

Difference-in-Differences with Compositional Changes

Pedro H. C. Sant'Anna*

Qi Xu†

December 26, 2024

Abstract

This paper studies difference-in-differences (DiD) setups with repeated cross-sectional data and potential compositional changes across time periods. We begin our analysis by deriving the efficient influence function and the semiparametric efficiency bound for the average treatment effect on the treated (ATT). We introduce nonparametric estimators that attain the semiparametric efficiency bound under mild rate conditions on the estimators of the nuisance functions, exhibiting a type of rate doubly robust (DR) property. Additionally, we document a trade-off related to compositional changes: We derive the asymptotic bias of DR DiD estimators that erroneously exclude compositional changes and the efficiency loss when one fails to correctly rule out compositional changes. We propose a nonparametric Hausman-type test for compositional changes based on these trade-offs. The finite sample performance of the proposed DiD tools is evaluated through Monte Carlo experiments and an empirical application. We consider extensions of our framework that accommodate double machine learning procedures with cross-fitting, and setups when some units are observed in both pre- and post-treatment periods. As a by-product of our analysis, we present a new uniform stochastic expansion of the local polynomial multinomial logit estimator, which may be of independent interest.

*Emory University. E-mail: pedro.santanna@emory.edu

†Department of Health Outcomes and Biomedical Informatics, University of Florida. E-mail: qixu@ufl.edu.

1 Introduction

Difference-in-differences (DiD) designs have been used widely for identifying and estimating causal effects with observational data. Identification in this research design typically relies on a conditional parallel trends assumption stipulating that conditional on a set of covariates, the average untreated outcomes among treated and comparison groups would have evolved “in parallel”. When one pairs this assumption with common support and no-anticipation assumptions, it is easy to establish that the average treatment effect on the treated (ATT) is nonparametrically identified when panel data is available. When one only observes repeated cross-sectional data, it is common to impose further a no-compositional change assumption, also known as the stationarity assumption. This is the case in the widely cited DiD procedures of Heckman, Ichimura and Todd (1997), Abadie (2005), Sant’Anna and Zhao (2020), and Callaway and Sant’Anna (2021), for example.

Although we have seen a lot of recent developments in DiD methods (see Roth, Sant’Anna, Bilinski and Poe, 2023 for an overview of recent DiD developments), little attention has been paid to understanding the importance and limitations of the no-compositional changes assumption. This paper aims to fill this gap by providing researchers with new tools that can be used when they doubt such an assumption and/or to test its plausibility.

Before discussing the paper’s contributions, it is worth stressing why ruling out compositional changes across periods can be restrictive in real empirical applications. Essentially, the no-compositional changes assumption requires one to sample observations from the same population across periods, which can be unrealistic in some scenarios. For example, Hong (2013) studies the effect of Napster on recorded music sales. He uses data from the 1996–2002 Interview Surveys of the Consumer Expenditure Survey. Over this period, the composition of internet users has changed substantially. The early adopters tend to be younger, richer, more educated, and technically savvy, whereas later adopters exhibit a higher diversity level in demographics. If one ignores such imbalances of group composition across time, the (negative) effect of Napster on music sales can be overestimated, as the decrease in the average music expenditure may be attributed to a post-Napster group with more households having low reservation prices for recorded music. Other applications also share this concern, as discussed below and in more detail in Section 6. Therefore, having causal inference tools that can assess if the findings are robust to compositional changes in the sample is of practical interest.

We begin our analysis by showing that one can identify the ATT in DiD setups without invoking the no-compositional changes assumption. In this scenario, we derive the efficient influence function (EIF) and the semiparametric efficiency bound for the ATT. We then form generic nonparametric estimators built on the EIF that can achieve the semiparametric efficient bound under mild smoothness conditions, a rate doubly robust (DR) property (Smucler, Rotnitzky and Robins, 2019). Heuristically, this rate DR property allows for a trade-off between the rate of convergence of the two nuisance estimators. It implies that nonparametric DiD estimators for the ATT based on the EIF are $\sqrt{n}$ -consistent and asymptotically normal even if

one of the outcome regression or generalized propensity score functions is very complex so long as the other is simple enough; this is weaker than requiring that the estimators for *both* nuisance functions converge sufficiently fast. These results are general and do not rely on a specific choice of estimators for nuisance functions. Nonetheless, they do not help us with practical inference procedures. For that, we use a local polynomial estimator for the outcome-regressions models and the local multinomial logit regression to estimate the generalized propensity score, the latter of which is fairly new in the DiD literature. Importantly, our nonparametric estimators can accommodate both discrete and continuous covariates,¹ and all tuning parameters are selected in a data-driven way via cross-validation.² Finally, we show that the estimator proposed by Sant’Anna and Zhao (2020) is no longer DR in this DiD setup with compositional changes. In fact, we show that even when all nuisance functions are correctly specified, the Sant’Anna and Zhao (2020)’s DR DiD estimand does not identify the ATT in this general setup. Overall, this first set of results highlights “the best” one can do in DiD setups with compositional changes.

Next, we tackle the problem of how much efficiency one may lose by not exploring the no-compositional change assumption when it is valid. To answer this question, we compare our derived semiparametric efficiency bound that does not impose the no-compositional changes assumption with the semiparametric efficiency bound derived by Sant’Anna and Zhao (2020) that fully exploits it. As expected, the extra layer of robustness comes at the cost of loss of efficiency. Heuristically speaking, the no-compositional change assumption allows one to pool the covariate data from all time periods, substantially increasing the effective sample size and the precision of the DiD estimator compared to the one that does not impose the no-compositional change assumption. Regarding the estimation of nuisance functions, we also note that, under the no-compositional changes assumption, one can use standard (binary) propensity score estimators. However, when one allows for compositional changes, one needs to use generalized propensity score estimators, as there are now four effective groups depending on the treatment group and the time the unit is observed.

In practice, determining whether compositional changes are a concern for a given empirical application is not always obvious. Specifically, it is unclear whether imposing a no-compositional change assumption will lead to biased ATT estimates. Using our previous results, we propose a nonparametric Hausman (1978)-type test for no-compositional changes. The test compares our nonparametric DiD estimator of the ATT, which is robust to compositional changes, with the nonparametric extension of Sant’Anna and Zhao (2020)’s DR DiD estimator, which assumes no compositional changes. We derive the large sample properties of the proposed test, which shows that it controls size asymptotically and is consistent against a

1 As a side contribution of this paper, we provide a new result on the uniform expansion of the local (multinomial) logit estimators, which accommodates both continuous and discrete variables. This result may be of independent interest.

2 Bandwidths are selected independently for each set of nuisance functions instead of directly for the ATT estimator. Nonetheless, the second-step asymptotic results remain unaffected, as the double-robust rate property of our proposed estimator ensures that using optimal bandwidths for each nuisance function still enables valid inference on the ATT estimator.

broad set of alternatives.

We demonstrate the practical appeal of our proposed DiD tools through Monte Carlo simulations and an empirical application that revisits Sequeira (2016). She leverages a quasi-experimental variation created by a large reduction in the average nominal tariff rate between South Africa and Mozambique in 2008 to study the causal effect of tariff rate reduction on trade costs and corruption behavior using a two-way fixed effects specification with covariates that implicitly imposes a no-compositional changes assumption, among other arguably unnecessary homogeneity assumptions. We use our nonparametric tests to assess the plausibility of the no-compositional changes assumption and fail to reject it at the usual significance levels. Our results support the conclusions by Sequeira (2016) that tariff liberalization decreases corruption, and our DR DiD estimates are similar to those in the original paper.

Finally, we consider some extensions of our framework. Although our paper primarily focuses on leave-one-out estimators for the nuisance functions, we also consider cross-fitted estimators, drawing on the modern double machine learning literature; see, e.g., Farrell (2015), Belloni, Chernozhukov, Fernández-Val and Hansen (2017), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2017), Colangelo and Lee (2023), and Kennedy (2023). An important difference between cross-fitted and leave-one-out estimators is that the former requires the number of folds J to be fixed as $n \rightarrow \infty$, while the latter allows it to grow with sample size (as $J = n$). As a result, these estimation procedures rely on different types of assumptions and require different proof strategies to establish their large-sample properties.

We also extend our analysis to applications where some units are observed in both pre- and post-treatment periods (a balanced panel data component), while the remaining units are observed in only one of the two periods (a cross-sectional component). However, our main results focus on sampling schemes without overlapping units, thus precluding these scenarios. To cover this practically relevant setup, we derive the EIF and the semiparametric efficiency bound for the ATT in setups that allow for this type of sampling scheme. Similar to the pure repeated cross-sectional setup, we discuss how one can build on the derived EIF to form nonparametric estimators that achieve the semiparametric efficiency bound.

Related literature: This article belongs to the extensive literature on semiparametric DiD methods. We refer the reader to Roth et al. (2023) for a synthesis of recent advances in the econometrics of DiD. Within this broad literature, the paper closest to ours is Sant’Anna and Zhao (2020), which proposes DR DiD estimators for the ATT and derives semiparametric efficiency bound for such estimators, too. In sharp contrast to us, though, all the results in Sant’Anna and Zhao (2020) rely on a no-compositional change assumption. Thus, our results complement theirs. Furthermore, Sant’Anna and Zhao (2020)’s theoretical results rely on parametric first-step estimators, while we accommodate nonparametric estimators. A perhaps side and minor contribution of our paper is establishing the statistical properties of Sant’Anna and Zhao (2020)’s DR DiD estimator with nonparametric estimates of the nuisance functions; see also Chang (2020).

Our paper also relates to the causal inference literature on compositional changes over

time. Hong (2013) develops a matching-based estimator under a “selection-on-observable”-type assumption, which is different and arguably stronger than our conditional parallel trends assumption. Hong (2013) also does not discuss efficiency issues as we do. Stuart, Huskamp, Duckworth, Simmons, Song, Chernew and Barry (2014) propose inverse probability weighted estimators for the ATT in DiD setups under compositional changes. In contrast to us, their estimator does not enjoy any DR property and may not attain the semiparametric efficiency bound. Nie, Lu and Wager (2019) is also interested in DiD estimators under compositional changes. Their estimator substantially differs from ours: they use meta-learners and cross-fitting to estimate nuisance functions, while our estimator is based on the EIF for the ATT. When treatment effects are heterogeneous, their estimators do not target the ATT but the ATE, which, in our context, is not identified. They do not consider tests for the no-compositional changes assumption as we do.

Finally, we contribute to the semiparametric two-stage estimation that depends on non-parametrically estimated functions. See, e.g., Newey (1994), Chen, Linton and Van Keilegom (2003), Chen, Hong and Tarozzi (2008), Akerberg, Chen, Hahn and Liao (2014), Rothe and Firpo (2019), among many others. Our results on local multinomial logit regression builds on Fan, Heckman and Wand (1995), Claeskens and Van Keilegom (2003), Li and Ouyang (2005), and Kong, Linton and Xia (2010). The novel result on the uniform expansion of the local multinomial logit estimator may be of independent interest. Our cross-fitting double machine learning results build on, among others, Belloni et al. (2017), Chernozhukov et al. (2017), Colangelo and Lee (2023), and Kennedy (2023) and extend their cross-sectional results under unconfoundedness to the DiD context.

Organization of the paper: Section 2 introduces the identification framework of the DiD parameter under compositional changes, presents the semiparametric efficiency results, and discusses the bias-variance trade-off of ruling out compositional changes. In Section 3, we present our nonparametric DR DiD estimators, discuss their large sample properties, and how to pick tuning parameters. Section 4 discusses a test for no-compositional changes. Monte Carlo simulations are provided in Section 5, and an empirical illustration is considered in Section 6. Section 7 discusses extensions. Proofs and additional results are reported in the Supplemental Appendix available [here](#).

2 Difference-in-Differences

2.1 Framework

This section describes our setup. We focus on the canonical two-period and two-group setup for conciseness and transparency. We have two time periods, $t = 0$, where no unit is exposed to the treatment, and time $t = 1$, where units in the group with $D = 1$ are exposed to treatment; here, D is a binary treatment indicator. We adopt the potential outcome notation where $Y_{it}(0)$ and $Y_{it}(1)$ denote the untreated and treated potential outcome for unit i at time t ,

respectively. Observed outcomes are given by $Y_{it} = D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0)$. We also assume that a k -dimensional vector of pre-treatment characteristics $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$ is available.

This paper considers the case where one has access to repeated cross-sectional data. To formalize this idea, let T_i be a dummy variable that takes value one if the observation i is observed only in the post-treatment period $t = 1$, and zero if observation i is only observed in the pre-treatment period $t = 0$. Define $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$, and let n_1 and n_0 be the sample sizes of the post-treatment and pre-treatment periods such that $n = n_1 + n_0$.

Assumption 1 (Sampling) The pooled data $\{Y_i, D_i, X_i, T_i\}_{i=1}^n$ consists of independent and identically distributed draws from the mixture distribution $P \in \mathcal{P}$:

$$\begin{aligned} \mathbb{P}(Y \leq y, D = d, X \leq x, T = t) &= t \cdot \mathbb{P}(T = 1) \cdot \mathbb{P}(Y_1 \leq y, D = d, X \leq x | T = 1) \\ &\quad + (1 - t) \cdot \mathbb{P}(T = 0) \mathbb{P}(Y_0 \leq y, D = d, X \leq x | T = 0), \end{aligned}$$

where $(y, d, x, t) \in \mathcal{Y} \times \{0, 1\} \times \mathcal{X} \times \{0, 1\}$.

Assumption 1 allows for different sampling schemes. For instance, it accommodates the binomial sampling scheme where an observation i is randomly drawn from either (Y_1, D, X) or (Y_0, D, X) with a fixed probability. It also accommodates the “conditional” sampling scheme where n_1 observations are sampled from (Y_1, D, X) , n_0 observations are sampled from (Y_0, D, X) and $\mathbb{P}(T = 1) = n_1/n$ (here, T is treated as fixed). Importantly, Assumption 1 does not impose that we are sampling from the same underlying distribution across time periods, implying that it is fully compatible with compositional changes (Hong, 2013). This is in contrast to most of the DiD literature. For example, Assumption 1(b) in Sant’Anna and Zhao (2020) explicitly imposes that $(D, X) \perp\!\!\!\perp T$; see also Heckman et al. (1997), and Abadie (2005) for other DiD procedures that rely on this stationarity condition.

As is typical in DiD setups, we are interested in the average treatment effect in time period $t = 1$ among the treated units,

$$ATT = \tau = \mathbb{E}[Y_1(1) | D = 1, T = 1] - \mathbb{E}[Y_1(0) | D = 1, T = 1]. \quad (2.1)$$

Given that the untreated potential outcome $Y_{i1}(0)$ is never observed for the treated units, we need to impose assumptions to uncover $\mathbb{E}[Y_1(0) | D = 1, T = 1]$ from the data. We make conditional parallel trends, no-anticipation, and strong overlap assumptions toward this goal. Let $\mathcal{S} \equiv \{0, 1\}^2$ and $\mathcal{S}_- \equiv \{(1, 0), (0, 1), (0, 0)\}$.

