}|^q + \sum_{k=2}^{R} \left(|\rho_{r_k,r_k,j} - \rho_{r_{k-1},r_{k-1},j}|^q + \sum_{t \geq r_k+1} |\rho_{r_k,t,j} - \rho_{r_k,t-1,j}|^q \right) \right], \quad (5.1)
\end{aligned}
$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p_N}$ collects all of the coefficients. (Refer again to Figure 1 for a graphical depiction of the penalty terms on the estimated treatment effects τ_{rt}.) We can more compactly express this as

$$
p(\boldsymbol{\beta}; q) = \|\boldsymbol{D}_N \boldsymbol{\beta}\|_q^q, \quad (5.2)
$$

where $\boldsymbol{D}_N \in \mathbb{R}^{p_N \times p_N}$ is a suitably defined differences matrix which we later define explicitly in (E.1).

Remark 5.1. *Our theory could easily be extended to allow a much broader class of penalties than the one penalty we specify in (5.1). Any choice of $\boldsymbol{D}_N$ that has all finite entries, is invertible, and is block diagonal with blocks of sizes not increasing in N would work for our theoretical results. Such modifications would lead to a change only in constant factors in our theory. For example, one could choose to only use fusion penalties for the treatment parameters τ_{rt} and ρ_{rt} and directly penalize all of the remaining parameters. For another example, if it is believed that treatment effects are more likely to be equal in time since treatment rather than within cohorts, one could replace our structure of treatment effect penalties with fusion penalties across treatment effects at the same time since treatment; see Figure 2 for a graphical depiction of such a structure. It is important that the chosen $\boldsymbol{D}_N$ makes Assumption $S(s_N)$, which we describe later in Section 6.1, plausible.*
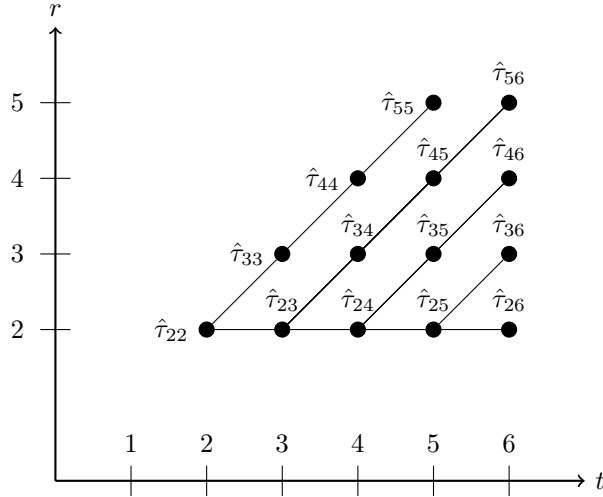
14

Figure 2: Another example of a possible valid penalty structure that mainly penalizes treatment effects across cohorts towards each other based on time since the start of treatment, as mentioned in Remark 5.1. Compare to Figure 1. All of our theoretical results for FETWFE would work with such a penalty structure up to a change in constants in the results.

It is much easier to see that these candidate restrictions are plausible due to the time-structured nature of the parameters than it is to know exactly which restrictions to choose. FETWFE allows practitioners to leverage this structure to improve estimation efficiency without requiring them to choose restrictions by hand, which could introduce bias if too many restrictions are chosen or compromise efficiency if too few restrictions are chosen.

5.3 Constructing the FETWFE Optimization Problem

Finally, we take care of some details. First we account for the random effects discussed in Section 2. As suggested by (2.1), in practice we will arrange the rows of $\tilde{\boldsymbol{Z}}$ to contain consecutive blocks of T rows for each observation and carry out a regression on $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{Z}}$ transformed by

$$\boldsymbol{G}_N := \sigma \boldsymbol{I}_N \otimes \boldsymbol{\Omega}^{-1/2} \in \mathbb{R}^{NT \times NT}, \tag{5.3}$$

where $\otimes$ denotes the Kronecker product (recall that $\boldsymbol{\Omega}$ was defined in Section 2). We will also center both the response $\boldsymbol{G}_N \tilde{\boldsymbol{y}}$ and each column of the covariates $\boldsymbol{G}_N \tilde{\boldsymbol{Z}}$, as is common in penalized regression. Denote these centered objects by $\boldsymbol{y}$ and $\boldsymbol{Z}$. Then our optimization problem is

$$\hat{\boldsymbol{\beta}}_N^{(q)}(\lambda_N) := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p_N}}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda_N \|\boldsymbol{D}_N \boldsymbol{\beta}\|_q^q \right\}, \tag{5.4}$$

where λ_N is a tuning parameter. For brevity, we will hereafter write $\hat{\boldsymbol{\beta}}_N^{(q)}(\lambda_N) = \hat{\boldsymbol{\beta}}^{(q)}$.

Remark 5.2. *Optimization problem (5.4) is convex if and only if $q \geq 1$, but we will be interested in solving this problem with $q \in (0,1)$ as well. In this case, finding the global optimum can in principle be difficult, but many heuristics exist to solve this optimization problem that seem to work well in practice; see, for example Huang et al. (2008, Section 4.1) and Breheny and Huang (2015). In Sections 7 and 8 we implement FETWFE by solving (5.4) with the **grpreg** R package (Breheny and Huang, 2009). Our simulation results seem to align with our theory, including the asymptotic normality of our estimators, suggesting that these heuristics seem to work well for FETWFE.*

5.4 Causal Estimators

As shown in Theorem A.5(a), under (CNAS), (LINS), and (CCTS)

$$\tau_{\text{CATT}}(r, t, \boldsymbol{x}) = \tau_{rt}^* + (\boldsymbol{x} - \mathbb{E}\left[\boldsymbol{X}_i \mid W_i = r\right])^\top \boldsymbol{\rho}_{rt}^*$$

15

for any fixed $\boldsymbol{x}$. Since under these assumptions (and other regularity conditions) FETWFE consistently estimates τ_{rt}^* and $\boldsymbol{\rho}_{rt}^*$, we can estimate this quantity by

$$\hat{\tau}_{\mathrm{CATT}}(r, t, \boldsymbol{x}) := \hat{\tau}_{rt}^{(q)} + \left(\boldsymbol{x} - \overline{\boldsymbol{X}}_r\right)^\top \hat{\rho}_{rt}^{(q)}, \qquad r \in \mathcal{R}, t \geq r, \tag{5.5}$$

where $\hat{\tau}_{rt}^{(q)}$ and $\hat{\rho}_{rt}^{(q)}$ are the estimators of τ_{rt} and ρ_{rt} from $\hat{\boldsymbol{\beta}}^{(q)}$ as defined in (5.4) and

$$\overline{\boldsymbol{X}}_r = \frac{1}{N_r} \sum_{\{i: W_i = r\}} \boldsymbol{X}_i$$

is the sample mean of the covariates for units observed in cohort r, with $N_r := \sum_{i=1}^N \mathbb{1}\{W_i = r\}$ the total number of units in cohort r. The forms of (4.1) and (4.5) then naturally suggest the estimators

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \hat{\tau}_{\mathrm{CATT}}(r, t, \boldsymbol{x})) \tag{5.6}$$

and

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \left(\frac{\hat{\pi}_r(\boldsymbol{x})}{\sum_{r' \in \mathcal{R}} \hat{\pi}_{r'}(\boldsymbol{x})}\right) \hat{\tau}_{\mathrm{CATT}}(r, t, \boldsymbol{x}), \tag{5.7}$$

where $\hat{\pi}_r(\boldsymbol{x})$ is an estimator of the propensity scores $\pi_r(\boldsymbol{x})$. We will discuss what properties this estimator $\hat{\pi}_r(\boldsymbol{x})$ should have in more detail in Section 6.1.

Finally we discuss estimating the average treatment effects marginalized over $\boldsymbol{X}_i$. As we showed in Theorem A.5(a) and (b), because the treatment effects are interacted with the covariates centered with respect to their cohort means, under either (CCTS) or (CTS) and (CIUN) we can estimate the average treatment effects using

$$\hat{\tau}_{\mathrm{ATT}}(r, t) := \hat{\tau}_{rt}^{(q)} \tag{5.8}$$

as proposed by Wooldridge (2021, Sections 5.3 and 6.3). In particular, we propose the estimators

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \hat{\tau}_{\mathrm{ATT}}(r, t) \tag{5.9}$$

and

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \frac{N_r}{N_\tau} \hat{\tau}_{\mathrm{ATT}}(r, t) \tag{5.10}$$

for estimands (4.8) and (4.9), respectively. Notice that estimator (5.10) avoids requiring the specification of an estimator $\hat{\boldsymbol{\pi}}(\boldsymbol{x})$ of the generalized propensity scores by instead using the observed cohort counts to estimate the marginal cohort assignment probabilities, as we alluded to at the end of Section 4.2.1.

6 Theoretical Results and Discussion

Before we present theoretical guarantees for FETWFE, we will state a sparsity assumption that we will require along with some additional regularity conditions. We will require fewer assumptions to prove consistency alone than we will to prove restriction selection consistency and asymptotic normality. We emphasize that the conceptually most important assumptions are (CCTSB), (CCTSA), and (CTSA), which validate the difference-in-differences approach; (LINS), which justifies a model that is linear in the covariates $\boldsymbol{X}_i$; and Assumption S(s_N), defined below, which assumes sparsity in a sense that matches our penalty (5.2). (See Remark 5.1 for some comments on how this can be made more flexible.)

6.1 Consistency

Assumption S(s_N): Assume the vector $\boldsymbol{D}_N\boldsymbol{\beta}_N^*$ (where the differences matrix $\boldsymbol{D}_N$ associated with our penalty (5.2) is defined in Equation E.1) is sparse, with indices $\mathcal{S} \subseteq [p_N]$ nonzero and the remaining indices in $[p_N] \setminus \mathcal{S}$ equal to 0, where $|\mathcal{S}| = s_N$.

Assumption S(s_N) allows us to have sparsity in the appropriate sense: the differences between parameters that we penalized in (5.2) are sparse, and irrelevant covariates have true coefficient equal to 0. That is, all but s_N of the restrictions that we expect to hold will in fact hold. Crucially, Assumption S(s_N) does not require us to know which restrictions to select, or even how many restrictions to select—the value of s_N does not need to be known. Our consistency and selection consistency theorems will allow s_N to grow with N subject to certain constraints in later assumptions.

The remaining regularity conditions for Theorem 6.1 are more routine. Before we describe them in detail, we briefly characterize them informally.

- Assumption (R1) requires the tails of the covariates to not be too heavy, but accommodates a wide variety of covariate distributions.

- Assumption (R2) requires the eigenvalues of the empirical Gram matrix to be reasonably well-behaved. In particular, it requires $\boldsymbol{Z}$ to have full column rank, so $p_N \leq NT$. Otherwise this is a mild assumption as long as (1) none of the covariates are very highly correlated and (2) none of the cohort propensity scores tend too closely to 0. For example, an overlap assumption that $\pi_r(\boldsymbol{x}) \geq \pi_{\text{low}}$ for some $\pi_{\text{low}} > 0$ for all $N \in \mathbb{N}$, $r \in \{0\} \cup \mathcal{R}$, and $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$ would help the plausibility of this assumption.

- In Assumption (R2), the assumptions on the penalty parameter λ_N are not of practical importance because in practice λ_N can be selected from a set of candidate values by a criterion like BIC. (We show in simulation studies in Section 7 that the theoretical guarantees from this section seem to hold in practice when λ_N is selected in this way.)

- Assumption (R3) is essentially a technical assumption. It does suggest that FETWFE may not work well if $\boldsymbol{D}_N\boldsymbol{\beta}_N^*$ has many very small entries, or a small number of very large entries.

In particular, these assumptions allow the number of covariates d_N and the number of relevant coefficients s_N to grow to infinity with N and allow for unbounded distributions of the covariates.

We now state and describe the assumptions.

Assumption (R1): For each $i \in [N]$, $\mathbb{E}[X_{ij}^4] < \infty$ for all $j \in [d_N]$ and $\mathbb{E}[u_{(it)}^4] < \infty$ for all $t \in [T]$. Further, $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\top \mid W_i = r]$. Finally, for some $K < \infty$,

$$\sup_{1 \leq N \leq \infty} \left\{ \frac{1}{p_N} \sum_{t=1}^{T} \sum_{j=1}^{p_N} \mathbb{E}\left[(\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(1t)j}^2\right] \right\} \leq K.$$

We show later (Lemma E.1) that $\boldsymbol{D}_N$ is invertible. The matrix $\boldsymbol{Z}\boldsymbol{D}_N^{-1}$ is of interest because it turns out this is the design matrix we will actually use for estimation; see the proof of Theorem C.1 for details. Because the singular values of $\boldsymbol{D}_N^{-1}$ are bounded away from 0 and from above for all N even if $d_N \to \infty$ (see Lemma F.1), this is qualitatively similar to making the same statement about $\boldsymbol{Z}$ itself. We discuss a closely related point in more detail below.

Before we state Assumption (R2), we will define some needed notation. For any matrix $\boldsymbol{A} \in \mathbb{R}^{N \times p_N}$, denote the empirical Gram matrix by $\hat{\boldsymbol{\Sigma}}(\boldsymbol{A})$:

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{A}) := \frac{1}{N}\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{p_N \times p_N}. \tag{6.1}$$

Notice that if $\boldsymbol{A}$ has centered columns then $\hat{\boldsymbol{\Sigma}}(\boldsymbol{A})$ is the estimated covariance matrix. Let e_{1N} and e_{2N} be the smallest and largest eigenvalues of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$. (Since $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$ is random, e_{1N} and e_{2N} are as well.)

Assumption (R2): Assume e_{1N} is almost surely upper-bounded by a constant for all N and $h'_N \xrightarrow{a.s.} 0$, where

$$h'_N := \sqrt{\frac{p_N + \lambda_N s_N}{N e_{1N}}} \tag{6.2}$$

and $\xrightarrow{a.s.}$ denotes almost sure convergence. Also, assume $\lim_{N \to \infty} \lambda_N \sqrt{\frac{s_N}{N}} = 0$.

Kock (2013) points out that (6.2) requires the design matrix to be full rank, so this assumption requires that p_N grows slower than N, λ_N does not grow too quickly, and e_{1N} does not vanish too quickly. (In practice, in finite samples λ_N can be selected by cross-validation or a criterion like BIC, similarly to the lasso.) Under the qualitative assumption that the eigenvalues of $\hat{\Sigma}(Z)$ and $\hat{\Sigma}(G_N \tilde{Z})$ are "close,"[4] (R2) roughly requires that $G_N \tilde{Z}$ is full rank, and we can provide sufficient conditions for this to hold. First, the singular values of G_N are bounded from above and below by positive constants for all N (see Lemma I.1 in the appendix), so $\tilde{Z}$ being full rank is enough for $G_N \tilde{Z}$ to be full rank. For $\tilde{Z}$ to be full rank, we require at least $d_N + 1$ observed units in each cohort (see Lemma I.2 in the appendix), all of the marginal cohort probabilities must be large enough relative to the growth rate of d_N as $N \to \infty$, and the distribution of X_i must be well-behaved as $N \to \infty$.

Assumption (R3): There exist constants $0 < b_0 < b_1 < \infty$ such that $b_0 \leq \min_{j \in \mathcal{S}} \{|(D_N \beta^*_N)_j|\} \leq \max_{j \in \mathcal{S}} \{|(D_N \beta^*_N)_j|\} \leq b_1$.

The lower bound in Assumption (R3) is essentially a signal strength assumption.

Note that we allow $s_N \to \infty$ and $p_N \to \infty$ as long as the above assumptions are satisfied.

We are now prepared to state our consistency theorem.

Theorem 6.1 (Consistency of FETWFE). *Assume that Assumptions (CNAS), (CCTSB), and (LINS) hold, as well as Assumptions (F1), (F2), $S(s_N)$, and (R1) - (R3). Let $q > 0$.*

(a) Suppose that either Assumptions (CTSA) and (CIUN) hold or Assumption (CCTSA) holds. Then for any set of finite constants $\{\psi_{rt}\}$,

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\hat{\tau}_{ATT}(r,t) - \tau_{ATT}(r,t) \right) \right| = \mathcal{O}_{\mathbb{P}} \left(\min\{h_N, h'_N\} \right)$$

and

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{N_r}{N_\tau} \hat{\tau}_{ATT}(r,t) - \tilde{\pi}_r \tau_{ATT}(r,t) \right) \right| = \mathcal{O}_{\mathbb{P}} \left(\min\{h_N, h'_N\} \right),$$

where

$$h_N := \frac{1}{e_{1N}} \sqrt{\frac{p_N}{N}} \tag{6.3}$$

and h'_N is as defined in (6.2).

(b) Suppose Assumption (CCTSA) holds and $d_N = d$ is fixed. For all N and all $r \in \{0\} \cup \mathcal{R}$, assume all of the eigenvalues of $\text{Cov}(X_i \mid W_i = r)$ are bounded between $\lambda_{min} > 0$ and $\lambda_{max} < \infty$. Then for any set of finite constants $\{\psi_{rt}\}$ and any fixed x in the support of X_i,

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\hat{\tau}_{CATT}(r,t,x) - \tau_{CATT}(r,t,x) \right) \right| = \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}).$$

---

[4]This can be made formal by noting that centering the columns of $G_N \tilde{Z}$ can be characterized as multiplication on the left by a centering matrix C, then bounding the singular values of C. See, for example, Section 8.1 of Ding (2024) for further details.

Further, if $\hat{\pi}_r(\boldsymbol{x})$ is an estimator of $\pi_r(\boldsymbol{x})$ that satisfies $|\hat{\pi}_r(\boldsymbol{x}) - \pi_r(\boldsymbol{x})| = \mathcal{O}_{\mathbb{P}}(a_N)$ for a decreasing sequence a_N for each $r \in \{0\} \cup \mathcal{R}$ and all $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$, then

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{\hat{\pi}_r(\boldsymbol{x})}{\sum_{r' \in \mathcal{R}} \hat{\pi}_{r'}(\boldsymbol{x})} \hat{\tau}_{CATT}(r, t, \boldsymbol{x}) - \tilde{\pi}_r(\boldsymbol{x}) \tau_{CATT}(r, t, \boldsymbol{x}) \right) \right|$$
$$= \mathcal{O}_{\mathbb{P}}(\min\{h_N, h_N'\} \vee a_N).$$

Proof. Provided in Section E of the appendix. $\square$

Theorem 6.1 shows that all of our causal estimators are consistent and characterizes their rates of consistency. In particular, it also shows that the classes of marginal parallel trends estimators (5.9) and (5.10) are consistent under either (CCTSA) or (CIUN) and (CTSA) for the marginal average treatment effects (4.8) and (4.9). Although we require (CCTSB) to hold regardless of whether (CIUN) holds, (CCTSB) also depends only on the observed outcomes and is testable. In the next section, we show that under mild assumptions FETWFE automatically provides a valid estimate of whether (CIUN) holds (by estimating whether the $\boldsymbol{\xi}_t^*$ coefficients all equal $\boldsymbol{0}$).

To interpret the rate of convergence $\min\{h_N, h_N'\}$, consider the upper bound h_N from (6.3). Suppose the minimum eigenvalue e_{1N} is bounded away from 0 with high probability, which happens under relatively mild assumptions in our $p_N \leq NT$ setting (see our more detailed discussion later in Section 6.3 about Assumption R6). Then $\mathcal{O}_{\mathbb{P}}(\min\{h_N, h_N'\}) = \mathcal{O}_{\mathbb{P}}\left(\sqrt{p_N/N}\right)$; see Lemma I.3 in the appendix.

In general we will focus more on $q \in (0, 1)$ as required by Theorems 6.2, 6.3, and 6.4, but we note that Theorem 6.1 holds for any $q > 0$. This is useful because the convex lasso ($q = 1$) and ridge ($q = 2$) optimization problems, for example, may be more tractable on very large data sets than the nonconvex bridge estimator with $q < 1$.

To give a concrete example of an estimator $\hat{\boldsymbol{\pi}}(\boldsymbol{x})$, in the setting of Theorem 6.1(b) where $d_N = d$ is fixed, the multinomial logit model is $1/\sqrt{N}$-consistent for the generalized propensity scores by standard maximum likelihood theory under correct specification. If this model is overly simplistic or d is large relative to N, regularized machine learning methods with theoretical convergence guarantees—like deep neural networks (Farrell et al., 2021, see Lemma 9c), random forests (Gao et al., 2022), the ℓ_1-penalized classifiers of Levy and Abramovich (2023), or the group lasso multinomial logit model of Farrell (2015)—would also work. On the other hand, since the cohorts have a natural ordering, we could add more structure by using an ordinal response estimator like the proportional odds model (McCullagh, 1980).

It may be possible to relax Assumption (LINS) so that linearity only has to hold approximately, along the lines of e.g. Condition ASTE in Belloni et al. (2014). (At a high level, we might expect that if approximate linearity holds so that the error due to misspecification is on the same order as the estimation error of the population least squares model, we can achieve convergence at the same rate as when the linearity assumption holds exactly.)

It may also be possible to relax the fixed d assumption in part (b) by using high-dimensional central limit theorems like those in Lopes (2022) and Chernozhukov et al. (2023), to name two recent examples among many papers containing such results.

Lastly, we point out that under Theorem 6.1 the generalized propensity scores and treatment effects can be estimated on the same data sets; that is, the full data set can be used for both. Intuitively, this is because these estimators are used for the distinct tasks of estimating the effects $\tau_{\text{CATT}}(r, t, \boldsymbol{x})$ and estimating the weights used to combine them in a weighted average. For technical details, see the proof of Theorem 6.1 in Appendix E.

In the appendix, we present an extension of Theorem 6.1, Theorem D.1, which proves consistency for a set of estimators that allows for much broader functions of the estimated cohort membership probabilities than the classes (5.10) and (5.7). This result includes, for example, all of the estimands from Callaway and Sant'Anna (2021) as special cases.

6.2 Restriction Selection Consistency

Again, we first state some additional needed regularity conditions and then present our selection consistency result, Theorem 6.2. Theorem 6.2 shows that with probability tending towards one FETWFE identifies the correct restrictions, resulting in fusing together parameters that equal each other, screening out irrelevant covariates, and improving estimation efficiency. At a high level, Assumptions (R4) and (R5) add the following requirements:

- Assumption (R4) is most plausible if the covariates have bounded support and if d_N is not increasing in N. Intuitively, this means that in finite samples results relying on Assumption (R4) will be most likely to hold when NT is reasonably large compared to p_N, though we show in the simulation studies in Section 7 that in practice the conclusion of Theorem 6.2 seems to hold even when N is not all that large and p_N is fairly close to NT.

- Assumption (R5) is a technical assumption that does not require much more than (R2). (R5) is easier to satisfy if d_N is increasing only slowly in N and if the sparsity s_N is fixed. (In finite samples, we would expect better results if p_N and the sparsity s_N are small relative to NT.)

Assumption (R4): There exists a positive constant $e_{\max} < \infty$ such that $e_{2N} \leq e_{\max}$ almost surely (recall from Section 6.1 that e_{2N} is the largest eigenvalue of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$).

A sufficient condition for Assumption (R4) is that the distribution of $\boldsymbol{X}_i$ has bounded support in the sense that $\|\boldsymbol{X}_i\|_2 \leq B$ almost surely for some $B < \infty$.

Assumption (R5):

$$\lambda_N \frac{e_{1N}^{2-q}}{\sqrt{N^q p_N^{2-q}}} \xrightarrow{a.s.} \infty.$$

We will briefly examine how Assumption (R5) interacts with Assumption (R2). For simplicity, assume the minimum eigenvalue $e_{1N} \geq c$ for some $c > 0$ with probability tending towards 1, so that we can ignore e_{1N} asymptotically (using a similar argument to the one in the proof of Lemma I.3 in the appendix). Then Assumption (R5) is equivalent to the deterministic condition

$$\lambda_N \frac{1}{N^{q/2} p_N^{1-q/2}} \to \infty.$$

If we also assume $s_N = s$ is fixed, we have from Assumption (R2) that

$$\frac{\lambda_N}{N} \to 0 \qquad \text{and} \qquad \lim_{N \to \infty} \frac{\lambda_N}{N^{1/2}} = 0.$$

We will require $q \in (0,1)$, and we see that under these assumptions for (R5) and (R2) to hold simultaneously we require $N^{q/2} p_N^{1-q/2}$ to grow slower than $N^{1/2}$; that is, p_N must grow slower than

$$\left(\frac{N^{1/2}}{N^{q/2}} \right)^{2/(2-q)} = N^{(1-q)/(2-q)}.$$

Even if e_{1N} is not bounded away from 0 with high probability, Assumption (R5) requires that e_{1N} does not vanish too quickly. As we discussed in reference to Assumption (R2), this requires that the probability of being in any one cohort does not vanish too quickly as $N \to \infty$. It also requires that λ_N not decrease too quickly.

Notice that other than these relatively mild regularity conditions on the eigenvalues of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$, we do not require a condition like the *irrepresentable condition* (Zhao and Yu, 2006) or *neighborhood stability condition* (Meinshausen and Bühlmann, 2006) that are required for selection consistency of the lasso. However, our restriction selection consistency result will only work for $q < 1$; we do not prove selection consistency of FETWFE with a lasso ($q = 1$) penalty.

Theorem 6.2 (Selection consistency). *For the FETWFE estimator $\hat{\boldsymbol{\beta}}^{(q)}$ defined in (5.4), define*

$$\hat{\mathcal{S}} := \left\{ j \in [p_N] : (\boldsymbol{D}_N \hat{\boldsymbol{\beta}}^{(q)})_j \neq 0 \right\}$$

to be the set of nonzero values in $\boldsymbol{D}_N \hat{\boldsymbol{\beta}}^{(q)}$. Recall from Assumption $S(s_N)$ that

$$\mathcal{S} = \{ j \in [p_N] : (\boldsymbol{D}_N \boldsymbol{\beta}_N^*)_j \neq 0 \}.$$

Assume that Assumptions (CNAS), (CCTSB), and (LINS) hold, as well as Assumptions (F1), (F2), $S(s_N)$, and (R1) - (R5). Assume $q \in (0, 1)$. Then $\lim_{N \to \infty} \mathbb{P}\left(\hat{\mathcal{S}} = \mathcal{S}\right) = 1$.

Proof. This is immediate from Theorem C.1(c) in the appendix. □

Theorem 6.2 shows that FETWFE fulfills its promise to identify the correct restrictions in accordance with our ℓ_q penalty structure, as discussed in Section 5. This serves our goal of improving estimation by avoiding unnecessarily estimating parameters that are in fact equal.

FETWFE also screens out irrelevant covariates and interactions. As mentioned in Section 3.2, if Assumptions (LINS) and (CIUN) both hold than the $\boldsymbol{\xi}_t^*$ coefficients must all equal 0. If this is true, in large enough samples FETWFE will estimate $\boldsymbol{\xi}_t^*$ to equal exactly $\mathbf{0}$ with high probability.

The proof of Theorem 6.2 relies on an extension of Theorem 2 of Kock (2013) (Theorem F.4 in the appendix) to show that not only does the bridge estimator exclude false selections, it also selects all of the relevant components of $\boldsymbol{D}_N \boldsymbol{\beta}_N^*$ with probability tending to 1. Theorem F.4 requires no added assumptions from Theorem 2 of Kock (2013).

One consequence of Theorem 6.2 is that if $\tau_{\mathrm{ATT}}(r, t) = 0$ for any r, then

$$\lim_{N \to \infty} \mathbb{P}\left(\hat{\tau}_{\mathrm{ATT}}(r, t) = 0\right) = 1.$$

The same is true of all of the other causal estimators. This is stronger than convergence in probability to 0, and this stronger notion of convergence turns out to be a key ingredient in the proof of Theorems 6.3 and 6.4, to come.

6.3 Asymptotic Normality and Oracle Property

Our asymptotic convergence results (Theorems 6.3 and 6.4 below) require one additional regularity condition, an eigenvalue condition which does not require much more than Assumptions (R2) and (R5).

Assumption (R6): The minimum eigenvalue of

$$\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)\right] \tag{6.4}$$

is greater than or equal to a fixed $\delta > 0$ for all N sufficiently large, and

$$\frac{\lambda_N}{e_{1N}\sqrt{N}} \in o_{\mathbb{P}}(1). \tag{6.5}$$

Further, for all N and all $r \in \{0\} \cup \mathcal{R}$, all of the eigenvalues of $\mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r)$ are bounded between $\lambda_{\min} > 0$ and $\lambda_{\max} < \infty$.

The assumption that the minimum eigenvalue of (6.4) is bounded away from 0 is fairly mild since we already require $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$ to be full rank and for its minimum eigenvalue to not vanish too quickly, and the minimum singular value of $\boldsymbol{D}_N$ is bounded from below by a positive constant; see Lemma F.1 in the appendix. Further, this assumption is stronger than necessary for the sake of interpretability; we only require a bound on the minimum eigenvalue of a different matrix (F.6) for all N. This condition holding for (6.5) is sufficient but not necessary.

Condition (6.5) is another assumption that requires e_{1N} to not vanish too quickly, which is qualitatively similar to Assumptions (R2) and (R5). Holding $s = s_N$ fixed (as is required in Theorems 6.3 and 6.4) and ignoring the distinction between almost sure convergence and convergence in probability, (6.5) is more restrictive on the growth rate of λ_N than (R2) since (R2) requires

$$\frac{\lambda_N}{Ne_{1N}} \to 0$$

while (6.5) requires

$$\frac{\lambda_N}{\sqrt{N}e_{1N}} \to 0.$$

However, if e_{1N} is bounded away from 0 with high probability, using a similar argument to the one in the proof of Lemma I.3 in the appendix, (6.5) is essentially equivalent to the other requirement in (R2) that $\lim_{N\to\infty} \lambda_N \sqrt{s_N/N} = 0$ (in the setting of Theorem 6.4 where $s_N = s$ is fixed). If the smallest eigenvalue of $\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})\right]$ is bounded away from 0, one can show that the asymptotic distribution for the minimum eigenvalue of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$ has support bounded away from 0 in our setting where $\boldsymbol{X}_i$ has finite fourth moments and $p_N \leq NT$ (Marchenko and Pastur, 1967). If we also assume that $\boldsymbol{X}_i$ is subgaussian (a weaker assumption than requiring $\boldsymbol{X}_i$ to have bounded support), under Theorem 5.39 from Vershynin (2012) e_{1N} is bounded away from 0 with high probability for finite N large enough relative to p_N.

6.3.1 Oracle Property

Next we show that under Assumptions (R1) - (R6) the estimators (5.9) and (5.10) are both oracle estimators (Fan and Li, 2001; Fan and Peng, 2004), converge at a $1/\sqrt{N}$ rate (faster than the guarantee from Theorem 6.1), and are asymptotically Gaussian.

Theorem 6.3 (Oracle Property of FETWFE). *Assume that Assumptions (CNAS), (CCTSB), and (LINS) hold, as well as Assumptions (F1), (F2), S(s) for a fixed s, and (R1) - (R6). Assume $q \in (0,1)$. Suppose that either Assumptions (CTSA) and (CIUN) hold or Assumption (CCTSA) holds.*

For an arbitrary set of finite constants $\{\psi_{rt}\}$, if at least one of the $\tau_{ATT}(r,t)$ with a nonzero ψ_{rt} is nonzero, then the sequence of random variables

$$\sqrt{NT} \sum_{r\in\mathcal{R}} \sum_{t=r}^{T} \psi_{rt}(\hat{\tau}_{ATT}(r,t) - \tau_{ATT}(r,t))$$

converges in distribution to a mean zero Gaussian random variables with variance that depends only on the s parameters of the model with all restrictions correctly identified.

Further, if $\hat{\tau}_{ATT}(r,t)$ and the probability ratios N_r/N_τ are estimated on two independent data sets of size N, then the sequence of random variables

$$\sqrt{NT} \sum_{r\in\mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{N_r}{N_\tau} \hat{\tau}_{ATT}(r,t) - \tilde{\pi}_r \tau_{ATT}(r,t) \right)$$

also converges in distribution to a mean zero Gaussian random variable with variance that depends only on the s parameters of the model with all restrictions correctly identified.

(If all of the $\tau_{ATT}(r,t)$ with nonzero ψ_{rt} terms equal 0, both sequences of random variables converge in probability to 0.)

Proof. Provided in Section E of the appendix. □

Theorems 6.3 and 6.4 leverage novel (to the best of our knowledge) extensions of Theorem 2 from Kock (2013) that establish finite-sample variance estimators, including when the weights in a linear combination of the estimated coefficients are themselves estimated; see Theorem F.5 in the appendix.

Theorem 6.3 (and, more directly, Theorem C.1(d) and (f) in the appendix) show that even if $p_N \to \infty$ FETWFE converges at a $1/\sqrt{N}$ rate and its asymptotic covariance matrix depends only on the s parameters of the correct model. That is, FETWFE estimates the model with the same asymptotic efficiency as an OLS-estimated ETWFE estimator that knows the $p_N - s$ correct restrictions. FETWFE leverages the possible restrictions in the ETWFE estimator to improve efficiency without requiring the practitioner to make assumptions on which restrictions to choose, which could induce asymptotic bias if the wrong restrictions are chosen or compromise asymptotic efficiency if not enough restrictions are chosen. This addresses the need to consider both bias and variance when estimating treatment effects (Gelman and Vákár, 2021).

In Remark 3.1 we mentioned that the inclusion of the parameters (3.4) in our regression allows for unbiased estimation of treatment effects under both (CCTS) and unconfoundedness of the untreated potential outcomes, but under unconfoundedness the parameters equal exactly 0 and are inefficient to include. However, Theorem 6.3 shows that when coefficients that equal 0 are included, FETWFE sets these coefficients equal to exactly 0 with high probability and estimates the nonzero coefficients with the same asymptotic efficiency as if those coefficients were omitted from the model. FETWFE is therefore a good model to choose if one wants to maximize robustness without compromising asymptotic efficiency. The same is true if (CIUN) holds and the regression estimands $\boldsymbol{\xi}_t^*$ all equal $\boldsymbol{0}$, or if any irrelevant covariates are included in the regression. So practitioners do not need to hesitate to include a possibly relevant covariate that could be helpful for the plausibility of Assumption (CCTSB) or (CCTS) as long as $\boldsymbol{Z}$ has full column rank.

6.3.2 Asymptotically Valid Confidence Intervals

Theorem 6.3 is conceptual and yields insight on the kind of behavior we can expect from FETWFE in large samples. The following result, which has the same assumptions as Theorem 6.3, is more practical and provides explicit formulas for variance estimators for asymptotically valid confidence intervals of treatment effect estimates.

Theorem 6.4 (Asymptotic Confidence Intervals for FETWFE). *Assume that Assumptions (CNAS), (CCTSB), and (LINS) hold, as well as Assumptions (F1), (F2), S(s) for a fixed s, and (R1) - (R6). Assume $q \in (0,1)$. Suppose that either Assumptions (CTSA) and (CIUN) hold or Assumption (CCTSA) holds. Let $\{\psi_{rt}\}$ be an arbitrary set of finite constants, and for all of the below results, assume at least one of the $\tau_{ATT}(r,t)$ corresponding to a nonzero ψ_{rt} is nonzero (otherwise, all of the below sequences of random variables converge in probability to 0).*

(a) For the estimator (5.9) it holds that

$$\sqrt{\frac{NT}{\hat{v}_N^{(C)}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt}(\hat{\tau}_{ATT}(r,t) - \tau_{ATT}(r,t)) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{v}_N^{(C)}$ is a finite-sample variance estimator defined in (E.7). (All variance estimators depend only on the known σ^2 and the observed data.)

(b) Suppose one set of data of size N is used to estimate the cohort probabilities N_r/N_τ and another independent data set of size N is used to estimate each $\hat{\tau}_{ATT}(r,t)$. Then for the estimator (5.10) it holds that

$$\sqrt{\frac{NT}{\hat{v}_N^{C;\hat{\pi}}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{N_r}{N_\tau} \hat{\tau}_{ATT}(r,t) - \tilde{\pi}_r \tau_{ATT}(r,t) \right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{v}_N^{C;\hat{\pi}}$ is a finite-sample variance estimator defined in (E.11).