Assumption 2 (Conditional parallel trends, no-anticipation, and overlap)

For some $\varepsilon > 0$, $(d, t) \in \mathcal{S}_-$, and for almost every $x \in \mathcal{X}$

- (i) $\mathbb{E}[Y_1(0) | D = 1, T = 1, X = x] - \mathbb{E}[Y_0(0) | D = 1, T = 0, X = x]$
 $= \mathbb{E}[Y_1(0) | D = 0, T = 1, X = x] - \mathbb{E}[Y_0(0) | D = 0, T = 0, X = x].$
- (ii) $\mathbb{E}[Y_0(0) | D = 1, T = 0, X = x] = \mathbb{E}[Y_0(1) | D = 1, T = 0, X = x].$
- (iii) $\mathbb{P}(D = 1, T = 1) > \varepsilon$ and $\mathbb{P}(D = d, T = t | X = x) \geq \varepsilon.$

Assumption 2(i) is the conditional parallel trends assumption (CPT) stating that conditioning on X , the average evolution of the untreated potential outcome is the same among the treated and untreated groups. This assumption allows for covariate-specific trends and does not restrict the trends among different covariate strata. Assumption 2(ii) is a no-anticipation assumption (NAA) stating that, on average, treated units do not act on the future treatment prior to its implementation (Abbring and van den Berg, 2003; Malani and Reif, 2015). Assumption 2(iii) is an overlap condition that guarantees that there are some treated units in the post-treatment period and that the covariates do not fully determine treatment status. This condition is crucial for guaranteeing nonparametric regular inference procedures (Khan and Tamer, 2010).

2.2 Identification and semiparametric efficiency bound

Under Assumptions 1 and 2, it is straightforward to show that the ATT is nonparametrically identified by the outcome regression estimand³

$$\tau = \tau_{or} \equiv \mathbb{E}[Y|D = 1, T = 1] - \mathbb{E}[m_{1,0}(X) + m_{0,1}(X) - m_{0,0}(X)|D = 1, T = 1], \quad (2.2)$$

where $m_{d,t}(x) = E[Y|D = d, T = t, X = x]$. Alternatively, it is also easy to show that one can identify the ATT using an inverse probability weighted estimand

$$\tau = \tau_{ipw} \equiv \mathbb{E}[(w_{1,1}(D, T) - w_{1,0}(D, T, X) - w_{0,1}(D, T, X) + w_{0,0}(D, T, X))Y], \quad (2.3)$$

where, for $(d, t) \in \mathcal{S}_-$

$$\begin{aligned} w_{1,1}(D, T) &= \frac{DT}{\mathbb{E}[DT]}, \\ w_{d,t}(D, T, X) &= \frac{I_{d,t} \cdot p(1, 1, X)}{p(d, t, X)} \bigg/ \mathbb{E} \left[\frac{I_{d,t} \cdot p(1, 1, X)}{p(d, t, X)} \right], \end{aligned} \quad (2.4)$$

$I_{d,t} = \mathbb{1}\{D = d, T = t\}$, and $p(d, t, x) = \mathbb{P}(D = d, T = t|X = x)$ is a so-called generalized propensity score. Notice that the weights in (2.4) are of the Hájek (1971)-type. This guarantees that all the weights sum up to one and typically results in more stable finite sample behavior; see, e.g., Millimet and Tchernis (2009); Busso, Dinardo and McCrary (2014); Sant'Anna and Zhao (2020).

From (2.2) and (2.3), it is clear that any linear combination of τ_{or} and τ_{ipw} also identifies the ATT under our assumptions. There are also many other potential estimands that make use of nonlinear combinations of the different terms in τ_{or} and τ_{ipw} and identify the ATT. From this simple observation, a natural question that arises is: How can we combine these two strategies to obtain an efficient estimator for the ATT? The next theorem addresses this question through the lens of semiparametric efficiency theory. Specifically, we derive the EIF for the ATT under Assumptions 1 and 2, as well as its semiparametric efficiency bound. This bound represents the

³ See Lemma A.1 in Appendix A for the formalization of these results.

maximum precision achievable in this context under the given assumptions. As so, it provides a benchmark that researchers can use to assess whether any given (regular) semiparametric DiD estimator for the ATT fully exploits the empirical content of Assumptions 1 and 2.⁴ Hereafter, let $\tau_{dr}(Y, X) = Y - (m_{1,0}(X) + (m_{0,1}(X) - m_{0,0}(X)))$ and $W = (Y, D, X, T)$. We also denote the ATT by τ .

Theorem 1 (Semiparametric efficiency bound) Suppose Assumptions 1 and 2 hold. Then, the EIF for τ is given by

$$\eta_{\text{eff}}(W) = w_{1,1}(D, T)(\tau_{dr}(Y, X) - \tau) + \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} w_{d,t}(D, T, X)(Y - m_{d,t}(X)), \quad (2.5)$$

where the weights are defined in (2.4). Furthermore, the semiparametric efficiency bound for the set of all regular estimators of τ is

$$\mathbb{E}[\eta_{\text{eff}}(W)^2] = \frac{1}{\mathbb{E}[DT]^2} \mathbb{E} \left[DT(\tau_{dr}(Y, X) - \tau)^2 + \sum_{(d,t) \in \mathcal{S}_-} \frac{I_{d,t} \cdot p(1, 1, X)^2}{p(d, t, X)^2} (Y - m_{d,t}(X))^2 \right].$$

Apart from providing an efficiency benchmark, Theorem 1 also provides us a template to construct efficient estimators for τ . That is, given that any influence function has a mean of zero, we can take the expected value of $\eta_{\text{eff}}(W)$ and isolate τ to get the following estimand for the ATT

$$\tau = \tau_{dr} \equiv \mathbb{E} \left[w_{1,1}(D, T)\tau_{dr}(Y, X) + \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} w_{d,t}(D, T, X)(Y - m_{d,t}(X)) \right]. \quad (2.6)$$

We discuss in Section 3 how to leverage (2.6) to nonparametrically estimate τ under mild assumptions.

2.3 Bias-Variance trade-off with respect to stationarity

All the estimands described in Section 2.2 account for compositional changes over time. As discussed the introduction, most DiD estimators typically assume no compositional changes *a priori*. A natural question then arises: How biased would these estimators be when they erroneously rule out compositional changes?

To tackle this question, we examine the bias of the semiparametrically efficient DiD estimator for the ATT proposed by Sant’Anna and Zhao (2020) that rule out compositional changes. Before diving into this analysis, we need to introduce some additional notation and clarify the assumptions, estimands, and other aspects of Sant’Anna and Zhao (2020)’s approach.

First, Sant’Anna and Zhao (2020) explicitly rules out compositional changes by relying on the following stationarity assumption.

4 To simplify exposition, we abstract from additional technical discussions related to the conditions to guarantee quadratic mean differentiability and their implications for the precise definition of EIF; see, e.g., Chapter 3 of Bickel, Klaassen, Ritov and Wellner (1998) for more details.

Assumption 3 (Stationarity) $(D, X) \perp\!\!\!\perp T$.

Intuitively, Assumption 3 enables researchers to pool covariates and treatment variables from both periods. As a result, under Assumption 3, it follows that $\mathbb{E}[D|X, T = 1] = \mathbb{E}[D|X] \equiv \tilde{p}(X)$, which also affects the definition of the “relevant” propensity score. Indeed, as discussed in Sant’Anna and Zhao (2020), one can identify the ATT under Assumption 3 using the IPW estimand

$$\tau_{ipw,ncc} \equiv \mathbb{E} \left[\left(w_{1,1}^{sz}(D, T, X) - w_{1,0}^{sz}(D, T, X) - w_{0,1}^{sz}(D, T, X) + w_{0,0}^{sz}(D, T, X) \right) Y \right], \quad (2.7)$$

where, for $t = 0, 1$,

$$\begin{aligned} w_{1,t}^{sz}(D, T, X) &= \frac{D \cdot \mathbb{1}\{T = t\}}{\mathbb{E}[D \cdot \mathbb{1}\{T = t\}]}, \\ w_{0,t}^{sz}(D, T, X) &= \frac{\tilde{p}(X)(1 - D) \cdot \mathbb{1}\{T = t\}}{1 - \tilde{p}(X)} \bigg/ \mathbb{E} \left[\frac{\tilde{p}(X)(1 - D) \cdot \mathbb{1}\{T = t\}}{1 - \tilde{p}(X)} \right]; \end{aligned} \quad (2.8)$$

see also Abadie (2005). When one compares the IPW weights in (2.7) with the IPW weights in (2.3), it is clear that $w_{1,1}^{sz}(\cdot) = w_{1,1}(\cdot)$ whereas the remaining three IPW weights differ. Under Assumption 3, one only needs to consider binary propensity score models, and use these simpler functions to construct IPW weights for untreated units (in both periods) in (2.8). When one allows for compositional changes such that Assumption 3 is potentially violated, the distribution of the covariates may vary over time. Thus, one needs to consider generalized propensity scores, as now one has four groups depending on the treatment group and the period that a unit is observed. This affects the IPW weights in (2.4) not only for the untreated units, but also for the treated units observed only in the pre-treatment period. This discussion sheds light on how Assumption 3 “simplifies” the construction of IPW estimators.

We note that Assumption 3 also allows one to leverage more data than (2.2) when constructing an outcome regression estimand. More precisely, under Assumption 3, one can identify the ATT using the following regression-based estimand⁵

$$\tau_{or,ncc} \equiv \mathbb{E}[Y | D = 1, T = 1] - \mathbb{E}[m_{1,0}(X) + m_{0,1}(X) - m_{0,0}(X) | D = 1]; \quad (2.9)$$

see, e.g., Sant’Anna and Zhao (2020). To see the difference between (2.9) and (2.2), note that the last term in (2.9) integrates the covariates X using the distribution of treated units from both periods, i.e., the pooled treated distribution. This is valid under Assumption 3. When one allows for compositional changes over time, as in (2.2), one needs to integrate over the covariate distribution among treated units observed in the post-treatment period only.

Just like in Section 2.2, one can combine $\tau_{or,ncc}$ and $\tau_{ipw,ncc}$ to get more efficient estimators under Assumption 3. Indeed, Sant’Anna and Zhao (2020) show that, under Assumptions 1, 2,

⁵ Under Assumption 3, one can also use the alternative regression-based estimand that leverages more data than $\tau_{or,ncc}$ defined as $\tau_{or,ncc,2} \equiv \mathbb{E}[m_{1,1}(X) - m_{1,0}(X) - m_{0,1}(X) + m_{0,0}(X) | D = 1]$. We do not discuss this further as we directly address estimators based on the EIF in (2.11).

and 3, the EIF for the ATT is given by

$$\eta_{sz}(W) = \frac{D}{\mathbb{E}[D]} \left(\tau(X) - \tau \right) + \sum_{(d,t) \in \mathcal{S}} (-1)^{(d+t)} w_{d,t}^{sz}(D, T, X) (Y - m_{d,t}(X)), \quad (2.10)$$

where $\tau(x) = (m_{1,1}(x) - m_{1,0}(x)) - (m_{0,1}(x) - m_{0,0}(x))$ is the conditional ATT. Based on (2.10), Sant'Anna and Zhao (2020) propose the following DR estimand for the ATT,

$$\tau_{sz} \equiv \mathbb{E} \left[\frac{D}{\mathbb{E}[D]} \tau(X) + \sum_{(d,t) \in \mathcal{S}} (-1)^{(d+t)} w_{d,t}^{sz}(D, T, X) (Y - m_{d,t}(X)) \right]. \quad (2.11)$$

Heuristically, τ_{sz} can be viewed as the analog of τ_{dr} when, on top of the identification assumptions, one assumes that there are no compositional changes. Compared to τ_{dr} , it is based on binary propensity scores and integrates the conditional ATT using data from both time periods.

The following proposition shows that τ_{sz} does not recover the ATT when Assumption 3 is potentially violated, i.e., under compositional changes. It also precisely quantifies the bias relative to τ_{sz} .

Proposition 1 Under Assumptions 1 and 2, we have that

$$\begin{aligned} \tau_{sz} - \tau_{dr} &= \sum_{(d,t) \in \mathcal{S}} (-1)^{(d+t)} \mathbb{E} \left[\left(\frac{D}{\mathbb{E}[D]} - \frac{DT}{\mathbb{E}[DT]} \right) m_{d,t}(X) \right] \\ &\quad + \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} \mathbb{E} \left[(w_{d,t}^{sz}(D, T, X) - w_{d,t}(D, T, X)) (Y - m_{d,t}(X)) \right] \\ &= \mathbb{E}[\tau(X)|D = 1] - \mathbb{E}[\tau(X)|D = 1, T = 1] \\ &= \mathbb{E}[\tau(X)|D = 1] - \tau. \end{aligned}$$

Proposition 1 provides bias decomposition for τ_{sz} when the stationarity assumption is not imposed. The first equality in Proposition 1 follows from a direct comparison between our proposed estimand for the ATT and the one proposed by Sant'Anna and Zhao (2020), while the second equality is a consequence of the law of iterated expectations.⁶ The third equality is due to the definition of ATT and Assumptions 1 and 2. These calculations show that Sant'Anna and Zhao (2020)'s DR DiD estimand for the ATT can be biased when Assumption 3 is violated. In contrast, our proposed estimand τ_{dr} is fully robust against compositional changes.

Proposition 1 also highlights that not all violations of Assumption 3 result in biases in ATT when using Sant'Anna and Zhao (2020)'s estimand. Although intuitive and simple, this insight seems to be new in the literature. Based on this observation, one can determine if violations of Assumption 3 lead to empirically relevant biases in the ATT by comparing nonparametric estimates based on τ_{sz} with those based on our proposed estimand τ_{dr} . This would detect only

⁶ Here, we are implicitly considering the case where there are no (global) model misspecifications, which aligns with the fully nonparametric approach we adopt. One can compute a similar bias decomposition when one adopts parametric working models for the nuisance functions, though the notation becomes much more cumbersome.

the “relevant” violations of Assumption 3 that affect the target parameter of interest. That is, it would concentrate power in the directions that one cares about in this context. We discuss this testing procedure in greater detail in Section 4.

At this point, one may also wonder what the price one pays for such robustness in terms of semiparametric efficiency. Specifically, how much efficiency one loses by using τ_{dr} when Assumption 3 holds but is not fully exploited. The next proposition compares the semiparametric efficiency bound derived in Theorem 1 with the one derived by Sant’Anna and Zhao (2020).

Proposition 2 (Efficiency loss under stationarity) Suppose that Assumptions 1, 2, and 3 hold. Then

$$\rho_{sz} \equiv \mathbb{E}[\eta_{\text{eff}}(W)^2] - \mathbb{E}[\eta_{sz}(W)^2] = \frac{1 - \mathbb{E}[T]}{\mathbb{E}[D] \mathbb{E}[T]} \mathbb{V}\text{ar}[\tau(X)|D = 1], \quad (2.12)$$

It is evident from Proposition 2 that our proposed estimator is asymptotically less efficient than the one proposed by Sant’Anna and Zhao (2020) when there are no compositional changes over time. The efficiency loss is greater if any of the following three quantities is larger: 1) the population ratio of the pre-treatment period vs. the post-treatment period, 2) the population proportion of the comparison group vs. the treated group, and 3) the expected variability of treatment effect heterogeneity among the treated. In the extreme case where the treatment effect on the treated is homogeneous, our ATT estimator would achieve the same efficiency level as the one that imposes stationarity *a priori*. However, we imagine this case is not very realistic.

Propositions 1 and 2 characterize a bias-variance trade-off. Although our proposed estimand for the ATT is robust against Assumption 3, there is an asymptotic efficiency loss of not exploiting Assumption 3 when it does hold. We revisit this trade-off in Section 4.

3 Estimation and inference

The results from Section 2.2 suggest one can estimate the ATT by building on the EIF derived in Theorem 1, as emphasized by (2.6). The results from Propositions 1 and 2 also suggest a testing procedure to assess whether compositional changes translate to biased ATT estimates. However, all the discussions so far have involved estimands that depend on unknown nuisance functions, and we have not yet discussed how one can estimate these to form feasible two-step estimators. This section discusses how to proceed when adopting a fully nonparametric approach, avoiding additional functional form assumptions.

We first present a generic result that emphasizes that estimators based on (2.6) possess a *rate* DR property, regardless of how the nuisance functions are estimated nonparametrically. Since we employ a nonparametric estimation procedure, model misspecifications are not a primary concern, at least asymptotically. This suggests that the traditional notion of DR estimators that leverage potentially misspecified parametric working models, is unsuitable for our procedure. However, we recognize that different nonparametric estimators may exhibit varying convergence

rates for the nuisance parameters. Instead of requiring that both nonparametric models for the nuisance parameters converge to their true values sufficiently fast (i.e., the differences are $o_p(n^{-1/4})$), our rate DR property relies on weaker conditions that allow for a trade-off between the convergence rates of the two nuisance estimators. Heuristically, the rate DR property implies that nonparametric DiD estimators for the ATT based on the EIF are $\sqrt{n}$ -consistent and asymptotically normal even if one of the outcome regression or generalized propensity score functions is very complex provided that the other is sufficiently simple. For more discussions on rate DR, see, e.g., Kennedy (2016, 2023), Smucler et al. (2019), Rotnitzky, Smucler and Robins (2021), Jordan, Wang and Zhou (2022), and Bonvini, Kennedy, Dukes and Balakrishnan (2024).

Although interesting and useful, this generic rate DR is agnostic about the choice of the nonparametric estimator for the nuisance functions, and, therefore, does not help us with practical inference procedures. Towards that end, we discuss how one can concretely estimate the generalized propensity score (PS) and outcome regression (OR) nuisance functions using local polynomials, even in the presence of discrete covariates. We then establish the large sample properties of our DR DiD two-step estimator for the ATT based on local polynomials. We provide a data-driven bandwidth selection method in Subsection 3.4. We defer the construction of the Hausman-type test for compositional changes to Section 4.

3.1 Rate doubly robust

Let $\hat{p}$, and $\hat{m}_{d,t}$ be generic estimators of p , and $m_{d,t}$, for $(d, t) \in \mathcal{S}_-$. Given these first-step estimators, our proposed two-step estimator for the ATT based on (2.6) is given by

$$\hat{\tau}_{dr} = \mathbb{E}_n \left[\hat{w}_{1,1}(D, T) \hat{\tau}(Y, X) + \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} \hat{w}_{d,t}(D, T, X) (Y - \hat{m}_{d,t}(X)) \right], \quad (3.1)$$

where $\hat{\tau}(Y, X) = Y - (\hat{m}_{1,0}(X) + (\hat{m}_{0,1}(X) - \hat{m}_{0,0}(X)))$, and, for $(d, t) \in \mathcal{S}_-$,

$$\hat{w}_{1,1}(D, T) = \frac{DT}{\mathbb{E}_n[DT]}, \quad (3.2)$$

$$\hat{w}_{d,t}(D, T, X) = \frac{I_{d,t} \cdot \hat{p}(1, 1, X)}{\hat{p}(d, t, X)} \bigg/ \mathbb{E}_n \left[\frac{I_{d,t} \cdot \hat{p}(1, 1, X)}{\hat{p}(d, t, X)} \right]. \quad (3.3)$$

We impose the following assumptions on the quality of nuisance function estimators. We let $\|f\|_{L_2} \equiv (\int f^2 d\mu)^{1/2}$ and $\|f\|_\infty \equiv \sup_{x \in \mathcal{X}} |f(x)|$ denote the L_2 - and sup-norm of a function f , respectively, and let $\mathbb{G}_n(\cdot)$ denote the empirical process $\sqrt{n}(\mathbb{E}_n - \mathbb{E})(\cdot)$.

Assumption 4 (Estimation of nuisance parameters)

1. The estimators $\hat{p}$ and $\hat{m}$ are uniformly convergent in the sense that

$$\|\hat{p}(\cdot, \cdot, \cdot) - p(\cdot, \cdot, \cdot)\|_\infty = o_p(1), \quad \max_{(d,t) \in \mathcal{S}_-} \|\hat{m}_{d,t}(\cdot) - m_{d,t}(\cdot)\|_\infty = o_p(1).$$

2. For $(d, t) \in \mathcal{S}_-$,

$$\begin{aligned}
(i) \quad & \mathbb{E}_n[(Y - m_{d,t}(X)) \cdot (\hat{w}_{d,t} - w_{d,t})(W)] = o_p(n^{-1/2}). \\
(ii) \quad & \mathbb{E}_n[(w_{1,1} - w_{d,t})(W) \cdot (\hat{m}_{d,t} - m_{d,t})(X)] = o_p(n^{-1/2}). \\
(iii) \quad & \mathbb{G}_n \left\{ I_{d,t} \cdot \left(\frac{\hat{p}(1, 1, X)}{\hat{p}(d, t, X)} - \frac{p(1, 1, X)}{p(d, t, X)} \right) \cdot (\hat{m}_{d,t} - m_{d,t})(X) \right\} = o_p(1). \\
(iv) \quad & \mathbb{G}_n [w_{d,t}(W) \cdot (\hat{m}_{d,t} - m_{d,t})(X)] = o_p(1). \\
(v) \quad & \mathbb{G}_n \left[I_{d,t} \cdot \left(\frac{\hat{p}(1, 1, X)}{\hat{p}(d, t, X)} - \frac{p(1, 1, X)}{p(d, t, X)} \right) \right] = o_p(1).
\end{aligned}$$

3. $\text{Var}[Y|D = 1, T = 1] < \infty$ and $\text{Var}[Y|D = \cdot, T = t, X = x]$ is bounded uniformly over $\mathcal{X}$, for $(d, t) \in \mathcal{S}_-$.

These high-level conditions ensure that the functional space containing the first-stage estimation error does not become excessively complex (see, e.g., Kennedy, Ma, McHugh and Small (2017)). The verification of these conditions depends on the type of first-step estimators employed. In our paper, we utilized leave-one-out kernel estimators and relied on arguments based on U-statistics to verify the conditions. Cross-fitting is another popular approach, particularly due to its compatibility with double machine learning estimators and high-dimensional feature spaces. The assumptions can be significantly simplified when using cross-fitting estimators. A detailed explanation of the assumption simplifications and their implications for the error rate is provided in Appendix 7.1.

Let $(r_n)_{n \geq 1}$ and $(s_n)_{n \geq 1}$ be positive sequences converging to zero such that

$$\begin{aligned}
\max_{(d,t) \in \mathcal{S}} \|\hat{p}(d, t, \cdot) - p(d, t, \cdot)\|_e &= O_p(r_n), \\
\max_{(d,t) \in \mathcal{S}_-} \|\hat{m}_{d,t}(\cdot) - m_{d,t}(\cdot)\|_e &= O_p(s_n),
\end{aligned}$$

where $e = L_2$ or ∞ .

Lemma 3.1 (Doubly Robust error rate with generic first step estimators) Suppose that $e = \infty$, and that Assumptions 1, 2, 4.1 4.2(i, ii), and 4.3 are satisfied. Then,

$$\hat{\tau}_{dr} - \tau = \frac{1}{n} \sum_{i=1}^n \eta_{\text{eff}}(W_i) + O_p(r_n s_n) + o_p(n^{-1/2}). \quad (3.4)$$

Furthermore, if Assumptions 4.2(iii)-(v) are also fulfilled, the equation (3.4) remains valid when $e = L_2$.

The lemma demonstrates that our estimator is doubly robust in terms of its convergence rate. The remaining term is the product of the error rates of the first-stage estimators. Due to the product structure, each estimator typically needs only to converge to its true value at a rate of $o(n^{-1/4})$ for the ATT estimator to converge at the parametric rate. This property also allows for a trade-off between precision in the two nuisance estimators.

In the following subsection, we present lower-level conditions for cases in which the nuisance functions are estimated nonparametrically using leave-one-out local polynomial estimators. This technique enables us to directly establish the conditions in Assumption 2 without relying on empirical process theory. This is desirable, as verifying the complexity of the space where local polynomial (logistic) estimators reside is not a trivial task.

3.2 Local polynomial estimation of nuisance functions

We first introduce the estimator for the PS functions. Conditional probability functions are naturally bounded within the unit interval. However, these bounds may not be respected when using linear probability models. As a nonparametric generalization of parametric multinomial logit regression, local multinomial logit regression enforces such bounds by design. Through extensive Monte Carlo simulations, Frölich (2006) demonstrates that the local multinomial logit estimator consistently outperforms local least squares, Klein–Spady, and Nadaraya–Watson estimators. Hence, we prefer this estimator over other nonparametric methods.

Without loss of generality, the generalized PS can be represented by a multinomial logistic transformation applied to a set of unknown functions $\{g_{d,t}(\cdot)\}_{(d,t) \in \mathcal{S}_-}$ as follows

$$p(d, t, x) = \frac{\exp(g_{d,t}(x))}{1 + \sum_{(d',t') \in \mathcal{S}_-} \exp(g_{d',t'}(x))},$$

for $(d, t) \in \mathcal{S}_-$, and $p(1, 1, x) = \left(1 + \sum_{(d',t') \in \mathcal{S}_-} \exp(g_{d',t'}(x))\right)^{-1}$. The representation is well-defined as long as the overlapping condition in Assumption 2 (iii) holds. Instead of imposing specific functional forms on $\{g_{d,t}(\cdot)\}_{(d,t) \in \mathcal{S}_-}$, the local multinomial logit estimator approximates these unknown functions locally using polynomials, which we will describe in detail below.

In accordance with the conventions of local polynomial estimation, we adopt the following notations as shorthand for common vector operators,

$$\mathbf{k} = (k_1, \dots, k_v), \quad |\mathbf{k}| = \sum_{\ell=1}^v k_\ell, \quad \mathbf{k}! = \prod_{\ell=1}^v k_\ell!, \quad x^{\mathbf{k}} = \prod_{\ell=1}^v x_\ell^{k_\ell},$$

$$f^{(\mathbf{k})}(x) = \frac{\partial^{\mathbf{k}} f(x)}{\partial x_1^{k_1} \cdot \partial x_2^{k_2} \cdots \partial x_v^{k_v}}, \quad \sum_{0 \leq |\mathbf{k}| \leq p} = \sum_{\ell=0}^p \sum_{k_1=0}^{\ell} \cdots \sum_{\substack{k_v=0 \\ k_1+\dots+k_v=\ell}}^{\ell}.$$

Furthermore, we define $n_k = \binom{k+\ell-1}{\ell-1}$ as the number of distinct ℓ -tuples $\mathbf{k}$ with $|\mathbf{k}| = k$. We arrange these n_k ℓ -tuples in a lexicographically-ordered sequence, prioritizing the last position, and denote the mapping from the rank in the ordered sequence to the corresponding ℓ -tuple as $\pi_k(\cdot)$.

Since our method accommodates both discrete and continuous covariates, we must distinguish between these types of variables. We assume that $x = (x_c, x_d)$, where x_c is a v_c -vector of continuous covariates, and x_d is the subvector of discrete variables. We also distinguish between ordered and unordered discrete variables. That is, $x_d = (x_u, x_o)$, where x_u is a v_u -vector of

unordered covariates and x_o is a v_o -vector of ordered covariates.

Now, for a generic function, $g : \mathcal{X} \rightarrow \mathbb{R}$, and a point, $x^* \in \mathcal{X}$, $g(\cdot)$ can be approximated in a neighborhood of x^* by a p -th order Taylor series with respect to the continuous variables, as

$$g(x) \approx \sum_{0 \leq |\mathbf{k}| \leq p} \frac{1}{\mathbf{k}!} g^{(\mathbf{k})}(x^*) (x_c - x_c^*)^{\mathbf{k}} = \underline{\mathbf{X}}(x_c^*)' \gamma_g(x^*),$$

where $\underline{\mathbf{X}}(x_c) = (\underline{\mathbf{X}}^{(0)'}(x_c), \dots, \underline{\mathbf{X}}^{(p)'}(x_c))'$ is a $N_p \times 1$ vector that contains the sorted $(X_c - x_c)^{\mathbf{k}}$, with $N_p \equiv \sum_{k=0}^p n_k$. The l -th entry of $\underline{\mathbf{X}}^{(k)}(x_c)$, denoted as $\underline{\mathbf{X}}^{(k,l)}(x_c)$, is equal to $(X_c - x_c)^{\pi_k(l)}$. The vector $\gamma_g(x) = (\gamma_g^{(0)'}(x), \dots, \gamma_g^{(p)'}(x))'$ is defined as the vector of lexicographically-ordered $g^{(\mathbf{k})}(x)/\mathbf{k}!$.

The local approximation is achieved through kernel smoothing. For continuous variables, we let the kernel function be denoted by $K^j(\mathbf{u})$, $j = ps, or$. It is a nonnegative function supported on $[-1, 1]^{v_c}$. Suppose $h > 0$ is a generic bandwidth parameter. We denote the scaled kernel function by $K_h(\mathbf{u}) = K(\mathbf{u}/h)/h^{v_c}$. We use the kernel function proposed by Li and Racine (2007) for discrete variables. This kernel function is defined as

$$L_\lambda(x_d, z_d) = \prod_{s=1}^{v_u} \lambda_u^{\mathbb{1}_{\{x_{u,s} \neq z_{u,s}\}}} \prod_{s=1}^{v_o} \lambda_o^{|x_{o,s} - z_{o,s}|}, \quad (3.5)$$

where $\lambda = (\lambda_u, \lambda_o) \in [0, 1]^2$ is a generic smoothing parameter. When $\lambda = 0$, the estimator reduces to the frequency estimator.⁷

For the j -th observation of covariates, X_j , our local polynomial (multinomial) logit estimator of γ , denoted by $\hat{\gamma}$, satisfies

$$\hat{\gamma}(X_j) \equiv (\hat{\gamma}'_{1,0}(X_j), \hat{\gamma}'_{0,1}(X_j), \hat{\gamma}'_{0,0}(X_j))' = \arg \max_{\gamma \in \mathbb{R}^{3N_p}} \frac{1}{n-1} \sum_{i \neq j}^n \ell(W_i, X_j; \gamma) \tilde{K}_{ps}(X_i; X_j, h, \lambda), \quad (3.6)$$

where $\tilde{K}_{ps}(X_i; X_j, h, \lambda) = K_h^{ps}(X_{c,i} - X_{c,j}) L_\lambda(X_d, X_{d,j})$ and the local likelihood function $\ell(w, x; \gamma)$ is defined as

$$\ell(w, x; \gamma) = \sum_{(d', t') \in \mathcal{S}_-} I_{d,t} \underline{\mathbf{X}}_p(x_c)' \gamma_{d,t} - \log \left(1 + \sum_{(d', t') \in \mathcal{S}_-} \exp(\underline{\mathbf{X}}_p(x_c)' \gamma_{d', t'}) \right).$$

Note that we have used a “leave-one-out” version of the local regression estimator for the construction of $\hat{\gamma}$, i.e., $\gamma(X_j)$ are estimated using every observation except the j -th. This technique, standard in the literature (Powell and Stoker, 1996; Powell, Stock and Stoker, 1989; Rothe and Firpo, 2019), serves to avoid a “leave-in” bias that is of first-order importance when estimating the ATT.