(c) *Suppose a single set of data of size N is used to estimate both the cohort probabilities N_r/N_τ and each $\hat{\tau}_{ATT}(r,t)$. Then the estimator (5.10) is asymptotically subgaussian: the sequence of random variables*

$$\sqrt{\frac{NT}{\hat{v}_N^{C,\ (cons);\hat{\pi}}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{N_r}{N_\tau} \hat{\tau}_{ATT}(r,t) - \tilde{\pi}_r \tau_{ATT}(r,t) \right)$$

converges in distribution to a mean-zero subgaussian random variable with variance at most 1, where $\hat{v}_N^{C,\ (cons);\hat{\pi}}$ is a finite-sample conservative variance estimator defined in (E.12).

Proof. Provided in Section E of the appendix. $\qquad\square$

Theorem 6.4 shows that we can form asymptotically valid $1 - \alpha$ confidence intervals with estimator $\hat{T}_N$ and variance estimator $\hat{v}_N$ with

$$\mathrm{CI}_N(\hat{v}_N; \alpha) := \left[\hat{T}_N - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{v}_N}{NT}}, \hat{T}_N + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{v}_N}{NT}} \right], \qquad (6.6)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution.

Theorem 6.4(b) requires data splitting for asymptotic normality of the estimator (5.10). However, the data used for estimating the marginal cohort probabilities do not need to include observed responses. In some domains like public health and medicine where difference-in-differences has many potential uses (Rothbard et al., 2023; Wing et al., 2018), such *unlabeled* data may be more widely available than *labeled* data (Candès et al., 2018, Section 1.3.2). In a setting like this, asymptotic normality can be achieved without reducing the amount of data used to estimate the model. (It is straightforward to adjust the variance estimator appropriately if the sample sizes of the two data sets are not equal.)

Theorem 6.4(c) does not require sample splitting, at the price of sacrificing asymptotic normality for mere asymptotic subgaussianity and necessitating a conservative variance estimator. Because a subgaussian random variable can be defined as a variable X satisfying $\mathbb{P}(|X| \geq x) \leq 2\exp\{-x^2/K^2\}$ for some $K > 0$ for all $x \geq 0$ (Vershynin, 2018, Proposition 2.5.2(a)), the tails of such a variable decay at least as quickly as a Gaussian random variable, so we might expect 95% confidence intervals to have similar asymptotic validity to the confidence intervals from part (b). In practice, we show in the simulation studies in Section 7 that these confidence intervals do seem to retain validity.

We cannot test the null hypothesis of treatment effects equalling 0 under Theorem 6.4, since these test statistics are not asymptotically normal if the estimated coefficients all equal 0; instead they converge in probability to 0 (even after scaling by $\sqrt{NT}$).

Theorem D.3 in the appendix is a generalization of both Theorems 6.3 and 6.4 for the classes of conditional average treatment effects estimators (5.6) and (5.7) in the fixed $d_N = d$ setting under the assumption that an asymptotically Gaussian estimator of the generalized propensity scores (like the multinomial logit or proportional odds models) is available and an extra split of the data is available to estimate the cohort covariate means.

7 Simulation Studies

To test the efficacy of FETWFE under our assumptions, we conduct simulation studies in R using the `simulator` package (Bien, 2016). We choose parameters that bear resemblance to the empirical application from Section III of Goodman-Bacon (2021) ($N = 51, T = 33, R = 12, d_N = 3$), which we explore in Section 8. We use fewer cohorts for simplicity of presentation, more covariates to increase p_N in compensation for the smaller R, and more units to avoid realizations where no units are assigned to a cohort by random chance. We generate data with $N = 120$ units, $T = 30$ time

periods, $R = 5$ cohorts entering at times $\{2, \ldots, 6\}$, and $d_N = 12$ features, which results in a total of $p_N = 2209$ covariates and $NT = 3600$ observations. To generate a $\boldsymbol{\beta}_N^*$ such that $\boldsymbol{D}_N \boldsymbol{\beta}_N^*$ is sparse, we generate a single random sparse $\boldsymbol{\theta}_N^* \in \mathbb{R}^{p_N}$, then transform this into a coefficient vector $\boldsymbol{\beta}_N^* = \boldsymbol{D}_N^{-1} \boldsymbol{\theta}_N^*$ to use across all simulations. We generate $\boldsymbol{\theta}_N^*$ by taking a p_N-vector of all 0 entries and setting each entry equal to 2 randomly with probability 0.1. We set the sign of each term randomly, but since each entry of $\boldsymbol{\beta}_N^*$ is a sum of terms in $\boldsymbol{\theta}_N^*$, we bias the signs to avoid treatment effects that are too close to 0: individual nonzero terms in $\boldsymbol{\theta}_N^*$ are positive with probability 0.6 and negative otherwise.

On each of 700 simulations, we randomly generate N independent realizations of $\boldsymbol{X}_{(i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_N})$ for the time-invariant covariates. We randomly assign treatments with probabilities $(\pi_0, \pi_1, \ldots, \pi_R) = (1/(R+1), \ldots, 1/(R+1))$, where

$$\pi_r := \begin{cases} \mathbb{P}(W_i = 0), & r = 0, \\ \mathbb{P}(W_i = r), & r \in \mathcal{R}, \end{cases} \tag{7.1}$$

are the marginal probabilities of treatment assignments. To ensure the model is estimable, in the rare instances where there are no untreated units or one cohort has no observed units, we draw another set of random assignments.

After drawing covariates and treatment assignments, we construct $\tilde{\boldsymbol{Z}}$. Then we draw N independent random effects $c_i \sim \mathcal{N}(0, 5)$ as well as NT independent noise terms $u_{(it)} \sim \mathcal{N}(0, 5)$ to generate $\boldsymbol{\epsilon}_{(i\cdot)} = c_i \boldsymbol{1}_T + \boldsymbol{u}_{(i\cdot)}$ and finally generate $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{Z}} \boldsymbol{\beta}_N^* + \boldsymbol{\epsilon}_{(\cdot\cdot)}$.

Besides FETWFE, we also consider three competitor methods. The first is ETWFE, estimated by OLS on $\tilde{\boldsymbol{Z}}$. We also consider bridge regression on $\boldsymbol{Z}$ directly penalizing $\boldsymbol{\beta}$ rather than penalizing $\boldsymbol{D}_N \boldsymbol{\beta}$ (BETWFE). For both FETWFE and BETWFE, we follow the experiments of Kock (2013) and use $q = 0.5$, selecting the penalty λ_N over a grid of 100 values (equally spaced on a logarithmic scale) using BIC. We implement bridge regression for both BETWFE and FETWFE using the R `grpreg` package (Breheny and Huang, 2009). Finally, we also consider a slightly more flexible version of (1.3), where we estimate the model

$$\tilde{y}_{it} = \nu_r + \gamma_t + \boldsymbol{X}\boldsymbol{\kappa} + \sum_{r \in \mathcal{R}} \tau_r \cdot \mathbb{1}\{W_i = r\} \mathbb{1}\{t \geq r\} + \epsilon_{it}$$

by OLS (TWFE_COVS). (Notice that adding unit-specific covariates to Model (1.3) induces collinearity. Fitting the model instead on cohort fixed effects, like ETWFE, also matches, e.g., Equation 3.2 in Callaway and Sant'Anna (2021). See also the discussion around Equation 2.5 in Sant'Anna and Zhao 2020.) The inclusion of covariates might make it seem more plausible that Assumptions (CNAS) and (CCTS) could hold in this model, though Wooldridge (2021, Section 6) points out that adding time-invariant covariates does not change the estimates of the treatment effects. Separate treatment effects for each cohort, while insufficient to avoid bias and capture the variation in treatment effects, should improve estimation relative to (1.3) and allows for the estimation of cohort-specific treatment effects to compare to the other methods.

7.1 Estimation Error

We start by evaluating the error of each method in estimating the average treatment effect (4.11). The assumptions of Theorem 6.1 are satisfied, so we expect FETWFE to estimate the treatment effects more accurately than ETWFE, which will suffer from high variance; BETWFE, which does not assume the correct form of sparsity; and TWFE_COVS, which is asymptotically biased. We estimate the average treatment effect using the FETWFE estimator (5.10) with ψ_{rt} from (4.6), as well as analogous estimators using the competitor estimated regression coefficients and observed cohort counts.

We calculate the squared error of each average treatment effect estimate for each method on each iteration. Boxplots of the results are displayed in Figure 3. We also provide the means and

standard errors for the squared errors of each method in Table 1. Table 2 contains p-values for paired one-tailed t-tests of the alternative hypothesis that the squared error for FETWFE is less than the squared error for each competitor method; all results show significantly better performance for FETWFE at the 0.05 significance level.

These results show that even if most of the restrictions we would naturally consider hold and we are only interested in overall average treatment effect (4.11), a simple model like TWFE_COVS is an unsuitable estimator. This aligns with the conclusions of previous works, as we discussed in the introduction. We see that ETWFE has better estimation error, though in practice it still does not perform as well as FETWFE because it estimates about ten times as many parameters as are needed, leading to imprecise estimates.
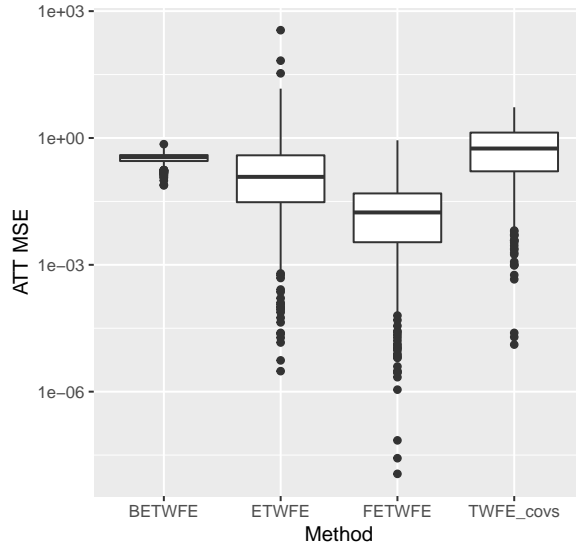

Figure 3: Boxplots of squared errors for each treatment effect estimate across all 700 simulations. Vertical axis is on a log scale.

Table 1: Means and standard errors for the squared error of the estimator of the ATT (4.11) for each method across all 700 simulations in the first simulation study from Section 7.

FETWFE	ETWFE	BETWFE	TWFE_COVS
0.0399 (0.00263)	1.19 (0.633)	0.339 (0.00336)	0.938 (0.0395)

Table 2: p-values from paired one-tailed t-tests of the alternative hypothesis that the FETWFE ATT estimator (5.10) has a lower squared estimation error for the ATT (4.11) than each competitor method calculated across all 700 simulations in the first simulation study from Section 7. (Low p-values indicate better predictive performance for FETWFE.)

ETWFE	BETWFE	TWFE_COVS
0.0343	1.24e-320	1.68e-86

In Appendix B, we also present estimation error results for the cohort-specific average treatment effects $\tau_{\text{ATT}}(r)$ as defined in (4.10) for each of the five cohorts, and estimation error for the coefficients of $\boldsymbol{\beta}_N^*$ corresponding to the interaction effects, $\boldsymbol{\rho}^*$. These results show that FETWFE outperforms the competitor methods for these estimands as well.

7.2 Restriction Selection Consistency

Next, we examine the extent to which we can trust FETWFE to select the correct restrictions, as we know it does with asymptotically high probability under Theorem 6.2. In the same experiment as before, on each simulation we calculate the percentage of restrictions FETWFE successfully identifies among the treatment parameters. In particular, we calculate the percentage of entries of $\boldsymbol{D\beta}_N^*$ that FETWFE successfully estimates as either zero or nonzero.

On average across the 700 simulations, FETWFE successfully decides whether to use a restriction or not in 95.0% of the p_N possible cases, with a standard error of 0.025%. Of the true restrictions (that is, the entries of $\boldsymbol{D\beta}_N^*$ that equal 0), on average 97.5% (0.021%) are correctly identified. This suggests that FETWFE is effective at identifying restrictions in practice, which contributes to FETWFE's success in estimating treatment effects that we saw in the previous section. (Figure 4 in Appendix B shows a boxplot of the proportion of restrictions successfully identified across all 700 simulations.)

To get a better sense of the asymptotics, we conduct an additional simulation study in the same way as the first one, but with $T = 5$,[5] $R = 3, d_N = 2$ (resulting in $p_N = 50$), and $N = 1200$ observations (resulting in 300 observations per cohort in expectation). The sparse $\boldsymbol{\theta}_N^*$ is generated as before, but with entries each equalling 0 with probability 0.5. Again, we conduct 700 simulations. In this second simulation study, FETWFE successfully decides whether to use a restriction or not in 99.3% (0.042%) of the p_N possible cases across the 700 simulations, and on average FETWFE identifies 98.5% (0.096%) of the true restrictions.

7.3 Asymptotic Distribution

Finally we investigate the asymptotic distribution of the FETWFE estimators. On each simulation, we estimate nominal asymptotic 95% confidence intervals for the cohort average treatment effects using (6.6) and the variance estimator from Theorem 6.4(a). In the first simulation study, we find coverage rates of 0.791 for the $r = 2$ cohort (standard error 0.015), 0.884 for the $r = 3$ cohort (standard error 0.012), 0.854 (0.013) for the $r = 4$ cohort, 0.803 (0.015) for the $r = 5$ cohort, and 0.850 (0.014) for the $r = 6$ cohort. We see that the coverage is unsatisfactory for FETWFE in this relatively small sample size, where in expectation 20 units are observed in each cohort.

We also estimate conservative confidence intervals for the average treatment effect with the conservative variance estimator from Theorem 6.4(c), yielding a coverage rate of 0.993 (0.001). That is, the conservative confidence intervals seem to achieve the desired coverage level in practice, even though the asymptotic distribution of the test statistic is not known to be Gaussian (merely subgaussian).

Lastly, we examine the non-conservative estimator from Theorem 6.4(b): on each simulation study, we generate N independent unlabeled samples where only cohort assignments (drawn from the same distribution as the N labeled units) are observed, and we use these to estimate asymptotically exact confidence intervals for the average treatment effect. The resulting coverage is 0.986 (0.001). The coverage may be better for the average treatment effect than the cohort average treatment effects because the number of observations that are relevant for estimating the overall average treatment effect is much larger than that for each of the cohort average treatment effects.

In the second simulation study, described previously in Section 7.2, we find coverage rates of 0.944 (0.009) for the $r = 2$ cohort, 0.941 (0.009) for the $r = 3$ cohort, and 0.930 (0.010) for the third cohort. The conservative confidence intervals for the average treatment effects have coverage 0.997 (0.001), and the asymptotically exact confidence intervals have coverage 0.973 (0.002). Overall, the coverage rates are close to the nominal 95% level in this finite N setting. This is especially promising because we are using heuristics to solve the FETWFE optimization problem, as discussed in Remark

---

[5]Notice that increasing T increases the number of observations while also increasing the number of treatment effects and interactions to estimate. So we might expect larger T to have a roughly neutral effect on the convergence properties, while greatly increasing the computational cost of simulations with large N.

5.2.

8   Empirical Application

We analyze the data from Stevenson and Wolfers (2006), as collected in the `divorce` data set from Flack and Jee (2020). Stevenson and Wolfers estimate the causal effect of unilateral or "no-fault" divorce laws on suicide rates among women. The data set consists of observations from $N = 51$ states (including the District of Columbia) for $T = 33$ years, along with demographic covariates. The response variable is the aggregate state suicide rate among women by year (as in the first column of Table 1 of Stevenson and Wolfers 2006). See Stevenson and Wolfers (2006) and Goodman-Bacon (2021, Section III) for more details on the data.

We remove states treated at time $T = 1$ from the data, leaving $N = 42$ observations. 5 states are never treated, and $R = 12$ cohorts are treated at various times. Similarly to the analysis in Goodman-Bacon (2021, Section IV(B)), we consider the time-varying covariates `murderrate` (state female homicide rate), `lnpersinc` (natural log of state personal income per capita), and `afdcrolls` (state welfare participation rate), including the values from time $t = 1$ as time-invariant controls. However, `murderrate` is missing for some observations at $t = 1$ so is omitted, leaving $d_N = 2$ covariates. In total, we have $NT = 1386$ observations and $p_N = 908$ coefficients to estimate.

We estimate the variances σ_c^2 (the variance of the unit random effects) and σ^2 (the variance of the added noise) using the estimators from Pesaran (2015, Section 26.5.1) with ridge regression. We use FETWFE with $q = 0.5$ to estimate the marginal average treatment effects, again leveraging the R `grpreg` package as in the simulation studies and selecting λ_N by BIC. We estimate that the overall average percent change to the suicide rate for women is -3.76%, which is similar to the estimate of -2.52% (standard error 1.09%) that Goodman-Bacon (2021) calculates when using similar controls. Applying Theorem 6.4(c), we estimate a conservative standard error of 4.70%, yielding a conservative 95% confidence interval of $(-12.97\%, 5.45\%)$ for the average treatment effect.

In Table 3 we present point estimates for the estimated average treatment effects for each of the cohorts. (We omit standard errors since the individual cohorts have very small sample sizes.) We estimate that four cohorts have non-zero treatment effects, with the 1970 cohort (consisting of California and Iowa) having by far the largest treatment effect estimate. We emphasize that these individual estimates are likely less reliable than the average treatment effect estimate since most cohorts have a small number of units.

Table 3: *Cohort average treatment effect estimates from the empirical application in Section 8, targeting estimand* $\tau_{ATT}(r)$ *as defined in (4.10).*

Cohort	$\hat{\tau}_{\mathrm{ATT}}(r)$
1969	0%
1970	-40.142%
1971	0%
1972	0%
1973	-3.466%
1974	0%
1975	0%
1976	-4.703%
1977	-5.338%
1980	0%
1984	0%
1985	0%

9   Conclusion

In this work, we have proposed the fused extended two-way fixed effects estimator for difference-in-differences with staggered adoptions. Our estimator leverages the fact that the extended two-way fixed effects estimator includes coefficients that are likely to be similar (for example, because of their proximity in time), but it selects appropriate restrictions automatically rather than relying on hand-selected restrictions, which could be biased. FETWFE retains the asymptotic unbiasedness (and normality) of ETWFE but leverages sparsity for large efficiency gains. In particular, we showed that our estimator is consistent, selection consistent, asymptotically Gaussian, and has an oracle property for several classes of heterogeneous marginal and conditional average treatment effects. Finally, we demonstrated the usefulness of FETWFE in practice in both simulation studies and an empirical application.

9.1   Future Work

Throughout the paper we have mentioned places where our work might be relatively straightforward to extend. Here we specify some other possible extensions. It would be useful to develop theory allowing for FETWFE to be estimated when the design matrix $\boldsymbol{Z}$ is deficient in column rank; for example, when $p_N > NT$. One could extend our method to this setting by suitably adapting the marginal bridge estimator of Kock (2013), which involves a screening step to remove irrelevant covariates before estimating a regression on the reduced design matrix in a second step. This allows for consistency and asymptotic normality of the second-step estimator under similar assumptions to those that we have used in this work. Alternatively, consistency (but not asymptotic normality) could be established under weaker assumptions in the $q = 1$ case by suitably adapting theory for lasso estimators like Corollary 2 from Negahban et al. (2012).

Another potentially useful extension would be relaxing the parallel trends assumption to allow for heterogeneous trends as described in Wooldridge (2021, Section 7.2). This requires adding extra parameters to the ETWFE estimator, but FETWFE is well-poised to adapt to this setting because it can automatically restrict to 0 any parameters that are added unnecessarily.

References

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

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

J. Abrevaya, Y.-C. Hsu, and R. P. Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

D. Arkhangelsky and G. Imbens. Causal models for longitudinal and panel data: A survey. Technical report, National Bureau of Economic Research, 2023.

E. Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects: Table 1. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, May 2016. doi: 10.1073/pnas. 1510489113.

S. Athey and G. W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

S. Athey and G. W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79, 2022.

S. Athey, D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt. Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data. *ArXiv e-prints*, Jan. 2018.

K. Ban and D. Kedagni. Generalized difference-in-differences models: Robust bounds. *arXiv preprint arXiv:2211.06710*, 2022.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

J. Bien. The simulator: an engine to streamline simulations. *arXiv preprint arXiv:1607.00021*, 2016.

A. Bilinski and L. A. Hatfield. Seeking evidence of absence: Reconsidering tests of model assumptions. *arXiv preprint arXiv:1805.03273*, 2018.

M. Blackwell, A. Glynn, H. Hilbig, and C. H. Phillips. Estimating controlled direct effects with panel data: An application to reducing support for discriminatory policies. 2024.

N. Bolger, K. S. Zee, M. Rossignac-Milon, and R. R. Hassin. Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, 148(4):601, 2019.

K. Borusyak and X. Jaravel. *Revisiting event study designs*. SSRN, 2018.

K. Borusyak, X. Jaravel, and J. Spiess. Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, page rdae007, 2024.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

J. E. Brand and Y. Xie. Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, 75(2):273–302, 2010.

R. Breen, S. Choi, and A. Holm. Heterogeneous causal effects and sample selection bias. *Sociological Science*, 2:351–369, 2015.

P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

P. Breheny and J. Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25:173–187, 2015.

N. Brown and K. Butts. A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models. Working Paper 1495, Economics Department, Queen's University, Nov. 2022. URL `https://ideas.repec.org/p/qed/wpaper/1495.html`.

C. Caetano, B. Callaway, S. Payne, and H. S. Rodrigues. Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903*, 2022.

B. Callaway. *Handbook of Labor, Human Resources and Population Economics*, chapter Difference-in-Differences for Policy Evaluation, pages 1–61. Springer International Publishing, 2023.

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

B. Callaway and T. Li. Policy evaluation during a pandemic. *Journal of Econometrics*, 236(1): 105454, 2023.

B. Callaway and P. Sant'Anna. Pre–testing in a did setup using the did package. *Vignette URL https://bcallaway11. github. io/did/articles/pre-testing. html*, 2022.

B. Callaway and P. H. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(3):551–577, 2018. ISSN 14679868. doi: 10.1111/rssb.12265. URL `https://rss-onlinelibrary-wiley-com.libproxy1.usc.edu/doi/pdf/10.1111/rssb.12265`.

N.-C. Chang. Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191, 2020.

V. Chernozhukov, D. Chetverikov, and Y. Koike. Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *The Annals of Applied Probability*, 33(3):2374–2425, 2023.

A. Chiu, X. Lan, Z. Liu, and Y. Xu. What to do (and not to do) with causal panel analysis under parallel trends: Lessons from a large reanalysis study. *Available at SSRN 4490035*, 2023.

C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, page 00491241221099552, 2022.

A. DasGupta. *Asymptotic theory of statistics and probability*, volume 180. Springer, 2008.

C. De Chaisemartin and X. d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

C. de Chaisemartin and X. D'Haultfœuille. Difference-in-differences for simple and complex natural experiments. *Available at SSRN*, November 2023. URL `https://ssrn.com/abstract=4487202`.

C. De Chaisemartin and X. d'Haultfoeuille. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, 26(3):C1–C30, 2023.

H. Dette. Difference-in-differences estimation under non-parallel trends. 2020.

P. Ding. Linear model and extensions. *arXiv preprint arXiv:2401.00649*, 2024.

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

H. Djebbari and J. Smith. Heterogeneous impacts in progresa. *Journal of Econometrics*, 145(1-2): 64–80, 2008.

P. B. Ellickson, W. Kar, and J. C. Reeder III. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, 2022.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. 2004.

Q. Fan, Y.-C. Hsu, R. P. Lieli, and Y. Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327, 2022.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

A. R. Feinstein, J. F. Rubinstein, and W. A. Ramshaw. Estimating prognosis with the aid of a conversational-mode computer program. *Annals of Internal Medicine*, 76(6):911–921, 1972.

E. Flack and E. Jee. *bacondecomp: Goodman-Bacon Decomposition*, 2020. URL `https://CRAN.R-project.org/package=bacondecomp`. R package version 0.1.1.

J.-P. Florens, J. J. Heckman, C. Meghir, and E. Vytlacil. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206, 2008.

J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

J. Frankovich, C. A. Longhurst, and S. M. Sutherland. Evidence-based medicine in the emr era. *N Engl J Med*, 365(19):1758–1759, 2011.

W. Gao, F. Xu, and Z.-H. Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022.

E. García-Portugués. *Notes for Nonparametric Statistics*. 2023. URL `https://bookdown.org/egarpor/NP-UC3M/`. Version 6.9.0. ISBN 978-84-09-29537-1.

E. Gavrilova, A. Langørgen, and F. Zoutman. Dynamic causal forests, with an application to payroll tax incidence in norway, 2023.

A. Gelman and M. Vákár. Slamming the sham: A bayesian model for adaptive adjustment with noisy control data. *Statistics in medicine*, 40(15):3403–3424, 2021.

D. Ghanem, P. H. C. Sant'Anna, and K. Wüthrich. Selection and parallel trends, 2024. URL `https://arxiv.org/abs/2203.09001`.

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

J. Grimmer, S. Messing, and S. J. Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

A. D. Güngör. Some bounds for the product of singular values. *Int. J. Contemp. Math. Sciences*, 2(26):1285–1292, 2007.

D. W. Ham and L. Miratrix. Benefits and costs of matching prior to a difference in differences analysis when parallel trends does not hold. *arXiv preprint arXiv:2205.08644*, 2022.

T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143:143, 2015.

J. Hatamyar, N. Kreif, R. Rocha, and M. Huber. Machine learning for staggered difference-in-differences and dynamic treatment effect heterogeneity, 2023. URL https://arxiv.org/abs/2310.11962.

J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4): 605–654, 1997.

J. J. Heckman, H. Ichimura, J. A. Smith, and P. E. Todd. Characterizing selection bias using experimental data, 1998.

P. Heiler and J. Mareckova. Shrinkage for categorical regressors. *Journal of Econometrics*, 223(1): 161–189, 2021.

D. J. Henderson and S. Sperlich. A complete framework for model-free difference-in-differences estimation. *Available at SSRN: https://ssrn.com/abstract=4300808*, (No. 15799), 2022. URL http://dx.doi.org/10.2139/ssrn.4300808.

G. J. Hitsch and S. Misra. Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957*, 2018.

H. Höfling, H. Binder, and M. Schumacher. A coordinate-wise optimization algorithm for the fused lasso. *arXiv preprint arXiv:1011.6409*, 2010.

R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511840371.

J. Huang, J. L. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36(2):587–613, 2008. ISSN 00905364. doi: 10.1214/009053607000000875.

K. Imai and I. S. Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2):467–490, 2019.

K. Imai and A. Strauss. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):1–19, 2011.

K. Imai, I. S. Kim, and E. H. Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2023.

G. Imbens and D. B. Rubin. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge University Press, 2015.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Y. Kaddoura and J. Westerlund. Estimation of panel data models with random interactive effects and multiple structural breaks when t is fixed. *Journal of Business & Economic Statistics*, pages 1–13, 2022.

A. Kahn-Lang and K. Lang. The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 38(3):613–620, 2020.

A. B. Kock. Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory*, 29(1):115–152, 2013.

S. Kwon. Optimal shrinkage estimation of fixed effects in linear panel data models. *EliScholar–A Digital Platform for Scholarly Publishing at Yal e*, page 1, 2021.

M. Lechner et al. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011.

E. L. Lehmann. *Elements of large-sample theory*. Springer, 1999.

L. Lesko. Personalized medicine: elusive dream or imminent reality? *Clinical Pharmacology & Therapeutics*, 81(6):807–816, 2007.

T. Levy and F. Abramovich. Generalization error bounds for multiclass sparse linear classifiers. *Journal of Machine Learning Research*, 24(151):1–35, 2023.

S. Lindner and K. J. McConnell. Difference-in-differences and matching on outcomes: a tale of two unobservables. *Health Services and Outcomes Research Methodology*, 19:127–144, 2019.

M. E. Lopes. Central limit theorem and bootstrap approximation in high dimensions: Near $1/n$ rates via implicit smoothing. *The Annals of Statistics*, 50(5):2492–2513, 2022.

C. F. Manski and J. V. Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244, 2018.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

P. Marx, E. Tamer, and X. Tang. Parallel trends and dynamic choices. *Journal of Political Economy Microeconomics*, 2(1):129–171, 2024.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.

B. D. Meyer. Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161, 1995.

S. L. Morgan. Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. *Sociology of education*, pages 341–374, 2001.

M. Murphy, S. Redding, and J. Twyman. *Handbook on personalized learning for states, districts, and schools*. IAP, 2016.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. 2012.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2021.

X. Nie, C. Lu, and S. Wager. Nonparametric heterogeneous treatment effect estimation in repeated cross sectional designs, 2021. URL https://arxiv.org/abs/1905.11622.

C. Park and E. T. Tchetgen. A universal difference-in-differences approach for causal inference. 2023. URL https://arxiv.org/abs/2212.13641.

J. H. Park and S. Yamauchi. Change-point detection and regularization in time series cross-sectional data analysis. *Political Analysis*, 31(2):257–277, 2023.

M. H. Pesaran. *Time Series and Panel Data Econometrics*. Number 9780198759980 in OUP Catalogue. Oxford University Press, 2015. URL https://ideas.repec.org/b/oxp/obooks/9780198759980.html.

A. Rambachan and J. Roth. A more credible approach to parallel trends. *Review of Economic Studies*, page rdad018, 2023.

J. Roth. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322, 2022.

J. Roth and P. H. Sant'Anna. When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747, 2023.

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

S. Rothbard, J. C. Etheridge, and E. J. Murray. A tutorial on applying the difference-in-differences method to health data. *Current Epidemiology Reports*, pages 1–11, 2023.

P. H. Sant'Anna and J. Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

V. Semenova and V. Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

A. M. Shaikh and P. Toulis. Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848, 2021.

B. Stevenson and J. Wolfers. Bargaining in the shadow of the law: Divorce laws and family distress. *The Quarterly Journal of Economics*, 121(1):267–288, 2006.

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

L. Sun and J. M. Shapiro. A linear panel model with heterogeneous coefficients and variation in exposure. *Journal of Economic Perspectives*, 36(4):193–204, 2022.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 58(1):267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1): 91–108, 2005.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011.

A. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN 9780521784504. URL https://books.google.com/books?id=UEuQEM5RjWgC.

R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108415194. URL `https://books.google.com/books?id=J-VjswEACAAJ`.

V. Viallon, S. Lambert-Lacroix, H. Höfling, and F. Picard. Adaptive generalized fused-lasso: Asymptotic properties and applications. 2013.

V. Viallon, S. Lambert-Lacroix, H. Hoefling, and F. Picard. On the robustness of the generalized fused lasso to prior specifications. *Statistics and Computing*, 26(1-2):285–301, 2016.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, jul 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1319839. URL `https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839`.

L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

C. Wing, K. Simon, and R. A. Bello-Gomez. Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, 39(1):453–469, 2018. doi: 10.1146/annurev-publhealth-040617-013507. URL `https://doi.org/10.1146/annurev-publhealth-040617-013507`. PMID: 29328877.

S. Winkelbeiner, S. Leucht, J. M. Kane, and P. Homan. Evaluation of differences in individual treatment response in schizophrenia spectrum disorders: a meta-analysis. *JAMA psychiatry*, 76 (10):1063–1073, 2019.

J. M. Wooldridge. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*, 2021.

Y. Xie, J. E. Brand, and B. Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

B. Xin, Y. Kawahara, Y. Wang, and W. Gao. Efficient generalized fused lasso and its application to the diagnosis of alzheimer's disease. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

J. Yang, D. Eckles, P. Dhillon, and S. Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*, 2020.

P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

M. Zimmert. Efficient difference-in-differences estimation with high-dimensional common trend confounding, 2020. URL `https://arxiv.org/abs/1809.01643`.

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006. ISSN 0162-1459. URL `https://doi.org/10.1198/016214506000000735`.

In Appendix A we investigate in more detail the relatively plausibility of conditional and marginal parallel trends, how Assumption (CIUN) interacts with these assumptions, and how we can achieve consistency of FETWFE for marginal average treatment effects under either form of parallel trends if (CIUN) holds. We present additional results from the simulation studies from Section 7 in Appendix B. We state Theorem C.1, which Theorems 6.1, 6.2, 6.3, and 6.4 all depend on, in Appendix C, and we prove it in Appendix F. In Appendix D we state Theorem D.1, which shows the consistency of FETWFE for broader classes of estimators than those considered in Theorem 6.1. Theorem D.2 similarly extends Theorem 6.4, and Theorem D.3 shows that two classes of conditional treatment effect estimators are asymptotically Gaussian and provides consistent variance estimators for constructing asymptotically valid confidence intervals, analogously to Theorem D.2 for marginal treatment effect estimators. We prove the results from Appendix D, along with Theorems 6.1, 6.3, and 6.4, in Appendix E. In Appendix G we provide proofs for results that were stated but not proven in Appendix A and supporting results used in Appendix E. We prove supporting results for Theorem C.1 in Appendix H. Finally, in Appendix I we state and prove some technical lemmas.

A   More on Parallel Trends and Assumption (CIUN)

In Section A.1 we discuss the plausibility of marginal and conditional parallel trends in relation to an assumption of marginal or conditional (respectively) independence of the treatment and the potential outcomes. The latter assumptions imply the former, which motivates difference-in-differences estimators like FETWFE as more plausible than identification strategies that rely on stronger untestable assumptions.

We point out that no such strict hierarchy exists to relate marginal and conditional parallel trends in Section A.2—there exist settings where either assumption can hold without the other holding. Investigating this sheds light on when conditional or marginal parallel trends may be more plausible. We later show (Theorem A.5) that conditional parallel trends is necessary to estimate conditional average treatment effects via FETWFE, but there are estimators that are consistent for marginal average treatment effects under either conditional or marginal parallel trends. So it is natural to wonder what conditions allow an estimator to be consistent for marginal average treatment effects under either assumption, and whether this might be possible for FETWFE. In Section A.3 we explore when this might hold, ultimately investigating Assumption (CIUN) in more detail and showing how it affects the consistency of regression (1.5). These ideas underlie the proof of Theorem 6.1.

A.1   Relationships Between Unconfounded Treatment Assignment, Marginally Randomized Treatment Assignment, and Parallel Trends Assumptions

In this section we present Proposition A.1, which relates the plausibility of several assumptions on the potential outcomes. Ultimately, this motivates difference-in-differences estimators like FETWFE as having more plausible untestable assumptions than methods that rely on unconfoundedness (like matching) or randomized assignment.

We consider an assumption of unconfoundedness of the untreated potential outcomes, which we define in this setting as

$$\tilde{y}_{it}(0) \perp\!\!\!\perp W_i \mid \boldsymbol{X}_i \qquad i \in [N], t \in \{1, \ldots, T\}. \tag{A.1}$$

This requires that treatment assignment is independent of the untreated potential outcomes conditional on the time-invariant (pre-treatment) covariates $\boldsymbol{X}_i$. Abadie (2005, Section 3.2.2) considered a similar assumption in the $T = 2$ case. As we discuss in more detail later in this section, it has previously been noted that the assumption of unconfoundedness (Imbens and Rubin, 2015, Def. 3.6) that is common in causal inference is a stronger assumption than (CCTS).

Sometimes unconfoundedness is taken to be a stronger assumption that all of the potential outcomes are independent of treatment status conditional on covariates; such a condition would

imply (A.1). Shaikh and Toulis (2021, Assumption 2.1) is one example of a work that uses such a stronger assumption in a setting with continuous treatment status.

We note that different notions of unconfoundedness have been considered in the difference-in-differences setting. For example, Imbens and Wooldridge (2009, Section 6.5.4) consider unconfoundedness in a two-period difference-in-differences setting with no covariates, where the lagged outcome is conditioned on rather than covariates. Similarly, Callaway and Li (2023) consider unconfoundedness conditional on several lagged time-varying variables. Ding and Li (2019) and Lindner and McConnell (2019) consider an ignorability condition where the untreated potential outcome is independent of treatment assignment status conditional on covariates and the lagged outcome.