Let $e_{\ell,k}$ denote an ℓ -dimensional vector in which the k -th element is set to one, while all

⁷ Here, we adopt the convention that $0^0 = 1$, ensuring the estimator remains well-defined even when $x_d = z_d$.

remaining elements are zero. Then, for a given $\hat{\gamma}$, the generalized PS can be approximated by⁸

$$\hat{p}(d, t, x) = \frac{\exp(e'_{N_p,1} \hat{\gamma}_{d,t}(x))}{1 + \sum_{(d',t') \in \mathcal{S}_-} \exp(e'_{N_p,1} \hat{\gamma}_{d',t'}(x))}, \quad (3.7)$$

for $(d, t) \in \mathcal{S}_-$, and $\hat{p}(1, 1, x) = 1 - \sum_{(d,t) \in \mathcal{S}_-} \hat{p}(d, t, x)$.

For OR models, we employ leave-one-out q -th order local polynomial least squares estimators. First, the local polynomial regression coefficients are estimated by solving the following equation:

$$\hat{\beta}_{d,t}(X_j) = \arg \min_{\beta \in \mathbb{R}^{N_p}} \frac{1}{n-1} \sum_{i \neq j}^n (Y_i - \mathbf{x}_{q,i}(X_{c,j})' \beta)^2 I_{d,t,i} \tilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}), \quad (3.8)$$

where $\tilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}) = K_{b_{d,t}}^{or}(X_{c,i} - X_{c,j}) L_{\vartheta_{d,t}}(X_d, X_{d,j})$, and $I_{d,t,i} = \mathbb{1}\{D_i = d, T_i = t\}$. Then, we estimate the OR functions by

$$\hat{m}_{d,t}(X_j) = e'_{N_q,1} \hat{\beta}_{d,t}(X_j), \quad (3.9)$$

for $(d, t) \in \mathcal{S}_-$.

We analyze the asymptotic behaviors of these local polynomial estimators in Appendix B. We provide results on the uniform convergence rate for the approximation error. In particular, we establish a uniform stochastic expansion for the local multinomial logit regression that is of independent interest.

Remark 1 The choice of polynomial order depends on considerations such as computational tractability and the trade-off between bias and variance properties. We adhere to the recommendation made by Fan et al. (1995) to employ odd-degree polynomial fits, as they simplify the analysis for the boundary bias when using symmetric kernel functions. We allow varying local polynomial orders for the PS and OR estimators and, in the case of the latter, for distinct treatment groups. This flexibility is desirable as the propensity score and conditional mean functions might display varying degrees of smoothness.

3.3 Asymptotic normality

With $\{\hat{m}_{d,t}\}_{(d,t) \in \mathcal{S}_-}$ given in (3.9), and $\hat{p}$ defined in (3.7), we can construct an estimator for τ_{dr} as shown in (3.1). In the following, we derive the large sample properties of the estimator $\hat{\tau}_{dr}$ by applying Lemma 3.1. To achieve this objective, we begin by presenting a set of regularity assumptions. Henceforth, we use $\mathcal{B}(x, \delta)$ to denote a ball centered at x with radius δ , and $\lambda_{\min}(A)$ to represent the smallest eigenvalue of a square matrix A .

Assumption 5 (Support, smoothness, integrability, kernel, and bandwidth conditions)

⁸ We abuse notation and denote the local polynomial estimators for the generalized propensity score as $\hat{p}$ and for the outcome regression as $\hat{m}$, which are the same as the generic estimators introduced in Section 3.1.

1. (i) $\mathcal{X} = \mathcal{X}_c \otimes \mathcal{X}_d$, where $\mathcal{X}_c$ is a compact subset of $\mathbb{R}^{v_c}$ and $\mathcal{X}_d$ is finite; (ii) For all $x_d \in \mathcal{X}_d$, $\mathbb{P}(X_d = x_d) > 0$, and the conditional probability density of X_c , $f_{X_c|X_d}(\cdot|x_d)$, is continuously differentiable and bounded away from zero on $\mathcal{X}_c$; (iii) There are positive constants κ_0 and κ_1 in $(0, 1]$ such that for any $x \in \mathcal{X}$ and all $\epsilon \in (0, \kappa_0]$, there exists a $x' \in \mathcal{X}$ satisfying, $x'_d = x_d$, and

$$\mathcal{B}(x', \kappa_1 \epsilon) \subset \mathcal{B}(x, \epsilon) \cap \mathcal{X}.$$

2. For all $x \in \mathcal{X}$, (i) $p(d, t, x)$ is $(p+1)$ -times continuously differentiable in x_c , with uniformly bounded derivatives, for $(d, t) \in \mathcal{S}$; (ii) $m_{d,t}(x)$ is $(q+1)$ -times continuously differentiable in x_c , with uniformly bounded derivatives, for $(d, t) \in \mathcal{S}_-$.
3. $\mathbb{E}[|Y|^\zeta | X, D, T] < \infty$ a.s. for some constant $\zeta > 2$.
4. For $j = ps$, or, (i) $K^j : [-1, 1]^{v_c} \rightarrow \mathbb{R}_+$; (ii) $K^j(\cdot)$ satisfies the Lipschitz condition, i.e. $|K^j(\mathbf{u}) - K^j(\mathbf{u}')| \leq L \|\mathbf{u} - \mathbf{u}'\|$ for some $L > 0$ and any $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^d$.
5. (i) $h = o(1)$; (ii) $\log n / (nh^{v_c+2p}) = o(1)$ and $\lambda/h^p = o(1)$; (iii) $h^{p+1} = o(n^{-1/4})$ and $\log n / (nh^{v_c}) = o(n^{-1/2})$. For $(d, t) \in \mathcal{S}_-$, (iv) $b_{d,t} = o(1)$; (v) $\log n / (n^{1-2/\zeta} b_{d,t}^{v_c}) = o(1)$; (vi) $b_{d,t}^{q+1} = o(n^{-1/4})$ and $\log n / (nb_{d,t}^{v_c}) = o(n^{-1/2})$; (vii) $\lambda, \vartheta_{d,t} = o(n^{-1/4})$.
6. With $\mathbf{Q}_j(x_c)$ defined in (B.6), $\inf_{x_c \in \mathcal{X}_c} \lambda_{\min}(\mathbf{Q}_j(x_c)) > 0$, for $j = p, q$.

A few remarks on the assumptions are in order. Assumption 5.1 indicates that our local polynomial estimator can handle discrete, categorical data. The final part of the condition, proposed by Fan and Guerre (2016), requires that the boundary of $\mathcal{X}$ is sufficiently dense for the first-stage estimators to exhibit good bias and variance properties near the boundary. Importantly, this has to hold for all covariates.⁹ Assumption 5.2 describes the standard smoothness condition for the nuisance functions. Assumption 5.3 is a regularity condition that controls the conditional moments of Y . Assumption 5.4 collects the regularity conditions on the kernel functions. We note that different kernels can be used for the propensity score and conditional mean models. In practice, the kernel $K(\cdot)$ typically takes a product form, that is, $K(\mathbf{u}) = \prod_{i=1}^{v_c} \mathcal{K}(u_i)$, where $\mathcal{K}(\cdot)$ can be chosen from several options, such as triangular, biweight, triweight, or Epanechnikov kernels. However, the Gaussian kernel is ruled out due to the restriction on compact support. Assumption 5.5 compiles the rate condition on the bandwidths. Assumptions 5.5 (ii) and (v) are imposed to ensure linear expansions of the local polynomial estimators hold uniformly over $\mathcal{X}$. When Y has finite moments of any order, such as when it has bounded support, Assumption 5.5 (v) is implied by Assumption 5.5 (vi). Assumptions 5.5 (iii), (vi), and (vii) specify rate conditions on the bias and stochastic part of the first-step estimation error.

⁹ If this denseness condition does not hold in practice—for example, if the data is very sparse near the boundary of covariates—it may be necessary to restrict the analysis to interior points and impose trimming when estimating treatment effects.

It is important to note that our estimator builds on the efficient influence function and, therefore, inherits a rate DR property. Without such a DR property, it would typically require more stringent rate conditions on the bias part, which can only be satisfied with higher-order kernel functions. Heuristically, this follows because one needs to ensure that the nonparametric estimator converges fast enough. See, for example, Newey (1994) and Lee (2018) for detailed discussion. The rate DR property relax this condition.

Remark 2 Rothe and Firpo (2019) provides a result that can be applied to weaken the rate conditions on the nuisance functions. They present higher-order expansions of semiparametric two-step DR estimators, demonstrating that if the first-step error's bias and the stochastic components are of order $o_p(n^{-1/6})$, and their product is of order $o_p(n^{-1/2})$, the resulting DR estimator achieves root- n consistency. To maintain focus, we will not delve into an in-depth discussion on this topic.

Theorem 2 (Asymptotic normality doubly robust estimator) Under Assumptions 1, 2, and 5, we have

$$\sqrt{n}(\hat{\tau}_{dr} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_{\text{eff}}(W_i) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \Omega_{dr}), \quad (3.10)$$

where $\Omega_{dr} = \mathbb{E}[\eta_{\text{eff}}(W)^2]$.

Theorem 2 states that $\hat{\tau}_{dr}$ is root- n consistent, and asymptotically normal. It also shows that the estimation error of the nuisance functions does not affect the asymptotic distribution of $\hat{\tau}_{dr}$. Furthermore, the asymptotic variance of $\hat{\tau}_{dr}$ is equal to the semiparametric efficiency bound.

The theorem can be applied to calculate confidence intervals for the ATT. To achieve this, we need an estimator of the asymptotic variance, Ω_{dr} . One approach to constructing such an estimator is by using empirical analogs of the influence function or through bootstrapping. Here, we focus on the first method, while a weighted bootstrap procedure that accommodates clustered inference is provided in Appendix C.6. Let

$$\hat{\eta}_{\text{eff}}(W) = \sum_{(d,t) \in \mathcal{S}_-} (-1)^{d+t} \hat{w}_{d,t}(D, T, X)(Y - \hat{m}_{d,t}(X)) + \hat{w}_{1,1}(D, T, X)(\hat{\tau}(Y, X) - \hat{\tau}_{dr}), \quad (3.11)$$

and $\hat{\Omega}_{dr} = \mathbb{E}_n[\hat{\eta}_{\text{eff}}(W)^2]$. Under mild regularity conditions, the consistency of $\hat{\Omega}_{dr}$ can be established, with its proof included in that of Theorem 3 presented in the following section.

3.4 Bandwidth selection

This subsection addresses the practical selection of bandwidth for the first-step local polynomial estimators. It is well-documented that smoothing parameters have a significant impact on balancing the trade-off between bias and variance. Although robustness checks employing multiple bandwidths can be useful, a reliable data-driven selection rule is often preferred. In the following, we outline two cross-validation procedures for choosing these tuning parameters.

Define the following two criterion functions

$$C_n^{ls}(h, \lambda, \{b_{d,t}, \vartheta_{d,t}\}_{(d,t) \in \mathcal{S}_-}) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{(d,t) \in \mathcal{S}} (I_{d,t,i} - \hat{p}(d, t, X_i))^2 + \sum_{(d,t) \in \mathcal{S}_-} I_{d,t,i} (Y_i - \hat{m}_{d,t}(X_i))^2 \right\}, \quad (3.12)$$

$$C_n^{ml}(h, \lambda, \{b_{d,t}, \vartheta_{d,t}\}_{(d,t) \in \mathcal{S}_-}) = \frac{1}{n} \sum_{i=1}^n \left\{ - \sum_{(d,t) \in \mathcal{S}} I_{d,t,i} \log(\hat{p}(d, t, X_i)) + \sum_{(d,t) \in \mathcal{S}_-} I_{d,t,i} (Y_i - \hat{m}_{d,t}(X_i))^2 \right\}. \quad (3.13)$$

The least-squares criterion, C_n^{ls} , is a standard choice in the kernel estimation literature. It is based on the sum of the least squares distances between the observed and leave-one-out fitted values for both PS and OR estimators. The second criterion, C_n^{ml} , replaces the PS estimator's least squares sum with that of the observed likelihood. This idea of using a likelihood-based criterion in local logistic estimation can be traced back to Staniswalis (1989).

The leave-one-out cross-validated bandwidths, $(\hat{h}^j, \hat{\lambda}^j, \{\hat{b}_{d,t}^j, \hat{\vartheta}_{d,t}^j\}_{(d,t) \in \mathcal{S}})$, minimizes C_n^j for $j = ls, ml$. In Appendix C.2, we analyze the mean integrated squared error (MISE) properties of the first-step estimators and derive the convergence rates of the optimal bandwidths. For local linear estimation (i.e., $p = q = 1$), Theorem C.1 shows that the optimal bandwidths ensure the rate conditions in Assumption 5.5 are satisfied if the number of continuous variables is less than 4. Notably, this result imposes no restrictions on the number of discrete variables.

Remark 3 Leave-one-out cross-validation can be computationally demanding when combined with local multinomial logit estimation. This is partly because, unlike local least squares regression, local multinomial logit regression does not have a closed-form solution. As a result, evaluating the criterion function requires solving n minimization problems, which can be time-consuming, especially for large datasets. To mitigate this computational burden, we propose using the rescaled cross-validation method introduced by Li, Li and Li (2021), as described in Appendix C.4. This method divides the data into training and validation sets and computes the multinomial logistic loss using only the validation data, significantly reducing the computation cost.

4 Testing for compositional changes

Propositions 1 and 2 reveal that our proposed estimator for the ATT is robust against compositional changes; however, it is less efficient than the DR DiD estimator proposed by Sant'Anna and Zhao (2020) when the covariate-stationarity assumption is correctly imposed. This trade-off suggests a nonparametric Hausman (1978)-type test for the absence of compositional changes can be constructed by comparing our proposed estimator with that of Sant'Anna and Zhao (2020). Although Sant'Anna and Zhao (2020) focus on parametric first-step estima-

tors for the nuisance parameters, we modestly extend their analysis by considering nonparametric first-step estimators in this section.