We also refer to an assumption of marginal independence of the untreated potential outcomes and treatment status,

$$\tilde{y}_{it}(0) \perp\!\!\!\perp W_i \qquad i \in [N], t \in \{1, \ldots, T\}. \tag{A.2}$$

Completely randomized assignment, an assumption considered by, for example, Athey and Imbens (2022, Assumption 1) and Roth and Sant'Anna (2023, Case 1), is sufficient for (A.2).

Lastly, we say that (CCTS) or (CTS) holds in a *transformation-invariant* sense if (3.1) or (3.2) (respectively) hold under any transformation $h(\tilde{y}_{it}(r))$ of the potential outcomes such that the expectations exist. Some works have expressed concern that parallel trends assumptions are fragile because they may hold only if the response is not transformed, or only under a specific transformation (Meyer, 1995; Athey and Imbens, 2006; Kahn-Lang and Lang, 2020). Roth and Sant'Anna (2023) investigate in detail when marginal parallel trends assumptions are transformation-invariant in the $T = 2$ setting, and they point out that their arguments extend straightforwardly to the arbitrary T setting as well as to conditional parallel trends assumptions. Proposition A.1 explicitly draws attention to an implication of their Propositions 3.1 and 3.2 when extended to these more general settings.

Proposition A.1. *Assume that the conditional expectations in (3.1) and (3.2) exist and are finite.*

(a) *Unconfoundedness of the untreated potential outcomes is a strictly stronger assumption than conditional parallel trends: (A.1) is sufficient for Assumption CCTS to hold in a transformation-invariant sense, but (CCTS) does not imply unconfoundedness of the untreated potential outcomes.*

(b) *Marginal independence of treatment status and the untreated potential outcomes is a strictly stronger assumption than marginal parallel trends: (A.2) is sufficient for Assumption CTS to hold in a transformation-invariant sense, but (CTS) does not imply (A.2).*

Proof. (a) Under unconfoundedness (A.1), for any $i \in [N]$, $t \in \{1, \ldots, T\}, r \in \mathcal{R}$ almost surely it holds that

$$\mathbb{E}[h(\tilde{y}_{(it)}(0)) \mid W_i, \boldsymbol{X}_i] = \mathbb{E}[h(\tilde{y}_{(it)}(0)) \mid \boldsymbol{X}_i] \tag{A.3}$$

for any arbitrary $h(\cdot)$ such that the above expectation exists. For (CCTS) to hold with transformational invariance, we need that for any $h(\cdot)$ such that the expectations exist and any $i \in [N]$, $t \in \{2, \ldots, T\}, r \in \mathcal{R}$, almost surely

$$\mathbb{E}[h(\tilde{y}_{(it)}(0)) - h(\tilde{y}_{(i1)}(0)) \mid W_i, \boldsymbol{X}_i] = \mathbb{E}[h(\tilde{y}_{(it)}(0)) - h(\tilde{y}_{(i1)}(0)) \mid \boldsymbol{X}_i].$$

But this is immediate from (A.3). We show that the reverse implication does not hold by constructing a counterexample. (CCTS) can be rearranged as

$$\mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i, \boldsymbol{X}_i] - \mathbb{E}[\tilde{y}_{(it)}(0) \mid \boldsymbol{X}_i] = \mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i, \boldsymbol{X}_i] - \mathbb{E}[\tilde{y}_{(i1)}(0) \mid \boldsymbol{X}_i] \ \forall t \in [T]. \tag{A.4}$$

Suppose that for all $t \in \{1, 2, \ldots, T\}$, it holds that $\mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i, \boldsymbol{X}_i] - \mathbb{E}[\tilde{y}_{(it)}(0) \mid \boldsymbol{X}_i] = f(W_i, \boldsymbol{X}_i)$ almost surely for an arbitrary $f(\cdot, \cdot)$ that is not identically 0. Then from (A.4) we see that (CCTS) will hold. However, unconfoundedness does not hold since (A.3) with $h(\cdot)$ as the identity function is violated.

(b) This proof is similar. Under marginal independence of treatment status and the untreated potential outcomes (A.2), for any $i \in [N]$, $t \in \{1, \ldots, T\}$, $r \in \mathcal{R}$ almost surely it holds that

$$\mathbb{E}[h(\tilde{y}_{(it)}(0)) \mid W_i] = \mathbb{E}[h(\tilde{y}_{(it)}(0))] \qquad (A.5)$$

for any $h(\cdot)$ such that the expectation holds. It is immediate that for any such $h(\cdot)$ and any $i \in [N]$, $t \in \{2, \ldots, T\}$, $r \in \mathcal{R}$, almost surely $\mathbb{E}[h(\tilde{y}_{(it)}(0)) - h(\tilde{y}_{(i1)}(0)) \mid W_i] = \mathbb{E}[h(\tilde{y}_{(it)}(0)) - h(\tilde{y}_{(i1)}(0))]$, so (CTS) holds with transformational invariance. On the other hand, (CTS) can be rearranged as

$$\mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i] - \mathbb{E}[\tilde{y}_{(it)}(0)] = \mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i] - \mathbb{E}[\tilde{y}_{(i1)}(0)] \qquad \forall t \in [T]. \qquad (A.6)$$

If for all $t \in \{1, 2, \ldots, T\}$ we have $\mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i] - \mathbb{E}[\tilde{y}_{(it)}(0)] = g(W_i)$ almost surely for an arbitrary $g(\cdot)$ that is not identically 0, then (A.6) holds but (A.5) with $h(\cdot)$ as the identity function is violated. So (CTS) holds but marginal independence of treatment status and the untreated potential outcomes does not.

□

Since unconfoundedness often seems more plausible than marginal independence of the potential outcomes and treatment status, Proposition A.1(a) and (b) considered together suggest an intuition that conditional parallel trends may often be more plausible than marginal parallel trends. Proposition A.1 also points out that difference-in-differences estimators are strictly more robust than identification strategies that rely on assumptions of completely randomized assignment or unconfoundedness.

Proposition A.1(a) verifies that (CCTS) is a strictly weaker assumption than unconfoundedness. As we mentioned earlier, this is a fairly straightforward observation that has been noted before; for example, Abadie (2005); Heckman et al. (1997), and Heckman et al. (1998) mention that this fact holds in the $T = 2$ case. Besides the previously mentioned intuition that Proposition A.1(a) and (b) provide that conditional parallel trends may be more plausible in many applications than marginal parallel trends, we also draw attention to the facts in Proposition A.1(a) to point out explicitly that difference-in-differences identification strategies like FETWFE that rely on (CCTS) are more robust than methods like matching on time-invariant covariates that rely on an unconfoundedness assumption (Abadie and Imbens, 2006; Imbens and Rubin, 2015).

The implications of Proposition A.1(b) are analogous, and previous works have also shown that random or "as good as random" treatment assignment is sufficient for (CTS) (Athey and Imbens, 2022; Roth and Sant'Anna, 2023; Ghanem et al., 2024).

A.2 Neither Conditional Nor Marginal Parallel Trends is Strictly More Plausible

Proposition A.1 showed that (CTS) and (CCTS) are strictly more plausible than (respectively) marginal independence of the potential outcomes and treatment and unconfoundedness. Theorem A.2 shows that (CTS) and (CCTS) cannot be related in this way—neither is strictly more plausible than the other.

Theorem A.2. *Assume that the conditional expectations in (3.1) and (3.2) exist and are finite. Neither of conditional parallel trends or marginal parallel trends is a relaxation of the other: Assumption CCTS (3.1) (conditional parallel trends) does not imply Assumption CTS (3.2) (marginal parallel trends), nor does (CTS) imply (CCTS).*

Proof. Provided in Appendix G.1.

□

We prove Theorem A.2 by providing counterexamples where one of the parallel trends assumptions holds but the other does not. We briefly provide some intuition behind the counterexamples here in order to highlight some of the relative strengths and weaknesses of each assumption.

(CCTS) may hold when the time-invariant covariates have a time-varying effect on the potential outcomes, while (CTS) will not hold. This seems plausible if external conditions change that change the relevance of some features for the response. Ham and Miratrix (2022) and Callaway and Karami (2023) touch on closely related issues, investigating settings where parallel trends does not hold because time-invariant confounders have time-varying effects on the potential outcomes. We also expand on related ideas in Section A.3.

On the other hand, (CCTS) requires parallel trends to hold on every subgroup of $\boldsymbol{X}_i$. But (CTS) can hold if this is violated and parallel trends holds merely on average across the distribution of the covariates. That is, for any $r \in \{0\} \cup \mathcal{R}$, the quantity

$$\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] - \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid \boldsymbol{X}_i] \tag{A.7}$$

may be positive on some subgroups of $\boldsymbol{X}_i$ if this is canceled out by (A.7) being negative on other subgroups such that (3.2) holds after marginalizing the left side of (3.1) across the distribution of $\boldsymbol{X}_i$ conditional on $W_i = r$ and the right side across the distribution of $\boldsymbol{X}_i$ conditional on $W_i = 0$. In this particular sense, (CTS) allows for more variation in the time trends across the distribution of $\boldsymbol{X}_i$ than (CCTS). We also emphasize that both (CCTS) and (CTS) allow for heterogeneity in the sense that the values of $\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i, \boldsymbol{X}_i]$ and $\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid \boldsymbol{X}_i]$ may vary with $\boldsymbol{X}_i$.

However, if (3.1) does not hold, (3.2) holding exactly after marginalizing (3.1) appropriately may seem somewhat implausible. Nonetheless, it could be the case that (3.1) is not close to holding, but (3.2) approximately holds. Relaxing (CTS) by allowing small violations has been considered in recent works on sensitivity analyses for difference-in-differences estimation (Manski and Pepper, 2018; Ban and Kedagni, 2022; Rambachan and Roth, 2023; Ghanem et al., 2024). That said, Ghanem et al. (2024, Section 7.2) point out that adjusting for covariates (that is, performing a sensitivity analysis under an assumption of conditional parallel trends that may not hold exactly) can result in smaller plausible ranges of marginal average treatment effects.

Overall, this suggests an intuition that (CCTS) may be somewhat more plausible in practice than (CTS), but since neither assumption nests the other, an ideal estimator would be asymptotically unbiased under either assumption. We discuss this in more detail in the next section.

A.3 Parallel Trends, Assumption (CIUN), and Regression (1.5)

Theorem A.2 motivates interest in estimators of marginal average treatment effects that are consistent under either conditional or marginal parallel trends. We have seen that Assumption (CIUN) is helpful for this—under the conditions of Theorem 6.1, FETWFE consistently estimates heterogeneous marginal average treatment effects under either conditional or marginal parallel trends if (CIUN) holds. In this section we fill in more detail about (CIUN), build some intuition for its role in the consistency of difference-in-differences estimators, and establish the consistency of regression (1.5) under (CIUN).

We first introduce a related assumption that, unlike (CIUN), is not directly testable.

Assumption (CIU) ("conditional independence of trends of the untreated potential outcomes"): Almost surely for all $r \in \{0\} \cup \mathcal{R}$, $i \in [N]$, and any $t \in \{2, \ldots, T\}$,

$$\mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r\right] = \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i\right].$$

(CIU) is similar to (CIUN), but it requires the trends in the untreated potential outcomes to be mean-independent of $\boldsymbol{X}_i$ for units in all cohorts, so it cannot be tested.

It is immediate that (CCTS) and (CTS) are equivalent under (CIU).

Proposition A.3. *Under (CIU), conditional parallel trends (CCTS) and marginal parallel trends (CTS) are equivalent.*

Proof. We can apply (CIU) on each side of (CTS) to immediately yield that for any $i \in [N]$, $r \in \mathcal{R}$, and any $t \geq 2$,

$$\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0]$$

$$\iff \quad \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i].$$

$\square$

One implication of Proposition A.3 is that any method that consistently estimates the marginal average treatment effects under either (CCTS) or (CTS) is consistent under both assumptions if (CIU) holds.

Although (CIUN) does not guarantee the equivalence of (CTS) and (CCTS), under (CIUN) conditional and marginal parallel trends are still connected.

Proposition A.4. *Suppose (CIUN) holds. Then conditional parallel trends (CCTS) implies marginal parallel trends (CTS).*

Proof. Applying (CIUN) to one side of (CCTS), we have that for any $i \in [N]$, $r \in \mathcal{R}$, and any $t \geq 2$,

$$\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i]$$

$$\iff \quad \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0]$$

$$\implies \mathbb{E}\left[\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i] \mid W_i = r\right] = \mathbb{E}\left[\mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0] \mid W_i = r\right]$$

$$\iff \quad \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r\right] = \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0].$$

$\square$

Proposition A.4 shows that any method that consistently estimates the marginal average treatment effects is also consistent under (CIUN) and (CCTS).

Notice that Assumptions (CIU) and (CIUN) are closely related to the time-varying effects of the covariates on the potential outcomes that we mentioned in our discussion of Theorem A.2 in Section A.2, which can cause marginal parallel trends to fail to hold even if conditional parallel trends holds. Proposition A.4 shows that (CIUN) rules out such settings.

Next we present a key result that underlies Theorem C.1, and in turn Theorems 6.1, 6.3, and 6.4.

Theorem A.5 (Assumption (CIUN) and the Consistency of Regression 1.5)**.** *Assume that (CNAS) and (3.6) from (LINS) hold.*

(a) *(Consistency under conditional parallel trends.) The following are equivalent: (i) conditional parallel trends after treatment (CCTSA) holds, and (ii) for all $r \in \mathcal{R}$ and all $t \geq r$, for the regression estimands τ_{rt}^* and $\boldsymbol{\rho}_{rt}^*$ defined in (3.6) it holds almost surely that the conditional average treatment effects $\tau_{CATT}(r, t, \boldsymbol{X}_i)$ defined in (1.6) equal $\tau_{rt}^* + (\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r])^\top \boldsymbol{\rho}_{rt}^*$.*

(b) *(Consistency under marginal parallel trends and CIUN.) If (CIUN) and marginal parallel trends after treatment (CTSA) hold, then for all $r \in \mathcal{R}$ and all $t \geq r$ the regression estimand τ_{rt}^* is equal to the marginal average treatment effect $\tau_{ATT}(r, t)$ from (1.1)..*

(c) *(Identifying the bias-inducing covariates under CTSA if CIUN does not hold.) Suppose (CTSA) and (3.5) from (LINS) hold. Then for any $r \in \mathcal{R}$ and $t \in \{r, \ldots, T\}$,*

$$\tau_{rt}^* = \tau_{ATT}(r, t) - (\mathbb{E}[\boldsymbol{X}_i \mid W_i = r] - \mathbb{E}[\boldsymbol{X}_i \mid W_i = 0])^\top \boldsymbol{\xi}_t^*.$$

Proof. The proofs of parts (b) and (c) are provided in Appendix G.1. Part (a) follows from Wooldridge (2021, Sections 6.3 and 6.4), but for illustrative purposes we provide a proof in our notation here. For all $r \in \mathcal{R}, t \geq r$ and any $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$,

$$
\tau_{rt}^* + \dot{\boldsymbol{x}}_r^\top \boldsymbol{\rho}_{rt}^*
$$

$$
\overset{(a)}{=} \mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(i1)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right] - \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right]
$$

$$
\overset{(b)}{=} \mathbb{E}[\tilde{y}_{(it)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}] - \Big(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}]
$$

$$
+ \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]\Big)
$$

$$
\overset{(c)}{=} \mathbb{E}[\tilde{y}_{(it)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}] - \mathbb{E}[\tilde{y}_{(it)}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}]
$$

$$
= \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x}), \tag{A.8}
$$

where in (a) we used (3.6) from (LINS), in (b) we used (CNAS), and (c) follows from rearranging (CCTSA).

$\square$

A.3.1   Discussion of Theorem A.5

Theorem A.5(a) and (b) show the consistency of regression (1.5): if (LINS), (CNAS), and (CIUN) hold, any estimator that is consistent for the regression estimands in (3.6) is also consistent for the marginal average treatment effects under either conditional parallel trends after treatment or marginal parallel trends after treatment. Theorem A.5(a) and (b) applies to ETWFE immediately under the conditions of Theorem 6.1(ii) in Wooldridge (2021), which shows that the ETWFE estimator of τ_{rt}^* is consistent. It also applies to FETWFE in light of Theorem C.1(b), which is used to prove Theorem 6.1.

Theorem A.5(c) shows that it is specifically the covariates $\boldsymbol{X}_i \mathbb{1}\{t = t'\}$ for $t' \in \{2, \ldots, T\}$ (which correspond to regression estimands $\boldsymbol{\xi}_t^*$) included in regression (1.5) that induce an asymptotic bias in the estimated marginal average treatment effects $\tau_{\mathrm{ATT}}(r, t')$ if (CIUN) does not hold. Suppose we assume that (3.5) from (LINS) holds and $\boldsymbol{X}_i$ and W_i are not mean-independent; in particular, for all $r \in \mathcal{R}$ we have $\mathbb{E}[\boldsymbol{X}_i \mid W_i = r] \neq \mathbb{E}[\boldsymbol{X}_i \mid W_i = 0]$[6]. Then if (CCTSA) and (CIUN) both do not hold, not only does Theorem A.5(a) show that there is no hope of estimating conditional average treatment effects, Theorem A.5(c) shows that in general including the covariates $\boldsymbol{X}_i \mathbb{1}\{t = t'\}$ in a regression induces bias in the estimated average treatment effects (since if (CIUN) does not hold and (3.5) from (LINS) does then $\boldsymbol{\xi}_t^* \neq \boldsymbol{0}$ for at least some t). This is true even if (CTS) holds and the marginal average treatment effects are estimable (by, for example, model (6.15) in Wooldridge 2021, which omits these covariates). That is, the covariates $\boldsymbol{X}_i \mathbb{1}\{t = t'\}$ for $t' \in \{2, \ldots, T\}$ are "bad controls," in the language of Cinelli et al. (2022). This observation aligns with the observation that Wooldridge (2021, Section 6.3) makes that including time-invariant covariates and their interactions with cohort dummies has no effect on the estimated marginal average treatment effects—it is specifically the inclusion of the covariate/time dummy interactions that changes the estimand for each τ_{rt}^*.

Notice that these covariates correspond exactly to the kind of effect of the covariates on the untreated trends that is ruled out by Assumptions (CIU) and (CIUN), as well as the time-varying effects of the covariates on the potential outcomes that we mentioned in our discussion of Theorem A.2 in Section A.2.

---

[6]Notice from Theorem A.5(c) that $\tau_{rt}^* = \tau_{\mathrm{ATT}}(r, t)$ if and only if either $\mathbb{E}[\boldsymbol{X}_i \mid W_i = r] = \mathbb{E}[\boldsymbol{X}_i \mid W_i = 0]$ or $\boldsymbol{\xi}_t^* = \boldsymbol{0}$. There is an interesting connection here: it is also true that (CCTS) implies (CTS) under either independence of $\boldsymbol{X}_i$ and W_i or (CIUN), which under (LINS) is equivalent to $\boldsymbol{\xi}_t^* = \boldsymbol{0}$. The latter fact is from Proposition A.4. Under independence of $\boldsymbol{X}_i$ and W_i, the distribution of $\boldsymbol{X}_i$ conditional on $W_i = r$ is identical to the marginal distribution of $\boldsymbol{X}_i$ for all $r \in \{0\} \cup \mathcal{R}$, so we can marginalize each side of (3.1) across the marginal distribution of $\boldsymbol{X}_i$ to yield (3.2).

However, if (3.5) holds, we see that under (CIUN) the coefficients $\boldsymbol{\xi}_t^*$ must all equal $\mathbf{0}$. In this case, Theorem A.5(c) shows that no harm is done to asymptotic unbiasedness by including those coefficients in the regression, and any estimator that is consistent for τ_{rt}^* from (3.6) is consistent for $\tau_{\mathrm{ATT}}(r, t)$. (This aligns with part (b), which shows that if (CIUN) holds then we can consistently estimate the marginal average treatment effects under (CTSA) with regression 1.5.)

As we showed in Theorem 6.3, under the required assumptions no harm is done to the asymptotic efficiency of FETWFE by including irrelevant covariates (as we mentioned earlier in Remark 3.1), like when (CIUN) holds and all of the $\boldsymbol{\xi}_t^*$ equal $\mathbf{0}$.

Theorem A.5(c) has similarities to Theorems 4.1 and 5.1 from Ham and Miratrix (2022), which analyze the $T = 2$, $\mathcal{R} = \{2\}$ and arbitrary T, $\mathcal{R} = \{T\}$ cases, respectively. In particular, their estimand $\mathbb{E}[\hat{\tau}_{\mathrm{DiD}}^X]$ matches our estimand τ_{rt}^* under our assumptions, and they also identify the bias of this estimand in order to investigate when conditional parallel trends might be more plausible than marginal parallel trends. Besides the fact that we consider a more general staggered adoptions (arbitrary $\mathcal{R}$) setting, they also consider different structural equations for the potential outcomes (for example, they omit interactions between the covariates and treatment status or cohort membership, though they do allow for time-varying coefficients on the covariates) and they consider a setting with a latent confounder.

B   Additional Results For Simulation Studies

Table 4 contains results for the cohort-specific squared estimation errors for the cohort-specific average treatment effects $\tau_{\mathrm{ATT}}(r)$ as defined in (4.10). We see that on average FETWFE outperforms every method for most cohorts, except that BETWFE outperforms FETWFE on average for cohort 2. This is because the true treatment effect for Cohort 2 is close to 0 (0.069), so BETWFE's direct penalization of the treatment coefficients directly towards 0 happens to be useful. In addition, this difference in the mean outcomes seems to be driven by outliers—we see from the boxplots in Figure 5 that in the median case FETWFE still outperforms BETWFE.

Table 5 contains p-values for paired one-tailed t-tests of the alternative hypothesis that FETWFE has lower squared estimation error than the competitor methods, showing that the differences are statistically significant at the 0.01 level in all settings except for the Cohort 2 BETWFE estimate and the last two ETWFE estimates, which are not significant because the ETWFE estimator has very high variance (see Table 4).

Table 6 contains results for the sum of squared errors of all methods' estimates of the treatment interaction coefficients ρ^*. We see that FETWFE outperforms every method. Table 7 contains p-values for paired one-tailed t-tests of the alternative hypothesis that FETWFE has lower squared estimation error than the competitor methods, showing that the differences are statistically significant at the 0.05 level.

Figure 4 depicts a boxplot of the proportion of the p_N possible restrictions that FETWFE selects correctly in the first simulation study. Figures 5, 6, 7, 8, and 9 display boxplots of these squared errors, similar to Figure 3 for the average treatment effect. These plots also indicate that FETWFE typically outperforms the other methods.

C   Theorem C.1

Theorem C.1 connects the bridge regression theory of Kock (2013) to the difference-in-differences theory of Wooldridge (2021). It also makes use of some extensions of the theory of Kock (2013) (Theorem F.4, Theorem F.5). Equipped with Theorem C.1, Theorems 6.1, 6.2, 6.3, and 6.4 are relatively straightforward to prove.

Theorem C.1. *(a) (Correct specification.) Suppose Assumptions (CNAS), (CCTSB), and (LINS)*

Table 4: Means and standard errors for the squared errors of estimates of $\tau_{ATT}(r)$ by each method for each cohort in the simulation study from Section 7.

	FETWFE	ETWFE	BETWFE	TWFE_COVS
Cohort 2	0.035 (0.006)	1.502 (0.312)	0.008 (0)	7.265 (0.163)
Cohort 3	0.037 (0.004)	0.943 (0.19)	0.425 (0.011)	1.881 (0.091)
Cohort 4	0.129 (0.009)	0.856 (0.179)	3.492 (0.016)	1.269 (0.062)
Cohort 5	0.093 (0.007)	4.12 (3.079)	4.794 (0.011)	1.685 (0.082)
Cohort 6	0.068 (0.005)	3.212 (2.016)	3.759 (0.016)	0.949 (0.049)

Table 5: p-values for paired one-tailed t-tests of the alternative hypothesis that the FETWFE estimate of $\tau_{ATT}(r)$ has lower squared error than the competitor methods for each cohort in the simulation study from Section 7.

	ETWFE	BETWFE	TWFE_COVS
Cohort 2	1.14e-05	1	8.2e-206
Cohort 3	8.06e-06	2.35e-145	1.85e-72
Cohort 4	0.000119	0	3.89e-63
Cohort 5	0.118	0	1.12e-68
Cohort 6	0.0795	0	8.71e-60

Table 6: Means and standard errors for the summed squared errors of estimates of the treatment interaction coefficients $\boldsymbol{\rho}^*$ by each method for each cohort in the simulation study from Section 7.

FETWFE	ETWFE	BETWFE
0.529 (0.00827)	38.1 (18.3)	4.82 (0.0119)

Table 7: p-values for paired one-tailed t-tests of the alternative hypothesis that the FETWFE estimates of the treatment interaction coefficients $\boldsymbol{\rho}^*$ have lower summed squared error than the competitor methods for each cohort in the simulation study from Section 7.

ETWFE	BETWFE
0.02	0
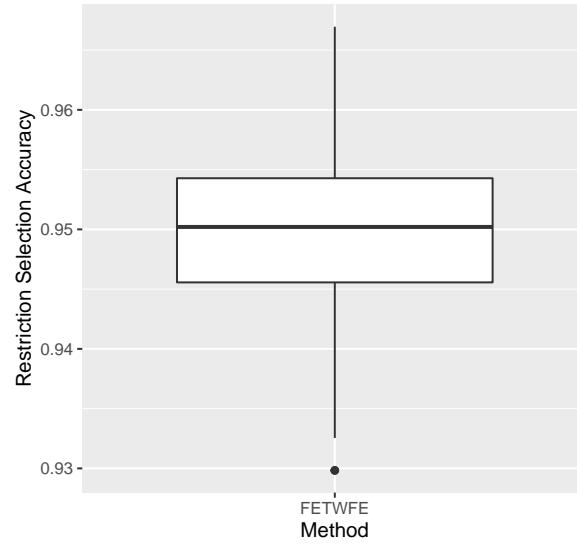
Figure 4: Boxplot displaying proportions of correct treatment effect restriction decisions correctly made by FETWFE across each of the 700 simulations from the first simulation study in Section 7.
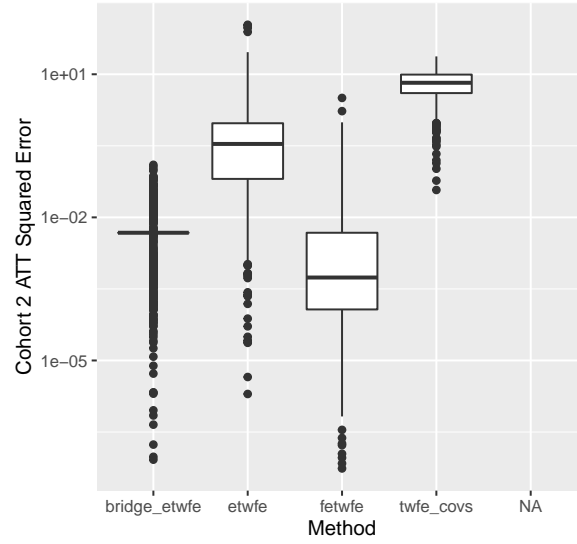
Figure 5: Boxplots of squared errors for each method's estimate of $\tau_{ATT}(2)$ across all 700 simulations. Vertical axis is on a log scale.
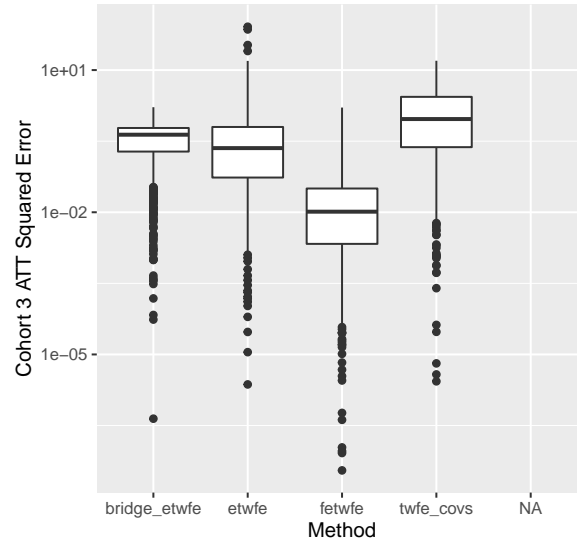
Figure 6: Boxplots of squared errors for each method's estimate of $\tau_{ATT}(3)$ across all 700 simulations. Vertical axis is on a log scale.
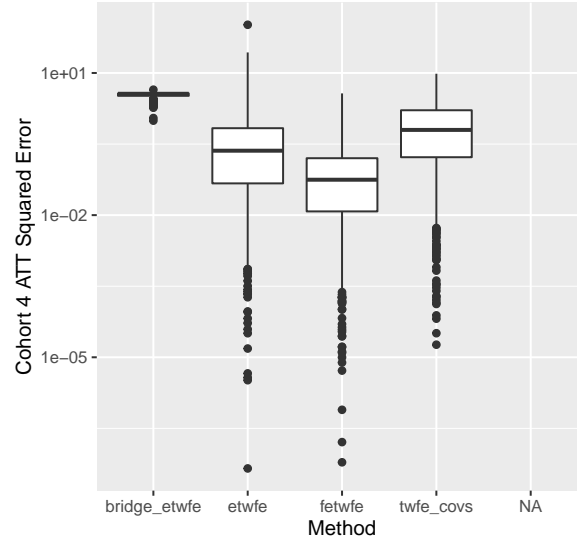
Figure 7: Boxplots of squared errors for each method's estimate of $\tau_{ATT}(4)$ across all 700 simulations. Vertical axis is on a log scale.
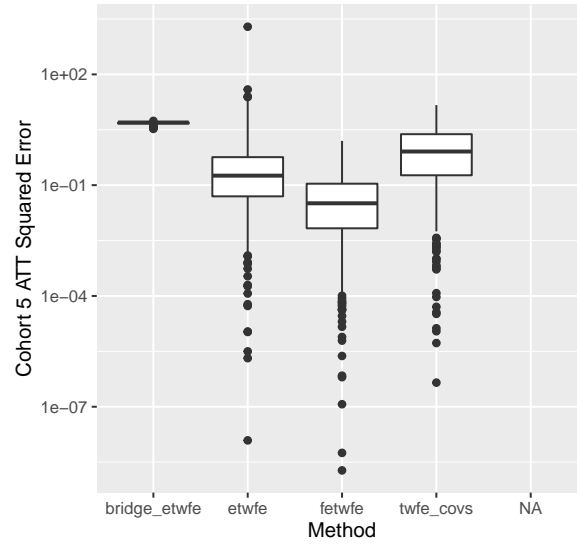
Figure 8: Boxplots of squared errors for each method's estimate of $\tau_{ATT}(5)$ across all 700 simulations. Vertical axis is on a log scale.
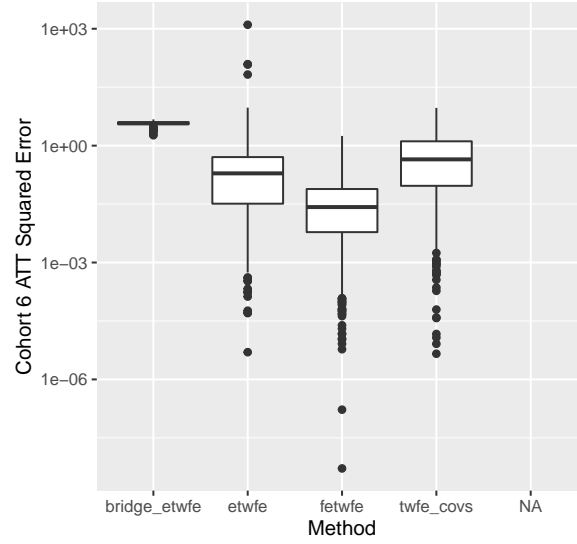
Figure 9: Boxplots of squared errors for each method's estimate of $\tau_{ATT}(6)$ across all 700 simulations. Vertical axis is on a log scale.

hold. Then for all $i \in [N]$, $t \in [T]$, $r \in \mathcal{R}$,

$$\mathbb{E}[\tilde{y}_{(it)} \mid W_i = 0, \boldsymbol{X}_i] = \eta^* + \gamma_t^* + \boldsymbol{X}_i^\top(\boldsymbol{\kappa}^* + \boldsymbol{\xi}_t^*) \qquad and$$

$$\mathbb{E}[\tilde{y}_{(it)} \mid W_i = r, \boldsymbol{X}_i] = \eta^* + \gamma_t^* + \nu_r^* + \boldsymbol{X}_i^\top(\boldsymbol{\kappa}^* + \boldsymbol{\xi}_t^* + \boldsymbol{\zeta}_r^*) + \mathbb{1}\{t \geq r\}(\tau_{rt}^* + \dot{\boldsymbol{X}}_{(ir)}^\top \boldsymbol{\rho}_{rt}^*),$$

where $\dot{\boldsymbol{X}}_{(ir)} = \boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]$, for identifiability we define the time-dependent parameters to equal 0 at $t = 1$, and recall that η^*, γ_t^*, etc. are the regression estimands defined in (LINS).

(b) *(Estimation consistency.) In addition to the assumptions from part (a), suppose Assumptions (F1), (F2), $S(s_N)$, and (R1) - (R3) hold. Then for any $q > 0$, $\|\hat{\boldsymbol{\beta}}^{(q)} - \boldsymbol{\beta}_N^*\|_2 = \mathcal{O}_\mathbb{P}(\min\{h_N, h_N'\})$, where h_N is defined in (6.3) and h_N' is defined in (6.2).*

(c) *(Selection consistency.) Define*

$$\boldsymbol{\theta}_N^* := \boldsymbol{D}_N \boldsymbol{\beta}_N^* \tag{C.1}$$

and $\hat{\boldsymbol{\theta}}^{(q)} := \boldsymbol{D}_N \hat{\boldsymbol{\beta}}^{(q)}$. In addition to the previous assumptions, assume that $q \in (0, 1)$ and Assumptions (R4) - (R5) hold. Then $\lim_{N\to\infty} \mathbb{P}\left(\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = \boldsymbol{0}\right) = 1$. (This also implies that $\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} \xrightarrow{p} \boldsymbol{0}$, $\lim_{N\to\infty} \mathbb{P}\left(\sqrt{NT}\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = \boldsymbol{0}\right) = 1$, and $\lim_{N\to\infty} \mathbb{P}\left(\sqrt{NT}\boldsymbol{v}_N^\top\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = 0\right) = 1$ for any sequence of finite vectors $\boldsymbol{v}_N \in \mathbb{R}^{p_N}$.)

Further, let $\hat{\mathcal{S}}_N := \{j : \hat{\theta}_j^{(q)} \neq 0\}$, and similarly let $\mathcal{S}_N \subseteq [p_N]$ be the set of components of $\boldsymbol{\theta}_N^*$ that are nonzero. Then

$$\lim_{N\to\infty} \mathbb{P}\left(\hat{\mathcal{S}}_N = \mathcal{S}_N\right) = 1.$$

(d) *(Asymptotic normality and oracle property.) In addition to the previous assumptions, assume Assumption (R6) holds and Assumption $S(s)$ holds for a fixed $s_N = s$, so $\mathcal{S}_N = \mathcal{S}$ is fixed as well. Let $\{\boldsymbol{\psi}_N\}_{N=1}^\infty$ be a sequence of real-valued vectors where for each $\boldsymbol{\psi}_N \in \mathbb{R}^{p_N}$, $(\boldsymbol{\psi}_N)_{\mathcal{S}} = \boldsymbol{\alpha} \in \mathbb{R}^s$ is fixed and finite and $(\boldsymbol{\psi}_N)_{[p_N]\setminus\mathcal{S}} = \boldsymbol{b}_N$, where $\{\boldsymbol{b}_N\}_{N=1}^\infty$ is any sequence of constants where each $\boldsymbol{b}_N \in \mathbb{R}^{p_N - s}$ contains all finite entries. Then if $\boldsymbol{\alpha} \neq \boldsymbol{0}$,*

$$\sqrt{NT}\boldsymbol{\psi}_N^\top(\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\alpha}^\top\left(\text{Cov}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)_{(\cdot\cdot)\mathcal{S}}\right)^{-1}\boldsymbol{\alpha}\right),$$

where $\text{Cov}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)_{(\cdot\cdot)\mathcal{S}}$ is the columns and rows of the covariance matrix corresponding to the nonzero components of $\boldsymbol{\theta}_N^*$.