Before detailing the test construction, we define the null and alternative hypotheses. Let $\mathbf{P}_0 = \{P \in \mathcal{P} : \text{Assumptions 1, 2, and 3 hold, and } \text{Var}_P[\tau(X) \mid D = 1] > 0\}$ and $\mathbf{P}_1 = \{P \in \mathcal{P} : \text{Assumptions 1 and 2 hold, and } \tau_{dr} \neq \tau_{sz}\}$, where τ_{dr} and τ_{sz} are defined in (2.6) and (2.11), respectively. Here, we aim to test

$$H_{0,P} : P \in \mathbf{P}_0 \quad \text{against} \quad H_{1,P} : P \in \mathbf{P}_1.$$

We index the hypotheses with P to emphasize the point-wise nature of our test—namely, that it applies to a fixed data-generating distribution $P \in \mathcal{P}$. Under the null hypothesis, Sant’Anna and Zhao (2020)’s DR DiD estimand coincides with our proposed estimand, so that $\tau_{sz} = \tau_{dr}$. In this setting, both estimands identify the ATT, though their estimator (asymptotically) achieves the semiparametrically efficient bound. Under $H_{1,P}$, Sant’Anna and Zhao (2020)’s DR DiD estimator is not consistent for the ATT, while ours remains consistent and is (asymptotically) semiparametric efficient. Finally, we note that for distributions satisfying Assumptions 1 and 2, the alternative hypothesis space constitutes only a subset of the complement of the null hypothesis space. This occurs because the two ATT estimators can still coincide, even when the stationarity assumption is violated. We concentrate on this subset of hypotheses to test deviations where the stationarity assumption influences the target parameter of interest.

To operationalize this testing procedure without invoking additional parametric assumptions, we require a nonparametric estimator for τ_{sz} , which in turn necessitates nonparametric estimators for the PS $\tilde{p}(\cdot)$ and the OR functions $m_{d,t}(\cdot)$, $(d, t) \in \mathcal{S}$. For the PS, we can use the local polynomial estimators from Section 3.2 to construct an estimator for $\tilde{p}(\cdot)$ as

$$\hat{\tilde{p}}(X) = \hat{p}(1, 1, X) + \hat{p}(1, 0, X),$$

where $\hat{p}(1, t, X)$ is given by (3.7). We can estimate the OR $m_{d,t}(\cdot)$ as in (3.9), noting that all four conditional mean functions must be estimated here (unlike the three in Section 3). Using these, we nonparametrically estimate τ_{sz} by

$$\hat{\tau}_{sz} \equiv \mathbb{E}_n \left[\frac{D}{\mathbb{E}_n[D]} \hat{\tau}(X) + \sum_{(d,t) \in \mathcal{S}} (-1)^{(d+t)} \hat{w}_{d,t}^{sz}(D, T, X) (Y - \hat{m}_{d,t}(X)) \right]. \quad (4.1)$$

where $\hat{\tau}(x) = (\hat{m}_{1,1}(x) - \hat{m}_{1,0}(x)) - (\hat{m}_{0,1}(x) - \hat{m}_{0,0}(x))$, and, for $t = 0, 1$,

$$\begin{aligned} \hat{w}_{1,t}^{sz}(D, T, X) &= \frac{D \cdot \mathbb{1}\{T = t\}}{\mathbb{E}_n[D \cdot \mathbb{1}\{T = t\}]}, \\ \hat{w}_{0,t}^{sz}(D, T, X) &= \frac{\hat{\tilde{p}}(X) (1 - D) \cdot \mathbb{1}\{T = t\}}{1 - \hat{\tilde{p}}(X)} \bigg/ \mathbb{E}_n \left[\frac{\hat{\tilde{p}}(X) (1 - D) \cdot \mathbb{1}\{T = t\}}{1 - \hat{\tilde{p}}(X)} \right]. \end{aligned}$$

Given this nonparametric estimator for τ_{sz} and our nonparametric estimator for τ_{dr} in (3.1),

the test statistic is defined as

$$\mathcal{T}_n = n\hat{V}_n^{-1} (\hat{\tau}_{dr} - \hat{\tau}_{sz})^2, \quad (4.2)$$

where

$$\hat{V}_n \equiv \mathbb{E}_n [(\hat{\eta}_{\text{eff}}(W) - \hat{\eta}_{sz}(W))^2],$$

with $\hat{\eta}_{\text{eff}}(W)$ defined in (3.11) and

$$\hat{\eta}_{sz}(W) \equiv \frac{D}{\mathbb{E}_n[D]} (\hat{\tau}(X) - \hat{\tau}_{sz}) + \sum_{(d,t) \in \mathcal{S}} (-1)^{(d+t)} \hat{w}_{d,t}^{sz}(D, T, X) (Y - \hat{m}_{d,t}(X)). \quad (4.3)$$

$\hat{V}_n$ estimates the variance of the difference between the two DiD estimators for the ATT. While an alternative variance estimator could be constructed using the variances of each DiD estimator, i.e., $\tilde{V}_n = \hat{\Omega}_{dr} - \hat{\Omega}_{sz}$, with $\hat{\Omega}_{dr} = \mathbb{E}_n[\hat{\eta}_{\text{eff}}(W)^2]$ and $\hat{\Omega}_{sz} = \mathbb{E}_n[\hat{\eta}_{sz}(W)^2]$, this approach may yield negative variance estimates in finite samples. Using $\hat{V}_n$ avoids this drawback.

In the following theorem, we characterize the asymptotic behavior of this statistic. Let $c_{1-\alpha}^*$ denote the $(1 - \alpha)$ -th quantile of the chi-squared distribution with one degree of freedom (i.e. χ_1^2).

Theorem 3 Suppose Assumption 5, and in addition, Assumptions 5.2(ii) and 5.5(iv)-(vii) are fulfilled for $(d, t) = (1, 1)$, for $P \in \mathcal{P}$. Then:

(a) Under the null hypothesis, $H_{0,P}$, we have $\hat{V}_n \xrightarrow{p} \rho_{sz} > 0$, and

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{T}_n \geq c_{1-\alpha}^*) = \alpha; \quad (4.4)$$

(b) Under the alternative hypothesis, $H_{1,P}$, we have

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{T}_n \geq c_{1-\alpha}^*) = 1. \quad (4.5)$$

The theorem states that the test controls size and is consistent. Although not discussed in detail here, it is easy to show that our test also has power against sequences of Pitman-type local alternatives that converge to the null at the parametric rate.

Remark 4 It is crucial to recognize that our test should be viewed as a “model validation” instead of a “model selection” procedure. For researchers concerned about the validity of Assumption 3, it may be tempting to perform a two-stage test. In the first stage, a Hausman specification test is used to “pretest” for the presence of compositional changes, and then, in the second stage, the usual t -test is conducted based on either $\hat{\tau}_{dr}$ or $\hat{\tau}_{sz}$, depending on the outcome of the Hausman-test. However, as demonstrated by Guggenberger (2010a), Guggenberger (2010b), and Roth (2022), such a model-selection procedure can lead to substantial size distortions when using standard inference methods.

5 Monte Carlo simulation study

In this section, we examine the finite sample properties of our proposed estimators and testing procedure. We conduct two main Monte Carlo experiments in this section. In the first experiment, there are compositional changes over time, so Assumption 3 is violated. In contrast, the second experiment adheres to this assumption, maintaining a joint distribution of covariates and treatment that is independent of treatment timing. For each design, we compare our nonparametric DR DiD estimator $\hat{\tau}_{dr}$ defined in (3.1), which is robust against compositional changes and semiparametrically efficient, with the nonparametric extension of Sant’Anna and Zhao (2020)’s estimator $\hat{\tau}_{sz}$ defined in (4.1), which assumes no compositional change, and with the estimates of the regression coefficients, τ_{fe} , associated with two-way fixed effect (TWFE) regression specifications of the type

$$Y = \alpha_1 + \alpha_2 T + \alpha_3 D + \tau_{fe}(T \cdot D) + \theta'X + \epsilon.$$

We consider two TWFE specifications: 1) a linear specification, where all the covariates X enter linearly, and 2) a saturated specification, where, in addition to the linear terms, quadratic terms of the continuous covariates and all the interactive terms of the covariates are also included. We include the TWFE specifications in our comparison set as they are prominent in empirical work.

We employ local linear ($p, q = 1$) kernel estimators for both the PS and OR functions. As described in Section 3.2, the PS is estimated using the local likelihood method with the (multinomial) logistic link function, whereas the OR is estimated using the local least squares estimator. We utilize the second-order Epanechnikov kernel for the continuous covariates, and the kernel given in (3.5) for discrete variables. Bandwidth selection methods are explored in detail later in this section.

Our main experiments involve a sample size of $n = 1000$, with each design undergoing 5,000 Monte Carlo replications. We evaluate the DiD estimators for the ATT using various metrics: average bias, median bias, root mean square error (RMSE), empirical 95% coverage probability, the average length of a 95% confidence interval, and the average of the plug-in estimator for the asymptotic variance. Confidence intervals are calculated using a normal approximation, with asymptotic variances estimated by their sample analogues. We also compute the semiparametric efficiency bound for each design to gauge the potential loss of efficiency/accuracy associated with using inefficient DiD estimators for the ATT. We perform a Hausman-type test as described in Section 4 under each design and report the empirical rejection rates.

In addition to these two main experiments, we further investigate the power properties of our Hausman-type test by examining its performance under a sequence of local alternatives. Finally, we evaluate the performance of our estimator under different bandwidth selection methods, comparing leave-one-out cross-validation (LOOCV), rescaled cross-validation (RCV), and a plug-in estimator.

5.1 Simulation 1: non-stationary covariate distribution

We first consider a scenario in which the stationarity condition is not satisfied. The DGP is described in Appendix C.3. Under this design, the covariate distribution does not exhibit time variation. However, the PS function is different in the two cross-sections. The mean absolute difference between $p^{s1}(1, 1, X)$ and $p^{s1}(1, 0, X)$, as well as between $p^{s1}(0, 1, X)$ and $p^{s1}(0, 0, X)$, are both approximately 0.125, with the maximum difference reaching up to 0.63.¹⁰ Consequently, we expect all estimators to produce biased results, except for $\hat{\tau}_{dr}$. In addition, the stationarity test is likely to reject the null hypothesis with high probability. The results in Table 1 support these claims.

Table 1: Monte Carlo results under compositional changes. Sample size: $n = 1,000$.

True value of ATT: 4.31. Semiparametric Efficiency Bound: 1753.6							
Two-way Fixed Effect Estimators							
	Spec.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{fe}$	Linear	-10.437	-10.445	10.933	10425.033	0.121	12.633
$\hat{\tau}_{fe}$	Saturated	-11.176	-11.206	11.579	8797.289	0.045	11.612
Nonparametric Doubly Robust DiD Estimators for the ATT							
	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{dr}$	ML	-0.009	-0.010	1.374	1838.495	0.949	5.304
$\hat{\tau}_{dr}$	LS	-0.013	-0.010	1.379	1848.848	0.949	5.314
$\hat{\tau}_{sz}$	ML	4.427	4.436	4.543	983.436	0.009	3.884
$\hat{\tau}_{sz}$	LS	4.427	4.435	4.543	983.746	0.009	3.884
Hausman-type test							
	CV Crit.	Avg. Test Stats.	Emp. Pow. (0.10)	Emp. Pow. (0.05)	Emp. Pow. (0.01)		
	ML	21.250	0.998	0.996	0.978		
	LS	21.199	0.998	0.995	0.976		

Note: Simulations based on 5,000 Monte Carlo experiments. $\hat{\tau}_{fe}$ the TWFE regression estimator, $\hat{\tau}_{dr}$ is our proposed nonparametric DR DiD estimator (3.1), and $\hat{\tau}_{sz}$ is the nonparametric DR DiD estimator (4.1) based on Sant’Anna and Zhao (2020). For TWFE regression, we use a linear specification, “Linear”, and a saturated specification, “Saturated”. For DR DiD estimators, the PS and the OR models are estimated using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the PS function is selected with the log-likelihood criterion, “ML”, and the least squares criterion, “LS”, respectively. Lastly, “Spec.”, “CV Crit.”, “Avg. Bias”, “Med. Bias”, “RMSE”, “Asy. Var.”, “Cover.”, and “CIL”, stand for the specification, cross-validation criterion, average simulated bias, median simulated bias, simulated root-mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. The Hausman-type test statistic is calculated based on (4.2). Columns “Avg. Test Stats.”, and “Emp. Pow. (α)” stand for the average test statistic, and empirical power of the test with a nominal size α , respectively. See the main text for further details.

First, results in Table 1 suggest that both $\hat{\tau}_{fe}$ and $\hat{\tau}_{sz}$ are severely biased under this DGP, while $\hat{\tau}_{dr}$ exhibits negligible bias on average. Moreover, among the three sets of estimators considered, only our proposed estimator attains the correct coverage rate. This result is robust to the bandwidth selection method. Notably, the performance of the TWFE does not improve with a fully-saturated specification, indicating that incorporating nonlinear terms into a TWFE regression does not generally help in identifying heterogeneous treatment effects. In terms of efficiency, it is worth noting that the asymptotic variance of $\hat{\tau}_{dr}$ is close to the semiparametric efficiency bound, which corroborates the findings of Theorem 2. Regarding the testing performance, our Hausman-type test can effectively distinguish between the two nonparametric DiD estimators with a high degree of certainty, which is in line with our theoretical finding.

¹⁰ See Appendix C.3 for a definition of the p^{s1} functions.

5.2 Simulation 2: stationary covariate distribution

We now slightly adjust the first design by taking the average of propensity scores over time while keeping all other aspects of the DGP constant. Specifically, we define

$$p^{s2}(d, t, X) = \mathbb{P}^{s1}(T = t) \cdot (p^{s1}(d, 1, X) + p^{s1}(d, 0, X)),$$

where $\mathbb{P}^{s1}(T = t) = \mathbb{E}[p^{s1}(1, t, X) + p^{s1}(0, t, X)]$. The treatment groups are then assigned based on the realization of a standard uniform random variable on the unit interval partitioned by $\{p^{s2}(d, t, X)\}_{(d,t) \in \mathcal{S}}$. Furthermore, the potential outcomes are determined by (C.15)–(C.17). Unlike the first DGP, both the covariate distribution and the propensity score function are stationary in this case. As a result, we anticipate that both $\hat{\tau}_{dr}$ and $\hat{\tau}_{sz}$ will be consistent for the true ATT. Furthermore, the empirical rejection rate of the Hausman-type test is expected to converge to the nominal sizes. The Monte Carlo results under this DGP are summarized in Table 2.

Table 2: Monte Carlo results under no compositional changes. Sample size: $n = 1,000$.