(e) *(Asymptotic normality with feasible variance estimator.) Maintain the assumptions from part (d). For a random matrix $\boldsymbol{A} \in \mathbb{R}^{NT \times p_N}$ with centered columns, define*

$$\widehat{\text{Cov}}(\boldsymbol{A}) := \frac{1}{NT}\boldsymbol{A}^\top\boldsymbol{A} \tag{C.2}$$

to be the sample covariance matrix of $\boldsymbol{A}$. For any $\mathcal{A}_N \subseteq [p_N]$ such that $\mathcal{A}_N \neq \emptyset$, consider $\widehat{\text{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{A}_N}\right)$, the estimated covariance matrix of the columns of $\boldsymbol{Z}\boldsymbol{D}_N^{-1}$ corresponding to $\mathcal{A}_N$. For sets $(\mathcal{A}_N)_{N=1}^\infty$, define $\boldsymbol{\alpha}(\mathcal{A}_N) \in \mathbb{R}^{|\mathcal{A}_N|}$ to be a fixed, finite vector for each unique $\mathcal{A}_N$, and define the sequence of p_N-vectors $\{\boldsymbol{\psi}_N(\mathcal{A}_N)\}_{N=1}^\infty$ to have components in the entries corresponding to $\mathcal{A}_N$ equal to $\boldsymbol{\alpha}(\mathcal{A}_N)$ and remaining entries $\boldsymbol{b}_n(\mathcal{A}_N)$ all finite (but otherwise arbitrary). Define the sequence of random variables $U_1(\mathcal{A}_N)$ to equal

$$\frac{1}{\sigma}\sqrt{\frac{NT}{\boldsymbol{\alpha}(\mathcal{A}_N)^\top\left(\widehat{\text{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{A}_N}\right)\right)^{-1}\boldsymbol{\alpha}(\mathcal{A}_N)}} \cdot \left(\boldsymbol{\psi}_N(\mathcal{A}_N)^\top\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\psi}_N(\mathcal{S})^\top\boldsymbol{\theta}_N^*\right)$$

if $\mathcal{A}_N \neq \emptyset$ and $\widehat{\text{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{A}_N}\right)$ is invertible and 0 otherwise. Then if $\boldsymbol{\alpha}(\mathcal{S}) \neq \boldsymbol{0}$,

$$U_1(\hat{\mathcal{S}}_N) \xrightarrow{d} \mathcal{N}(0, 1).$$

(f) (*Asymptotic normality when $\boldsymbol{\alpha}$ is estimated independently from $\hat{\boldsymbol{\theta}}^{(q)}$.*) *Maintain the assumptions from part (e). For sets $(\mathcal{A}_N)_{N=1}^{\infty}$, let $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N) \in \mathbb{R}^{|\mathcal{A}_N|}$ be a sequence of random variables for each $\mathcal{A}_N$ with the properties that for a given data set, $\mathcal{A} = \mathcal{A}'$ implies $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}) = \hat{\boldsymbol{\alpha}}_N(\mathcal{A}')$ and*

$$\sqrt{NT}(\boldsymbol{\theta}_N^*)^{\top} (\hat{\boldsymbol{\alpha}}_N(\mathcal{S}) - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(\mathbf{0}_s, v_{\psi}(\boldsymbol{\alpha}))$$

for some $v_{\psi}(\boldsymbol{\alpha}) > 0$. Define the sequence of random vectors $\{\hat{\boldsymbol{\psi}}_N(\mathcal{A}_N)\}_{N=1}^{\infty}$ to have components in $\mathcal{A}_N$ equal to $\hat{\boldsymbol{\alpha}}(\mathcal{A}_N)$ and components $\hat{\boldsymbol{b}}_N(\mathcal{A}_N)$ in the components corresponding to $[p_N] \setminus \mathcal{A}_N$ that are all almost surely finite for all N. Define the sequence of random variables

$$U_2(\mathcal{A}_N) := \sqrt{NT} \cdot \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}_N)^{\top} \hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\psi}_N(\mathcal{S})^{\top} \boldsymbol{\theta}_N^* \right).$$

Assume $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ and $\hat{\boldsymbol{\theta}}^{(q)}$ are estimated on independent data sets of size N. Then if $\boldsymbol{\alpha} \neq \mathbf{0}$,

$$U_2(\hat{\mathcal{S}}_N) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\alpha}^{\top} \left(\mathrm{Cov}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{S}}\right)\right)^{-1} \boldsymbol{\alpha} + v_{\psi}(\boldsymbol{\alpha})\right).$$

(g) (*Asymptotic normality when $\boldsymbol{\alpha}$ is estimated independently from $\hat{\boldsymbol{\theta}}^{(q)}$ with feasible variance estimator.*) *Maintain the assumptions from part (f). Define the sequence of random variables $U_3(\mathcal{A}_N)$ to equal*

$$\sqrt{\frac{NT}{\hat{v}_N(\mathcal{A}_N)}} \cdot \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}_N)^{\top} \hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\psi}_N(\mathcal{S})^{\top} \boldsymbol{\theta}_N^* \right)$$

if $\mathcal{A}_N \neq \emptyset$, $\hat{\boldsymbol{\alpha}}(\mathcal{A}_N) \neq \mathbf{0}$, and $\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{A}_N}\right)$ is invertible and 0 otherwise, where

$$\hat{v}_N(\mathcal{A}_N) := \sigma^2 \hat{\boldsymbol{\alpha}}(\mathcal{A}_N)^{\top} \left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{A}_N}\right) \right)^{-1} \hat{\boldsymbol{\alpha}}(\mathcal{A}_N) + \hat{v}_{\psi}(\boldsymbol{\alpha}) \tag{C.3}$$

and $\hat{v}_{\psi}(\boldsymbol{\alpha})$ is an estimator of $v_{\psi}(\boldsymbol{\alpha})$ that satisfies $\hat{v}_{\psi}(\boldsymbol{\alpha}) \xrightarrow{p} v_{\psi}(\boldsymbol{\alpha})$. Assume $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ and $\hat{\boldsymbol{\theta}}^{(q)}$ are estimated on independent data sets of size N. Then if $\boldsymbol{\alpha} \neq \mathbf{0}$,

$$U_3(\hat{\mathcal{S}}_N) \xrightarrow{d} \mathcal{N}(0, 1).$$

(h) (*Asymptotic subgaussianity when $\boldsymbol{\alpha}$ is estimated on the same data set as $\hat{\boldsymbol{\theta}}^{(q)}$ with feasible variance estimator.*) *Maintain the assumptions from part (g). Define the sequence of random variables $U_4(\mathcal{A}_N)$ to equal*

$$\sqrt{\frac{NT}{\hat{v}_N^{(cons)}(\mathcal{A}_N)}} \cdot \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}_N)^{\top} \hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\psi}_N(\mathcal{S})^{\top} \boldsymbol{\theta}_N^* \right)$$

if $\mathcal{A}_N \neq \emptyset$, $\hat{\boldsymbol{\alpha}}(\mathcal{A}_N) \neq \mathbf{0}$, and $\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{A}_N}\right)$ is invertible and 0 otherwise, where $\hat{v}_N^{(cons)}(\mathcal{A}_N)$ is defined as

$$\sigma^2 \hat{\boldsymbol{\alpha}}(\mathcal{A}_N)^{\top} \left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{A}_N}\right) \right)^{-1} \hat{\boldsymbol{\alpha}}(\mathcal{A}_N)$$

$$+ 2\sqrt{\sigma^2 \hat{\boldsymbol{\alpha}}(\mathcal{A}_N)^{\top} \left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{A}_N}\right) \right)^{-1} \hat{\boldsymbol{\alpha}}(\mathcal{A}_N)\hat{v}_{\psi}(\boldsymbol{\alpha})} + \hat{v}_{\psi}(\boldsymbol{\alpha}). \tag{C.4}$$

Assume $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ and $\hat{\boldsymbol{\theta}}^{(q)}$ are estimated on the same data set of size N. Then if $\boldsymbol{\alpha} \neq \mathbf{0}$, $U_4(\hat{\mathcal{S}}_N)$ converges in distribution to a mean-zero subgaussian random variable with variance at most 1.

Proof. Provided in Section F. $\qquad\qquad\square$

D   Extensions of Main Results

Theorem 6.1 is an immediate consequence of the following more general result that allows for estimands that include broader classes of functions of the marginal cohort membership probabilities or propensity scores.

Theorem D.1 (Consistency of FETWFE; extension of Theorem 6.1)**.** *Assume that Assumptions (CNAS), (CCTSB), and (LINS) hold, as well as Assumptions (F1), (F2), $S(s_N)$, and (R1) - (R3). Let $q > 0$.*

(a) Suppose that either Assumptions (CTSA) and (CIUN) hold or Assumption (CCTSA) holds. Then for any set of finite constants $\{\psi_{rt}\}$,

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\hat{\tau}_{ATT}(r,t) - \tau_{ATT}(r,t) \right) \right| = \mathcal{O}_{\mathbb{P}} \left(\min\{h_N, h'_N\} \right)$$

and for any differentiable function $\boldsymbol{f} : [0,1]^{R+1} \to \mathbb{R}^R$ that is pointwise finite and has a Jacobian $\nabla \boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$ at $(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ that exists and is nonzero and finite,

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) \hat{\tau}_{ATT}(r,t) - f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \tau_{ATT}(r,t) \right) \right| = \mathcal{O}_{\mathbb{P}} \left(\min\{h_N, h'_N\} \right),$$

where $\hat{\pi}_r := N_r/N$ for all $r \in \mathcal{R}$, h'_N is defined in (6.2), and h_N is defined in (6.3).

(b) Suppose Assumption (CCTSA) holds and $d_N = d$ is fixed. For all N and all $r \in \{0\} \cup \mathcal{R}$, assume all of the eigenvalues of $\mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r)$ are bounded between $\lambda_{min} > 0$ and $\lambda_{max} < \infty$. Then for any set of finite constants $\{\psi_{rt}\}$ and any fixed $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$,

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\hat{\tau}_{CATT}(r,t,\boldsymbol{x}) - \tau_{CATT}(r,t,\boldsymbol{x}) \right) \right| = \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}).$$

Further, for any $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$, let $\hat{\boldsymbol{\pi}}(\boldsymbol{x})$ be an estimator for $\pi_r(\boldsymbol{x})$ satisfying $|\hat{\pi}_r(\boldsymbol{x}) - \pi_r(\boldsymbol{x})| = \mathcal{O}_{\mathbb{P}}(a_N)$ for a decreasing sequence a_N for all $r \in \{0\} \cup \mathcal{R}$, and let $\boldsymbol{a}_r$ and $\boldsymbol{b}_r \in \mathbb{R}^{R+1}$ be finite vectors satisfying $\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x}) > 0$ and $\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x}) > 0$ for any $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$. Then

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} \hat{\tau}_{CATT}(r,t,\boldsymbol{x}) - \frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \tau_{CATT}(r,t,\boldsymbol{x}) \right) \right|$$
$$= \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\} \vee a_N).$$

Proof. Provided in Section E. $\square$

Theorem D.2 is a similar generalization of Theorem 6.4.

Theorem D.2 (Asymptotic Confidence Intervals for FETWFE; extension of Theorem 6.4)**.** *Assume that Assumptions (CNAS), (CCTSB), and (LINS) hold, as well as Assumptions (F1), (F2), $S(s)$ for a fixed s, and (R1) - (R6). Assume $q \in (0,1)$. Suppose that either Assumptions (CTSA) and (CIUN) hold or Assumption (CCTSA) holds. Let $\{\psi_{rt}\}$ be an arbitrary set of finite constants, and for all of the below results, assume at least one of the $\tau_{ATT}(r,t)$ corresponding to a nonzero ψ_{rt} is nonzero (otherwise, all of the below sequences of random variables converge in probability to 0).*

(a) *For the estimator (5.9) it holds that*

$$\sqrt{\frac{NT}{\hat{v}_N^{(C)}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt}(\hat{\tau}_{ATT}(r,t) - \tau_{ATT}(r,t)) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{v}_N^{(C)}$ is a finite-sample variance estimator defined in (E.7). (All variance estimators depend only on the known σ^2 and the observed data.)

(b) *Suppose one set of data of size N is used to estimate the cohort probabilities N_r/N_τ and another independent data set of size N is used to estimate each $\hat{\tau}_{ATT}(r,t)$. Then for any differentiable function $\boldsymbol{f} : [0,1]^{R+1} \to \mathbb{R}^R$ that is pointwise finite and has a Jacobian $\nabla \boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$ at $(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ that exists and is nonzero and finite, it holds that*

$$\sqrt{\frac{NT}{\hat{v}_N^{C;\hat{\pi}}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(f_r(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R}) \hat{\tau}_{ATT}(r,t) - f_r(\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_R}) \tau_{ATT}(r,t) \right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{v}_N^{C;\hat{\pi}}$ is a finite-sample variance estimator defined in (E.11).

(c) *Suppose a single set of data of size N is used to estimate both the cohort probabilities N_r/N_τ and each $\hat{\tau}_{ATT}(r,t)$. Then for any differentiable function $\boldsymbol{f} : [0,1]^{R+1} \to \mathbb{R}^R$ that is pointwise finite and has a Jacobian $\nabla \boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$ at $(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ that exists and is nonzero and finite, the estimator (5.10) is asymptotically subgaussian: the sequence of random variables*

$$\sqrt{\frac{NT}{\hat{v}_N^{C,\ (cons);\hat{\pi}}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(f_r(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R}) \hat{\tau}_{ATT}(r,t) - f_r(\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_R}) \tau_{ATT}(r,t) \right)$$

converges in distribution to a mean-zero subgaussian random variable with variance at most 1, where $\hat{v}_N^{C,\ (cons);\hat{\pi}}$ is a finite-sample conservative variance estimator defined in (E.12).

Proof. Provided in Section E. $\qquad\square$

Theorem D.3 shows the asymptotic normality of two classes of conditional treatment effect estimators, analogously to Theorem D.2, and provides consistent variance estimators for constructing asymptotically valid confidence intervals. It also provides sufficient assumptions for the FETWFE estimators of conditional average treatment effects to converge at $1/\sqrt{N}$ rates.

Theorem D.3 (Asymptotic Confidence Intervals for conditional average treatment effects in FETWFE). *Assume that Assumptions (CNAS), (CCTS), and (LINS) hold, as well as Assumptions (F1), (F2), S(s) for a fixed s, and (R1) - (R6). Assume $q \in (0,1)$, and assume $d_N = d$ is fixed. Let $\{\psi_{rt}\}$ be an arbitrary set of finite constants, and for both of the below results, assume at least one of the $\tau_{CATT}(r,t,\boldsymbol{x})$ corresponding to a nonzero ψ_{rt} is nonzero (otherwise, all of the below sequences of random variables converge in probability to 0).*

(a) *For $\hat{\tau}_{CATT}(r,t,\boldsymbol{x})$ as defined in (5.5), suppose one set of data of size N is used to estimate each cohort conditional covariate mean $\overline{\boldsymbol{X}}_r$ and another independent data set of size N is used to estimate each $\hat{\tau}_{rt}^{(q)}$ and $\hat{\rho}_{rt}^{(q)}$ via the FETWFE regression. Then for any set of finite constants $\{\psi_{rt}\}_{r \in \mathcal{R}, t \in \{r,\ldots,T\}}$ and any fixed $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$, it holds that*

$$\sqrt{\frac{NT}{\hat{v}_N^{(C;CATT)}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt}(\hat{\tau}_{CATT}(r,t,\boldsymbol{x}) - \tau_{CATT}(r,t,\boldsymbol{x})) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{v}_N^{(C;CATT)}$ is a finite-sample variance estimator defined in (E.13).

51

(b) Assume we have an estimator $(\hat{\pi}_0(\boldsymbol{x}), \hat{\pi}_{r_1}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}))$ of the generalized propensity scores $(\pi_0(\boldsymbol{x}), \pi_{r_1}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}))$ that satisfies for any fixed $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$

$$\sqrt{N} \left(\hat{\pi}_0(\boldsymbol{x}) - \pi_0(\boldsymbol{x}), \hat{\pi}_{r_1}(\boldsymbol{x}) - \pi_{r_1}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) - \pi_{r_R}(\boldsymbol{x})\right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}_\pi)$$

for some finite positive definite $\boldsymbol{\Sigma}_\pi \in \mathbb{R}^{(R+1) \times (R+1)}$. Assume we also have available a consistent estimator $\hat{\boldsymbol{\Sigma}}_\pi$ that converges entrywise in probability to $\boldsymbol{\Sigma}_\pi$. Suppose one set of data of size N is used to estimate each cohort conditional covariate mean $\overline{\boldsymbol{X}}_r$, another independent data set of size N is used to estimate each $\hat{\tau}_{rt}^{(q)}$ and $\hat{\rho}_{rt}^{(q)}$ via the FETWFE regression, and a third independent data set is used to estimate the generalized propensity scores. Then for any differentiable function $f : [0,1]^{R+1} \to \mathbb{R}^R$ that is pointwise finite and has a Jacobian $\nabla \boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$ at $(\pi_0(\boldsymbol{x}), \pi_{r_1}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}))$ that exists and is nonzero and finite, it holds that

$$\sqrt{\frac{NT}{\hat{v}_N^{(C;CATT,\pi)}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \big(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right) \hat{\tau}_{CATT}(r, t, \boldsymbol{x})$$

$$- f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \tau_{CATT}(r, t, \boldsymbol{x}) \big) \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\hat{v}_N^{(C;CATT,\pi)}$ is a finite-sample variance estimator defined in (E.16).

Proof. Provided in Section E. □

It is trivial to extend Theorem D.3 to show that these classes of conditional average treatment effect estimators enjoy oracle properties analogous to those of Theorem 6.3. Theorem D.3 applies immediately to the classes of conditional average treatment effects estimators (5.6) and (5.7) by the same argument we used to prove Theorem 6.1 since the Jacobian $\nabla \boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$

$$\nabla \boldsymbol{f} := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \frac{\sum_{r' \in \mathcal{R} \setminus r_1} \pi_{r'}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} & \frac{-\pi_{r_2}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} & \cdots & \frac{-\pi_{r_R}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} \\ \frac{-\pi_{r_1}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} & \frac{\sum_{r' \in \mathcal{R} \setminus r_2} \pi_{r'}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} & \cdots & \frac{-\pi_{r_R}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-\pi_{r_1}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} & \frac{-\pi_{r_2}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} & \cdots & \frac{\sum_{r' \in \mathcal{R} \setminus r_R} \pi_{r'}(\boldsymbol{x})}{\left(\sum_{r' \in \mathcal{R}} \pi_{r'}(\boldsymbol{x})\right)^2} \end{pmatrix}$$

at $(\pi_0(\boldsymbol{x}), \pi_{r_1}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}))$ and exists and is nonzero and finite by Assumption (F2).

E   Proofs of Main Results

We present two lemmas we will use in the following proofs.

Lemma E.1. *The number of treatment effects $\mathfrak{W}$ is at most $(T-1)^2$. Further, after re-ordering the columns of $\tilde{\boldsymbol{Z}}$, we can write $\boldsymbol{D}_N$ as*

$$\text{diag} \left(\boldsymbol{D}^{(1)}(R), \boldsymbol{D}^{(1)}(T-1), \boldsymbol{I}_{d_N}, \boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(1)}(R), \boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(1)}(T-1), \boldsymbol{D}^{(2)}(\mathcal{R}), \boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(2)}(\mathcal{R})\right), \text{ (E.1)}$$

where $\boldsymbol{D}^{(1)}(t) \in \mathbb{R}^{t \times t}$ is defined in (G.4) and $\boldsymbol{D}^{(2)}(\mathcal{R}) \in \mathbb{R}^{\mathfrak{W} \times \mathfrak{W}}$ is defined in (G.5). Further, $\boldsymbol{D}_N$ is invertible, and $\boldsymbol{D}_N^{-1}$ can be written as

$$\text{diag} \Big(\boldsymbol{D}^{(1)}(R)^{-1}, ((\boldsymbol{D}^{(1)}(T-1))^{-1}, \boldsymbol{I}_{d_N}, \boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(1)}(R)^{-1}, \boldsymbol{I}_{d_N} \otimes ((\boldsymbol{D}^{(1)}(T-1))^{-1},$$

$$(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}, \boldsymbol{I}_{d_N} \otimes (\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}\Big). \text{ (E.2)}$$

Finally, both $\boldsymbol{D}_N$ and $\boldsymbol{D}_N^{-1}$ consists of blocks of size at most $(T-1)^2 \times (T-1)^2$. For $\boldsymbol{D}_N$ the entries of these blocks are either -1, 0, or 1; and for $\boldsymbol{D}_N^{-1}$ the entries of these blocks are either 0 or 1. Both matrices equal 0 elsewhere.

Lemma E.2. *Under Assumption (R3),* $\max_{j \in p_N}\{|(\boldsymbol{\beta}_N^*)_j|\} \le (T-1)^2 b_1$.

Proofs of both results are provided in Appendix G.2.

E.1 Proofs of Theorems D.1 and 6.1

We start by proving Theorem D.1, which leads almost immediately to a proof of Theorem 6.1.

Proof of Theorem D.1. (a) Observe that under our assumptions our model is correctly specified due to Theorem C.1(a). Recall that due to Theorem A.5, under our assumptions the regression estimands correspond to our causal estimands.

Examining our estimator $\hat{\tau}_{\mathrm{ATT}}(r,t)$ for the average treatment effect for a unit in cohort r at time t from (5.8), we see that we can then apply Theorem C.1(b) to immediately obtain that for every $r \in \mathcal{R}$ and $t \ge r$,

$$|\hat{\tau}_{\mathrm{ATT}}(r,t) - \tau_{\mathrm{ATT}}(r,t)| = \left|\hat{\tau}_{rt}^{(q)} - \tau_{rt}^*\right| \le \|\hat{\boldsymbol{\beta}}^{(q)} - \boldsymbol{\beta}_N^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\min\{h_N, h_N'\}\right). \tag{E.3}$$

Similarly, for any deterministic fixed linear combination of average treatment effects, using (E.3) and Theorem A.5 we have that for our class of estimators (5.9) and estimands (4.8),

$$\left|\sum_{r \in \mathcal{R}}\sum_{t=r}^{T}\psi_{rt}\hat{\tau}_{\mathrm{ATT}}(r,t) - \sum_{r \in \mathcal{R}}\sum_{t=r}^{T}\psi_{rt}\tau_{\mathrm{ATT}}(r,t)\right| \le \sum_{r \in \mathcal{R}}\sum_{t=r}^{T}|\psi_{rt}|\,|\hat{\tau}_{\mathrm{ATT}}(r,t) - \tau_{\mathrm{ATT}}(r,t)|$$
$$= \mathcal{O}_{\mathbb{P}}\left(\min\{h_N, h_N'\}\right),$$

where we used the fact that T is fixed (and upper bounds R).

Next we characterize the rate of convergence of each $f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \dots, \hat{\pi}_{r_R})$. Because by assumption $\boldsymbol{f}(\cdot)$ has a finite Jacobian $\nabla\boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$ at $(\pi_0, \pi_{r_1}, \dots, \pi_{r_R})$ that exists and is nonzero and finite, we have by the asymptotic normality of the maximum likelihood estimator $\left(\frac{N_0}{N}, \frac{N_{r_1}}{N}, \frac{N_{r_2}}{N}, \dots, \frac{N_{r_R}}{N}\right)$ (Theorem 14.6 in Wasserman 2004) and the delta method (Theorem 3.1 in van der Vaart 2000) that

$$\sqrt{N}\left(f(\hat{\pi}_0, \hat{\pi}_{r_1}, \dots, \hat{\pi}_{r_R}) - f(\pi_0, \pi_{r_1}, \dots, \pi_{r_R})\right) \xrightarrow{d} \mathcal{N}(0, \nabla\boldsymbol{f}^\top\boldsymbol{\Sigma}_M\nabla\boldsymbol{f})$$
$$\implies \sqrt{N}\left(f_{r'}(\hat{\pi}_0, \hat{\pi}_{r_1}, \dots, \hat{\pi}_{r_R}) - f_{r'}(\pi_0, \pi_{r_1}, \dots, \pi_{r_R})\right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{e}_{r'}^\top\nabla\boldsymbol{f}^\top\boldsymbol{\Sigma}_M\nabla\boldsymbol{f}\boldsymbol{e}_{r'}),$$

for any $r' \in [R]$, where

$$\boldsymbol{\Sigma}_M := \begin{pmatrix} \pi_0(1-\pi_0) & -\pi_0\pi_{r_1} & \cdots & -\pi_0\pi_{r_R} \\ -\pi_0\pi_{r_1} & \pi_{r_1}(1-\pi_{r_1}) & \cdots & -\pi_{r_1}\pi_{r_R} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_0\pi_{r_R} & -\pi_{r_1}\pi_{r_R} & \cdots & \pi_{r_R}(1-\pi_{r_R}) \end{pmatrix} \in \mathbb{R}^{(R+1)\times(R+1)} \tag{E.4}$$

and $\boldsymbol{e}_r \in \mathbb{R}^R$ is the selector vector with a 1 in the rth entry and zeroes elsewhere. Therefore by Theorem 2.4(i) in van der Vaart (2000) and Proposition 1.8(iii) in García-Portugués (2023), we have that for all $r \in [R]$

$$\sqrt{N}\,|f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \dots, \hat{\pi}_{r_R}) - f_r(\pi_0, \pi_{r_1}, \dots, \pi_{r_R})| = \mathcal{O}_{\mathbb{P}}(1)$$
$$\implies |f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \dots, \hat{\pi}_{r_R}) - f_r(\pi_0, \pi_{r_1}, \dots, \pi_{r_R})| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N}}\right). \tag{E.5}$$

Finally, for our class of estimators (5.10) and estimands (4.9), we have

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) \hat{\tau}_{\mathrm{ATT}}(r,t) - \psi_{rt} f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \tau_{\mathrm{ATT}}(r,t) \right|$$

$$= \left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \big[f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \left(\hat{\tau}_{\mathrm{ATT}}(r,t) - \tau_{\mathrm{ATT}}(r,t) \right) \right.$$

$$+ \left(f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) - f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \right) \tau_{\mathrm{ATT}}(r,t)$$

$$\left. + \left(f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) - f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \right) \left(\hat{\tau}_{\mathrm{ATT}}(r,t) - \tau_{\mathrm{ATT}}(r,t) \right) \big] \right|$$

$$\leq \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \big(f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \left| \hat{\tau}_{\mathrm{ATT}}(r,t) - \tau_{\mathrm{ATT}}(r,t) \right| +$$

$$\left| f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) - f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \right| \tau_{\mathrm{ATT}}(r,t)$$

$$+ \left| f_r(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) - f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R}) \right| \left| \hat{\tau}_{\mathrm{ATT}}(r,t) - \tau_{\mathrm{ATT}}(r,t) \right| \big)$$

$$\overset{(a)}{=} \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}) + \mathcal{O}_{\mathbb{P}}(1/\sqrt{N}) + \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}) \mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$$

$$\overset{(b)}{=} \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}),$$

where in (a) we used that all the $f_r(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ and $\tau_{\mathrm{ATT}}(r,t)$ are finite, (E.3), and (E.5); and in (b) we used that $1/\sqrt{N} = \mathcal{O}(h_N)$ since e_{1N} is upper-bounded by a constant and p_N is nondecreasing.

(b) Again, using Theorem A.5,

$$\tau_{\mathrm{CATT}}(r,t,\boldsymbol{x}) = \tau_{rt}^* + (\boldsymbol{\rho}_{rt}^*)^\top \left(\boldsymbol{x} - \mathbb{E}\left[\boldsymbol{X}_{(1t)\cdot} \mid W_i = r \right] \right) = \tau_{rt}^* + (\boldsymbol{\rho}_{rt}^*)^\top \left(\boldsymbol{x} - \boldsymbol{\mu}_r \right),$$

where $\boldsymbol{\mu}_r := \mathbb{E}\left[\boldsymbol{X}_{(it)\cdot} \mid W_i = r \right]$. Using this and (5.5) we have

$$|\hat{\tau}_{\mathrm{CATT}}(r,t,\boldsymbol{x}) - \tau_{\mathrm{CATT}}(r,t,\boldsymbol{x})|$$

$$= \left| \hat{\tau}_{rt}^{(q)} - \tau_{rt}^* + (\boldsymbol{x} - \overline{\boldsymbol{X}}_r)^\top \hat{\boldsymbol{\rho}}_{rt} - (\boldsymbol{x} - \boldsymbol{\mu}_r)^\top \boldsymbol{\rho}_{rt}^* \right|$$

$$= \left| \hat{\tau}_{rt}^{(q)} - \tau_{rt}^* + (\boldsymbol{x} - \boldsymbol{\mu}_r)^\top (\hat{\boldsymbol{\rho}}_{rt} - \boldsymbol{\rho}_{rt}^*) + (\boldsymbol{\mu}_r - \overline{\boldsymbol{X}}_r)^\top (\hat{\boldsymbol{\rho}}_{rt} - \boldsymbol{\rho}_{rt}^*) - (\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r)^\top \boldsymbol{\rho}_{rt}^* \right|$$

$$\leq \left| \hat{\tau}_{rt}^{(q)} - \tau_{rt}^* \right| + \left| (\boldsymbol{x} - \boldsymbol{\mu}_r)^\top (\hat{\boldsymbol{\rho}}_{rt} - \boldsymbol{\rho}_{rt}^*) \right| + \left| (\boldsymbol{\mu}_r - \overline{\boldsymbol{X}}_r)^\top (\hat{\boldsymbol{\rho}}_{rt} - \boldsymbol{\rho}_{rt}^*) \right| + \left| (\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r)^\top \boldsymbol{\rho}_{rt}^* \right|$$

$$\overset{(c)}{\leq} \left| \hat{\tau}_{rt}^{(q)} - \tau_{rt}^* \right| + \|\boldsymbol{x} - \boldsymbol{\mu}_r\|_2 \|\hat{\boldsymbol{\rho}}_{rt} - \boldsymbol{\rho}_{rt}^*\|_2 + \|\boldsymbol{\mu}_r - \overline{\boldsymbol{X}}_r\|_2 \|\hat{\boldsymbol{\rho}}_{rt} - \boldsymbol{\rho}_{rt}^*\|_2 + \|\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r\|_2 \|\boldsymbol{\rho}_{rt}^*\|_2$$

$$\overset{(d)}{=} \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}) + \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}) + \mathcal{O}_{\mathbb{P}}\left(1/\sqrt{N}\right) \mathcal{O}_{\mathbb{P}}\left(\min\{h_N, h'_N\}\right) + \mathcal{O}_{\mathbb{P}}\left(1/\sqrt{N}\right)$$

$$= \mathcal{O}_{\mathbb{P}}(\min\{h_N, h'_N\}),$$

where in (c) we applied Cauchy-Schwarz and (d) follows from Theorem C.1, the fact that $\|\boldsymbol{\rho}_{rt}^*\|_2$ is fixed and bounded since d is fixed and each coefficient is finite due to Lemma E.2, and Lemma E.3:

Lemma E.3. *For each $r \in \{0\} \cup \mathcal{R}$, let $\boldsymbol{\psi}_r \in \mathbb{R}^d$ be constant vectors, with $\boldsymbol{\psi}_r \neq \mathbf{0}$ for at least one $r = r^*$. For all N and all $r \in \{0\} \cup \mathcal{R}$, assume all of the eigenvalues of $\mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r)$ are bounded between $\lambda_{min} > 0$ and $\lambda_{max} < \infty$. Then*

$$\sqrt{N} \sum_{r \in \{0\} \cup \mathcal{R}} \boldsymbol{\psi}_r^\top \left(\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r \right) \overset{d}{\to} \mathcal{N}(0, v_R),$$

where

$$v_R := \sum_{r \in \{0\} \cup \mathcal{R}} \frac{\boldsymbol{\psi}_r^\top \mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r)\boldsymbol{\psi}_r}{\mathbb{P}(W_i = r)}.$$

Further, for any $r \in \{0\} \cup \mathcal{R}$,

$$\|\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r\|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N}}\right).$$

Proof. Provided in Appendix G.2. $\qquad\square$

Next, for any deterministic fixed set of constants $\{\psi_{rt}\}$, the above result gives us

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \hat{\tau}_{\mathrm{CATT}}(r, t, \boldsymbol{x}) - \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x}) \right|$$

$$\leq \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} |\psi_{rt}| \, |\hat{\tau}_{\mathrm{CATT}}(r, t, \boldsymbol{x}) - \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})|$$

$$= \mathcal{O}_{\mathbb{P}}(\min\{h_N, h_N'\}). \tag{E.6}$$

Finally, we have

$$\left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} \right) \hat{\tau}_{\mathrm{CATT}}(r, t, \boldsymbol{x}) - \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x}) \right|$$

$$= \left| \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left[\left(\frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \right) (\hat{\tau}_{\mathrm{CATT}}(r, \boldsymbol{x}, t) - \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})) \right. \right.$$

$$+ \left[\frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} - \frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \right] \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})$$

$$\left. \left. + \left[\frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} - \frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \right] (\hat{\tau}_{\mathrm{CATT}}(r, \boldsymbol{x}, t) - \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})) \right] \right|$$

$$\leq \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left[\frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} |\hat{\tau}_{\mathrm{CATT}}(r, \boldsymbol{x}, t) - \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})| \right.$$

$$+ \left| \frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} - \frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \right| \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})$$

$$\left. + \left| \frac{\boldsymbol{a}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} - \frac{\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})} \right| |\hat{\tau}_{\mathrm{CATT}}(r, \boldsymbol{x}, t) - \tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})| \right]$$

$$\overset{(e)}{\leq} \mathcal{O}_{\mathbb{P}}(\min\{h_N, h_N'\}) + \mathcal{O}_{\mathbb{P}}(a_N) + \mathcal{O}_{\mathbb{P}}(a_N)\mathcal{O}_{\mathbb{P}}(\min\{h_N, h_N'\})$$

$$\overset{(f)}{=} \mathcal{O}_{\mathbb{P}}(\min\{h_N, h_N'\} \vee a_N),$$

where in (f) we used Proposition 1.8 from García-Portugués (2023) and in (e) we used (E.6), the fact that $|\psi_{rt}|$ is bounded for any fixed $\boldsymbol{x}$ for all $r \in \mathcal{R}$ due to Lemma E.2, the fact that $|\tau_{\mathrm{CATT}}(r, t, \boldsymbol{x})|$ and $\boldsymbol{a}_r^\top \boldsymbol{\pi}(\boldsymbol{x})/\boldsymbol{b}_r^\top \boldsymbol{\pi}(\boldsymbol{x})$ are bounded for any fixed $\boldsymbol{x}$, and Lemma E.4:

Lemma E.4. *For any* $\boldsymbol{x}$ *in the support of* $\boldsymbol{X}_i$, *let* $\hat{\boldsymbol{\pi}}(\boldsymbol{x})$ *be an estimator for* $\pi_r(\boldsymbol{x})$ *satisfying* $|\hat{\pi}_r(\boldsymbol{x}) - \pi_r(\boldsymbol{x})| = \mathcal{O}_{\mathbb{P}}(a_N)$ *for a decreasing sequence* a_N *for all* $r \in \{0\} \cup \mathcal{R}$. *Let* $\boldsymbol{a}$ *and* $\boldsymbol{b} \in \mathbb{R}^{R+1}$ *be finite vectors satisfying* $\boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x}) > 0$ *and* $\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) > 0$ *for any* $\boldsymbol{x}$ *in the support of* $\boldsymbol{X}_i$. *Then*

$$\left| \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} - \frac{\boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x})} \right| = \mathcal{O}_{\mathbb{P}}(a_N).$$

Proof. Provided in Appendix G.2. □

The proof is complete.