True value of ATT: 9.13. Semiparametric Efficiency Bound: 796.8							
Two-way Fixed Effect Estimators							
	Spec.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{fe}$	Linear	-10.649	-10.672	11.106	9907.607	0.087	12.325
$\hat{\tau}_{fe}$	Saturated	-10.563	-10.617	10.946	7924.684	0.048	11.026
Nonparametric Doubly Robust DiD Estimators for the ATT							
	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{dr}$	ML	-0.007	-0.020	1.323	1721.037	0.946	5.133
$\hat{\tau}_{dr}$	LS	-0.010	-0.027	1.328	1732.416	0.946	5.139
$\hat{\tau}_{sz}$	ML	-0.015	-0.024	0.958	926.689	0.953	3.771
$\hat{\tau}_{sz}$	LS	-0.016	-0.024	0.958	926.821	0.953	3.771
Hausman-type test							
	CV Crit.	Avg. Test Stats.	Emp. Size (0.10)	Emp. Size (0.05)	Emp. Size (0.01)		
	ML	1.045	0.108	0.055	0.009		
	LS	1.045	0.107	0.056	0.009		

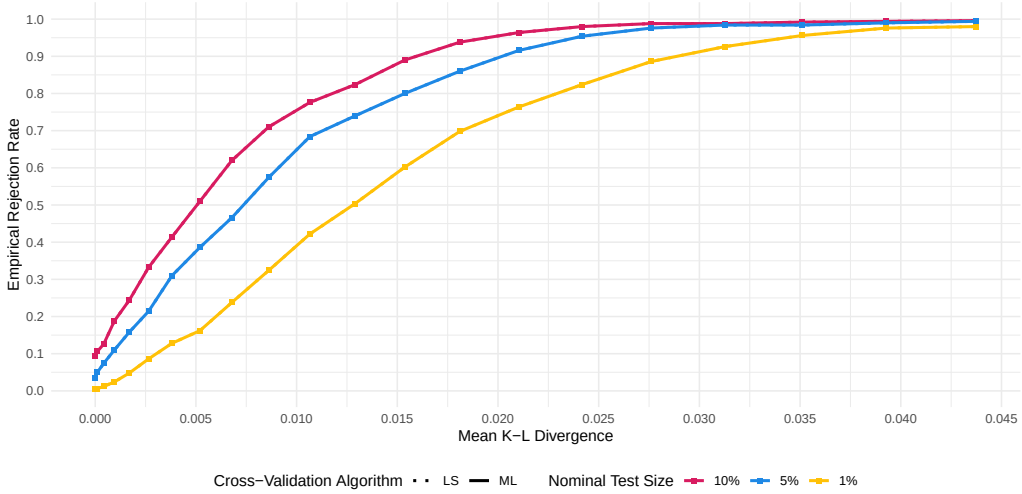
Note: Simulations based on 5,000 Monte Carlo experiments. $\hat{\tau}_{fe}$ the TWFE regression estimator, $\hat{\tau}_{dr}$ is our proposed nonparametric DR DiD estimator (3.1), and $\hat{\tau}_{sz}$ is the nonparametric DR DiD estimator (4.1) based on Sant’Anna and Zhao (2020). For TWFE regression, we use a linear specification, “Linear”, and a saturated specification, “Saturated”. For DR DiD estimators, the PS and the OR models are estimated using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the PS function is selected with the log-likelihood criterion, “ML”, and the least squares criterion, “LS”, respectively. Lastly, “Spec.”, “CV Crit.”, “Avg. Bias”, “Med. Bias”, “RMSE”, “Asy. Var.”, “Cover.”, and “CIL”, stand for the specification, cross-validation criterion, average simulated bias, median simulated bias, simulated root-mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. The Hausman-type test statistic is calculated based on (4.2). Columns “Avg. Test Stats.”, and “Emp. Size (α)” stand for the average test statistic, and empirical size of the test with a nominal size α , respectively. See the main text for further details.

In contrast to the results presented in Table 1, both $\hat{\tau}_{dr}$ and $\hat{\tau}_{sz}$ exhibit minimal bias, and their confidence intervals achieve nominal coverage. Their performance is consistently good across different bandwidth selection methods. The TWFE estimators, however, continue to show substantial bias and achieve nearly negligible coverage, despite having much wider confidence intervals compared to the DR DiD estimators. This occurs because the true treatment effects are heterogeneous, but TWFE specifications do not account for that (i.e., the models are misspecified). In terms of efficiency, the asymptotic variance of $\hat{\tau}_{sz}$ is reasonably close to the

semiparametric efficiency bound. The asymptotic variance of $\hat{\tau}_{dr}$ is, on average, 2.2 times larger than the semiparametric efficiency bound (that imposes no-compositional changes), which is still significantly lower than that of the TWFE estimators. Given that Assumption 3 holds for this DGP, the null hypothesis is true. The empirical rejection frequency of our Hausman-type test is nearly identical to its nominal value, highlighting the desirable properties of this testing procedure.

5.3 Test power

Figure 1: Power curves of the Hausman-type test. Sample Size: $n = 1,000$.



Notes: Simulations are based on 500 Monte Carlo experiments. The bandwidth for the PS function is selected using LOOCV with two criteria: the log-likelihood criterion (“ML”) and the least-squares criterion (“LS”). The “mean K-L divergence” refers to the population mean of the conditional Kullback-Leibler divergence between the propensity scores under the local alternatives and the null hypothesis.

In this subsection, we examine the power properties of the proposed Hausman-type test. Specifically, we consider a sequence of deviations from the null hypothesis, $\{H_{1,j}\}_{j=1}^{20}$, where the PS functions gradually transition from those in Simulation 2 towards those in Simulation 1. The propensity score functions under these alternatives are generated as follows:

$$p^{\delta_j^{alt}}(d, t, X) = \delta_j^{alt} p^{s1}(d, t, X) + (1 - \delta_j^{alt}) p^{s2}(d, t, X), \quad (5.1)$$

where $\delta_j^{alt} = 0.05 * j$, for $j = 1, \dots, 20$. All other aspects of the DGP remain the same as in Simulation 2. The deviation of the sequence of alternative PS functions can be quantified using the mean Kullback-Leibler (K-L) divergence between $p^{\delta^{alt}}$ and p^{s2} ,

$$D_{KL}(p^{\delta^{alt}} || p^{s2}) = \mathbb{E} \left[\sum_{(d,t) \in \mathcal{S}} p^{\delta^{alt}}(d, t, X) \log \left(\frac{p^{\delta^{alt}}(d, t, X)}{p^{s2}(d, t, X)} \right) \right]. \quad (5.2)$$

Figure 1 presents the power curves for the sequence of local alternatives. It shows that, even when the mean K-L divergence is as small as 0.015, the empirical rejection frequency of the 5% test is 0.79 (1%: 0.59, 10%: 0.88). Additionally, the empirical power does not vary based

on the cross-validation criterion used for the PS function: the curves are on top of each other, making them virtually indistinguishable. This demonstrates that the proposed test exhibits robust power properties, even under small deviations from the stationarity assumption.

5.4 Bandwidth choices

In this subsection, we present Monte Carlo simulation results comparing three different bandwidth selection methods: LOOCV (described in Section 3.4), RCV (detailed in Appendix C.4), and the plug-in estimator (described in Appendix C.5).

Table 3 shows that the plug-in bandwidth estimator, which is based on a frequency approach for discrete covariates, exhibits significant bias, elevated variance, and lower empirical test power compared to the other two methods. This is likely due to data sparsity in each stratum when multiple discrete covariates are present.

In contrast, LOOCV and RCV show no noticeable difference in bias across the three DGPs, yet LOOCV demonstrates higher precision, with lower RMSE and shorter confidence intervals. In the non-stationary setting with $\delta^{alt} = 0$, LOOCV reduces RMSE by as much as 10% and length of 95% confidence interval by 4.3% relative to RCV. In terms of hypothesis testing, LOOCV attains an empirical size closer to the nominal level under the null, and achieves higher empirical power in the non-stationary scenario with $\delta^{alt} = 1$.

Table 3: Monte Carlo results comparing bandwidth selection methods. Sample size: $n = 1,000$.

Non-parametric DR DiD Estimators and Hausman-type Test									
Non-Stationary Covariate Distribution with $\delta^{alt} = 0$									
	Avg. Bias	Med. Bias	RMSE	Cover.	CIL	Avg. Test Stats.	Emp. Rej. (0.10)	Emp. Pow. (0.05)	Emp. Pow. (0.01)
LOOCV	0.01	-0.021	1.312	0.968	5.306	21.309	1	0.999	0.98
RCV	-0.008	-0.028	1.459	0.961	5.543	21.228	0.994	0.991	0.971
Plug-in	-2.764	-2.684	13.122	0.939	49.689	1.85	0.238	0.155	0.054
Non-Stationary Covariate Distribution with $\delta^{alt} = 0.5$									
	Avg. Bias	Med. Bias	RMSE	Cover.	CIL	Avg. Test Stats.	Emp. Rej. (0.10)	Emp. Pow. (0.05)	Emp. Pow. (0.01)
LOOCV	0.014	0.01	1.363	0.953	5.258	6.211	0.742	0.627	0.378
RCV	-0.029	-0.08	1.434	0.968	5.423	6.497	0.745	0.643	0.44
Plug-in	-1.49	-1.882	12.52	0.953	48.172	1.364	0.167	0.085	0.025
Stationary Covariate Distribution with $\delta^{alt} = 1$									
	Avg. Bias	Med. Bias	RMSE	Cover.	CIL	Avg. Test Stats.	Emp. Rej. (0.10)	Emp. Pow. (0.05)	Emp. Pow. (0.01)
LOOCV	0.021	-0.027	1.268	0.958	5.131	1.027	0.104	0.052	0.004
RCV	-0.053	-0.109	1.282	0.96	5.229	1.06	0.109	0.055	0.009
Plug-in	-0.392	-0.424	12.091	0.95	46.657	1.212	0.136	0.076	0.019

Note: Simulations based on 1,000 Monte Carlo experiments based on the three DGPs in Section 5.3, with $\delta^{alt} = 0, 0.5, 1$. Columns “Avg. Bias”, “Med. Bias”, “RMSE”, “Cover.”, and “CIL” refer to the average simulated bias, median simulated bias, simulated root-mean-squared errors, 95% coverage probability, and 95% confidence interval length for our proposed nonparametric DR DiD estimator, $\hat{\tau}_{dr}$, as defined in (3.1). Bandwidth selection is performed via leave-one-out cross-validation (“LOOCV”), rescaled cross-validation (“RCV”), or a plug-in estimator (“Plug-in”), with the first two methods using the log-likelihood criterion in (3.13). The Hausman-type test statistics are computed according to (4.2). Columns “Avg. Test Stats.”, and “Emp. Size (α)” denote the average test statistic, and empirical size of the test with a nominal size α , respectively.

6 Empirical illustration: the effect of tariff reduction on corruption

In this section, we revisit a study from Sequeira (2016) on the effect of import tariff liberalization on corruption patterns. Prior to the phaseout of high tariffs between South Africa and Mozambique, bribery payment was pervasive, often used to dodge tariff taxes. According to Sequeira and Djankov (2014), bribery payments can be found in approximately 80% of all shipment records in a random sample of tracked shipments before a tariff rate reduction in 2008.

This tariff change is the result of a long-standing trade agreement between South Africa and Mozambique. The agreement, the Southern African Development Community Trade Protocol, was signed in 1996. The protocol established a timeline for import tariff reductions between 2001 and 2015. The most significant reduction occurred in 2008, with the average nominal rate decreasing by 5%. The effect of such a tariff liberalization scheme is considerable, as both the likelihood and the amount of bribe payments experienced a significant decline following the phaseout.

To investigate the causal relationship between tariff rate reduction and changes in bribery patterns, Sequeira (2016) leverages a quasi-experimental variation induced by trade protocol: Not all products were subject to the change in tariff rate during the analysis period, enabling products unaffected by the tariff changes to serve as a control group. It is thus possible to utilize the DiD design to analyze how tariff rate changes affect bribe patterns along trade routes.

Sequeira (2016) collects data on the bribe payment along the trade routes between the two countries from 2007 to 2013. This data set has a repeated cross-section structure. Sequeira (2016) mainly considers the following two TWFE regressions:

$$\begin{aligned}
 \text{(Linear)} \quad y_{it} &= \gamma_1 TCC_i \times Post + \mu Post + \gamma_2 TCC_i + \beta_2 BT_i + \Gamma_i + p_i + w_t + \delta_i + \epsilon_{it}, \\
 \text{(Interactive)} \quad y_{it} &= \gamma_1 TCC_i \times Post + \mu Post + \gamma_2 TCC_i + \beta_2 BT_i + \Gamma_i + \Gamma_i \times Post \\
 &\quad + p_i + w_t + \delta_i + \epsilon_{it},
 \end{aligned}$$

where TCC_i and BT_i denote Tariff Change Category and Baseline Tariff, respectively, and y_{it} is one of the measurements of bribery payments for shipment i in period t . TCC is the treatment indicator, which takes value one if the product shipped experienced a tariff reduction in 2008, and zero otherwise. The post-treatment period indicator, $Post$, is equal to one for the years following 2008. BT refers to the tariff rates before 2008. A vector of covariates, Γ , industry, year, and clearing agent fixed effects, p, ω, δ , are also included in the regressions. The interactive specification differs from the linear one by an interaction of $Post$ and the covariates, Γ .

Sequeira (2016) focuses on interpreting γ_1 in both specifications as an estimate of the ATT. However, this interpretation might not be valid when treatment effects are heterogeneous (Meyer, 1995; Abadie, 2005). Our proposed DR DiD estimator, $\hat{\tau}_{dr}$, and the one based on Sant’Anna and Zhao (2020), $\hat{\tau}_{sz}$, could be better suited for the task of identifying and con-

sistently estimating the ATT in the present context. In what follows, we estimate the ATT using our proposed DR DiD estimator and compare the results to those obtained by Sequeira (2016).

To achieve this, we first estimate the PS and OR functions based on local linear logistic regression and local linear OLS, respectively. Following Sequeira (2016), we consider four different outcome measures: a binary variable denoting if a bribe is paid, the logarithmic form, $\log(x + 1)$, of the amount of bribe payment, the logarithmic form of the amount of bribe paid as a share of the value of the shipment, and as a share of the weight of the shipment, respectively. Across all four specifications, we include the following common covariates: baseline tariff rate, dummy variables for whether the shipper is a large firm, whether the product is perishable, differentiated, an agricultural good, whether the shipments are pre-inspected at origin, monitored, and originates from South Africa. Additionally, we include the day of arrival during the week and the terminal where the cargo was cleared. Our procedures allow for these covariate-specific trends, so the CPT Assumption 2(i) holds only after accounting for these observed characteristics. To avoid weak-overlap problems, we truncate PS estimates below 0.01.

Table 4 summarizes our results. For each estimator, we report both the unclustered standard errors based on asymptotic approximation (in parentheses) and the cluster-robust standard errors based on the bootstrap procedure in Algorithm C.2 (in brackets), where we cluster at the four-digit HS code level as in Sequeira (2016). Likewise, we conduct two sets of Hausman-type tests – one using unclustered influence functions based on (4.2) and the other that accounts for clustering using a bootstrap procedure given in Algorithm C.3.