□

Proof of Theorem 6.1. In light of Theorem D.1, to prove part (a) it is enough to note that

$$
f_r(\hat\pi_0, \hat\pi_{r_1}, \ldots, \hat{\pi_{r_R}}) = \begin{pmatrix} \frac{\hat\pi_{r_1}}{\sum_{r\in\mathcal{R}} \hat\pi_r} \\ \frac{\hat\pi_{r_2}}{\sum_{r\in\mathcal{R}} \hat\pi_r} \\ \vdots \\ \frac{\hat\pi_{r_R}}{\sum_{r\in\mathcal{R}} \hat\pi_r} \end{pmatrix}
$$

(where $\hat\pi_r = N_r/N$) is pointwise (that is, almost surely) finite and show that its Jacobian at $(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ exists and is nonzero and finite. The Jacobian at $(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ is

$$
\begin{pmatrix} 0 & 0 & \cdots & 0 \\ \frac{\sum_{r\in\mathcal{R}\setminus r_1} \hat\pi_r}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} & -\frac{\hat\pi_{r_2}}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} & \cdots & -\frac{\hat\pi_{r_R}}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} \\ -\frac{\hat\pi_{r_1}}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} & \frac{\sum_{r\in\mathcal{R}\setminus r_2} \hat\pi_r}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} & \cdots & -\frac{\hat\pi_{r_R}}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\hat\pi_{r_1}}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} & -\frac{\hat\pi_{r_2}}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} & \cdots & \frac{\sum_{r\in\mathcal{R}\setminus r_R} \hat\pi_r}{\left(\sum_{r\in\mathcal{R}} \hat\pi_r\right)^2} \end{pmatrix} \in \mathbb{R}^{(R+1)\times R};
$$

notice it is nonzero as required.

Part (b) is immediate from Theorem D.1(b) using $a_r = e_r$, the selector vector with a 1 in the rth position and zeroes elsewhere, and $b_r = (1, 1, \ldots, 1)^\top$. □

E.2 Proofs of Theorems D.2, 6.3, and 6.4

Next we prove Theorem D.2. Then we will use this result to prove Theorem 6.4, and finally Theorem 6.3.

Proof of Theorem D.2. By Theorem A.5, the average treatment effect $\tau_{\text{ATT}}(r, t)$ is equal to the regression estimand τ_{rt}^*. We will prove Theorem D.2 by applying Theorem C.1(e), (g), and (h) with particular choices of vectors $\boldsymbol\psi_N$ and $\hat{\boldsymbol\psi}_N$.

(a) Let $i(r, t)$ be the index of the row of $\boldsymbol{D}_N$ corresponding to $\hat\tau_{rt}^{(q)}$ (that is, the column of $\tilde{\boldsymbol{Z}}$ corresponding to this regression coefficient), so that

$$
\hat\tau_{rt}^{(q)} = (\boldsymbol{D}_N^{-1})_{i(r,t)\cdot}\hat{\boldsymbol\theta}^{(q)}
$$

and

$$
\tau_{rt}^* = (\boldsymbol{D}_N^{-1})_{i(r,t)\cdot}\boldsymbol\theta_N^*.
$$

We can see from Lemma E.1 that each row of $\boldsymbol{D}_N^{-1}$ contains at most $(T-1)^2$ nonzero entries, all of the nonzero entries equal 1, and the entries corresponding to the s relevant features will stay fixed even if $p_N \to \infty$. So the vector $\boldsymbol\psi_N = \left((\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}\right)^\top$ is a valid choice for Theorem C.1(d) - (h). In fact, any fixed, finite linear combination of rows $(\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}$ is a valid choice.

We therefore have from Theorem C.1(e) that if at least one of the nonzero components of $(\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}$ corresponds to one of the s features in $\mathcal{S}$,

$$\sqrt{\frac{NT}{\hat{v}_N^{\text{ATT};r,t}}}\left((\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}\hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}\boldsymbol{\theta}_N^*\right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\Longleftrightarrow \qquad \sqrt{\frac{NT}{\hat{v}_N^{\text{ATT};r,t}}}\left(\hat{\tau}_{\text{ATT}}(r,t) - \tau_{\text{ATT}}(r,t)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\hat{v}_N^{\text{ATT};r,t} := \sigma^2(\boldsymbol{D}_N^{-1})_{i(r,t),\hat{\mathcal{S}}}\left(\widehat{\text{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\hat{\mathcal{S}}}\right)\right)^{-1}((\boldsymbol{D}_N^{-1})_{i(r,t),\hat{\mathcal{S}}})^\top,$$

$\widehat{\text{Cov}}(\cdot)$ is defined in (C.2) and $(\boldsymbol{D}_N^{-1})_{i(r,t),\hat{\mathcal{S}}} \in \mathbb{R}^{|\hat{\mathcal{S}}|}$ is a row vector that contains the $|\hat{\mathcal{S}}|$ components of the $i(r,t)$ row of $\boldsymbol{D}_N^{-1}$ corresponding to the selected features. Similarly, for any deterministic fixed set of constants $\{\psi_{rt}\}$, since

$$\sum_{r\in\mathcal{R}}\sum_{t=r}^T \psi_{rt}\tau_{rt}^* = \sum_{r\in\mathcal{R}}\sum_{t=r}^T \psi_{rt}(\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}\boldsymbol{\theta}_N^*,$$

we can choose $\boldsymbol{\psi}_N^{(\text{C})} := \sum_{r\in\mathcal{R}}\sum_{t=r}^T \psi_{rt}(\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}^\top$ to establish that

$$\sqrt{\frac{NT}{\hat{v}_N^{(\text{C})}}}\left((\boldsymbol{\psi}_N^{(\text{C})})^\top(\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*)\right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\Longleftrightarrow \qquad \sqrt{\frac{NT}{\hat{v}_N^{(\text{C})}}}\sum_{r\in\mathcal{R}}\sum_{t=r}^T \psi_{rt}(\hat{\tau}_{\text{ATT}}(r,t) - \tau_{\text{ATT}}(r,t)) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\hat{v}_N^{(\text{C})} := \sigma^2\left((\boldsymbol{\psi}_N^{(\text{C})})_{\hat{\mathcal{S}}}\right)^\top\left(\widehat{\text{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\hat{\mathcal{S}}}\right)\right)^{-1}(\boldsymbol{\psi}_N^{(\text{C})})_{\hat{\mathcal{S}}}. \tag{E.7}$$

(b) We will prove the convergence of the results with estimated cohort probabilities by applying Theorem C.1(g). Consider the random vector

$$\hat{\boldsymbol{\psi}}_N^{(\text{C};\,\hat{\pi})} := \sum_{r\in\mathcal{R}} f_r(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R})\sum_{t=r}^T \psi_{rt}(\boldsymbol{D}_N^{-1})_{i(r,t),\cdot} \in \mathbb{R}^{p_N} \tag{E.8}$$

and define

$$\boldsymbol{\psi}_N^{(\text{C};\,\pi)} := \sum_{r\in\mathcal{R}} f_r(\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_R})\sum_{t=r}^T \psi_{rt}(\boldsymbol{D}_N^{-1})_{i(r,t),\cdot} \in \mathbb{R}^{p_N},$$

so that

$$(\hat{\boldsymbol{\psi}}_N^{(\text{C};\,\hat{\pi})})^\top\hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{\psi}_N^{(\text{C};\,\pi)})^\top\boldsymbol{\theta}_N^*$$

$$= \sum_{r\in\mathcal{R}}\sum_{t=r}^T \psi_{rt}\left(f_r(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R})\hat{\tau}_{\text{ATT}}(r,t) - f_r(\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_R})\tau_{\text{ATT}}(r,t)\right).$$

We have assumed that the $(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R})$ values (and therefore $\hat{\boldsymbol{\psi}}_N^{(\text{C};\,\hat{\pi})}$) are estimated on an independent data set from the one used to estimate $\hat{\boldsymbol{\theta}}^{(q)}$. It remains to show that

$$\sqrt{NT}(\boldsymbol{\theta}^*)_{\mathcal{S}}^\top\left((\hat{\boldsymbol{\psi}}_N^{(\text{C};\,\hat{\pi})})_{\mathcal{S}} - (\boldsymbol{\psi}_N^{(\text{C};\,\pi)})_{\mathcal{S}}\right)$$

57

is asymptotically normal and to identify a consistent estimator for the asymptotic variance. First we establish asymptotic normality. Because by assumption $\boldsymbol{f}(\cdot)$ has a finite Jacobian $\nabla \boldsymbol{f} \in \mathbb{R}^{(R+1) \times R}$ at $(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})$ that exists and is nonzero and finite, using the asymptotic normality of the maximum likelihood estimator $\left(\frac{N_0}{N}, \frac{N_{r_1}}{N}, \frac{N_{r_2}}{N}, \ldots, \frac{N_{r_R}}{N}\right)$ (Theorem 14.6 in Wasserman 2004) and the delta method (Theorem 3.1 in van der Vaart 2000) that

$$\sqrt{N}\left(\boldsymbol{f}(\hat{\pi}_0, \hat{\pi}_{r_1}, \ldots, \hat{\pi}_{r_R}) - \boldsymbol{f}(\pi_0, \pi_{r_1}, \ldots, \pi_{r_R})\right) \xrightarrow{d} \mathcal{N}(0, \nabla \boldsymbol{f}^\top \boldsymbol{\Sigma}_M \nabla \boldsymbol{f}),$$

where $\boldsymbol{\Sigma}_M$ was defined in (E.4). Then

$$\sqrt{NT}(\boldsymbol{\theta}^*)_{\mathcal{S}}^\top \left((\hat{\boldsymbol{\psi}}_N^{(C;\ \hat{\pi})})_{\mathcal{S}} - (\boldsymbol{\psi}_N^{(C;\ \pi)})_{\mathcal{S}}\right)$$

$$= \sqrt{NT}(\boldsymbol{\theta}_N^*)_{\mathcal{S}}^\top \left(\sum_{r \in \mathcal{R}} [f_r(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R}) - f_r(\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_R})] \sum_{t=r}^{T} \psi_{rt}(\boldsymbol{D}_N^{-1})_{i(r,t),\mathcal{S}}\right)$$

$$= \sqrt{NT}(\boldsymbol{\theta}_N^*)_{\mathcal{S}}^\top \left(\boldsymbol{M}^\top [\boldsymbol{f}(\hat{\pi}_{r_1}, \hat{\pi}_{r_2}, \ldots, \hat{\pi}_{r_R}) - \boldsymbol{f}(\pi_{r_1}, \pi_{r_2}, \ldots, \pi_{r_R})]\right)$$

$$\xrightarrow{d} \mathcal{N}(0, \boldsymbol{M}^\top \nabla \boldsymbol{f}^\top \boldsymbol{\Sigma}_M \nabla \boldsymbol{f} \boldsymbol{M}),$$

where

$$\boldsymbol{M} := \begin{pmatrix} \sum_{t=r_1}^{T} \psi_{r_1 t}(\boldsymbol{D}_N^{-1})_{i(r_1,t),\mathcal{S}} \\ \sum_{t=r_2}^{T} \psi_{r_2 t}(\boldsymbol{D}_N^{-1})_{i(r_2,t),\mathcal{S}} \\ \vdots \\ \sum_{t=r_R}^{T} \psi_{r_R t}(\boldsymbol{D}_N^{-1})_{i(r_R,t),\mathcal{S}} \end{pmatrix} \in \mathbb{R}^{R \times s}.$$

Further, we can estimate this covariance matrix by $\hat{\boldsymbol{V}}(\mathcal{A}_N)$ defined for any $\mathcal{A}_N \subseteq [p_N]$ by

$$\hat{\boldsymbol{V}}(\mathcal{A}_N) = T \cdot \hat{\boldsymbol{M}}(\mathcal{A}_N)^\top \nabla \hat{\boldsymbol{f}}^\top \hat{\boldsymbol{\Sigma}}_M \nabla \hat{\boldsymbol{f}} \hat{\boldsymbol{M}}(\mathcal{A}_N), \tag{E.9}$$

where

$$\hat{\boldsymbol{M}}(\mathcal{A}_N) := \begin{pmatrix} \sum_{t=r_1}^{T} \psi_{r_1 t}(\boldsymbol{D}_N^{-1})_{i(r_1,t),\mathcal{A}_N} \\ \sum_{t=r_2}^{T} \psi_{r_2 t}(\boldsymbol{D}_N^{-1})_{i(r_2,t),\mathcal{A}_N} \\ \vdots \\ \sum_{t=r_R}^{T} \psi_{r_R t}(\boldsymbol{D}_N^{-1})_{i(r_R,t),\mathcal{A}_N} \end{pmatrix} \in \mathbb{R}^{R \times |\mathcal{A}_N|},$$

$$\hat{\boldsymbol{\Sigma}}_M := \frac{1}{N^2} \begin{pmatrix} N_0(N - N_0) & -N_0 N_{r_1} & \cdots & -N_0 N_{r_R} \\ -N_0 N_{r_1} & N_{r_1}(N - N_{r_1}) & \cdots & -N_{r_1} N_{r_R} \\ \vdots & \vdots & \ddots & \vdots \\ -N_0 N_{r_R} & -N_{r_1} N_{r_R} & \cdots & N_{r_R}(N - N_{r_R}) \end{pmatrix} \in \mathbb{R}^{(R+1) \times (R+1)}, \tag{E.10}$$

and $\nabla \hat{\boldsymbol{f}}$ is defined by replacing each instance of π_r in $\nabla \boldsymbol{f}$ with its estimator $\hat{\pi}_r$. Using the consistency of $\hat{\boldsymbol{\theta}}^{(q)}$ and the sample covariance matrix, the continuous mapping theorem yields that

$$(\hat{\boldsymbol{\theta}}^{(q)})_{\mathcal{S}}^\top \hat{\boldsymbol{V}}(\mathcal{S})\hat{\boldsymbol{\theta}}_{\mathcal{S}}^{(q)} \xrightarrow{p} T(\boldsymbol{\theta}_N^*)_{\mathcal{S}}^\top (\nabla \boldsymbol{\psi}_N^{(C;\ \pi)})_{\mathcal{S}}^\top \boldsymbol{\Sigma}_M (\nabla \boldsymbol{\psi}_N^{(C;\ \pi)})_{\mathcal{S}}(\boldsymbol{\theta}_N^*)_{\mathcal{S}},$$

where $\boldsymbol{\Sigma}_M$ was defined in (E.4). It follows from Theorem C.1(g) that if at least one of the

nonzero components of $\boldsymbol{\psi}_N^{(\mathrm{C};\,\pi)}$ corresponds to one of the s features in $\mathcal{S}$, we have

$$\sqrt{\frac{NT}{\hat{v}_N^{\mathrm{C};\hat{\pi}}}}\left((\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})^\top\hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{\psi}_N^{(\mathrm{C};\,\pi)})^\top\boldsymbol{\theta}_N^*\right) \overset{d}{\to} \mathcal{N}(0,1)$$

$$\Longleftrightarrow \quad \sqrt{\frac{NT}{\hat{v}_N^{\mathrm{C};\hat{\pi}}}}\sum_{r\in\mathcal{R}}\sum_{t=r}^T \psi_{rt}\left(f_r(\hat{\pi}_{r_1},\hat{\pi}_{r_2},\ldots,\hat{\pi}_{r_R})\hat{\tau}_{\mathrm{ATT}}(r,t) - f_r(\pi_{r_1},\pi_{r_2},\ldots,\pi_{r_R})\tau_{\mathrm{ATT}}(r,t)\right)$$

$$\overset{d}{\to} \mathcal{N}(0,1),$$

where

$$\hat{v}_N^{\mathrm{C};\hat{\pi}} := \sigma^2(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})_{\hat{\mathcal{S}}}^\top\left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\hat{\mathcal{S}}}\right)\right)^{-1}(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})_{\hat{\mathcal{S}}} + (\hat{\boldsymbol{\theta}}^{(q)})_{\hat{\mathcal{S}}}^\top\hat{\boldsymbol{V}}(\hat{\mathcal{S}})(\hat{\boldsymbol{\theta}}^{(q)})_{\hat{\mathcal{S}}}. \qquad \text{(E.11)}$$

(c) Part (b) follows from the same argument as part (a) but applying Theorem C.1(h) to yield the stated convergence when using the conservative variance estimator

$$\hat{v}_N^{\mathrm{C,\ (cons)};\hat{\pi}}$$
$$:= \sigma^2(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})_{\hat{\mathcal{S}}}^\top\left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\hat{\mathcal{S}}}\right)\right)^{-1}(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})_{\hat{\mathcal{S}}} + (\hat{\boldsymbol{\theta}}^{(q)})_{\hat{\mathcal{S}}}^\top\hat{\boldsymbol{V}}(\hat{\mathcal{S}})(\hat{\boldsymbol{\theta}}^{(q)})_{\hat{\mathcal{S}}}$$
$$+ 2\sqrt{\sigma^2(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})_{\hat{\mathcal{S}}}^\top\left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\hat{\mathcal{S}}}\right)\right)^{-1}(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})})_{\hat{\mathcal{S}}}\cdot(\hat{\boldsymbol{\theta}}^{(q)})_{\hat{\mathcal{S}}}^\top\hat{\boldsymbol{V}}(\hat{\mathcal{S}})(\hat{\boldsymbol{\theta}}^{(q)})_{\hat{\mathcal{S}}}}. \qquad \text{(E.12)}$$

(Recall that the random vector $\hat{\boldsymbol{\psi}}_N^{(\mathrm{C};\,\hat{\pi})}$ was defined in Equation E.8 and the covariance matrix estimator $\hat{\boldsymbol{V}}(\hat{\mathcal{S}})$ was defined in Equation E.9.)

$\square$

Proof of Theorem 6.4. Part (a) is immediate from Theorem D.2(a). To prove parts (b) and (c), in light of Theorem D.2(b) and (c) it suffices to show that $(N_{r_1}/N_\tau,\ldots,N_{r_R}/N_\tau)$ is almost surely finite and has a Jacobian at $(\pi_0,\pi_{r_1},\ldots,\pi_{r_R})$ that exists and is nonzero and finite, which we already showed in the proof of Theorem 6.1.

$\square$

Proof of Theorem 6.3. In the case where the treatment effects are nonzero, this result follows from the same arguments that we used in the proof of Theorem D.2, but applying Theorem C.1(d) and (f). The oracle properties persist even in the case that the treatment effects equal 0 because, as noted in Theorem 6.2, in this case the random variables scaled by $\sqrt{NT}$ converge in probability (and distribution) to 0. $\square$

E.3 Proof of Theorem D.3

(a) From Theorem A.5, under our assumptions the regression estimands correspond to our causal estimands. Similarly to the proof of Theorem D.2, we will prove the result by applying Theorem C.1(g) and Lemma E.3. Let $i(r,t,\boldsymbol{x})$ denote the indices of d rows of $\boldsymbol{D}_N$ corresponding to $\hat{\rho}_{rt}^{(q)}$ (that is, the columns of $\tilde{\boldsymbol{Z}}$ corresponding to the indicator variables for those regression coefficients), so that

$$\boldsymbol{\rho}_{rt}^* = (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot}\boldsymbol{\theta}_N^*,$$

and for any $r\in\mathcal{R}$, $t\geq r$, and $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$,

$$\tau_{\mathrm{CATT}}(r,t,\boldsymbol{x}) = \tau_{rt}^* + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i\mid W_i=r])^\top\boldsymbol{\rho}_{rt}^*$$
$$= (\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}\boldsymbol{\theta}_N^* + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i\mid W_i=r])^\top(\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot}\boldsymbol{\theta}_N^*.$$

Define

$$\boldsymbol{\psi}_N^{(\text{C;CATT})} := \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left((\boldsymbol{D}_N^{-1})_{i(r,t),\cdot} + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r])^\top (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot} \right)$$

and

$$\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT})} := \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left((\boldsymbol{D}_N^{-1})_{i(r,t),\cdot} + (\boldsymbol{x} - \overline{\boldsymbol{X}}_r)^\top (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot} \right),$$

so

$$(\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT})})^\top \hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{\psi}_N^{(\text{C;CATT})})^\top \boldsymbol{\theta}_N^* = \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left(\hat{\tau}_{\text{CATT}}(r, t, \boldsymbol{x}) - \tau_{\text{CATT}}(r, t, \boldsymbol{x}) \right).$$

Since $\hat{\boldsymbol{\theta}}^{(q)}$ and each $\overline{\boldsymbol{X}}_r$ are estimated on two independent data sets, $\hat{\boldsymbol{\theta}}^{(q)}$ is independent of $\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT})}$. So to apply Theorem C.1(g) we need to establish the asymptotic normality of $\sqrt{NT} \left(\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT})} - \boldsymbol{\psi}_N^{(\text{C;CATT})} \right)^\top \boldsymbol{\theta}_N^*$ and identify a consistent estimator for its asymptotic variance. Applying Lemma E.3 we see that

$$= \sqrt{NT} \left(\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT})} - \boldsymbol{\psi}_N^{(\text{C;CATT})} \right)^\top \boldsymbol{\theta}_N^*$$

$$= \sqrt{NT} \left(\sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \left((\mathbb{E}[\boldsymbol{X}_i \mid W_i = r] - \overline{\boldsymbol{X}}_r)^\top (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot} \right) \right) \boldsymbol{\theta}_N^*$$

$$= \sqrt{NT} \left(\sum_{r \in \mathcal{R}} \sum_{t=r}^{T} -\psi_{rt} \left((\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot} \boldsymbol{\theta}_N^* \right)^\top \left(\overline{\boldsymbol{X}}_r - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r] \right) \right)$$

$$\xrightarrow{d} \mathcal{N} \left(0, v_R^{\text{CATT}} \right),$$

where

$$v_R^{\text{CATT}} := T \sum_{r \in \mathcal{R}} \frac{(\boldsymbol{\psi}_r^{\text{CATT}})^\top \text{Cov}(\boldsymbol{X}_i \mid W_i = r) \boldsymbol{\psi}_r^{\text{CATT}}}{\mathbb{P}(W_i = r)}$$

for

$$\boldsymbol{\psi}_r^{\text{CATT}} := \sum_{t=r}^{T} \psi_{rt} \left((\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot} \boldsymbol{\theta}_N^* \right)^\top = \sum_{t=r}^{T} \psi_{rt} \boldsymbol{\rho}_{rt}^*, \qquad r \in \mathcal{R}.$$

Notice that each v_R^{CATT} is strictly positive and finite because each $\text{Cov}(\boldsymbol{X}_i \mid W_i = r)$ is assumed to be finite and have a strictly positive minimum eigenvalue under Assumption (R6) and each $\mathbb{P}(W_i = r)$ is strictly positive under Assumption (F2). Also, by the consistency of $\hat{\boldsymbol{\theta}}^{(q)}$ and each N_r/N and $\widehat{\text{Cov}}(\boldsymbol{X}_i \mid W_i = r)$, along with the continuous mapping theorem, we have that

$$T \sum_{r \in \mathcal{R}} \frac{(\hat{\boldsymbol{\psi}}_r^{\text{CATT}})^\top \widehat{\text{Cov}}(\boldsymbol{X}_i \mid W_i = r) \hat{\boldsymbol{\psi}}_r^{\text{CATT}}}{N_r/N} \xrightarrow{p} v_R^{\text{CATT}}$$

where

$$\hat{\boldsymbol{\psi}}_r^{\text{CATT}} := \sum_{t=r}^{T} \psi_{rt} \hat{\boldsymbol{\rho}}_{rt}^{(q)}, \qquad r \in \mathcal{R}.$$

Then from Theorem C.1(g) we have

$$\sqrt{\frac{NT}{\hat{v}_N^{(\mathrm{C;CATT})}}} \left((\hat{\boldsymbol{\psi}}_N^{(\mathrm{C;CATT})})^\top \hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{\psi}_N^{(\mathrm{C;CATT})})^\top \boldsymbol{\theta}_N^*\right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\iff \quad \sqrt{\frac{NT}{\hat{v}_N^{(\mathrm{C;CATT})}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \left(\hat{\tau}_{\mathrm{CATT}}(r,t,\boldsymbol{x}) - \tau_{\mathrm{CATT}}(r,t,\boldsymbol{x})\right) \xrightarrow{d} \mathcal{N}(0,1)$$

for

$$\hat{v}_N^{(\mathrm{C;CATT})} := \sigma^2 (\hat{\boldsymbol{\psi}}_N^{(\mathrm{C;CATT})})^\top \left(\widehat{\mathrm{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot,\cdot)}\hat{\mathcal{S}}\right)\right)^{-1} \hat{\boldsymbol{\psi}}_N^{(\mathrm{C;CATT})}$$

$$+ T \sum_{r \in \mathcal{R}} \frac{(\hat{\boldsymbol{\psi}}_r^{\mathrm{CATT}})^\top \widehat{\mathrm{Cov}}(\boldsymbol{X}_i \mid W_i = r)\hat{\boldsymbol{\psi}}_r^{\mathrm{CATT}}}{N_r/N}. \tag{E.13}$$

(b) Observe that

$$\sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \tau_{\mathrm{CATT}}(r,t,\boldsymbol{x})$$

$$= \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \left((\boldsymbol{D}_N^{-1})_{i(r,t),\cdot}^\top\right.$$

$$\left. + (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot}^\top \left(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]\right)\right) \boldsymbol{\theta}_N^*$$

$$= \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) (\boldsymbol{\psi}_{rt}^*)^\top \boldsymbol{\theta}_N^*$$

$$= (\boldsymbol{\psi}_N^{(\mathrm{C;CATT},\pi)})^\top \boldsymbol{\theta}_N^*$$

where

$$\boldsymbol{\psi}_{rt}^* := (\boldsymbol{D}_N^{-1})_{i(r,t),\cdot} + (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot}^\top \left(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]\right) \in \mathbb{R}^p \tag{E.14}$$

and

$$\boldsymbol{\psi}_N^{(\mathrm{C;CATT},\pi)} := \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \boldsymbol{\psi}_{rt}^* \in \mathbb{R}^p.$$

Define the estimators

$$\hat{\boldsymbol{\psi}}_{rt}^* := (\boldsymbol{D}_N^{-1})_{i(r,t),\cdot} + (\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot}^\top \left(\boldsymbol{x} - \overline{\boldsymbol{X}}_r\right) \tag{E.15}$$

and

$$\hat{\boldsymbol{\psi}}_N^{(\mathrm{C;CATT},\pi)} := \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right) \hat{\boldsymbol{\psi}}_{rt}^*$$

and notice that

$$(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C;CATT},\pi)})^\top \hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{\psi}_N^{(\mathrm{C;CATT},\pi)})^\top \boldsymbol{\theta}_N^*$$

$$= \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \Big(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right) \hat{\tau}_{\mathrm{CATT}}(r,t,\boldsymbol{x})$$

$$- f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \tau_{\mathrm{CATT}}(r,t,\boldsymbol{x})\Big).$$

Since by assumption each $\hat{\pi}_r(\boldsymbol{x})$ and $\overline{\boldsymbol{X}}_r$ is independent of $\hat{\boldsymbol{\theta}}^{(q)}$, $\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)}$ is independent of $\hat{\boldsymbol{\theta}}^{(q)}$, and the result follows from applying Theorem C.1(g) if we can show that

$$\sqrt{NT}(\boldsymbol{\theta}_N^*)^\top \left(\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)} - \boldsymbol{\psi}_N^{(\text{C;CATT},\pi)} \right)$$

is asymptotically normal and identify a consistent estimator for the asymptotic variance. Showing this is conceptually straightforward—we can decompose $\sqrt{NT}(\boldsymbol{\theta}_N^*)^\top \left(\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)} - \boldsymbol{\psi}_N^{(\text{C;CATT},\pi)} \right)$ into a term that converges in probability to 0, a term that has an asymptotic distribution depending on $\sqrt{N}\left(\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r\right)$ (asymptotically normal due to Lemma E.3), and a term that depends on the $\hat{\pi}_r(\boldsymbol{x})$ predictions (asymptotically normal by assumption and application of the delta method). Then the sum converges to a normal distribution because of the independence of each $\hat{\pi}_r(\boldsymbol{x})$ and $\overline{\boldsymbol{X}}_r$ and Slutsky's theorem.

However, the calculations are sufficiently tedious that we defer the work to the proof of Lemma E.5.

Lemma E.5. *Under the assumptions of Theorem D.3(b),*

$$\sqrt{NT}(\boldsymbol{\theta}_N^*)^\top \left(\hat{\boldsymbol{\psi}}_N^{(C;CATT,\pi)} - \boldsymbol{\psi}_N^{(C;CATT,\pi)} \right) \xrightarrow{d} \mathcal{N}\left(0, v_R^{(C;CATT,\pi)} \right)$$

for

$$v_R^{(C;CATT,\pi)} := \sum_{r \in \mathcal{R}} \frac{(\boldsymbol{\psi}_r^{CATT;\ prob})^\top \text{Cov}(\boldsymbol{X}_i \mid W_i = r)\boldsymbol{\psi}_r^{CATT;\ prob}}{\mathbb{P}(W_i = r)} + T \cdot \boldsymbol{\psi}_\pi^\top \nabla \boldsymbol{f}^\top \boldsymbol{\Sigma}_\pi \nabla \boldsymbol{f} \boldsymbol{\psi}_\pi,$$

$\boldsymbol{\psi}_r^{CATT;\ prob}$ *is defined in (G.9), and* $\boldsymbol{\Sigma}_\pi$ *was defined in the statement of Theorem D.3. Further,* $\hat{v}_R^{(C;CATT,\pi)}$, *defined in (G.10), is a consistent estimator for* $v_R^{(C;CATT,\pi)}$.

Proof. Provided in Appendix G.2. □

Then from Lemma E.5 and Theorem C.1(g) we have

$$\sqrt{\frac{NT}{\hat{v}_N^{(\text{C;CATT},\pi)}}} \left((\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)})^\top \hat{\boldsymbol{\theta}}^{(q)} - (\boldsymbol{\psi}_N^{(\text{C;CATT},\pi)})^\top \boldsymbol{\theta}_N^* \right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\Longleftrightarrow \quad \sqrt{\frac{NT}{\hat{v}_N^{(\text{C;CATT},\pi)}}} \sum_{r \in \mathcal{R}} \sum_{t=r}^T \psi_{rt} \big(f_r\left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right) \hat{\tau}_{\text{CATT}}(r,t,\boldsymbol{x})$$

$$- f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \tau_{\text{CATT}}(r,t,\boldsymbol{x}) \xrightarrow{d} \mathcal{N}(0,1).$$

for

$$\hat{v}_N^{(\text{C;CATT},\pi)} := \sigma^2 (\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)})^\top \left(\widehat{\text{Cov}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot \cdot)\hat{\mathcal{S}}} \right) \right)^{-1} \hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)}$$

$$+ \hat{v}_R^{(\text{C;CATT},\pi)}. \tag{E.16}$$

F Proof of Theorem C.1

Before we prove the individual statements, we outline our general approach. Notice that the assumptions leading to equation (6.33) in Wooldridge (2021, Section 6.3) are satisfied, so we have for any $i \in [N]$

$$\mathbb{E}\left[\tilde{\boldsymbol{y}}_{(i \cdot)} \mid W_i, \boldsymbol{X}_i \right] = \tilde{\boldsymbol{Z}}_{(i \cdot)} \boldsymbol{\beta}_N^*, \tag{F.1}$$

where the subscript $(i\cdot)$ refers to the T observations corresponding to unit i. (This verifies part (a) of Theorem C.1, though we also provide a proof in our notation below.) We propose estimating the bridge regression (5.4),

$$\hat{\boldsymbol{\beta}}^{(q)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{pN}}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda_N \|\boldsymbol{D}_N \boldsymbol{\beta}\|_q^q \right\}.$$

We proved in Lemma E.1 that the differences matrix $\boldsymbol{D}_N$ is invertible. We can therefore solve (5.4) through the reparameterization $\hat{\boldsymbol{\theta}}^{(q)} = \boldsymbol{D}_N \hat{\boldsymbol{\beta}}^{(q)}$, solving for

$$\hat{\boldsymbol{\theta}}^{(q)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{pN}}{\arg\min} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{D}_N^{-1}\boldsymbol{\theta}\|_2^2 + \lambda_N \|\boldsymbol{\theta}\|_q^q \right\}$$

using standard bridge regression. Then $\hat{\boldsymbol{\beta}}^{(q)} = \boldsymbol{D}_N^{-1}\hat{\boldsymbol{\theta}}^{(q)}$ is identical to the solution to (5.4) (Tibshirani and Taylor, 2011, Section 3).

Having cast (5.4) as a standard bridge regression optimization problem, our results will follow from Theorem 1, Lemma 3, and Theorem 2(ii) from Kock (2013) if we show the assumptions are met in our setting where

$$
\begin{aligned}
\tilde{\boldsymbol{Y}}_{iN} &= \tilde{\boldsymbol{y}}_{(i\cdot)}, \\
\tilde{\boldsymbol{X}}_{iN} &= \tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1}, \\
\beta_0 &= \boldsymbol{\theta}_N^*, \\
\boldsymbol{c}_{iN} &= c_i \boldsymbol{1}_T, \\
\tilde{\boldsymbol{\epsilon}}_{iN} &= \boldsymbol{u}_{(i\cdot)}, \\
\boldsymbol{Y}_{iN} &= \boldsymbol{y}_{(i\cdot)}, \\
\boldsymbol{X}_{iN} &= \boldsymbol{Z}_{(i\cdot)}\boldsymbol{D}_N^{-1}, \\
\boldsymbol{X}_N &= \boldsymbol{Z}\boldsymbol{D}_N^{-1}, \\
\boldsymbol{W}_N &= (\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{S}}, \\
\boldsymbol{\Sigma}_{1N} &= \hat{\boldsymbol{\Sigma}}((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{S}}), \\
\boldsymbol{\epsilon}_{iN} &= \sigma\boldsymbol{\Omega}^{-1/2}\boldsymbol{\epsilon}_{(i\cdot)}, \\
\gamma &= q, \quad \text{and} \\
k_N &= s_N,
\end{aligned}
$$

where the left side of each equation uses the notation from Kock (2013) and the right side uses the notation in our setting. We focus on the random effects setting of Kock (2013) because we cannot perform the forward orthogonal deviations transform since $\tilde{\boldsymbol{Z}}$ (and $\tilde{\boldsymbol{Z}}\boldsymbol{D}_N^{-1}$) contain (linear combinations of) time-invariant covariates and cohort indicators.

Now we are prepared to prove each statement.

(a) As mentioned earlier, under our assumptions correct specification is immediate from Equation (6.33) in Wooldridge (2021, Section 6.3). We also provide a derivation for the correct specifica-

tion here in our notation. Observe that for any $i \in [N]$,

$$\mathbb{E}\left[\tilde{y}_{(i1)} \mid W_i, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$= \mathbb{1}\{W_i = 0\}\mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] + \sum_{r \in \mathcal{R}} \mathbb{1}\{W_i = r\}\mathbb{E}\left[\tilde{y}_{(i1)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(a)}{=} \mathbb{1}\{W_i = 0\}\mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] + \sum_{r \in \mathcal{R}} \mathbb{1}\{W_i = r\}\mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(b)}{=} \mathbb{1}\{W_i = 0\}\left(\eta^* + \boldsymbol{x}^\top \boldsymbol{\kappa}^*\right) + \sum_{r \in \mathcal{R}} \mathbb{1}\{W_i = r\}\left(\eta^* + \nu_r^* + \boldsymbol{x}^\top\left(\boldsymbol{\kappa}^* + \boldsymbol{\zeta}_r^*\right)\right)$$

$$= \eta^* + \boldsymbol{x}^\top \boldsymbol{\kappa}^* + \sum_{r \in \mathcal{R}} \mathbb{1}\{W_i = r\}\left(\nu_r^* + \boldsymbol{x}^\top \boldsymbol{\zeta}_r^*\right), \tag{F.2}$$

where in (a) we used Assumption (CNAS) and in (b) we used (3.3) and (3.4) from (LINS). Next, for any $t \in \{2, \ldots, T\}$ and i such that $W_i = 0$,

$$\mathbb{E}\left[\tilde{y}_{(it)} \mid W_i = 0, \boldsymbol{X}_i\right]$$

$$= \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right] + \mathbb{E}\left[\tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(c)}{=} \eta^* + \gamma_t^* + \boldsymbol{x}^\top(\boldsymbol{\xi}_t^* + \boldsymbol{\kappa}^*),$$

where in (c) we used (3.5) from (LINS) and (F.2). Then for any i such that $W_i = r \in \mathcal{R} \setminus \{2\}$ and $t \in \{2, \ldots, r-1\}$,

$$\mathbb{E}\left[\tilde{y}_{(it)} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$= \mathbb{E}\left[\tilde{y}_{(i1)} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right] + \mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(i1)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(e)}{=} \mathbb{E}\left[\tilde{y}_{(i1)} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right] + \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(f)}{=} \mathbb{E}\left[\tilde{y}_{(i1)} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right] + \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(g)}{=} \eta^* + \nu_r^* + \boldsymbol{x}^\top(\boldsymbol{\kappa}^* + \boldsymbol{\zeta}_r^* + \boldsymbol{\xi}_t^*) + \gamma_t^*,$$

where in (e) we used (CNAS), in (f) we used (CCTSB), and in (g) we used (F.2) and (3.5) from (LINS). Finally, for any i such that $W_i = r \in \mathcal{R}$ and $t \in \{r, \ldots, T\}$,

$$\mathbb{E}\left[\tilde{y}_{(it)} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$= \mathbb{E}\left[\tilde{y}_{(i1)} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right] + \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\quad + \mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(i1)}(r) \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}\right] - \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}\right]$$

$$\overset{(h)}{=} \eta^* + \nu_r^* + \boldsymbol{x}^\top(\boldsymbol{\kappa}^* + \boldsymbol{\zeta}_r^* + \boldsymbol{\xi}_t^*) + \gamma_t^* + \tau_{rt}^* + \dot{\boldsymbol{x}}_r \boldsymbol{\rho}_{rt}^*, \tag{F.3}$$

where in (h) we used (F.2) and (3.5) and (3.6) from (LINS). Therefore regression (1.5) is correctly specified and (F.1) holds.

(b) First we will show that Assumptions (FE1) - (FE3), (RE4), (A1), (A3), (A5), and (A6) from Kock (2013) are satisfied. Then we will show that the conclusion of Theorem 1 in Kock (2013) about $\boldsymbol{\theta}_N^*$ leads to our statement about $\boldsymbol{\beta}_N^*$. We will use three lemmas:

Lemma F.1. *The smallest and largest singular values of $\boldsymbol{D}_N$ are bounded as follows: $\sigma_{max}(\boldsymbol{D}_N) \leq 3$ and $\sigma_{min}(\boldsymbol{D}_N) \geq \frac{1}{T\sqrt{2T}}$.*

Lemma F.2. *Under Assumption (R1) it holds that*

$$\mathbb{E}\left[(\tilde{\boldsymbol{Z}}_{(i \cdot)} \boldsymbol{D}_N^{-1})_{(it)j}^4\right] \leq \kappa_4 \qquad \forall i \in [N], j \in [p_N], t \in [T]$$

for some finite $\kappa_4 > 0$.

Lemma F.3. *Under the assumptions of Theorem C.1(b),*

$$\rho_{1N} = \lambda_{min}\left(\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)\right) \geq \frac{1}{9}e_{1N} \qquad and$$

$$\rho_{2N} = \lambda_{max}\left(\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)\right) \leq e_{2N} \cdot 2T^3.$$

The proofs of all three of these lemmas are provided in Appendix H. Now we will verify one assumption at a time.

- In our notation, Assumption (FE1) from Kock (2013) requires that $(\tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}, c_i, \boldsymbol{u}_{(i\cdot)})_{i=1}^N$ are iid. This is satisfied by our Assumption (F2) since $\tilde{\boldsymbol{Z}}_{(i\cdot)}.$ is a deterministic, invertible function of $(W_i, \boldsymbol{X}_i)$, and $\boldsymbol{D}_N^{-1}$ is also deterministic and invertible conditional on $(W_i, \boldsymbol{X}_i)$.

- Assumption (FE2) requires that $\mathbb{E}[(\tilde{\boldsymbol{Z}}\boldsymbol{D}_N^{-1})_{(it)j}^4]$ and $\mathbb{E}[u_{(it)}^4]$ are finite for all $i \in [N], j \in [d_N], t \in [T]$. We assumed the latter in Assumption (R1). We proved that the former is satisfied in Lemma F.2.

- Next, Assumption (FE3) requires that

$$\mathbb{E}\left[\boldsymbol{u}_{(i\cdot)} \mid \tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}, c_i\right] = \boldsymbol{0} \qquad and$$

$$\text{Var}\left[\boldsymbol{u}_{(i\cdot)} \mid \tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}, c_i\right] = \sigma^2\boldsymbol{I}_T, \qquad i \in [N],$$

and Assumption (RE4) requires

$$\mathbb{E}\left[c_i \mid \tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}\right] = 0,$$

$$\text{Var}\left[c_i \mid \tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}\right] = \sigma_c^2,$$

and that σ and σ_c^2 are known and finite. Our Assumption (F1) is sufficient for this because conditioning on $W_i, \boldsymbol{X}_i$ is the same as conditioning on $\tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}$ by the following argument. Because $\boldsymbol{D}_N^{-1}$ is invertible, the σ-algebra generated by $\tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1}$, $\sigma(\tilde{\boldsymbol{Z}}_{(i\cdot)}.\boldsymbol{D}_N^{-1})$, contains the same information as $\sigma(\tilde{\boldsymbol{Z}}_{(i\cdot)}.)$. Similarly, for each $(W_i, \boldsymbol{X}_i)$ there is only one valid $\tilde{\boldsymbol{Z}}_{(i\cdot)}.$, and every valid $\tilde{\boldsymbol{Z}}_{(i\cdot)}.$ corresponds to a unique $(W_i, \boldsymbol{X}_i)$—that is, the mapping from $(W_i, \boldsymbol{X}_i)$ to $\tilde{\boldsymbol{Z}}_{(i\cdot)}.$ is also invertible, so $\sigma(W_i, \boldsymbol{X}_i) = \sigma(\tilde{\boldsymbol{Z}}_{(i\cdot)}.)$.

- Our Assumptions (R1), (R2), and (R3) exactly match assumptions (A1), (A3), and (A5), respectively, from Kock (2013) (in our setting where we assume T is fixed).

- Finally, Assumption (A6) in Kock (2013) requires that

$$\frac{p_N + \lambda_N s_N}{N\rho_{1N}} \xrightarrow{a.s.} 0,$$

where ρ_{1N} is the smallest eigenvalue of the random matrix $\hat{\boldsymbol{\Sigma}}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)$. Under Lemma F.3 the smallest eigenvalue of $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z})$, e_{1N}, is within a constant factor of ρ_{1N}, so our Assumption (R2) suffices to satisfy this condition.

We have shown that the assumptions of Theorem 1 in Kock (2013) are satisfied, so we have $\|\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(\min\{h_N, h_N'\}\right)$. Finally, part (2) is proven because $\left\|\hat{\boldsymbol{\beta}}^{(q)} - \boldsymbol{\beta}_N^*\right\|_2$ is within a constant factor of $\|\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*\|_2$:

$$\|\hat{\boldsymbol{\beta}}^{(q)} - \boldsymbol{\beta}_N^*\|_2 = \left\|\boldsymbol{D}_N^{-1}\left(\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*\right)\right\|_2 \leq \left\|\boldsymbol{D}_N^{-1}\right\|_{op}\left\|\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*\right\|_2 \leq T\sqrt{2T}\left\|\hat{\boldsymbol{\theta}}^{(q)} - \boldsymbol{\theta}_N^*\right\|_2,$$

where in the last step we applied Lemma F.1.

(c) In light of the results from part (a), part (c) follows from Theorem F.4 below, which extends Lemma 3 from Kock (2013), under our assumptions. In particular, we will show that our Assumptions (R4) and (R5) suffice for Kock (2013)'s Assumptions (A4) and (A7), as well as the needed part of Assumption (A2) for Lemma 3, in Kock (2013).

Theorem F.4 (Extension of Lemma 3 from Kock 2013). *Under the assumptions of Lemma 3 from Kock (2013),*

$$\lim_{N\to\infty} \mathbb{P}\left(\hat{\mathcal{S}} = \mathcal{S}\right) = 1.$$

Proof. Provided in Appendix H. □

Remark F.1. *It appears that the strategy for the proof of of Theorem F.4 could be used to analogously extend Lemma 2, and Theorem 2(i), in Huang et al. (2008).*

Again we verify one assumption at a time.

- Although Lemma 3 from Kock (2013) requires their Assumption (A2) in the statement of the theorem, examining the proof we see that a lower bound on the minimum eigenvalue is not needed—the only thing that is needed is an almost sure upper bound on the maximum eigenvalue of $\hat{\boldsymbol{\Sigma}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{S}}\right)$, the empirical Gram matrix of the columns of $\boldsymbol{Z}\boldsymbol{D}_N^{-1}$ corresponding to the relevant features in $\boldsymbol{\theta}_N^*$. (In the proof, note the use of τ_2 on p. 140, and note that τ_1 is not needed.) Our Assumption (R4) is sufficient for this, since we assume the maximum eigenvalue of the empirical Gram matrix for $\tilde{\boldsymbol{Z}}$, e_{2N} is finite, which ensures that the maximum eigenvalue of the Gram matrix for the full $\boldsymbol{Z}\boldsymbol{D}_N^{-1}$ is finite due to Lemma F.3. This maximum eigenvalue upper-bounds the maximum eigenvalue of the submatrix $\hat{\boldsymbol{\Sigma}}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot\cdot)\mathcal{S}}\right)$.
- Our Assumption (R5) matches Assumption (A4) in Kock (2013) in our setting where $T_N = T$ is fixed.
- Finally, we show that Assumption (A7) in Kock (2013) is satisfied under our Assumption (R4). We need to show that

$$e_{1N}\sqrt{\frac{e_{2N}}{p_N}} = \mathcal{O}_p(1).$$

But since $e_{1N} \leq e_{2N}$, almost surely

$$e_{1N}\sqrt{\frac{e_{2N}}{p_N}} \leq \sqrt{\frac{e_{2N}^3}{p_N}} \leq \sqrt{\frac{e_{\max}^3}{p_N}} = \mathcal{O}\left(\frac{1}{\sqrt{p_N}}\right),$$

so $e_{1N}\sqrt{e_{2N}/p_N} = \mathcal{O}_p(1)$.

So the first part of the result is now immediate. It only remains to show that

$$\lim_{N\to\infty} \mathbb{P}\left(\sqrt{NT}\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = \mathbf{0}\right) = 1.$$

From part (c), we have that $\lim_{n\to\infty} \mathbb{P}\left(\hat{\boldsymbol{\theta}}_{\mathcal{S}^C}^{(q)} = \mathbf{0}\right) = 1$, so using $\{\sqrt{NT}\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = 0\} = \{\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = 0\}$,

$$\lim_{N\to\infty} \mathbb{P}\left(\sqrt{NT}\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = 0\right) = \lim_{N\to\infty} \mathbb{P}\left(\hat{\boldsymbol{\theta}}_{\mathcal{S}^c}^{(q)} = 0\right) = 1.$$

(d) To prove parts (d) - (h), we will present an extension of Theorem 2(ii) from Kock (2013) that yields our result when in applied in our setting. Then we will show that the needed assumptions are satisfied to conclude the proof. In particular, our extension (a) obtains the

asymptotic convergence of linear combinations of the coefficients of both relevant and irrelevant features; (b) establishes the asymptotic convergence of a statistic formed using the estimated Gram matrix on the selected set of coefficients, not the population covariance matrix of the true active set; (c) is like (a) when $\boldsymbol{\psi}_N$ is estimated rather than known and is independent of $\hat{\beta}_N$; (d) is like (b) when $\boldsymbol{\psi}_N$ is estimated rather than known; and (e) is like (d) when $\hat{\beta}_N$ and $\hat{\boldsymbol{\psi}}_N$ are estimated on the same data set, so they are dependent.

Theorem F.5 (Extension of Theorem 2(ii) from Kock 2013). *Suppose the assumptions of Theorem 2(ii) from Kock (2013) are satisfied.*

(a) *(Asymptotic normality when coefficients equal to 0 are included.) Let $\{\boldsymbol{\psi}_N\}_{N=1}^{\infty}$ be a sequence of real-valued vectors where $\boldsymbol{\psi}_N \in \mathbb{R}^{p_N}$ has the structure $(\boldsymbol{\alpha}^\top, \boldsymbol{b}_N^\top)^\top$ where $\boldsymbol{\alpha} \in \mathbb{R}^k$ is a fixed, finite, nonzero vector and $\{\boldsymbol{b}_N\}_{N=1}^{\infty}$ is any sequence of constants where each $\boldsymbol{b}_N \in \mathbb{R}^{p_N-k}$ contains all finite entries. Then*

$$\sqrt{NT}\boldsymbol{\psi}_N^\top(\hat{\beta}_N - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}\left[\Sigma_{1N}\right]\right)^{-1} \boldsymbol{\alpha}\right).$$

(b) *(Asymptotic normality when coefficients equal to 0 are included and the covariance matrix and selected set are estimated.) Define a function $\boldsymbol{\alpha}$ that maps any $\mathcal{A} \in \mathcal{P}(\mathbb{N})$ to a fixed vector $\boldsymbol{\alpha}(\mathcal{A}) \in (\mathbb{R} \setminus \{0\})^{|\mathcal{A}|}$. Define the sequence of vectors $\{\boldsymbol{\psi}_N(\mathcal{A}_N)\}_{N=1}^{\infty}$ to have components in $\mathcal{A}_N$ equal to $\boldsymbol{\alpha}(\mathcal{A}_N)$ and components $\{\boldsymbol{b}_N(\mathcal{A}_N)\}_{N=1}^{\infty}$ in the components corresponding to $[p_N] \setminus \mathcal{A}_N$ that are all finite for all N (but are otherwise arbitrary). Define the sequence of random variables $U_b(\mathcal{A})$ to equal*

$$\frac{1}{\sigma}\sqrt{\frac{NT}{\boldsymbol{\alpha}(\mathcal{A})^\top \left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot\cdot)\mathcal{A}})\right)^{-1} \boldsymbol{\alpha}(\mathcal{A})}} \cdot \left(\boldsymbol{\psi}_N(\mathcal{A})^\top \left(\hat{\beta} - \beta_0\right)\right),$$

if $\mathcal{A} \neq \emptyset$ and $\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot\cdot)\mathcal{A}})$ is invertible and 0 otherwise, where $\hat{\beta}$ is the bridge estimator of Kock (2013, Section 3). Let $\hat{\mathcal{S}}_N := \{j : \hat{\beta}_{Nj} \neq 0\}$ be the estimated selected set on a data set of size N, and suppose $\beta_0 \neq \boldsymbol{0}$. Then if $\boldsymbol{\alpha}(\mathcal{S}) \neq \boldsymbol{0}$, the sequence $\{U_b(\hat{\mathcal{S}}_N)\}$ converges in distribution to a standard Gaussian random variable.

(c) *(Asymptotic normality when coefficients equal to 0 are included and the weights are estimated.) For any $\mathcal{A} \in \mathcal{P}(\mathbb{N})$, suppose that $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ is a random vector with the property that for a given data set, $\mathcal{A} = \mathcal{A}'$ implies $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}) = \hat{\boldsymbol{\alpha}}_N(\mathcal{A}')$. Define the sequence of random vectors $\{\hat{\boldsymbol{\psi}}_N(\mathcal{A}_N)\}_{N=1}^{\infty}$ to have components in $\mathcal{A}_N$ equal to $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ and components $\{\hat{\boldsymbol{b}}_N(\mathcal{A}_N)\}_{N=1}^{\infty}$ in the components corresponding to $[p_N] \setminus \mathcal{A}_N$ that are all almost surely finite for all N (but are otherwise arbitrary). Assume that for any fixed $\beta_0 \in \mathbb{R}^{p_N}$ with the k entries in the $\mathcal{S}$ positions finite and the remaining entries all equal to 0,*

$$\sqrt{NT}\beta_0^\top \left(\hat{\boldsymbol{\psi}}_N(\mathcal{S}) - \boldsymbol{\psi}_N(\mathcal{S})\right) \xrightarrow{d} \mathcal{N}(0, v_\psi(\beta_0, \mathcal{S})).$$

Define the sequence of random variables $U_c(\mathcal{A})$ to equal

$$\sqrt{NT}\left(\hat{\boldsymbol{\psi}}_N(\mathcal{A})^\top \hat{\beta}_N - \boldsymbol{\psi}_N(\mathcal{A})\beta_0\right),$$

where $\hat{\beta}_N$ is the bridge estimator of Kock (2013, Section 3). Assume $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ is independent of $\hat{\beta}_N$. Let $\hat{\mathcal{S}}_N := \{j : \hat{\beta}_{Nj} \neq 0\}$ be the estimated selected set on a data set of size N, and suppose $\beta_0 \neq \boldsymbol{0}$. Then

$$U_c(\hat{\mathcal{S}}_N) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}\left[\Sigma_{1N}\right]\right)^{-1} \boldsymbol{\alpha} + v_\psi(\beta_0, \mathcal{S})\right).$$

(d) (Asymptotic normality when coefficients equal to 0 are included and the weights, covariance matrix, and selected set are estimated.) Define $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ and the sequence of random vectors $\{\hat{\boldsymbol{\psi}}_N(\mathcal{A}_N)\}_{N=1}^{\infty}$ as in part (c), and assume they have the same properties. Define the sequence of random variables $U_d(\mathcal{A})$ to equal

$$\sqrt{\frac{NT}{\hat{v}_N(\mathcal{A})}} \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A})^{\top}\hat{\beta}_N - \boldsymbol{\psi}_N(\mathcal{A})^{\top}\beta_0\right)$$

if $\mathcal{A} \neq \emptyset, \hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(..)\mathcal{A}})$ is invertible, and $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}) \neq \boldsymbol{0}$; and equals 0 otherwise; where $\hat{\beta}_N$ is the bridge estimator of Kock (2013, Section 3),

$$\hat{v}_N(\mathcal{A}) := \sigma^2\hat{\boldsymbol{\alpha}}(\mathcal{A})^{\top}\left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot)\mathcal{A}})\right)^{-1}\hat{\boldsymbol{\alpha}}(\mathcal{A}) + \hat{v}_{\psi}(\hat{\beta}_N, \mathcal{A}), \tag{F.4}$$

and $\hat{v}_{\psi}(\hat{\beta}_N, \mathcal{A})$ is an estimator of $v_{\psi}(\beta_0, \mathcal{A})$ that satisfies $\hat{v}_{\psi}(\hat{\beta}_N, \mathcal{S}) \xrightarrow{p} v_{\psi}(\beta_0, \mathcal{S})$. Assume $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ is independent of $\hat{\beta}_N$, and suppose $\beta_0 \neq \boldsymbol{0}$. Then $U_d(\hat{\mathcal{S}}_N)$ converges in distribution to a standard Gaussian random variable.

(e) (Asymptotic subgaussianity when coefficients equal to 0 are included and the weights, covariance matrix, and selected set are estimated.) Let the sequence of random variables $U_e(\mathcal{A})$ be defined the same was as $U_d(\mathcal{A})$, except that we assume that $\hat{\boldsymbol{\alpha}}_N(\mathcal{A}_N)$ is estimated on the same data as $\hat{\beta}_N$, so they are not independent, and we use a different variance estimator:

$$U_e(\mathcal{A}) := \sqrt{\frac{NT}{\hat{v}_N^{(cons)}(\mathcal{A})}} \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A})^{\top}\hat{\beta}_N - \boldsymbol{\psi}_N(\mathcal{A})^{\top}\beta_0\right)$$

(unless $\hat{v}_N^{(cons)}(\mathcal{A})$ is not well-defined or equals 0, in which case $U_e(\mathcal{A}) = 0$), where

$$\begin{aligned}\hat{v}_N^{(cons)}(\mathcal{A}) := & \ \sigma^2\hat{\boldsymbol{\alpha}}(\mathcal{A})^{\top}\left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot)\mathcal{A}})\right)^{-1}\hat{\boldsymbol{\alpha}}(\mathcal{A}) + \hat{v}_{\psi}(\hat{\beta}_N, \mathcal{A}) \\ & + 2\sqrt{\sigma^2\hat{\boldsymbol{\alpha}}(\mathcal{A})^{\top}\left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot)\mathcal{A}})\right)^{-1}\hat{\boldsymbol{\alpha}}(\mathcal{A}) \cdot \hat{v}_{\psi}(\hat{\beta}_N, \mathcal{A})}\end{aligned} \tag{F.5}$$

is a conservative variance estimator. Assume again that $\hat{v}_{\psi}(\hat{\beta}_N, \mathcal{S}) \xrightarrow{p} v_{\psi}(\beta_0, \mathcal{S})$ and $\beta_0 \neq \boldsymbol{0}$. Then $U_e(\hat{\mathcal{S}}_N)$ converges in distribution to a mean-zero subgaussian random variable with variance at most 1.

Proof. Provided in Appendix H. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Remark F.2. *It appears that the strategies used to prove Theorem F.5 could be used to analogously extend Theorem 2(ii) in Huang et al. (2008).*

It only remains to show that the assumptions of Theorem F.5 are satisfied in our setting. First, we note that Kock (2013) states on p. 123 that a fixed T is enough to satisfy the last two assumptions of Theorem 2(ii). We have already shown that Assumptions (FE1), (FE2), (FE3), (RE4), (A1), (A3), (A4), (A5), (A6), and (A7) in Kock (2013) are satisfied. Due to Theorem 3(i) from Kock (2013), our fixed T setting ensures that the uniform integrability assumption of Theorem 2(ii) from Kock (2013) is satisfied. We have assumed fixed sparsity in Assumption S(s), and on p. 122 Kock (2013) explains why our fixed T setting satisfies all of the remaining assumptions except for (A2).

We conclude by showing that the conditions implied by Assumption (A2) in Kock (2013) are satisfied. Kock (2013) makes an assumption that the minimum eigenvalue of the empirical Gram matrix of the columns of $\boldsymbol{Z}\boldsymbol{D}_N^{-1}$ corresponding to the nonzero features in $\boldsymbol{\theta}_N^*$,

$$\frac{1}{NT}\left((\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{S}}\right)^{\top}(\boldsymbol{Z}\boldsymbol{D}_N^{-1})_{(\cdot)\mathcal{S}},$$

is bounded away from 0 almost surely. Call this minimum eigenvalue e_{1SN}. This assumption is used in two places, both in the proof of Theorem 2(ii).

(a) On p. 142, this assumption is used to establish that

$$\frac{q}{2}\frac{1}{e_{1SN}}\left(\frac{b_0}{2}\right)^{q-1}\lambda_N\sqrt{\frac{s_N}{NT}} \xrightarrow{p} 0$$

$$\iff \frac{1}{e_{1SN}}\frac{\lambda_N}{\sqrt{N}} \xrightarrow{p} 0$$

$$\iff \frac{1}{e_{1SN}} \in o_{\mathbb{P}}\left(\frac{\sqrt{N}}{\lambda_N}\right),$$

where the first equivalence follows because the removed quantities are constant under our assumptions (specifically, that T and s_N are fixed). Note that our assumption (6.5) ensures that this condition holds for the full empirical Gram matrix due to Lemma F.3. The minimum eigenvalue of the full empirical Gram matrix lower bounds e_{1SN}, so (6.5) is sufficient for this condition to hold.

(b) The assumption is used again on p. 144 of Kock (2013), where it is sufficient for the minimum eigenvalue of the (population) matrix

$$\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}\left((\boldsymbol{ZD}_N^{-1})_{(\cdot\cdot)\mathcal{S}}\right)\right] \tag{F.6}$$

to be bounded away from 0. This is a submatrix of (6.4), so the minimum eigenvalue of (6.4) is a lower bound for the minimum eigenvalue of $\mathbb{E}\left[\hat{\boldsymbol{\Sigma}}\left((\boldsymbol{ZD}_N^{-1})_{(\cdot\cdot)\mathcal{S}}\right)\right]$. Therefore our assumption in (R6) that the minimum eigenvalue of (6.4) is bounded away from 0 is enough to satisfy this assumption.

So our Assumption (R6), in combination with all of the previous assumptions, is enough to apply Theorem F.5, and the proof is complete.

G Proofs of Results Stated in the Appendix

In Appendix G.1 we prove results that were stated but not proven in Appendix A. In Appendix G.2 we prove supporting results for our main theorems.

G.1 Proofs of Results from Appendix A

Proof of Theorem A.2. First we provide a counterexample to show that (CCTS) does not imply (CTS). This counterexample is inspired by discussions in Sections 2 and 3 of Callaway and Karami (2023) that touch on closely related issues. For simplicity, let $T = 2$ and $\mathcal{R} = \{2\}$. Assume that treatment status is not independent of the covariates, so $\mathbb{E}[\boldsymbol{X}_i \mid W_i = 2] \neq \mathbb{E}[\boldsymbol{X}_i \mid W_i = 0]$. Consider a setting where (3.5) from Assumption (LINS) holds with $\boldsymbol{\xi}_2^* \neq \boldsymbol{0}$:

$$\mathbb{E}\left[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i\right] = \gamma_2^* + \boldsymbol{X}_i^\top\boldsymbol{\xi}_2^* \qquad a.s.,$$

and the same condition holds for the untreated potential outcomes of the treated units:

$$\mathbb{E}\left[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 2, \boldsymbol{X}_i\right] = \gamma_2^* + \boldsymbol{X}_i^\top\boldsymbol{\xi}_2^* \qquad a.s.$$

(CCTS) holds since $\mathbb{E}[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 2, \boldsymbol{X}_i] = \mathbb{E}[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i]$ almost surely. But for $r \in \{0, 2\}$,

$$\mathbb{E}[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r] = \mathbb{E}\left[\mathbb{E}\left[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i\right] \mid W_i = r\right]$$
$$= \gamma_2^* + \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]^\top\boldsymbol{\xi}_2^*,$$

so $\mathbb{E}[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 2] \neq \mathbb{E}[\tilde{y}_{(i2)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0]$ since $\mathbb{E}[\boldsymbol{X}_i \mid W_i = 2] \neq \mathbb{E}[\boldsymbol{X}_i \mid W_i = 0]$.

Next we show an example where (CTS) holds but (CCTS) does not. Let $\mathcal{R} = \{2, \ldots, T\}$, and suppose $\boldsymbol{X}_i$ has a discrete distribution, only taking on values $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. To simplify notation, define $G_{it} := \tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0)$ for all $t \in \{2, \ldots, T\}$. Then (CTS) is equivalent to

$$\mathbb{E}[G_{it} \mid W_i] = \mathbb{E}[G_{it}], \qquad i \in [N], t \in \{2, \ldots, T\}$$
$$\Longleftrightarrow \qquad \mathbb{E}[G_{it} \mid W_i = 0] = \mathbb{E}[G_{it} \mid W_i = r], \qquad r \in \mathcal{R}, i \in [N], t \in \{2, \ldots, T\} \qquad \text{(G.1)}$$

and (CCTS) is equivalent to, for all $i \in [N]$ and $t \in \{2, \ldots, T\}$,

$$\mathbb{E}[G_{it} \mid W_i, \boldsymbol{X}_i] = \mathbb{E}[G_{it} \mid \boldsymbol{X}_i]$$
$$\Longleftrightarrow \qquad \mathbb{E}[G_{it} \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_j] = \mathbb{E}[G_{it} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_j], \qquad j \in \{1, 2\}, r \in \mathcal{R}. \qquad \text{(G.2)}$$

Our goal is to construct an example where (G.1) holds and (G.2) does not hold. Suppose that for some $a > 0$, for all $i \in [N]$, $t \in \{2, \ldots, T\}$

$$\begin{aligned}
\mathbb{E}[G_{it} \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_1] &= -a, \\
\mathbb{E}[G_{it} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_1] &= -ra, \qquad r \in \mathcal{R}, \\
\mathbb{E}[G_{it} \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_2] &= a, \qquad \text{and} \\
\mathbb{E}[G_{it} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_2] &= ra, \qquad r \in \mathcal{R}.
\end{aligned}$$

By inspection, one can see that (G.2) does not hold. We are done if we show that (G.1) does. Suppose W_i and $\boldsymbol{X}_i$ are independent: let $\mathbb{P}(W_i = w, \boldsymbol{X}_i = \boldsymbol{x}_j) = 1/(2[R+1]) = 1/(2T)$ for $w \in \{0\} \cup \mathcal{R}$ and $j \in \{1, 2\}$. We have

$$\begin{aligned}
&\mathbb{E}[G_{it} \mid W_i = 0] \\
&= \frac{1}{\mathbb{P}(W_i = 0)} \big(\mathbb{E}[G_{it} \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_1] \mathbb{P}(W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_1) \\
&\qquad + \mathbb{E}[G_{it} \mid W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_2] \mathbb{P}(W_i = 0, \boldsymbol{X}_i = \boldsymbol{x}_2) \big) \\
&= T \left(-a \cdot \frac{1}{2T} + a \cdot \frac{1}{2T} \right) \\
&= 0
\end{aligned}$$

and for any $r \in \mathcal{R}$,

$$\begin{aligned}
&\mathbb{E}[G_{it} \mid W_i = r] \\
&= \frac{1}{\mathbb{P}(W_i = r)} \big(\mathbb{E}[G_{it} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_1] \mathbb{P}(W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_1) \\
&\qquad + \mathbb{E}[G_{it} \mid W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_2] \mathbb{P}(W_i = r, \boldsymbol{X}_i = \boldsymbol{x}_2) \big) \\
&= T \left(-ra \cdot \frac{1}{2T} + ra \cdot \frac{1}{2T} \right) \\
&= 0.
\end{aligned}$$

$\square$

Proof of Theorem A.5(b) and (c). For any $r \in \mathcal{R}$ and $t \in \{r, \dots, T\}$,

$$\tau_{rt}^*$$

$$= \mathbb{E}[\tau_{rt}^* + (\boldsymbol{X}_i - \mathbb{E}\,[\boldsymbol{X}_i \mid W_i = r])^\top \boldsymbol{\rho}_{rt}^* \mid W_i = r]$$

$$\overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(i1)}(r) \mid W_i = r, \boldsymbol{X}_i\right] \mid W_i = r\right] - \mathbb{E}\left[\mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i\right] \mid W_i = r\right]$$

$$\overset{(b)}{=} \mathbb{E}\left[\mathbb{E}[\tilde{y}_{(it)}(r) \mid W_i = r, \boldsymbol{X}_i] \mid W_i = r\right] - \mathbb{E}\left[\left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i]\right.\right.$$

$$\left.\left. + \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i]\right) \mid W_i = r\right] \qquad\qquad (G.3)$$

$$\overset{(c)}{=} \mathbb{E}\left[\mathbb{E}[\tilde{y}_{(it)}(r) \mid W_i = r, \boldsymbol{X}_i] \mid W_i = r\right] - \mathbb{E}\left[\left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i]\right.\right.$$

$$\left.\left. + \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0]\right) \mid W_i = r\right]$$

$$= \mathbb{E}\left[\tilde{y}_{(it)}(r) \mid W_i = r\right] - \left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r] + \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0\right]\right)$$

$$\overset{(d)}{=} \mathbb{E}\left[\tilde{y}_{(it)}(r) \mid W_i = r\right] - \mathbb{E}\left[\tilde{y}_{(it)}(0) \mid W_i = r\right]$$

$$= \tau_{\mathrm{ATT}}(r,t),$$

where in (a) we used (3.6) from (LINS), in (b) we used (CNAS), in (c) we used (CIUN), and in (d) we used (CTSA). This proves part (b).

Next we prove part (c). Observe that under (3.5) and (3.6) from (LINS), for any $r \in \mathcal{R}$ and $t \in \{r, \dots, T\}$, starting from (G.3) we have

$$\tau_{rt}^*$$

$$= \mathbb{E}\left[\mathbb{E}[\tilde{y}_{(it)}(r) \mid W_i = r, \boldsymbol{X}_i] \mid W_i = r\right] - \mathbb{E}\left[\left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i]\right.\right.$$

$$\left.\left. + \mathbb{E}[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i]\right) \mid W_i = r\right]$$

$$\overset{(e)}{=} \mathbb{E}\left[\mathbb{E}[\tilde{y}_{(it)}(r) \mid W_i = r, \boldsymbol{X}_i] \mid W_i = r\right] - \mathbb{E}\left[\left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r, \boldsymbol{X}_i]\right.\right.$$

$$\left.\left. + \gamma_t^* + \boldsymbol{X}_i^\top \boldsymbol{\xi}_t^*\right) \mid W_i = r\right]$$

$$= \mathbb{E}\left[\tilde{y}_{(it)}(r) \mid W_i = r\right] - \left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r] + \gamma_t^* + \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]^\top \boldsymbol{\xi}_t^*\right)$$

$$\overset{(f)}{=} \mathbb{E}\left[\tilde{y}_{(it)}(r) \mid W_i = r\right] - \left(\mathbb{E}[\tilde{y}_{(i1)}(0) \mid W_i = r]\right.$$

$$\left. + \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]\right]\right)$$

$$= \mathbb{E}\left[\tilde{y}_{(it)}(r) - \tilde{y}_{(i1)}(0) \mid W_i = r\right] - \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0\right]$$

$$\qquad + \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0\right] - \mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i = \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]\right]$$

$$\overset{(g)}{=} \tau_{\mathrm{ATT}}(r,t) + \mathbb{E}\left[\mathbb{E}\left[\tilde{y}_{(it)}(0) - \tilde{y}_{(i1)}(0) \mid W_i = 0, \boldsymbol{X}_i\right] \mid W_i = 0\right] - \left(\gamma_t^* + \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]^\top \boldsymbol{\xi}_t^*\right)$$

$$\overset{(h)}{=} \tau_{\mathrm{ATT}}(r,t) + \mathbb{E}\left[\gamma_t^* + \boldsymbol{X}_i^\top \boldsymbol{\xi}_t^* \mid W_i = 0\right] - \left(\gamma_t^* + \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]^\top \boldsymbol{\xi}_t^*\right)$$

$$= \tau_{\mathrm{ATT}}(r,t) + \gamma_t^* + \mathbb{E}\left[\boldsymbol{X}_i \mid W_i = 0\right]^\top \boldsymbol{\xi}_t^* - \left(\gamma_t^* + \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]^\top \boldsymbol{\xi}_t^*\right)$$

$$= \tau_{\mathrm{ATT}}(r,t) + \left(\mathbb{E}\left[\boldsymbol{X}_i \mid W_i = 0\right] - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]\right)^\top \boldsymbol{\xi}_t^*,$$

where in (e), (f), and (h) we used (3.5) from (LINS) and (g) uses an equality we found under (CTSA) in the proof of part (b) above as well as (3.5) from (LINS).

$$\square$$

G.2 Proofs of Supporting Results for Main Theorems

Proof of Lemma E.1. First, $\mathfrak{W} \leq (T-1)^2$ because treatment effects are interactions between the $R \leq T-1$ cohort indicators and (at most) $T-1$ treatment times.