We first observe that the point estimates are negative for all measures of bribery payment, consistent with the findings of Sequeira (2016). The results based on the two DR DiD methods are generally close to the TWFE estimates with the interactive specification. For instance, we find that a tariff reduction reduces the probability of paying a bribe by 28 to 43 percentage points, depending on the specific estimator used. The result is statistically and economically significant at the usual levels. Tariff reduction also seems to lead to a decrease in bribery.¹¹ The magnitude of the causal effects based on the weighted results, on the other hand, is more mixed.¹² Results based on the TWFE and DR DiD with no-compositional changes estimators suggest that tariff reduction leads to a statistically significant reduction in the average log of the ratio between bribery payment and shipment values of similar magnitude, while our proposed DR DiD estimator that is robust to compositional changes suggests a twice-as-large effect. When the log of the ratio between bribery payment and tonnage is considered, both nonparametric DR DiD estimators report large yet insignificant (at 95% level) ATT estimates.

11 Some of local linear OR estimates were a bit sensitive to bandwidth choice. This is arguably due to the limited number of observations within certain strata. To improve the stability of cross-validation, we impose a common bandwidth across all four treatment groups for each type of covariates.

12 We avoid attaching a precise interpretation of these log transformations due to the issues raised by Chen and Roth (2023).

Table 4: Difference-in-differences estimation results for Sequeira (2016)

Estimator/Outcome	Prob(bribe)	Log(1 + bribe)	Log(1 + bribe/shpt.val.)	Log(1 + bribe/shpt.tonn.)
TWFE - Linear Spec.	-0.429 (0.083) [0.131]	-3.748 (0.724) [1.064]	-0.011 (0.003) [0.003]	-1.914 (0.341) [0.496]
TWFE - Interactive Spec.	-0.296 (0.082) [0.124]	-2.928 (0.746) [0.917]	-0.010 (0.004) [0.004]	-1.597 (0.402) [0.457]
DR DiD $\hat{\tau}_{sz}$ (no-compositional changes)	-0.275 (0.067) [0.096]	-2.542 (0.636) [0.773]	-0.014 (0.005) [0.006]	-0.918 (0.451) [0.492]
DR DiD $\hat{\tau}_{dr}$ (robust to compositional changes)	-0.307 (0.084) [0.109]	-2.888 (0.798) [0.915]	-0.027 (0.010) [0.014]	-1.131 (0.602) [0.635]
Hausman-tests for no-compositional changes				
Unclustered p -value	0.270	0.199	0.084	0.601
Clustered p -value	0.338	0.238	0.175	0.643

Notes: Same data used by Sequeira (2016). The results represent the estimated ATT of tariff rate reduction on bribery payment behavior. Columns 2 through 5 denote estimates for dependent variables representing whether a bribe is paid, the logarithmic form, $\log(x + 1)$, of the amount of bribe paid, the logarithmic form of the amount of bribe paid as a share of the value of the shipment, and as a share of the weight of the shipment, respectively. We compare four different DiD estimators for the ATT: 1. the two-way fixed effect estimator based on specifications in Column (1) of Tables 8-11 in Sequeira (2016); 2. the two-way fixed effect estimator based on Column (2) from Tables 8-11 in Sequeira (2016); 3. DR DiD estimator based on (4.1), and 4. DR DiD estimator based on (3.1). The same set of covariates is used for the last two estimators. See the main text for further details on the covariates. Continuous variables are re-scaled between 0 and 1, and then added in with binary variables. For DR DiD estimators, the PS and the OR models are estimated nonparametrically, using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the local linear logistic regression is selected with the log-likelihood criterion. Numbers in the parentheses are unclustered standard errors based on asymptotic approximation. Numbers in brackets refer to standard errors clustered at the level of four-digit HS code. Cluster-robust standard errors are calculated following Algorithm C.2 with 9999 bootstrap draws. Hausman-tests are calculated based on (4.2). The clustered p -values are calculated following the bootstrap procedure in Algorithm C.3 with 9999 bootstrap draws. To avoid weak-overlap problems, we truncate PS estimates below 0.01.

The results of the Hausman-type test displayed at the bottom of Table 4 suggest that we lack statistical evidence against the assumption of no-compositional changes, especially when one clusters the standard errors.

In sum, our results support the conclusion of Sequeira (2016) that tariff liberalization decreases corruption. Our DR DiD estimates suggest the size of the effects is approximately the same as that of the original paper, indicating that ruling out treatment effect heterogeneity and compositional changes are not of primary concern in this particular application.

7 Extensions

We conclude the paper by considering two extensions of our main results: the use of cross-fitted first-step estimators, and the analysis of setups with rotating panel data structures where some units are observed in both pre- and post-treatment periods.

7.1 Cross-fitted first-step estimators

We describe a cross-fitting procedure for generic first-step estimators. Let J be a fixed positive integer such that $J < N$, and assume for simplicity that $n_J = n/J$ is an integer. Randomly split the dataset into J equal groups (folds) of size n_J . Denote the set of indices for the j -th group by $\mathcal{J}_j$, and let $\mathcal{J}_{-j} = \{1, \dots, n\} \setminus \mathcal{J}_j$ represent the indices for all observations except those in the j -th group. For each j , construct the first-step estimators $(\hat{p}_j, \{\hat{m}_{d,t,j}\}_{(d,t) \in \mathcal{S}_-})$ using

data from all folds except the j -th fold, i.e., $\{W_i\}_{i \in \mathcal{I}_{-j}}$.

The cross-fitted doubly robust estimator, $\hat{\tau}_{dr,J}^{cf}$, is then defined as:

$$\hat{\tau}_{dr,J}^{cf} = \frac{1}{n} \sum_{j=1}^J \sum_{i \in \mathcal{I}_j} \left\{ \hat{w}_{1,1}(D_i, T_i) \hat{\tau}_j(Y_i, X_i) + \sum_{(d,t) \in \mathcal{S}_{-}} (-1)^{(d+t)} \hat{w}_{d,t,j}(D_i, T_i, X_i) (Y_i - \hat{m}_{d,t,j}(X_i)) \right\}, \quad (7.1)$$

where $\hat{\tau}_j(Y, X) = Y + \sum_{(d,t) \in \mathcal{S}_{-}} (-1)^{d+t} \hat{m}_{d,t,j}(X)$ and $\hat{w}_{d,t,j}(X) = \frac{I_{d,t} \hat{p}_j(1,1,X)}{\hat{p}_j(d,t,X)} \Big/ \left(J^{-1} \sum_{j=1}^J \mathbb{E}_{n,j} \left[\frac{I_{d,t} \hat{p}_j(1,1,X)}{\hat{p}_j(d,t,X)} \right] \right)$. Here, $\mathbb{E}_{n,j}[\cdot]$ represents the sample average over the observations in j -th fold.

Let $(r_{n,d,t})_{n \geq 1}$, $(s_{n,d,t})_{n \geq 1}$, and $(\varepsilon'_n)_{n \geq 1}$, be sequences of positive constants approaching 0. We make the following low-level assumptions regarding the cross-fitted estimators.

Assumption 6 (Cross-fitted nuisance estimators)

- For any $j \in \{1, \dots, J\}$, the nuisance estimators $(\hat{p}_j, \{\hat{m}_{d,t,j}\}_{(d,t) \in \mathcal{S}_{-}})$ constructed using $(W_i)_{i \in \mathcal{I}_{-j}}$ belong to the realization set $\mathcal{J}_n \equiv \mathcal{J}_n^p \times \mathcal{J}_n^m$ with probability no less than $1 - \varepsilon'_n$. The sets $\mathcal{J}_n^p$ and $\mathcal{J}_n^m$ include the true nuisance functions and satisfy the following constraints:

- (i) $\sup_{\tilde{p} \in \mathcal{J}_n^p} \|\tilde{p}(\cdot, \cdot, \cdot) - p(\cdot, \cdot, \cdot)\|_{L_2} \leq r_{n,d,t}$.
- (ii) $\sup_{\tilde{m} \in \mathcal{J}_n^m} \|\tilde{m}_{d,t}(\cdot) - m_{d,t}(\cdot)\|_{L_2} \leq s_{n,d,t}$, for $(d, t) \in \mathcal{S}_{-}$.
- (iii) $r_{n,d,t} \cdot s_{n,d,t} = o(n^{-1/2})$ and $r_{n,1,1} \cdot s_{n,d,t} = o(n^{-1/2})$, for $(d, t) \in \mathcal{S}_{-}$.
- (iv) $0 < \inf_{x \in \mathcal{X}} |\tilde{p}(d, t, x)| < \sup_{x \in \mathcal{X}} |\tilde{p}(d, t, x)| < 1$, for $(d, t) \in \mathcal{S}$.

- $\text{Var}[Y|D=1, T=1] < \infty$ and $\text{Var}[Y|D=\cdot, T=t, X=x]$ is bounded uniformly over $\mathcal{X}$, for $(d, t) \in \mathcal{S}_{-}$.

Lemma 7.1 (Doubly robust error rate with cross-fitted estimators)

Suppose that Assumptions 1, 2, and 6 are satisfied. Then,

$$\sqrt{n} \left(\hat{\tau}_{dr,J}^{cf} - \tau \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta_{\text{eff}}(W_i) + o_p(1) \xrightarrow{d} N(0, \Omega_{dr}). \quad (7.2)$$

Due to the cross-fitting procedure, Assumption 4.2 simplifies to Assumption 6.1. This assumption now requires that the cross-fitted nuisance estimators converge to their true values in mean square. However, similar to the generic case, Lemma 7.1 still imposes the requirement that the product of first-stage approximation errors converge at a rate faster than $O_p(n^{-1/2})$.

Compared to verifying convergence rates for a generic estimator, establishing the mean convergence rate in this specific context is relatively straightforward. Established results provide L_2 -rate conditions for a wide range of nonparametric and machine learning estimators, including kernel and series estimators, as well as methods like Lasso, ridge regression, random forests, boosted trees, deep neural networks, and their ensembles. Although not formalized in this

paper, Lemma 7.1 can be extended to accommodate high-dimensional confounders, where the dimensionality of the nuisance functions increases with the sample size. This setting violates traditional assumptions, such as Donsker properties, on the complexity of the nuisance function space. For a comprehensive discussion of these challenges, refer to Chernozhukov et al. (2017) and references therein.

It is important to note that cross-fitting in the current setup can be challenging when the data size is limited. Since the data for each of the four treatment groups must be split into J folds, larger values of J result in fewer observations per fold, thereby increasing estimation error. Another important caveat is that the semiparametric efficiency results may not hold in high-dimensional covariate spaces, a topic we do not cover in this paper. For a detailed discussion of this limitation, see, for example, Jankova and Van De Geer (2018).

7.2 Overlapping cross section

While the primary focus of this paper is on repeated cross-sectional data where units are observed exclusively in either the pre-treatment or post-treatment period, as per Assumption 1, we recognize that some practical settings involve overlapping cross-sections. For example, surveys like the Current Population Survey (CPS) and the Consumer Expenditure Survey (CEX) employ rotating panel designs. In these surveys, a fraction of the respondents contribute to longitudinal data, appearing in both the pre-and post-treatment periods, while the remaining respondents are observed in only one of the two periods.

In the CPS, households are surveyed for four consecutive months, excluded for the following eight months, and then surveyed again for four more months. Similarly, the CEX follows a rotating panel design where housing units are interviewed once per quarter for four consecutive quarters before being replaced. These designs result in datasets with a mix of panel data and repeated cross-sectional data, where some units overlap across periods, while others are unique to specific periods.

Compared to our assumed sampling process, this structure introduces an additional layer of complexity due to the mixture of unit types. Specifically, let $R = 1$ indicate a panel subject (a unit with data from both pre- and post-treatment periods). The observed data is now $W^{oc} \equiv (RY_0, RY_1, (1 - R)Y, (1 - R)T, D, X, R)$, which is described by the following mixture distribution:

$$\begin{aligned} \mathbb{P}((1 - R)Y \leq y, RY_1 \leq y_1, RY_0 \leq y_0, (1 - R)T = t, D = d, R = r, X \leq x) = \\ r \cdot \mathbb{P}(R = 1) P_p(y_1, y_0, d, x) + (1 - r) \cdot \mathbb{P}(R = 0) P_{rc}(y, d, t, x), \end{aligned} \quad (7.3)$$

where

$$\begin{aligned} P_p(y_1, y_0, d, x) &= \mathbb{P}(Y_1 \leq y_1, Y_0 \leq y_0, D = d, X \leq x | R = 1), \\ P_{rc}(y, d, t, x) &= \mathbb{P}(Y \leq y, D = d, T = t, X \leq x | R = 0). \end{aligned}$$

The ATT under this setup is defined as

$$\tau^{oc} \equiv \mathbb{P}(R = 1) \mathbb{E}[Y_1(1) - Y_1(0)|D = 1, R = 1] + \mathbb{P}(R = 0) \mathbb{E}[Y_1(1) - Y_1(0)|D = 1, T = 1, R = 0].$$

The following conditions are imposed to identify the ATT.

Assumption 7 (Identification assumptions under overlapping cross sections)

1. The observed data $\{W_i^{oc}\}_{i=1}^n$ consists of *i.i.d.* draws from the mixture distribution defined in (7.3).
2. For some $\varepsilon > 0$, $(d, t) \in \mathcal{S}_-$, and for almost every $x \in \mathcal{X}$,

$$\begin{aligned} (p-i) \quad & \mathbb{E}[Y_1(0) - Y_0(0)|D = 1, R = 1, X = x] = \mathbb{E}[Y_1(0) - Y_0(0)|D = 0, R = 1, X = x]. \\ (p-ii) \quad & \mathbb{P}(D = 1|R = 1) > \varepsilon, \text{ and } \mathbb{P}(D = 1|R = 1, X = x) \leq 1 - \varepsilon. \\ (rc-i) \quad & \mathbb{E}[Y_1(0)|D = 1, T = 1, R = 0, X = x] - \mathbb{E}[Y_0(0)|D = 1, T = 0, R = 0, X = x] \\ & = \mathbb{E}[Y_1(0)|D = 0, T = 1, R = 0, X = x] - \mathbb{E}[Y_0(0)|D = 0, T = 0, R = 0, X = x]. \\ (rc-ii) \quad & \mathbb{E}[Y_0(0)|D = 1, T = 0, R = 0, X = x] = \mathbb{E}[Y_0(1)|D = 1, T = 0, R = 0, X = x]. \\ (rc-iii) \quad & \mathbb{P}(D = 1, T = 1|R = 0) > \varepsilon, \text{ and } \mathbb{P}(D = d, T = t|R = 0, X = x) \geq \varepsilon. \end{aligned}$$

This new identification assumption amounts to combining the conditions imposed on the panel units (Assumption 2 in Sant'Anna and Zhao (2020)) with those imposed on the cross-sectional units (Assumption 2).