Next, for any $t \in \mathbb{N}$, define

$$
\boldsymbol{D}^{(1)}(t) := \begin{pmatrix}
1 & -1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & -1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & -1 \\
0 & 0 & 0 & 0 & \cdots & 0 & 1
\end{pmatrix} \in \mathbb{R}^{t \times t}. \tag{G.4}
$$

Note that this is the linear transformation we apply to several groups of coefficients in $\boldsymbol{\beta}$. For example, for the vector of cohort fixed effects if $\boldsymbol{\nu} \in \mathbb{R}^R$,

$$
\boldsymbol{D}^{(1)}(R)\boldsymbol{\nu} = \begin{pmatrix}
\nu_2 - \nu_1 \\
\vdots \\
\nu_R - \nu_{R-1} \\
\nu_R
\end{pmatrix},
$$

which are the transformed coefficients to which we apply an ℓ_q penalty. We similarly penalize $\boldsymbol{\gamma}$ as well as $\boldsymbol{\zeta}_j$, the coefficients for the interactions of the cohort fixed effects and covariate j (separately for each of the d covariates) and $\boldsymbol{\xi}_j$, the coefficients for the interactions of the time fixed effects and covariate j. So if we rearrange the columns of $\tilde{\boldsymbol{Z}}$ corresponding to the cohort-covariate interactions so that all of the cohort interactions with the base effect of feature $j = 1$ come first, then the cohort interactions with feature 2, and so on, we can write this $Rd_N \times Rd_N$ block diagonal part of $\boldsymbol{D}_N$ as $\boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(1)}(R)$, where $\otimes$ denotes the Kronecker product. We can likewise rearrange the columns of $\tilde{\boldsymbol{Z}}$ and write the part of $\boldsymbol{D}_N$ corresponding to the time-covariate interactions as $\boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(1)}(T-1)$.

We penalize the coefficients $\boldsymbol{\kappa}$ directly, so this part of $\boldsymbol{D}$ can be represented simply as $\boldsymbol{I}_{d_N}$.

Next we consider the matrix $\boldsymbol{D}^{(2)}(\mathcal{R})$ we use to transform the base treatment effects $\boldsymbol{\tau}$ for ℓ_1 penalization. The matrix $\boldsymbol{D}^{(2)}(\mathcal{R})$ has a similar structure to $\boldsymbol{D}^{(1)}(t)$, except that we directly penalize each cohort's first treatment effect, not the last one, and we penalize each cohort's first treatment effect towards the previous cohort's first treatment effect. So denoting $\mathcal{R} = (r_1, r_2, \ldots, r_R)$, $\boldsymbol{D}^{(2)}(\mathcal{R})$ has a nearly block diagonal structure

$$
\boldsymbol{D}^{(2)}(\mathcal{R}) := \begin{pmatrix}
\tilde{\boldsymbol{D}}^{(1)}(1)^\top & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{U}(r_1) & \tilde{\boldsymbol{D}}^{(1)}(2)^\top & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{U}(r_2) & \tilde{\boldsymbol{D}}^{(1)}(3)^\top & \cdots & \boldsymbol{0} & \boldsymbol{0} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{U}(r_R) & \tilde{\boldsymbol{D}}^{(1)}(R)^\top
\end{pmatrix} \in \mathbb{R}^{\mathfrak{W} \times \mathfrak{W}}, \tag{G.5}
$$

where $\tilde{\boldsymbol{D}}(x) := \boldsymbol{D}^{(1)}(T - r_x + 1)$ and

$$
\boldsymbol{U}(w) := \begin{pmatrix}
-1 & \boldsymbol{0}_{1 \times (T-w)} \\
\boldsymbol{0}_{(T-w) \times 1} & \boldsymbol{0}_{(T-w) \times (T-w)}
\end{pmatrix} \in \mathbb{R}^{(T-w+1) \times (T-w+1)}.
$$

Then $\boldsymbol{D}^{(2)}(\mathcal{R})\boldsymbol{\tau}$ yields the transformed treatment effects to which we apply an ℓ_q penalty, and similarly, for every $j \in [d_N]$, $\boldsymbol{D}^{(2)}(\mathcal{R})\boldsymbol{\rho}_j$ yields the transformed coefficients of the interactions between the jth centered covariate and the treatment effects. Putting this all together, after reordering the columns appropriately we can write $\boldsymbol{D}_N$ as (E.1).

Next we show that these matrices are invertible by explicitly presenting their inverses. One can verify by matrix multiplication that

$$\boldsymbol{D}^{(1)}(t)^{-1} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{t \times t}, \tag{G.6}$$

Similarly,

$$(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1} = \begin{pmatrix} \tilde{\boldsymbol{D}}^{-1}(1)^\top & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \tilde{\boldsymbol{U}}(r_1, r_2) & \tilde{\boldsymbol{D}}^{-1}(2)^\top & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{\boldsymbol{U}}(r_1, r_R) & \tilde{\boldsymbol{U}}(r_2, r_R) & \cdots & \tilde{\boldsymbol{U}}(r_{R-1}, r_R) & \tilde{\boldsymbol{D}}^{-1}(R)^\top \end{pmatrix}, \tag{G.7}$$

where

$$\tilde{\boldsymbol{U}}(w_1, w_2) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{(T-w_2+1) \times (T-w_1+1)}.$$

Finally, using the properties of block diagonal matrices and the fact that $\boldsymbol{D}^{(1)}(t)$ and $\boldsymbol{D}^{(2)}(\mathcal{R})$ are invertible, we have from (E.1) that $\boldsymbol{D}_N^{-1}$ is

$$\text{diag}\left(\boldsymbol{D}^{(1)}(R)^{-1}, ((\boldsymbol{D}^{(1)}(T-1))^{-1}, \boldsymbol{I}_{d_N}, \boldsymbol{I}_{d_N} \otimes \boldsymbol{D}^{(1)}(R)^{-1}, \boldsymbol{I}_{d_N} \otimes ((\boldsymbol{D}^{(1)}(T-1))^{-1}, \right.$$
$$\left. (\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}, \boldsymbol{I}_{d_N} \otimes (\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1} \right),$$

matching the expression from (E.2).

Lastly, we show the claimed block diagonal structure of $\boldsymbol{D}_N$ and $\boldsymbol{D}_N^{-1}$. Examining the expressions for $\boldsymbol{D}_N$ and $\boldsymbol{D}_N^{-1}$, it is clear that they are both block diagonal with blocks of size $R \leq T-1$, $T-1$, and $\mathfrak{W} \leq (T-1)^2$. We have shown in this proof that all of the blocks contain entries equal to 0, -1, or 1.

$\square$

Proof of Lemma E.2. By Lemma E.1, $\boldsymbol{D}_N^{-1}$ has a block diagonal structure with blocks of size at most $(T-1)^2 \times (T-1)^2$ containing entries that equal either 0 or 1, so $\|\boldsymbol{D}_N^{-1}\|_\infty \leq (T-1)^2$, where for a matrix $\|\cdot\|_\infty$ denotes the maximum absolute row sum. Therefore

$$\|\boldsymbol{\beta}_N^*\|_\infty = \|\boldsymbol{D}_N^{-1}\boldsymbol{\theta}_N^*\|_\infty \leq \|\boldsymbol{D}_N^{-1}\|_\infty \|\boldsymbol{\theta}_N^*\|_\infty \leq (T-1)^2 b_1,$$

where in the last step we used Assumption (R3). $\square$

Proof of Lemma E.3. To prove the result we will apply Theorem 5.1.7 in Lehmann (1999), which we reproduce here:

Theorem G.1 (Theorem 5.1.7 from Lehmann (1999)). *Let $\boldsymbol{a} = (a_1, \ldots, a_k)$ be a vector of constants, and let $\{\boldsymbol{U}^{(N)}\}$ and $\{\boldsymbol{A}^{(N)}\}$ be sequences of random variables, with $\boldsymbol{U}^{(N)} = (U_1^{(N)}, \ldots, U_k^{(N)})$ and $\boldsymbol{A}^{(N)} = (A_1^{(N)}, \ldots, A_k^{(N)})$. Suppose the sequence of random variables $T_N = \sum_{j=1}^k a_j U_j^{(N)}$ converges in distribution to a random variable T, the random variables $A_j^{(N)} \xrightarrow{p} a_j$ for each $j \in \{1, \ldots, k\}$, and the random variables $U_j^{(N)} = \mathcal{O}_{\mathbb{P}}(1)$. Then the sequence of random variables $\sum_{j=1}^k A_j^{(n)} U_j^{(N)}$ converges in distribution to T.*

Given the joint probability distribution $\{(W_i, \boldsymbol{X}_i)\}_{i=1}^N$, consider the iid random variables Z_i defined by

$$Z_i := \sum_{r \in \{0\} \cup \mathcal{R}} \frac{\mathbb{1}\{W_i = r\}}{\mathbb{P}(W_i = r)} \boldsymbol{\psi}_r^\top \left(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r]\right)$$

$$= \sum_{r \in \{0\} \cup \mathcal{R}} \frac{\mathbb{1}\{W_i = r\}}{\mathbb{P}(W_i = r)} \boldsymbol{\psi}_r^\top \left(\boldsymbol{X}_i - \boldsymbol{\mu}_r\right).$$

We have

$$\mathbb{E}[Z_i] = \mathbb{E}[\mathbb{E}[Z_i \mid W_i]] = \mathbb{E}\left[\frac{1}{\mathbb{P}(W_i)} \boldsymbol{\psi}_{W_i}^\top \left(\mathbb{E}[\boldsymbol{X}_i \mid W_i] - \mathbb{E}[\boldsymbol{X}_i \mid W_i]\right)\right] = \mathbb{E}[0] = 0$$

and

$$\mathrm{Var}(Z_i) = \mathbb{E}[Z_i^2] = \mathbb{E}\left[\mathbb{E}\left[Z_i^2 \mid W_i\right]\right]$$

$$= \mathbb{E}\left[\frac{1}{(\mathbb{P}(W_i))^2} \boldsymbol{\psi}_{W_i}^\top \mathrm{Cov}(\boldsymbol{X}_i \mid W_i) \boldsymbol{\psi}_{W_i}\right]$$

$$= \sum_{r \in \{0\} \cup \mathcal{R}} \frac{1}{\mathbb{P}(W_i = r)^2} \boldsymbol{\psi}_r^\top \mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r) \boldsymbol{\psi}_r \cdot \mathbb{P}(W_i = r)$$

$$= v_R.$$

We can lower-bound v_R uniformly across choices of $\boldsymbol{\psi}_r$: since $\boldsymbol{\psi}_r \neq \boldsymbol{0}$ for at least one $r = r^*$ and by assumption the smallest eigenvalue of $\mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r^*)$ is bounded below by some $\lambda_{\min} > 0$, we have

$$v_R \geq \frac{\boldsymbol{\psi}_{r^*}^\top \mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r^*) \boldsymbol{\psi}_{r^*}}{\mathbb{P}(W_i = r^*)} \geq \frac{\lambda_{\min} \|\boldsymbol{\psi}_{r^*}\|_2^2}{1} > 0.$$

Similarly, since the largest eigenvalue of $\mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r) \leq \lambda_{\max}$ for some finite $\lambda_{\max}$ for all $r \in \{0\} \cup \mathcal{R}$ and $\min_{r \in \{0\} \cup \mathcal{R}}\{\mathbb{P}(W_i = r)\} > 0$ by Assumption (R1), we can upper-bound v_R by a constant. So by the central limit theorem, for any choice of finite $\{\boldsymbol{\psi}_r\}$ with at least one $\boldsymbol{\psi}_{r^*} \neq \boldsymbol{0}$,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N Z_i \xrightarrow{d} \mathcal{N}(0, v_R)$$

$$\iff \frac{1}{\sqrt{N}} \sum_{r \in \{0\} \cup \mathcal{R}} \sum_{\{i: W_i = r\}} \frac{1}{\mathbb{P}(W_i = r)} \boldsymbol{\psi}_r^\top \left(\boldsymbol{X}_i - \boldsymbol{\mu}_r\right) \xrightarrow{d} \mathcal{N}(0, v_R)$$

$$\iff \frac{1}{\sqrt{N}} \sum_{r \in \{0\} \cup \mathcal{R}} \frac{N_r}{\mathbb{P}(W_i = r)} \boldsymbol{\psi}_r^\top \left(\frac{1}{N_r} \sum_{\{i: W_i = r\}} \boldsymbol{X}_i - \boldsymbol{\mu}_r\right) \xrightarrow{d} \mathcal{N}(0, v_R)$$

$$\iff \sum_{r \in \{0\} \cup \mathcal{R}} \sqrt{N} \frac{N_r/N}{\mathbb{P}(W_i = r)} \boldsymbol{\psi}_r^\top \left(\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r\right) \xrightarrow{d} \mathcal{N}(0, v_R). \qquad \text{(G.8)}$$

Now consider the random variables

$$\boldsymbol{A}^{(N)} := \left(\pi_0 \cdot \frac{N}{\sum_{i=1}^N \mathbb{1}\{W_i = 0\}}, \pi_{r_1} \cdot \frac{N}{\sum_{i=1}^N \mathbb{1}\{W_i = r_1\}}, \ldots, \pi_{r_R} \cdot \frac{N}{\sum_{i=1}^N \mathbb{1}\{W_i = r_R\}}\right)$$

$$= \left(\pi_0 \cdot \frac{N}{N_0}, \pi_{r_1} \cdot \frac{N}{N_{r_1}}, \ldots, \pi_{r_R} \cdot \frac{N}{N_{r_R}}\right).$$

By the continuous mapping theorem $\boldsymbol{A}^{(N)} \xrightarrow{p} (1, 1, \ldots, 1)$. Further, let

$$\boldsymbol{U}^{(N)} := \left(\sum_{i=1}^{N} \frac{\mathbb{1}\{W_i = 0\}}{\pi_0} \boldsymbol{\psi}_0^\top \left(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i \mid W_i = 0] \right), \sum_{i=1}^{N} \frac{\mathbb{1}\{W_i = r_1\}}{\pi_{r_1}} \boldsymbol{\psi}_{r_1}^\top \left(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r_1] \right), \right.$$
$$\left. \cdots, \sum_{i=1}^{N} \frac{\mathbb{1}\{W_i = r_R\}}{\pi_{r_R}} \boldsymbol{\psi}_{r_R}^\top \left(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r_R] \right) \right).$$

Since from (G.8) we have

$$\sum_{r \in \{0\} \cup \mathcal{R}} 1 \cdot \frac{1}{\sqrt{N}} U_r^{(N)} \xrightarrow{d} \mathcal{N}(0, v_R),$$

we have from Theorem G.1

$$\sum_{r \in \{0\} \cup \mathcal{R}} A_r^{(N)} \cdot \frac{1}{\sqrt{N}} U_r^{(N)} \xrightarrow{d} \mathcal{N}(0, v_R)$$

$$\iff \qquad \sum_{r \in \{0\} \cup \mathcal{R}} \sqrt{\frac{N}{v_R}} \boldsymbol{\psi}_r^\top \left(\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r \right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Finally we prove the last claim. It follows from the above that

$$N \| \overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r \|_2^2$$

converges in distribution to a generalized chi-squared random variable, so by Theorem 2.4(i) in van der Vaart (2000) and Proposition 1.8(iii) in García-Portugués (2023) we have

$$N \| \overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r \|_2^2 = \mathcal{O}_\mathbb{P}(1)$$

$$\implies \qquad \| \overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r \|_2^2 = \mathcal{O}_\mathbb{P}\left(\frac{1}{N} \right)$$

$$\implies \qquad \| \overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r \|_2 = \mathcal{O}_\mathbb{P}\left(\frac{1}{\sqrt{N}} \right).$$

$\square$

Proof of Lemma E.4. For any $r \in \mathcal{R}$,

$$\left| \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} - \frac{\boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x})}{\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x})} \right|$$

$$= \left| \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \cdot \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) - \boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x}) \boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x})} \right|$$

$$= \left| \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \cdot \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) - \boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x}) \cdot \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) + \boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x}) \cdot \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) - \boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x}) \boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x})} \right|$$

$$= \left| \frac{\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) \cdot \boldsymbol{a}^\top \left(\hat{\boldsymbol{\pi}}(\boldsymbol{x}) - \boldsymbol{\pi}(\boldsymbol{x}) \right) + \boldsymbol{a}^\top \boldsymbol{\pi}(\boldsymbol{x}) \cdot \boldsymbol{b}^\top \left(\boldsymbol{\pi}(\boldsymbol{x}) - \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \right)}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x})} \right|$$

$$= \frac{1}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} \mathcal{O}_\mathbb{P}(a_N),$$

where in the last step we used our consistency assumption and the finiteness of $\boldsymbol{a}$ and $\boldsymbol{b}$. We are done if we can show that

$$\frac{1}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} = \mathcal{O}_\mathbb{P}(1).$$

That is, we want to show that for any $\epsilon > 0$ there exists a finite $M_\epsilon > 0$ and $N_\epsilon > 0$ such that for all $N > N_\epsilon$,

$$\mathbb{P}\left(\frac{1}{\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x})} > M_\epsilon\right) < \epsilon \qquad \Longleftrightarrow \qquad \mathbb{P}\left(\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) < \frac{1}{M_\epsilon}\right) < \epsilon.$$

Fix $\epsilon > 0$. By assumption, for any $\boldsymbol{x}$ in the support of $\boldsymbol{X}_i$ we have $|\boldsymbol{b}^\top(\hat{\boldsymbol{\pi}}(\boldsymbol{x}) - \boldsymbol{\pi}(\boldsymbol{x}))| = \mathcal{O}_\mathbb{P}(a_N)$, so there exists a finite $M_\epsilon^{(1)} > 0$ and $N_\epsilon^{(1)} > 0$ such that for all $N > N_\epsilon^{(1)}$,

$$\mathbb{P}\left(\frac{|\boldsymbol{b}^\top(\hat{\boldsymbol{\pi}}(\boldsymbol{x}) - \boldsymbol{\pi}(\boldsymbol{x}))|}{a_N} > M_\epsilon^{(1)}\right) < \epsilon \qquad \Longrightarrow \qquad \mathbb{P}\left(\boldsymbol{b}^\top \hat{\boldsymbol{\pi}}(\boldsymbol{x}) < \boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) - M_\epsilon^{(1)} a_N\right) < \epsilon.$$

Since by assumption $\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) > 0$, the result holds for $N_\epsilon \geq N_\epsilon^{(1)}$ large enough so that for all $N > N_\epsilon$ it holds that $\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) - M_\epsilon^{(1)} a_N > 0$ (which is assured since a_N is decreasing), and $M_\epsilon := 1/(\boldsymbol{b}^\top \boldsymbol{\pi}(\boldsymbol{x}) - M_\epsilon^{(1)} a_{N_\epsilon})$. $\qquad\square$

Proof of Lemma E.5. Observe that

$$\hat{\boldsymbol{\psi}}_N^{(\text{C;CATT},\pi)} - \boldsymbol{\psi}_N^{(\text{C;CATT},\pi)}$$

$$= \sum_{r\in\mathcal{R}} \sum_{t=r}^T f_r\left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right) \psi_{rt} \hat{\boldsymbol{\psi}}_{rt}^* - \sum_{r\in\mathcal{R}} f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \sum_{t=r}^T \psi_{rt} \boldsymbol{\psi}_{rt}^*$$

$$= \sum_{r\in\mathcal{R}} \sum_{t=r}^T f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \psi_{rt} \left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^*\right)$$

$$+ \sum_{r\in\mathcal{R}} \sum_{t=r}^T \left(f_r\left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right) - f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right)\right) \psi_{rt} \boldsymbol{\psi}_{rt}^*$$

$$+ \sum_{r\in\mathcal{R}} \sum_{t=r}^T \left(f_r\left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right) - f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right)\right) \psi_{rt} \left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^*\right),$$

where $\hat{\boldsymbol{\psi}}_{rt}^*$ was defined in (E.15) and $\hat{\boldsymbol{\psi}}_{rt}^*$ was defined in (E.14). We will show that after scaling by $\sqrt{NT}$ and multiplying by $(\boldsymbol{\theta}_N^*)^\top$, terms of the first two forms converge in distribution to independent Gaussian random variables, and terms of the third form converge in probability to 0.

- Applying Lemma E.3, we have

$$\sqrt{NT}(\boldsymbol{\theta}_N^*)^\top \sum_{r\in\mathcal{R}} f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \sum_{t=r}^T \psi_{rt}\left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^*\right)$$

$$= \sqrt{NT} \sum_{r\in\mathcal{R}} f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \sum_{t=r}^T -\psi_{rt}(\boldsymbol{\theta}_N^*)^\top \left((\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot}^\top \left(\overline{\boldsymbol{X}}_r - \boldsymbol{\mu}_r\right)\right)$$

$$\xrightarrow{d} \mathcal{N}(0, v_R^{\text{CATT; prob}}),$$

where

$$v_R^{\text{CATT; prob}} := T \cdot \sum_{r\in\mathcal{R}} \frac{(\boldsymbol{\psi}_r^{\text{CATT; prob}})^\top \text{Cov}(\boldsymbol{X}_i \mid W_i = r) \boldsymbol{\psi}_r^{\text{CATT; prob}}}{\mathbb{P}(W_i = r)}$$

for

$$\boldsymbol{\psi}_r^{\text{CATT; prob}} := f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \sum_{t=r}^T \psi_{rt}(\boldsymbol{D}_N^{-1})_{i(r,t,\boldsymbol{x}),\cdot} \boldsymbol{\theta}_N^*$$

$$= f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right) \sum_{t=r}^T \psi_{rt} \boldsymbol{\rho}_{rt}^*, \qquad r \in \mathcal{R}. \tag{G.9}$$

- Next, using the assumed joint asymptotic normality of $(\hat{\pi}_0(\boldsymbol{x}), \hat{\pi}_{r_1}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}))$ and the delta method (Theorem 3.1 in van der Vaart 2000), we have for each $r \in \mathcal{R}$

$$\sqrt{NT} \big(f_{r_1} \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right) - f_{r_1} \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right), \ldots,$$
$$f_{r_R} \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right) - f_{r_R} \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \big)$$
$$\xrightarrow{d} \mathcal{N} \left(\boldsymbol{0}, T \cdot \nabla \boldsymbol{f}^\top \boldsymbol{\Sigma}_\pi \nabla \boldsymbol{f} \right)$$

Therefore

$$\sqrt{NT} (\boldsymbol{\theta}_N^*)^\top \sum_{r \in \mathcal{R}} \left(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right) - f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \right) \sum_{t=r}^{T} \psi_{rt} \boldsymbol{\psi}_{rt}^*$$

$$= \sqrt{NT} \sum_{r \in \mathcal{R}} \left(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right) - f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \right) \sum_{t=r}^{T} \psi_{rt} (\boldsymbol{\theta}_N^*)^\top \boldsymbol{\psi}_{rt}^*$$

$$\xrightarrow{d} \mathcal{N} \left(0, T \cdot \boldsymbol{\psi}_\pi^\top \nabla \boldsymbol{f}^\top \boldsymbol{\Sigma}_\pi \nabla \boldsymbol{f} \boldsymbol{\psi}_\pi \right)$$

for

$$\boldsymbol{\psi}_\pi := \left(\sum_{t=r_1}^{T} \psi_{r_1 t} (\boldsymbol{\theta}_N^*)^\top \boldsymbol{\psi}_{r_1 t}^*, \sum_{t=r_2}^{T} \psi_{r_2 t} (\boldsymbol{\theta}_N^*)^\top \boldsymbol{\psi}_{r_2 t}^*, \ldots, \sum_{t=r_R}^{T} \psi_{r_R t} (\boldsymbol{\theta}_N^*)^\top \boldsymbol{\psi}_{r_R t}^* \right)$$

$$= \left(\sum_{t=r_1}^{T} \psi_{r_1 t} \left(\boldsymbol{\tau}_{r_1 t}^* + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r_1])^\top \boldsymbol{\rho}_{r_1 t}^* \right), \ldots, \right.$$

$$\left. \sum_{t=r_R}^{T} \psi_{r_R t} \left(\boldsymbol{\tau}_{r_R t}^* + (\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}_i \mid W_i = r_R])^\top \boldsymbol{\rho}_{r_R t}^* \right) \right).$$

- Finally, the above two results, along with Theorem 2.4(i) in van der Vaart (2000) and Proposition 1.8(iii) in García-Portugués (2023), yield that for any $r \in \mathcal{R}$

$$\left(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right) - f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \right) (\boldsymbol{\theta}_N^*)^\top \sum_{t=r}^{T} \left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^* \right)$$

$$= \mathcal{O}_\mathbb{P} \left(\frac{1}{\sqrt{N}} \right) \mathcal{O}_\mathbb{P} \left(\frac{1}{\sqrt{N}} \right) = \mathcal{O}_\mathbb{P} \left(\frac{1}{N} \right),$$

so

$$\sqrt{NT} (\boldsymbol{\theta}_N^*)^\top \sum_{r \in \mathcal{R}} \big(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right)$$

$$- f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \big) \sum_{t=r}^{T} \psi_{rt} \left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^* \right)$$

$$= \sqrt{NT} \sum_{r \in \mathcal{R}} \sum_{t=r}^{T} \psi_{rt} \big(f_r \left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x}) \right)$$

$$- f_r \left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x}) \right) \big) (\boldsymbol{\theta}_N^*)^\top \left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^* \right)$$

$$= \sqrt{N} \mathcal{O}_\mathbb{P} \left(\frac{1}{N} \right),$$

and

$$\sqrt{NT}(\boldsymbol{\theta}_N^*)^\top \sum_{r\in\mathcal{R}} \big(f_r\left(\hat{\pi}_{r_1}(\boldsymbol{x}), \hat{\pi}_{r_2}(\boldsymbol{x}), \ldots, \hat{\pi}_{r_R}(\boldsymbol{x})\right)$$

$$- f_r\left(\pi_{r_1}(\boldsymbol{x}), \pi_{r_2}(\boldsymbol{x}), \ldots, \pi_{r_R}(\boldsymbol{x})\right)\big) \sum_{t=r}^{T} \psi_{rt}\left(\hat{\boldsymbol{\psi}}_{rt}^* - \boldsymbol{\psi}_{rt}^*\right) \xrightarrow{p} 0.$$

Putting this together and using the assumed independence, we have

$$\sqrt{NT}(\boldsymbol{\theta}_N^*)^\top \left(\hat{\boldsymbol{\psi}}_N^{(\mathrm{C;CATT},\pi)} - \boldsymbol{\psi}_N^{(\mathrm{C;CATT},\pi)}\right) \xrightarrow{d} \mathcal{N}\left(0, v_R^{(\mathrm{C;CATT},\pi)}\right)$$

for

$$v_R^{(\mathrm{C;CATT},\pi)} := T \cdot \sum_{r\in\mathcal{R}} \frac{(\boldsymbol{\psi}_r^{\mathrm{CATT;\ prob}})^\top \mathrm{Cov}(\boldsymbol{X}_i \mid W_i = r)\boldsymbol{\psi}_r^{\mathrm{CATT;\ prob}}}{\mathbb{P}(W_i = r)} + T \cdot \boldsymbol{\psi}_\pi^\top \nabla \boldsymbol{f}^\top \boldsymbol{\Sigma}_\pi \nabla \boldsymbol{f} \boldsymbol{\psi}_\pi.$$

Finally, consider the plug-in estimator

$$\hat{v}_R^{(\mathrm{C;CATT},\pi)} := T \cdot \sum_{r\in\mathcal{R}} \frac{(\hat{\boldsymbol{\psi}}_r^{\mathrm{CATT;\ prob}})^\top \widehat{\mathrm{Cov}}(\boldsymbol{X}_i \mid W_i = r)\hat{\boldsymbol{\psi}}_r^{\mathrm{CATT;\ prob}}}{N_r/N_\tau} + T \cdot \hat{\boldsymbol{\psi}}_\pi^\top \nabla \hat{\boldsymbol{f}}^\top \hat{\boldsymbol{\Sigma}}_\pi \nabla \hat{\boldsymbol{f}} \hat{\boldsymbol{\psi}}_\pi, \text{ (G.10)}$$

where

$$\hat{\boldsymbol{\psi}}_r^{\mathrm{CATT;\ prob}} := \frac{\hat{\pi}_r(\boldsymbol{x})}{\sum_{r'\in\mathcal{R}} \hat{\pi}_{r'}(\boldsymbol{x})} \sum_{t=r}^{T} \psi_{rt}\hat{\boldsymbol{\rho}}_{rt}^{(q)}, \qquad r\in\mathcal{R},$$

$$\hat{\boldsymbol{\psi}}_\pi := \left(\sum_{t=r_1}^{T} \psi_{r_1 t}\left(\hat{\boldsymbol{\tau}}_{r_1 t}^{(q)} + \left(\boldsymbol{x} - \overline{\boldsymbol{X}}_{r_1}\right)^\top \hat{\boldsymbol{\rho}}_{r_1 t}^{(q)}\right), \ldots,\right.$$

$$\left.\sum_{t=r_R}^{T} \psi_{r_R t}\left(\hat{\boldsymbol{\tau}}_{r_R t}^{(q)} + \left(\boldsymbol{x} - \overline{\boldsymbol{X}}_{r_R}\right)^\top \hat{\boldsymbol{\rho}}_{r_R t}^{(q)}\right)\right),$$

and $\nabla \hat{\boldsymbol{f}}$ is defined by replacing each instance of $\pi_r(\boldsymbol{x})$ in $\nabla \boldsymbol{f}$ with its estimator $\hat{\pi}_r(\boldsymbol{x})$. Notice that $\hat{v}_R^{(\mathrm{C;CATT},\pi)} \xrightarrow{p} v_R^{(\mathrm{C;CATT},\pi)}$ by the consistency of the individual terms and the continuous mapping theorem. $\square$

H Proofs of Lemmas From Proof of Theorem C.1

Proof of Lemma F.1. Notice from (E.1) and the expressions in the proof of Lemma E.1 that the ℓ_1 norm of each row or column of $\boldsymbol{D}_N$ is at most 3, so we have

$$\sigma_{\max}(\boldsymbol{D}_N) \leq \sqrt{\|\boldsymbol{D}_N\|_1 \|\boldsymbol{D}_N\|_\infty} \leq \sqrt{3 \cdot 3} = 3,$$

where for matrices $\|\cdot\|_1$ denotes the maximum absolute column sum of a matrix and $\|\cdot\|_\infty$ denotes the maximum absolute row sum. Now we lower bound $\sigma_{\min}(\boldsymbol{D}_N)$. We have that for any matrix $\boldsymbol{A}$

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{1}{\sqrt{\|\boldsymbol{A}^{-1}\|_1 \|\boldsymbol{A}^{-1}\|_\infty}}. \tag{H.1}$$

Recall our expression for $\boldsymbol{D}_N^{-1}$ from (E.2). The singular values of a block diagonal matrix are the union of the singular values of the individual matrices on the diagonal, so it is enough to lower bound the singular values of each of these individual matrices using (H.1) and then take the minimum of all of the lower bounds.

Using the expression for $\boldsymbol{D}^{(1)}(t)^{-1}$ from (G.6) and (H.1),

$$\|\boldsymbol{D}^{(1)}(t)^{-1}\|_1 \|\boldsymbol{D}^{(1)}(t)^{-1}\|_\infty \leq t^2 \qquad \implies \qquad \sigma_{\min}\left(\boldsymbol{D}^{(1)}(t)^{-1}\right) \geq 1/t. \tag{H.2}$$

Therefore the maximum column ℓ_1 norm of $(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}$ is $\mathfrak{W}$, the ℓ_1 norm of the first column of $(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}$.

Similarly, using the expression for $(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}$ from (G.7), we see that the maximum row ℓ_1 norm, in the bottom block of $(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}$, is at most $R - 1 + T - r_R + 1 \leq R + 1$, because each of the $R - 1$ blocks of $\tilde{\boldsymbol{U}}$ matrices has a single 1 in each row and $\left(\boldsymbol{D}^{(1)}(T - r_R + 1)^{-1}\right)^\top$ has a row of $T - r_R + 1$ ones. So

$$\|(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}\|_1 \|(\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}\|_\infty \leq \mathfrak{W}(T + R)$$
$$\implies \qquad \sigma_{\min}\left((\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}\right) \geq \frac{1}{\sqrt{\mathfrak{W}(T + R)}}.$$

Using the fact that $\mathfrak{W} \leq (T - 1)^2$ from Lemma E.1, we have

$$\sigma_{\min}\left((\boldsymbol{D}^{(2)}(\mathcal{R}))^{-1}\right) \geq \frac{1}{(T - 1)\sqrt{T + R}} \geq \frac{1}{(T - 1)\sqrt{2T - 1}}. \tag{H.3}$$

Finally, using (H.2) and (H.3), $T \geq 2$, and the fact that the smallest singular value of the identity matrix is 1, we have

$$\sigma_{\min}(\boldsymbol{D}_N) \geq \min\left\{\frac{1}{R}, \frac{1}{T - 1}, 1, \frac{1}{(T - 1)\sqrt{2T - 1}}\right\} \geq \frac{1}{(T - 1)\sqrt{2T - 1}} \geq \frac{1}{T\sqrt{2T}}.$$

$\square$

Proof of Lemma F.2. We will first show that $\boldsymbol{X}_i$ having finite fourth moments under Assumption (R1) is enough for $\tilde{\boldsymbol{Z}}_{(i\cdot)}$ to have finite fourth moments as well. We will show that each entry of $\tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1}$ is a bounded linear combination of a finite number of elements of $\tilde{\boldsymbol{Z}}_{(i\cdot)}$, and likewise for $\boldsymbol{G}_N\tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1}$. Then the finite moments will follow using our Assumption (R1), Cauchy-Schwarz, and Hölder's inequality.

Besides $\boldsymbol{X}_i$, the entries of $\tilde{\boldsymbol{Z}}_i$ consist of indicator variables and interactions of the indicator variables with themselves and $\boldsymbol{X}_i$. All of these terms are upper-bounded by the greater of 1 or $\boldsymbol{X}_i$. Using this, our assumption that $\mathbb{E}[X_{(i)j}^4]$ is finite for all $j \in [p_N]$, and the fact that $\boldsymbol{X}_i$ is time-invariant, we have that for any fixed N, $\max_{j \in [d_N], t \in [T]}\{\mathbb{E}[Z_{(it)j}^4]\}$ is finite.

Now we consider the entries of $\tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1} \in \mathbb{R}^{NT \times p_N}$. The columns of $\tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1}$ are linear combinations of the columns of $\tilde{\boldsymbol{Z}}_{(i\cdot)}$; in particular, the jth column of $\tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1}$ is $\tilde{\boldsymbol{Z}}_{(i\cdot)}(\boldsymbol{D}_N^{-1})_{\cdot j} \in \mathbb{R}^{NT}$, where $(\boldsymbol{D}_N^{-1})_{\cdot j}$ is the jth column of $\boldsymbol{D}_N^{-1}$. From (E.2) and the expressions in the proof of Lemma E.1 we see that $\boldsymbol{D}_N^{-1}$ contains only entries that equal 1 or 0, so for any $t \in [T]$ and $j \in [p_N]$,

$$(\tilde{\boldsymbol{Z}}_{(i\cdot)}\boldsymbol{D}_N^{-1})_{(1t)j}^4 = \left[\left(\tilde{\boldsymbol{Z}}_{(i\cdot)}(\boldsymbol{D}_N^{-1})_{\cdot j}\right)_{(1t)}\right]^4 = \left(\sum_{\{j':(\boldsymbol{D}_N^{-1})_{j'j} \neq 0\}} Z_{(1t)j'}\right)^4.$$

From (E.2) and the expressions in the proof of Lemma E.1 we have that $\boldsymbol{D}_N^{-1}$ has a block diagonal structure with blocks of size at most $\mathfrak{W} \leq (T - 1)^2$ (an upper bound we established in Lemma E.1), so this sum contains at most $(T - 1)^2$ terms before expansion. By the multinomial theorem, the number of terms after expansion is therefore at most

$$\binom{4 + (T - 1)^2 - 1}{(T - 1)^2 - 1}.$$

T is fixed and finite, so the number of terms after expansion is finite as well. Therefore if we can show that each summand after expansion has finite 4th moment, we have shown that each entry of $\tilde{\boldsymbol{Z}}_{(i \cdot)} . \boldsymbol{D}_N^{-1}$ has finite 4th moment.

This sum when expanded contains terms of the form $\tilde{Z}_{(1t)j'}^4$, $\tilde{Z}_{(1t)j'}^2 \tilde{Z}_{(1t)j''}^2$ for $j' \neq j''$, $\tilde{Z}_{(1t)j'}^3 \tilde{Z}_{(1t)j''}$, $\tilde{Z}_{(1t)j'}^2 \tilde{Z}_{(1t)j''} \tilde{Z}_{(1t)j'''}$ and $\tilde{Z}_{(1t)j'} \tilde{Z}_{(1t)j''} \tilde{Z}_{(1t)j'''} \tilde{Z}_{(1t)j''''}$ (disregarding constant factors). We demonstrate that each of these terms has finite fourth moment below using Cauchy-Schwarz and Hölder's inequality:

- We have already shown that $\mathbb{E}[\tilde{Z}_{(1t)j'}^4]$ is finite.

- $\mathbb{E}[\tilde{Z}_{(1t)j'}^2 \tilde{Z}_{(1t)j''}^2] \leq \sqrt{\mathbb{E}[\tilde{Z}_{(1t)j'}^4] \mathbb{E}[\tilde{Z}_{(1t)j''}^4]}$, and terms of this form are finite using the finite fourth moments of $\tilde{\boldsymbol{Z}}_{(i \cdot)}$..

- Recall that by Hölder's inequality, for any $p, q > 1$ such that $1/p + 1/q = 1$, if X and Y are random variables with $(\mathbb{E}|X|^p)^{1/p}$ and $(\mathbb{E}|Y|^q)^{1/q}$ finite then

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}.$$

Choosing $p = 4/3$ and $q = 4$, we have

$$
\begin{aligned}
\mathbb{E}[\tilde{Z}_{(1t)j'}^3 \tilde{Z}_{(1t)j''}] \leq & \ \mathbb{E}[|\tilde{Z}_{(1t)j'}|^3 |\tilde{Z}_{(1t)j''}|] \leq \left(\mathbb{E}\left[(|\tilde{Z}_{(1t)j'}|^3)^{4/3} \right] \right)^{3/4} \left(\mathbb{E}\left[|\tilde{Z}_{(1t)j''}|^4 \right] \right)^{1/4} \\
= & \ \left(\mathbb{E}|\tilde{Z}_{(1t)j'}|^4 \right)^{3/4} \left(\mathbb{E}|\tilde{Z}_{(1t)j''}|^4 \right)^{1/4},
\end{aligned}
$$

which is finite.

- $\mathbb{E}[\tilde{Z}_{(1t)j'}^2 \tilde{Z}_{(1t)j''} \tilde{Z}_{(1t)j'''}] \leq \sqrt{\mathbb{E}[\tilde{Z}_{(1t)j'}^4]} \sqrt{\mathbb{E}[\tilde{Z}_{(1t)j''}^2 \tilde{Z}_{(1t)j'''}^2]}$, and we already showed that terms of this form are finite.

- $\mathbb{E}[\tilde{Z}_{(1t)j'} \tilde{Z}_{(1t)j''} \tilde{Z}_{(1t)j'''} \tilde{Z}_{(1t)j''''}] \leq \sqrt{\mathbb{E}[\tilde{Z}_{(1t)j'}^2 \tilde{Z}_{(1t)j''}^2] \mathbb{E}[\tilde{Z}_{(1t)j'''}^2 \tilde{Z}_{(1t)j''''}^2]}$, and we already showed that terms of this form are finite.

$\qquad\square$

Proof of Lemma F.3. Recall that for any matrix $\boldsymbol{A}$, the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ are the squares of the singular values of $\boldsymbol{A}$. So $\lambda_{\min}\left(\frac{1}{NT} \left(\boldsymbol{Z} \boldsymbol{D}_N^{-1} \right)^\top \boldsymbol{Z} \boldsymbol{D}_N^{-1} \right) = \sigma_{\min}\left(\boldsymbol{Z} \boldsymbol{D}_N^{-1} \right)^2 /(NT)$ and

$$\lambda_{\max}\left(\frac{1}{NT} \left(\boldsymbol{Z} \boldsymbol{D}_N^{-1} \right)^\top \boldsymbol{Z} \boldsymbol{D}_N^{-1} \right) = \sigma_{\max}\left(\boldsymbol{Z} \boldsymbol{D}_N^{-1} \right)^2 /(NT),$$

and it is enough to bound the singular values of $\boldsymbol{Z} \boldsymbol{D}_N^{-1}$, which we can do with Lemma H.1:

Lemma H.1. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be conformable real matrices where $\boldsymbol{B}$ has full column rank.*

(a) (Theorem 2 in Güngör 2007.)

$$\sigma_{min}\left(\boldsymbol{A} \boldsymbol{B} \right) \geq \sigma_{min}\left(\boldsymbol{A} \right) \sigma_{min}\left(\boldsymbol{B} \right).$$

(b) (Theorem 3.3.4 in Horn and Johnson (1991).)

$$\sigma_{max}\left(\boldsymbol{A} \boldsymbol{B} \right) \leq \sigma_{max}\left(\boldsymbol{A} \right) \sigma_{max}\left(\boldsymbol{B} \right).$$

We bounded the singular values of the deterministic $\boldsymbol{D}_N^{-1}$ in Lemma F.1. Using that and the fact that $\frac{1}{NT}\sigma_{\min}(\boldsymbol{Z})^2 = \lambda_{\min}\left(\frac{1}{NT}\boldsymbol{Z}^\top\boldsymbol{Z}\right) = e_{1N}$, we have

$$
\begin{aligned}
\lambda_{\min}\left(\frac{1}{NT}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)^\top \boldsymbol{Z}\boldsymbol{D}_N^{-1}\right) &\geq \frac{1}{NT}\sigma_{\min}(\boldsymbol{Z})^2 \,\sigma_{\min}\left(\boldsymbol{D}_N^{-1}\right)^2 \\
&\geq \lambda_{\min}\left(\frac{1}{NT}\tilde{\boldsymbol{Z}}^\top\tilde{\boldsymbol{Z}}\right)\left(\frac{1}{3}\right)^2 \\
&\geq \frac{1}{9}e_{1N},
\end{aligned}
$$

where we used that from Lemma F.1,

$$
\sigma_{\max}(\boldsymbol{D}_N) \leq 3 \qquad \Longleftrightarrow \qquad \sigma_{\min}(\boldsymbol{D}_N^{-1}) \geq \frac{1}{3}.
$$

Similarly,

$$
\lambda_{\max}\left(\frac{1}{NT}\left(\boldsymbol{Z}\boldsymbol{D}_N^{-1}\right)^\top \boldsymbol{Z}\boldsymbol{D}_N^{-1}\right) \leq \lambda_{\max}\left(\frac{1}{NT}\tilde{\boldsymbol{Z}}^\top\tilde{\boldsymbol{Z}}\right)\left(T\sqrt{2T}\right)^2 \leq e_{2N}\cdot 2T^3,
$$

where we used that from Lemma F.1,

$$
\sigma_{\min}(\boldsymbol{D}_N) \geq \frac{1}{T\sqrt{2T}} \qquad \Longleftrightarrow \qquad \sigma_{\max}(\boldsymbol{D}_N^{-1}) \leq T\sqrt{2T}.
$$

$\square$

Proof of Theorem F.4. For clarity, we use the notation of Kock (2013). To extend Lemma 3 from Kock (2013), we use a similar strategy to the one used to prove Theorem 11.3(d) in Hastie et al. (2015). Since under Assumption (A5) from Kock (2013) we have $\min_{1\leq j\leq k_N}\{|\beta_{10j}|\} \geq b_0$, it is enough to show that with high probability $\hat{\beta}_{2N} = \boldsymbol{0}$ and $\|\hat{\beta}_N - \beta_N\|_\infty \leq b_0/2$. That is,

$$
\begin{aligned}
\left\{\hat{\beta}_{2N} = \boldsymbol{0}\right\} \cap \left\{\left\|\hat{\beta}_N - \beta_N\right\|_\infty \leq \frac{b_0}{2}\right\} &\subseteq \left\{\hat{\mathcal{S}} = \mathcal{S}\right\} \\
\Longrightarrow \qquad \left\{\hat{\mathcal{S}} \neq \mathcal{S}\right\} &\subseteq \left\{\hat{\beta}_{2N} \neq \boldsymbol{0}\right\} \cup \left\{\left\|\hat{\beta}_N - \beta_N\right\|_\infty > \frac{b_0}{2}\right\} \\
\overset{(a)}{\Longrightarrow} \qquad \mathbb{P}\left(\hat{\mathcal{S}} \neq \mathcal{S}\right) &\leq \mathbb{P}\left(\left\{\hat{\beta}_{2N} \neq \boldsymbol{0}\right\} \cup \left\{\left\|\hat{\beta}_N - \beta_N\right\|_2 > \frac{b_0}{2}\right\}\right) \\
&\leq \mathbb{P}\left(\hat{\beta}_{2N} \neq \boldsymbol{0}\right) + \mathbb{P}\left(\left\|\hat{\beta}_N - \beta_N\right\|_2 > \frac{b_0}{2}\right),
\end{aligned}
$$

where in (a) we used $\|\cdot\|_\infty \leq \|\cdot\|_2$. Lemma 3 from Kock (2013) establishes that

$$
\lim_{N\to\infty} \mathbb{P}\left(\hat{\beta}_{2N} \neq \boldsymbol{0}\right) = 0,
$$

and Theorem 1 in Kock (2013) (the assumptions of which are satisfied under the assumptions of Lemma 3 in Kock 2013) establishes the consistency of $\hat{\beta}_N$, which is sufficient for

$$
\lim_{N\to\infty} \mathbb{P}\left(\left\|\hat{\beta}_N - \beta_N\right\|_2 > \frac{b_0}{2}\right) = 0.
$$

So

$$
0 \leq \lim_{N\to\infty} \mathbb{P}\left(\hat{\mathcal{S}} \neq \mathcal{S}\right) \leq \lim_{N\to\infty} \mathbb{P}\left(\hat{\beta}_{2N} \neq \boldsymbol{0}\right) + \lim_{N\to\infty} \mathbb{P}\left(\left\|\hat{\beta}_N - \beta_N\right\|_2 > \frac{b_0}{2}\right) = 0,
$$

proving the result.

$\square$

Proof of Theorem F.5. (a) Again, for clarity we use the notation of Kock (2013). Since

$$\sqrt{NT_N}\psi_N^\top(\hat{\beta}_N - \beta_0) = \sqrt{NT_N}\boldsymbol{\alpha}^\top(\hat{\beta}_{1N} - \beta_{10}) + \sqrt{NT_N}\boldsymbol{b}_N^\top(\hat{\beta}_{2N} - \beta_{20})$$
$$= \sqrt{NT_N}\boldsymbol{\alpha}^\top(\hat{\beta}_{1N} - \beta_{10}) + \sqrt{NT_N}\boldsymbol{b}_N^\top\hat{\beta}_{2N},$$

if we can show that

$$\sqrt{NT_N}\boldsymbol{b}_N^\top\hat{\beta}_{2N} \xrightarrow{p} 0$$

then the result follows from Theorem 2(ii) from Kock (2013) (in particular, Equation 7.5 in the proof) and Slutsky's theorem. Theorem 2(i) from Kock (2013) establishes that

$$\lim_{N\to\infty} \mathbb{P}\left(\hat{\beta}_{2N} = \mathbf{0}\right) = 1.$$

Since for any N

$$\left\{\hat{\beta}_{2N} = \mathbf{0}\right\} \subseteq \left\{\sqrt{NT_N}\sum_{j=1}^{p_N-k}\left|(\boldsymbol{b}_N)_j(\hat{\beta}_{2N})_j\right| = 0\right\},$$

we have

$$\mathbb{P}\left(\hat{\beta}_{2N} = \mathbf{0}\right) \leq \mathbb{P}\left(\sqrt{NT_N}\sum_{j=1}^{p_N-k}\left|(\boldsymbol{b}_N)_j(\hat{\beta}_{2N})_j\right| = 0\right),$$

so

$$1 \geq \lim_{N\to\infty} \mathbb{P}\left(\sqrt{NT_N}\sum_{j=1}^{p_N-k}\left|(\boldsymbol{b}_N)_j\hat{\beta}_{2N}\right| = 0\right) \geq \lim_{N\to\infty} \mathbb{P}\left(\hat{\beta}_{2N} = \mathbf{0}\right) = 1,$$

establishing

$$\lim_{N\to\infty} \mathbb{P}\left(\sqrt{NT_N}\sum_{j=1}^{p_N-k}\left|(\boldsymbol{b}_N)_j(\hat{\beta}_{2N})_j\right| = 0\right) = 1.$$

Since this notion of convergence is stronger than convergence in probability, it follows that

$$\sqrt{NT_N}\sum_{j=1}^{p_N-k}(\boldsymbol{b}_N)_j(\hat{\beta}_{2N})_j \xrightarrow{p} 0,$$

proving the result.

(b) Since by the consistency of the sample covariance matrix and the continuous mapping theorem

$$\frac{1}{\sigma}\sqrt{\frac{T}{\boldsymbol{\alpha}(\mathcal{S})^\top\left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot\cdot)\mathcal{S}})\right)^{-1}\boldsymbol{\alpha}(\mathcal{S})}} \xrightarrow{p} \frac{1}{\sigma}\sqrt{\frac{T}{\boldsymbol{\alpha}(\mathcal{S})^\top(\lim_{N\to\infty}\mathbb{E}[\Sigma_{1N}])^{-1}\boldsymbol{\alpha}(\mathcal{S})}},$$

by part (a) and Slutsky's theorem we have that $U_b(\mathcal{S})$ converges in distribution to a standard Gaussian random variable. Further, $\mathbb{P}(U_b(\hat{\mathcal{S}}_N) - U_b(\mathcal{S}) = 0) \geq \mathbb{P}(\hat{\mathcal{S}}_N = \mathcal{S})$, so from Theorem F.4 we have

$$1 \geq \lim_{N\to\infty} \mathbb{P}(U_b(\hat{\mathcal{S}}_N) - U_b(\mathcal{S}) = 0) \geq \lim_{N\to\infty} \mathbb{P}(\hat{\mathcal{S}}_N = \mathcal{S}) = 1,$$

which implies $U_b(\hat{\mathcal{S}}_N) - U_b(\mathcal{S}) \xrightarrow{p} 0$. Therefore by Slutsky's theorem,

$$U_b(\hat{\mathcal{S}}_N) = U_b(\mathcal{S}) + \left(U_b(\hat{\mathcal{S}}_N) - U_b(\mathcal{S})\right) \xrightarrow{d} \mathcal{N}(0,1).$$

(c) We begin by decomposing our sequence of random variables:

$$U_c(\mathcal{A}) = \sqrt{NT}\left(\hat{\boldsymbol{\psi}}_N(\mathcal{A})^\top \hat{\beta} - \boldsymbol{\psi}_N(\mathcal{A})^\top \beta_0\right)$$

$$= \sqrt{NT}\left(\boldsymbol{\psi}_N(\mathcal{A})^\top \hat{\beta} - \boldsymbol{\psi}_N(\mathcal{A})^\top \beta_0 + \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}) - \boldsymbol{\psi}_N(\mathcal{A})\right)^\top \hat{\beta}\right)$$

$$= \sqrt{NT}\left(\boldsymbol{\psi}_N(\mathcal{A})^\top \left(\hat{\beta} - \beta_0\right) + \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}) - \boldsymbol{\psi}_N(\mathcal{A})\right)^\top \beta_0\right.$$

$$\left. + \left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}) - \boldsymbol{\psi}_N(\mathcal{A})\right)^\top \left(\hat{\beta} - \beta_0\right)\right). \tag{H.4}$$

We examine these terms one at a time. By part (a),

$$\sqrt{NT}\boldsymbol{\psi}_N(\mathcal{S})^\top(\hat{\beta}_N - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}\left[\Sigma_{1N}\right]\right)^{-1} \boldsymbol{\alpha}\right).$$

By assumption, for any $\beta_0 \in \mathbb{R}^{p_N}$ with k entries nonzero and finite and the remaining entries equal to 0,

$$\sqrt{NT}\left(\hat{\boldsymbol{\psi}}_N(\mathcal{S}) - \boldsymbol{\psi}_N(\mathcal{S})\right)^\top \beta_0 \xrightarrow{d} \mathcal{N}(0, v_\psi(\beta_0, \mathcal{S})).$$

Finally, similarly to Kock (2013) we partition $\hat{\boldsymbol{\psi}}_N(\mathcal{A})$ into component $\hat{\boldsymbol{\psi}}_{1N}(\mathcal{A})$ that corresponds to the nonzero coefficients $\beta_{10} \in \mathbb{R}^k$ and component $\hat{\boldsymbol{\psi}}_{2N}(\mathcal{A})$ corresponding to $\beta_{20} = \mathbf{0} \in \mathbb{R}^{p_N - k}$, and likewise for $\boldsymbol{\psi}_N(\mathcal{A})$. Then we have

$$\left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}) - \boldsymbol{\psi}_N(\mathcal{A})\right)^\top \left(\hat{\beta} - \beta_0\right)$$

$$= \left(\hat{\boldsymbol{\psi}}_{1N}(\mathcal{A}) - \boldsymbol{\psi}_{1N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{1N} - \beta_{10}\right) + \left(\hat{\boldsymbol{\psi}}_{2N}(\mathcal{A}) - \boldsymbol{\psi}_{2N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{2N} - \beta_{20}\right)$$

$$= \left(\hat{\boldsymbol{\psi}}_{1N}(\mathcal{A}) - \boldsymbol{\psi}_{1N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{1N} - \beta_{10}\right) + \left(\hat{\boldsymbol{\psi}}_{2N}(\mathcal{A}) - \boldsymbol{\psi}_{2N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{2N}\right).$$

For the first term, applying Theorem 2.4(i) in van der Vaart (2000) and Proposition 1.8(iii) in García-Portugués (2023) termwise we have that

$$\left(\hat{\boldsymbol{\psi}}_{1N}(\mathcal{A}) - \boldsymbol{\psi}_{1N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{1N} - \beta_{10}\right) = \mathcal{O}_\mathbb{P}\left(\frac{1}{\sqrt{N}}\right) \mathcal{O}_\mathbb{P}\left(\frac{1}{\sqrt{N}}\right) = \mathcal{O}_\mathbb{P}\left(\frac{1}{N}\right)$$

$$\implies \quad \sqrt{NT}\left(\hat{\boldsymbol{\psi}}_{1N}(\mathcal{A}) - \boldsymbol{\psi}_{1N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{1N} - \beta_{10}\right) = \mathcal{O}_\mathbb{P}\left(\frac{1}{\sqrt{N}}\right).$$

For the second term, by a similar argument to the one we used in part (a), Theorem 2(i) from Kock (2013) gives us that the $\sqrt{NT}\left(\hat{\boldsymbol{\psi}}_{2N}(\mathcal{A}) - \boldsymbol{\psi}_{2N}(\mathcal{A})\right)^\top \left(\hat{\beta}_{2N}\right) \xrightarrow{p} 0$. Putting this together,

$$\sqrt{NT}\left(\hat{\boldsymbol{\psi}}_N(\mathcal{A}) - \boldsymbol{\psi}_N(\mathcal{A})\right)^\top \left(\hat{\beta} - \beta_0\right) \xrightarrow{p} 0.$$

Putting this all together, using Slutsky's theorem and the independence of $\hat{\beta}$ and $\hat{\boldsymbol{\psi}}_N(\mathcal{S})$ we have from (H.4) that

$$U_c(\mathcal{S}) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}\left[\Sigma_{1N}\right]\right)^{-1} \boldsymbol{\alpha} + v_\psi(\beta_0, \mathcal{S})\right).$$

Finally, the same argument we used in part (b) yields $U_c(\hat{\mathcal{S}}) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}\left[\Sigma_{1N}\right]\right)^{-1} \boldsymbol{\alpha} + v_\psi(\beta_0, \mathcal{S}))$.

(d) Using the consistency of the sample covariance matrix, the assumed consistency of $\hat{\boldsymbol{\alpha}}(\mathcal{S})$, the assumed consistency of $\hat{v}_\psi(\hat{\beta}_N, \mathcal{S})$, and the continuous mapping theorem,

$$
\frac{1}{\sqrt{\hat{v}_N(\mathcal{S})}}
$$

$$
= \frac{1}{\sqrt{\hat{\boldsymbol{\alpha}}(\mathcal{S})^\top \left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot\cdot)\mathcal{S}}) \right)^{-1} \hat{\boldsymbol{\alpha}}(\mathcal{S}) + \hat{v}_\psi(\hat{\beta}_N, \mathcal{S})}}
$$

$$
\xrightarrow{p} \frac{1}{\sqrt{\boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}[\Sigma_{1N}] \right)^{-1} \boldsymbol{\alpha} + v_\psi(\beta_0, \mathcal{S})}},
$$

so part (c) and Slutsky's theorem yield that $U_d(\mathcal{S}) = \frac{1}{\sqrt{\hat{v}_N(\mathcal{S})}} U_c(\mathcal{S}) \xrightarrow{d} \mathcal{N}(0, 1)$, where our finite-sample variance estimator $\hat{v}_N(\mathcal{S})$ was defined in (F.4). Finally, the same argument we used in part (b) yields $U_d(\hat{\mathcal{S}}) \xrightarrow{d} \mathcal{N}(0, 1)$.

(e) If $\hat{\beta}_N$ and $\hat{\boldsymbol{\psi}}_N(\mathcal{S})$ are estimated on the same data set, they are not independent. So by the same argument we used in part (c) we still have

$$
\sqrt{NT} \left(\hat{\boldsymbol{\psi}}_N(\mathcal{S})^\top (\hat{\beta}_N - \beta_0) + \left(\hat{\boldsymbol{\psi}}_N(\mathcal{S}) - \boldsymbol{\psi}_N(\mathcal{S}) \right)^\top \beta_0 \right)
$$

$$
\xrightarrow{d} \mathcal{N} \left(0, \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}[\Sigma_{1N}] \right)^{-1} \boldsymbol{\alpha} \right) + \mathcal{N} \left(0, v_\psi(\beta_0, \mathcal{S}) \right),
$$

but because the two Gaussian random variables are neither independent nor do we know that they are jointly Gaussian, we cannot say as much about the distribution of their sum. However, due to Lemma H.2 we do know that they are subgaussian.

Lemma H.2. *If X and Y are mean-zero subgaussian random variables, then $X + Y$ is mean zero and subgaussian.*

Proof. Provided in Appendix I. $\qquad\square$

Further, since for any two random variables with finite variance by Cauchy-Schwarz

$$
\begin{aligned}
\mathrm{Var}(X + Y) &= \mathrm{Var}(X) + 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y) \\
&\leq \mathrm{Var}(X) + 2\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} + \mathrm{Var}(Y),
\end{aligned}
$$

we have that the above sequence of random variables converges in distribution to a subgaussian random variable with variance at most

$$
\sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}[\Sigma_{1N}] \right)^{-1} \boldsymbol{\alpha} + 2\sqrt{\sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}[\Sigma_{1N}] \right)^{-1} \boldsymbol{\alpha} \cdot v_\psi(\beta_0, \mathcal{S})} + v_\psi(\beta_0, \mathcal{S}).
$$

Using the assumed consistency of $\hat{v}_\psi(\hat{\beta}_N, \mathcal{S})$, we have by the continuous mapping theorem that for $\hat{v}_N^{(\mathrm{cons})}(\mathcal{A})$ as defined in (F.5),

$$
\hat{v}_N^{(\mathrm{cons})}(\mathcal{S})
$$

$$
= \sigma^2 \hat{\boldsymbol{\alpha}}(\mathcal{S})^\top \left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot\cdot)\mathcal{S}}) \right)^{-1} \hat{\boldsymbol{\alpha}}(\mathcal{S}) + \hat{v}_\psi(\hat{\beta}_N, \mathcal{S}) + 2\sqrt{\sigma^2 \hat{\boldsymbol{\alpha}}(\mathcal{S})^\top \left(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X}_{(\cdot\cdot)\mathcal{S}}) \right)^{-1} \hat{\boldsymbol{\alpha}}(\mathcal{S}) \cdot \hat{v}_\psi(\hat{\beta}_N, \mathcal{S})}
$$

$$
\xrightarrow{p} \sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}[\Sigma_{1N}] \right)^{-1} \boldsymbol{\alpha} + v_\psi(\beta_0, \mathcal{S}) + 2\sqrt{\sigma^2 \boldsymbol{\alpha}^\top \left(\lim_{N\to\infty} \mathbb{E}[\Sigma_{1N}] \right)^{-1} \boldsymbol{\alpha} \cdot v_\psi(\beta_0, \mathcal{S})}.
$$

84

Therefore by Slutsky's theorem, $U_e(\mathcal{S})$ converges in distribution to a mean-zero subgaussian random variable with variance at most 1. Finally, once again a similar argument to the one we used in part (b) yields that $U_e(\hat{\mathcal{S}})$ converges in distribution to the same random variable as $U_e(\mathcal{S})$.

$\square$

I Statements and Proofs of Other Results

Proof of Lemma H.2. The sum $X + Y$ has mean zero due to linearity of expectation. To show subgaussianity, it is enough to show that $\mathbb{E} \exp \{t(X + Y)\} \leq \exp \{K^2 t^2\}$ for some finite K^2 for all $t \in \mathbb{R}$ (Vershynin, 2018, Proposition 2.5.2). Because X and Y are subgaussian, we have that $\mathbb{E} [\exp \{tX\}] \leq \exp \{K_X^2 t^2\}$ and $\mathbb{E} [\exp \{tY\}] \leq \exp \{K_Y^2 t^2\}$ for all $t \in \mathbb{R}$. So using these facts and Cauchy-Schwarz, we have for any $t \in \mathcal{R}$

$$
\begin{aligned}
\mathbb{E} \exp \{t(X + Y)\} = & \ \mathbb{E} [\exp \{tX\} \exp \{tY\}] \\
\leq & \ \sqrt{\mathbb{E} [\exp \{2tX\}] \mathbb{E} [\exp \{2tY\}]} \\
\leq & \ \sqrt{\exp \{K_X^2 \cdot 4t^2\} \exp \{K_Y^2 \cdot 4t^2\}} \\
= & \ \exp \{2t^2 (K_X^2 + K_Y^2)\},
\end{aligned}
$$

and the result is proven with $K^2 := 2 (K_X^2 + K_Y^2)$.

$\square$

Lemma I.1. *The smallest and largest singular values of $\boldsymbol{G}_N$ are bounded as follows: $\sigma_{max}(\boldsymbol{G}_N) \leq \sqrt{2}$ and $\sigma_{min} (\boldsymbol{G}_N) \geq \sigma / \sqrt{T \sigma_c^2 + \sigma^2}$.*

Proof of Lemma I.1. We will prove the result in a similar way to the proof of Lemma F.1. By properties of the Kronecker product, it is enough to bound the singular values of $\boldsymbol{\Omega}$. We have

$$
\sigma_{\max}(\boldsymbol{\Omega}) \leq \sqrt{\|\boldsymbol{\Omega}\|_1 \|\boldsymbol{\Omega}\|_\infty} \leq \sqrt{[T \cdot \sigma_c^2 + \sigma^2]^2} = T \cdot \sigma_c^2 + \sigma^2,
$$

where for matrices $\|\cdot\|_1$ denotes the maximum absolute column sum of a matrix and $\|\cdot\|_\infty$ denotes the maximum absolute row sum. We will use (H.1) to lower-bound $\sigma_{\min}(\boldsymbol{\Omega})$. First we will find $\boldsymbol{\Omega}^{-1}$. Observe that we can express $\boldsymbol{\Omega}$ as

$$
\boldsymbol{\Omega} = \sigma^2 \boldsymbol{I}_T + \boldsymbol{v}\boldsymbol{v}^\top
$$

where $\boldsymbol{v} = (\sigma_c, \ldots, \sigma_c) \in \mathbb{R}^T$. Therefore we can invert $\boldsymbol{\Omega}$ using the Sherman-Morrison formula:

$$
\begin{aligned}
\boldsymbol{\Omega}^{-1} = & \ \frac{1}{\sigma^2} \boldsymbol{I}_T - \frac{\frac{1}{\sigma^2} \boldsymbol{I}_T \boldsymbol{v}\boldsymbol{v}^\top \frac{1}{\sigma^2} \boldsymbol{I}_T}{1 + \boldsymbol{v}^\top \frac{1}{\sigma^2} \boldsymbol{I}_T \boldsymbol{v}} = \frac{1}{\sigma^2} \left(\boldsymbol{I}_T - \frac{\frac{1}{\sigma^2} \boldsymbol{v}\boldsymbol{v}^\top}{1 + \frac{1}{\sigma^2} \boldsymbol{v}^\top \boldsymbol{v}} \right) = \frac{1}{\sigma^2} \left(\boldsymbol{I}_T - \frac{\boldsymbol{v}\boldsymbol{v}^\top}{\sigma^2 + T\sigma_c^2} \right) \\
= & \ \frac{1}{\sigma^2} \begin{pmatrix} 1 - \frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} & -\frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} & \cdots & -\frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} \\ -\frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} & 1 - \frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} & \cdots & -\frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} & -\frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} & \cdots & 1 - \frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} \end{pmatrix}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\sigma_{\min} (\boldsymbol{\Omega}) \geq & \ \frac{1}{\sqrt{\|\boldsymbol{\Omega}^{-1}\|_1 \|\boldsymbol{\Omega}^{-1}\|_\infty}} \geq \sigma^2 \left(1 + T \frac{\sigma_c^2}{\sigma^2 + T\sigma_c^2} \right)^{-1} = \sigma^2 \left(\frac{\sigma^2 + 2T\sigma_c^2}{\sigma^2 + T\sigma_c^2} \right)^{-1} \\
= & \ \frac{\sigma^2 (\sigma^2 + T\sigma_c^2)}{\sigma^2 + 2T\sigma_c^2} \geq \frac{\sigma^2 (\sigma^2 + T\sigma_c^2)}{2 (\sigma^2 + T\sigma_c^2)} = \frac{\sigma^2}{2}.
\end{aligned}
$$

Finally, using these results we have $\sigma_{\min}\left(\boldsymbol{\Omega}^{-1/2}\right) \geq 1/\sqrt{T\sigma_c^2 + \sigma^2}$ and $\sigma_{\max}\left(\boldsymbol{\Omega}^{-1/2}\right) \leq \sqrt{2}/\sigma$. By properties of the Kronecker product, these bounds hold for $\boldsymbol{G}_N = \sigma \boldsymbol{I}_T \otimes \boldsymbol{\Omega}^{-1/2}$ as well after multiplying by σ.

$\square$

Lemma I.2. *Suppose one treated cohort has fewer than $d_N + 1$ units. Then $\tilde{\boldsymbol{Z}}$ does not have full column rank.*

Proof of Lemma I.2. The treatment effect dummies can be seen as interaction effects between the cohort dummies and (a subset of) the time dummies. In particular, the treatment effect for cohort r at time t is equal to the column for the rth cohort dummy, but with entries for all times except t set equal to 0. The $T - r + 1$ treatment dummies for cohort r are mutually orthogonal and span a $(T - r + 1)$-dimensional subspace, and they occupy an $N_r(T - r + 1)$-dimensional subspace, where N_r is the number of units in cohort r.

Because the covariates (like the cohort indicators) are time-invariant, the time-covariate interactions are closely related to the treatment effect dummies. The covariate-time interactions for cohort r lie in the same $N_r(T - r + 1)$-dimensional subspace as the treatment effects for cohort r and are either linearly dependent or span a subspace of dimension $d_N(T - r + 1)$. $\tilde{\boldsymbol{Z}}$ is rank deficient in the former setting, so we assume the latter. Then the treatment dummies and covariate-time interactions for cohort r span a subspace of dimension $(d_N + 1)(T - r + 1)$ within this $N_r(T - r + 1)$-dimensional subspace. So if $N_r < d_N + 1$, these columns must be linearly dependent, and $\tilde{\boldsymbol{Z}}$ does not have full column rank.

$\square$

Lemma I.3. *Suppose e_{1N} is bounded away from 0 with high probability: for some finite $N^*, c > 0$, for all $N > N^*$ it holds that $\mathbb{P}(e_{1N} < c) < a_N$ for some vanishing sequence $\{a_N\}$. Let $\{X_N\}$ be a sequence of random variables, and suppose that $X_N = \mathcal{O}_{\mathbb{P}}(\frac{1}{e_{1N}}\sqrt{p_N/N})$. Then $X_N = \mathcal{O}_{\mathbb{P}}(\sqrt{p_N/N})$.*

Proof of Lemma I.3. By assumption, for every $\epsilon > 0$ there exists a finite $M_\epsilon^{(1)} > 0$ and $N_\epsilon^{(1)} > 0$ such that for all $N > N_\epsilon^{(1)}$,

$$\mathbb{P}\left(\left|\frac{e_{1N}X_N}{\sqrt{p_N/N}}\right| > M_\epsilon^{(1)}\right) < \epsilon.$$

Now for all $N > \max\{N^*, N^{(1)}_{\epsilon/2}\}$,

$$\mathbb{P}\left(\left|\frac{X_N}{\sqrt{p_N/N}}\right| > \frac{M^{(1)}_{\epsilon/2}}{c}\right)$$

$$= \mathbb{P}\left(\left|\frac{X_N}{\sqrt{p_N/N}}\right| > \frac{M^{(1)}_{\epsilon/2}}{c} \mid e_{1N} \geq c\right)\mathbb{P}(e_{1N} \geq c) + \mathbb{P}\left(\left|\frac{X_N}{\sqrt{p_N/N}}\right| > \frac{M^{(1)}_{\epsilon/2}}{c} \mid e_{1N} < c\right)\mathbb{P}(e_{1N} < c)$$

$$\leq \mathbb{P}\left(\left|\frac{X_N}{\sqrt{p_N/N}}\right| > \frac{M^{(1)}_{\epsilon/2}}{c} \mid e_{1N} \geq c\right)\mathbb{P}(e_{1N} \geq c) + \mathbb{P}(e_{1N} < c)$$

$$\leq \mathbb{P}\left(\left|\frac{X_N}{\sqrt{p_N/N}}\right| > \frac{M^{(1)}_{\epsilon/2}}{e_{1N}} \mid e_{1N} \geq c\right)\mathbb{P}(e_{1N} \geq c) + \mathbb{P}(e_{1N} < c)$$

$$= \mathbb{P}\left(\left|\frac{e_{1N}X_N}{\sqrt{p_N/N}}\right| > M^{(1)}_{\epsilon/2} \mid e_{1N} \geq c\right)\mathbb{P}(e_{1N} \geq c) + \mathbb{P}(e_{1N} < c)$$

$$\leq \mathbb{P}\left(\left|\frac{e_{1N}X_N}{\sqrt{p_N/N}}\right| > M^{(1)}_{\epsilon/2} \mid e_{1N} \geq c\right)\mathbb{P}(e_{1N} \geq c) + \mathbb{P}\left(\left|\frac{e_{1N}X_N}{\sqrt{p_N/N}}\right| > M^{(1)}_{\epsilon/2} \mid e_{1N} < c\right)\mathbb{P}(e_{1N} < c)$$

$$\quad + \mathbb{P}(e_{1N} < c)$$

$$= \mathbb{P}\left(\left|\frac{e_{1N}X_N}{\sqrt{p_N/N}}\right| > M^{(1)}_{\epsilon/2}\right) + \mathbb{P}(e_{1N} < c)$$

$$< \frac{\epsilon}{2} + a_N.$$

Since a_N is vanishing, there exists a finite $N^{(2)}_\epsilon$ so that $a_N \leq \epsilon/2$ for all $N > N^{(2)}_\epsilon$. Then we have established that for any $\epsilon > 0$, for $M_\epsilon := M^{(1)}_{\epsilon/2}/c$ and $N_\epsilon := \max\{N^*, N^{(1)}_{\epsilon/2}, N^{(2)}_\epsilon\}$, for all $N > N_\epsilon$ it holds that

$$\mathbb{P}\left(\left|\frac{X_N}{\sqrt{p_N/N}}\right| > M_\epsilon\right) < \epsilon.$$

$\square$