The EIF and semiparametric efficiency bound can be derived using arguments analogous to those in Theorem 1. Before presenting our results, we introduce the following quantities:

$$\Delta Y = Y_1 - Y_0,$$

$$p_1(x) = \mathbb{P}(D = 1|R = 1, X = x), \quad p_0(d, t, x) = \mathbb{P}(D = d, T = t|R = 0, X = x),$$

$$w_1^p(D) = \frac{D}{\mathbb{E}[D|R = 1]}, \quad w_0^p(D, X) = \frac{(1 - D) \cdot p_1(X)}{1 - p_1(X)} \bigg/ \mathbb{E} \left[\frac{(1 - D) \cdot p_1(X)}{1 - p_1(X)} \bigg| R = 1 \right],$$

$$m_{d,\Delta}^p(X) = \mathbb{E}[\Delta Y|D = d, R = 1, X], \quad \tau_p^{oc} = \mathbb{E}[m_{1,\Delta}^p(X) - m_{0,\Delta}^p(X)|D = 1, R = 1],$$

$$w_{1,1}^{rc}(D, T) = \frac{DT}{\mathbb{E}[DT|R = 0]}, \quad w_{d,t}^{rc}(D, T, X) = \frac{I_{d,t} \cdot p_0(1, 1, X)}{p_0(d, t, X)} \bigg/ \mathbb{E} \left[\frac{I_{d,t} \cdot p_0(1, 1, X)}{p_0(d, t, X)} \bigg| R = 0 \right],$$

$$m_{d,t}^{rc}(X) = \mathbb{E}[Y|D = d, T = t, R = 0, X], \quad \tau_{rc}^{oc}(Y, X) = Y - \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} m_{d,t}^{rc}(X),$$

$$\tau_{rc}^{oc} = \mathbb{E}[\tau_{rc}^{oc}(Y, X)|D = 1, T = 1, R = 0].$$

In addition, let

$$\begin{aligned} \eta_p(Y_1, Y_0, D, X) &= \{w_1^p(D) (m_{1,\Delta}^p(X) - m_{0,\Delta}^p(X) - \tau_p^{oc}) \\ &\quad + w_1^p(D)(\Delta Y - m_{1,\Delta}^p(X)) - w_0^p(D, X)(\Delta Y - m_{0,\Delta}^p(X))\}, \\ \eta_{rc}(Y, D, T, X) &= w_{1,1}^{rc}(D, T)(\tau_{rc}^{oc}(Y, X) - \tau_{rc}^{oc}) + \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} w_{d,t}^{rc}(D, T, X)(Y - m_{d,t}^{rc}(X)), \end{aligned}$$

$$\eta_r(R) = (R - \mathbb{E}[R]) \cdot (\tau_p^{oc} - \tau_{rc}^{oc}),$$

and

$$\begin{aligned} V_p^{oc} &= \mathbb{E}[D|R=1]^{-2} \cdot \mathbb{E}\left[D(m_{1,\Delta}^p(X) - m_{0,\Delta}^p(X) - \tau_p^{oc})^2\right. \\ &\quad \left.+ D(\Delta Y - m_{1,\Delta}^p(X))^2 + \frac{(1-D) \cdot p_1(X)^2}{(1-p_1(X))^2} (\Delta Y - m_{0,\Delta}^p(X))^2 | R=1\right], \\ V_{rc}^{oc} &= \mathbb{E}[DT|R=0]^{-2} \cdot \mathbb{E}\left[DT(\tau_{rc}^{oc}(Y, X) - \tau_{rc}^{oc})^2\right. \\ &\quad \left.+ \sum_{(d,t) \in \mathcal{S}_-} \frac{I_{d,t} \cdot p_0(1, 1, X)^2}{p_0(d, t, X)^2} (Y - m_{d,t}^{rc}(X))^2 \middle| R=0\right], \\ V_r^{oc} &= \mathbb{V}\text{ar}[R] \cdot (\tau_p^{oc} - \tau_{rc}^{oc})^2. \end{aligned}$$

Theorem 4 (Semiparametric efficiency bound under overlapping cross sections)

Suppose Assumption 7 holds.

(a) The EIF for the ATT, τ^{or} is given by,

$$\eta_{oc}(W^{oc}) = R \cdot \eta_p(Y_1, Y_0, D, X) + (1 - R) \cdot \eta_{rc}(Y, D, T, X) + \eta_r(R).$$

(b) Furthermore, the semiparametric efficiency bound for the set of all regular estimators of τ^{oc} is

$$\mathbb{E}[\eta_{oc}(W^{oc})^2] = \mathbb{E}[R] \cdot V_p^{oc} + \mathbb{E}[1 - R] \cdot V_{rc}^{oc} + V_r^{oc}.$$

The EIF comprises three parts. The first two correspond, respectively, to the EIF for panel-only data (as derived in Proposition 1(a) of Sant'Anna and Zhao (2020)) and the EIF for cross-sectional-only data (as derived in Theorem 1 of our paper). The third component directly quantifies the relative effect of assignment to the panel units on average and reflects the efficiency cost associated with combining the two data types.

Interestingly, the semiparametric efficiency bound decomposes into three distinct terms, each reflecting the independent contribution of a different component of the EIF. Notably, when the two data sources share the same true ATT (e.g., when there are no compositional shifts between different sampling cohorts over time), the third term vanishes.

Leveraging this result, we propose the following DR estimand for the ATT:

$$\begin{aligned} \tau_{dr}^{oc} &= \mathbb{E}\left[R \cdot (w_1^p(D) (m_{1,\Delta}^p(X) - m_{0,\Delta}^p(X))\right. \\ &\quad \left.+ w_1^p(D)(\Delta Y - m_{1,\Delta}^p(X)) - w_0^p(D, X)(\Delta Y - m_{0,\Delta}^p(X)))\right. \\ &\quad \left.+ (1 - R) \cdot \left(w_{1,1}^{rc}(D, T) \tau^{rc}(Y, X) + \sum_{(d,t) \in \mathcal{S}_-} (-1)^{(d+t)} w_{d,t}^{rc}(D, T, X)(Y - m_{d,t}^{rc}(X))\right)\right]. \end{aligned}$$

Since τ_{dr}^{oc} is based on the EIF, it is straightforward to show that plug-in estimators for the ATT based on τ_{dr}^{oc} inherits the same rate doubly robust property as established in Theorem 2.

Adapting our paper’s arguments in Theorem 2 to show these results is straightforward. We omit the details for brevity.

References

- Abadie, Alberto**, “Semiparametric Difference-in-Difference Estimators,” *Review of Economic Studies*, 2005, 72, 1–19.
- Abbring, Jaap H. and Gerard J. van den Berg**, “The nonparametric identification of treatment effects in duration models,” *Econometrica*, 2003, 71 (5), 1491–1517.
- Ackerberg, Daniel, Xiaohong Chen, Jinyong Hahn, and Zhipeng Liao**, “Asymptotic Efficiency of Semiparametric Two-step GMM,” *The Review of Economic Studies*, 2014, 81 (3), 919–943,.
- Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen**, “Program Evaluation and Causal Inference With High-Dimensional Data,” *Econometrica*, 2017, 85 (1), 233–298.
- Bickel, Peter J., Chris A.J. Klaassen, Ya’acov Ritov, and Jon A. Wellner**, *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer-Verlag, 1998.
- Bonvini, Matteo, Edward H. Kennedy, Oliver Dukes, and Sivaraman Balakrishnan**, “Doubly-robust inference and optimality in structure-agnostic models with smoothness,” *arXiv:2405.08525*, 2024.
- Busso, Matias, John Dinardo, and Justin McCrary**, “New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators,” *The Review of Economics and Statistics*, 2014, 96 (5), 885–895.
- Callaway, Brantly and Pedro HC Sant’Anna**, “Difference-in-differences with multiple time periods,” *Journal of Econometrics*, 2021, 225 (2), 200–230.
- Chang, Neng-Chieh**, “Double/debiased machine learning for difference-in-differences models,” *The Econometrics Journal*, 2020, 23 (2), 177–191.
- Chen, Jiafeng and Jonathan Roth**, “Log-like? Identified ATEs defined with zero-valued outcomes are (arbitrarily) scale-dependent,” *Working Paper*, 2023.
- Chen, Xiaohong, Han Hong, and Alessandro Tarozi**, “Semiparametric efficiency in GMM models with auxiliary data,” *The Annals of Statistics*, apr 2008, 36 (2), 808–843.
- , **Oliver Linton, and Ingrid Van Keilegom**, “Estimation of semiparametric models when the criterion function is not smooth,” *Econometrica*, 2003, 71 (5), 1591–1608.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins**, “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, jun 2017, pp. 1–71.
- Claeskens, Gerda and Ingrid Van Keilegom**, “Bootstrap confidence bands for regression curves and their derivatives,” *The Annals of Statistics*, 2003, 31 (6), 1852–1884.
- Colangelo, Kyle and Ying-Ying Lee**, “Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments,” *arXiv:2004.03036v8*, 2023.

- Fan, Jianqing, Nancy E Heckman, and Matt P Wand**, “Local polynomial kernel regression for generalized linear models and quasi-likelihood functions,” *Journal of the American Statistical Association*, 1995, *90* (429), 141–150.
- Fan, Yangin and Emmanuel Guerre**, “Multivariate local polynomial estimators: Uniform boundary properties and asymptotic linear representation,” in “Essays in Honor of Aman Ullah,” Emerald Group Publishing Limited, 2016.
- Farrell, Max H.**, “Robust inference on average treatment effects with possibly more covariates than observations,” *Journal of Econometrics*, 2015, *189* (1), 1–23.
- Frölich, Markus**, “Non-parametric regression for binary dependent variables,” *The Econometrics Journal*, 2006, *9* (3), 511–540.
- Guggenberger, Patrik**, “The impact of a Hausman pretest on the asymptotic size of a hypothesis test,” *Econometric Theory*, 2010, *26* (2), 369–382.
- , “The impact of a Hausman pretest on the size of a hypothesis test: The panel data case,” *Journal of Econometrics*, 2010, *156* (2), 337–343.
- Hájek, J.**, “Discussion of ‘An essay on the logical foundations of survey sampling, Part I’, by D. Basu,” in V. P. Godambe and D. A. Sprott, eds., *Foundations of Statistical Inference*, Toronto: Holt, Rinehart, and Winston, 1971.
- Hausman, Jerry A.**, “Specification tests in econometrics,” *Econometrica*, 1978, pp. 1251–1271.
- Heckman, James J., Hidehiko Ichimura, and Petra Todd**, “Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme,” *The Review of Economic Studies*, 1997, *64* (4), 605–654.
- Hong, Seung-Hyun**, “Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes,” *Journal of Applied Econometrics*, mar 2013, *28* (2), 297–324.
- Jankova, Jana and Sara Van De Geer**, “Semiparametric efficiency bounds for high-dimensional models,” *The Annals of Statistics*, 2018, *46* (5), 2336–2359.
- Jordan, Michael, Yixin Wang, and Angela Zhou**, “Empirical Gateaux Derivatives for Causal Inference,” in S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., *Advances in Neural Information Processing Systems*, Vol. 35 Curran Associates, Inc. 2022, pp. 8512–8525.
- Kennedy, Edward H.**, “Semiparametric Theory and Empirical Processes in Causal Inference,” in Hua He, Pan Wu, and Ding-Geng (Din) Chen, eds., *Statistical Causal Inferences and Their Applications in Public Health Research*, New York: Springer, 2016, pp. 141–167.
- , “Semiparametric Doubly Robust Targeted Double Machine Learning: A Review,” *arXiv:2203.06469*, 2023.
- Kennedy, Edward H, Zongming Ma, Matthew D McHugh, and Dylan S Small**, “Non-parametric methods for doubly robust estimation of continuous treatment effects,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2017, *79* (4), 1229–1245.
- Khan, Shakeeb and Elie Tamer**, “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 2010, *78* (6), 2021–2042.

- Kong, Efang, Oliver Linton, and Yingcun Xia**, “Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model,” *Econometric Theory*, 2010, *26* (5), 1529–1564.
- Lee, Ying Ying**, “Efficient propensity score regression estimators of multivalued treatment effects for the treated,” *Journal of Econometrics*, 2018, *204* (2), 207–222.
- Li, Degui, Qi Li, and Zheng Li**, “Nonparametric quantile regression estimation with mixed discrete and continuous data,” *Journal of Business & Economic Statistics*, 2021, *39* (3), 741–756.
- Li, Qi and Desheng Ouyang**, “Uniform convergence rate of kernel estimation with mixed categorical and continuous data,” *Economics Letters*, 2005, *86* (2), 291–296.
- and **Jeffrey Scott Racine**, *Nonparametric econometrics: theory and practice*, Princeton, New Jersey: Princeton University Press, 2007.
- Malani, Anup and Julian Reif**, “Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform,” *Journal of Public Economics*, 2015, *124*, 1–17.
- Meyer, Bruce D.**, “Natural and Quasi-Experiments in Economics,” *Journal of Business & Economic Statistics*, 1995, *13* (2), 151–161.
- Millimet, Daniel L. and Rusty Tchernis**, “On the specification of propensity scores, with applications to the analysis of trade policies,” *Journal of Business & Economic Statistics*, 2009, *27* (3), 397–415.
- Newey, Whitney K.**, “The asymptotic variance of semiparametric estimators,” *Econometrica*, 1994, *62* (6), 1349–1382.
- Nie, Xinkun, Chen Lu, and Stefan Wager**, “Nonparametric Heterogeneous Treatment Effect Estimation in Repeated Cross Sectional Designs,” *arXiv preprint arXiv:1905.11622*, 2019.
- Powell, James L and Thomas M Stoker**, “Optimal bandwidth choice for density-weighted averages,” *Journal of Econometrics*, 1996, *75* (2), 291–316.
- , **James H Stock, and Thomas M Stoker**, “Semiparametric estimation of index coefficients,” *Econometrica: Journal of the Econometric Society*, 1989, pp. 1403–1430.
- Roth, Jonathan**, “Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends,” *American Economic Review: Insights*, 2022, *4* (3), 305–322.
- , **Pedro H. C. Sant’Anna, Alyssa Bilinski, and John Poe**, “What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature,” *Journal of Econometrics*, 2023, *Forthcoming*.
- Rothe, Christoph and Sergio Firpo**, “Properties of doubly robust estimators when nuisance functions are estimated nonparametrically,” *Econometric Theory*, 2019, *35* (5), 1048–1087.
- Rotnitzky, Andrea, Emiliano Smucler, and James M. Robins**, “Characterization of parameters with a mixed bias property,” *Biometrika*, March 2021, *108* (1), 231–238.
- Sant’Anna, Pedro H. C. and Jun Zhao**, “Doubly robust difference-in-differences estimators,” *Journal of Econometrics*, 2020, *219* (1), 101–122.
- Sequeira, Sandra**, “Corruption, trade costs, and gains from tariff liberalization: Evidence from Southern Africa,” *American Economic Review*, 2016, *106* (10), 3029–63.
- and **Simeon Djankov**, “Corruption and firm behavior: Evidence from African ports,” *Journal of International Economics*, 2014, *94* (2), 277–294.

- Smucler, Ezequiel, Andrea Rotnitzky, and James M. Robins**, “A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts,” *arXiv:1904.03737*, 2019.
- Staniswalis, Joan G**, “The kernel estimate of a regression function in likelihood-based models,” *Journal of the American Statistical Association*, 1989, *84* (405), 276–283.
- Stuart, Elizabeth A, Haiden A. Huskamp, Kenneth Duckworth, Jeffrey Simmons, Zirui Song, Michael E. Chernew, and Colleen L Barry**, “Using propensity scores in difference-in-differences models to estimate the effects of a policy change,” *Health Services and Outcomes Research Methodology*, dec 2014, *14* (4), 166–182.