The Econometrics Journal

Econometrics Journal (2020), volume **23**, pp. 177–191.
doi: 10.1093/ectj/utaa001

Double/debiased machine learning for difference-in-differences models

NENG-CHIEH CHANG

**Department of Economics, University of California Los Angeles, 315 Portola Plaza, Los Angeles, CA 90095, USA.*
Email: nengchiehchang@g.ucla.edu

First version received: 7 June 2019; final version accepted: 25 September 2019.

Summary: This paper provides an orthogonal extension of the semiparametric difference-in-differences estimator proposed in earlier literature. The proposed estimator enjoys the so-called Neyman orthogonality (Chernozhukov et al., 2018), and thus it allows researchers to flexibly use a rich set of machine learning methods in the first-step estimation. It is particularly useful when researchers confront a high-dimensional data set in which the number of potential control variables is larger than the sample size and the conventional nonparametric estimation methods, such as kernel and sieve estimators, do not apply. I apply this orthogonal difference-in-differences estimator to evaluate the effect of tariff reduction on corruption. The empirical results show that tariff reduction decreases corruption in large magnitude.

Keywords: *Difference-in-differences*, *high-dimensional data*, *causal inference*, *machine learning*.

JEL codes: *C1*.

1. INTRODUCTION

The difference-in-differences (DiD) estimator has been used widely in empirical economics to evaluate causal effects when there exists a natural experiment with a treated group and an untreated group. By comparing the variation over time in an outcome variable between the treated group and the untreated group, the DiD estimator can be used to calculate the effect of treatment on the outcome variable. Applications of DiD include but are not limited to studies of the effects of immigration on laboru markets (Card, 1990), the effects of minimum wage law on wages (Card and Krueger, 1994), the effect of tariff liberalisation on corruption (Sequeira, 2016), the effect of household income on children's personalities (Akee et al., 2018), and the effect of corporate tax on wages (Fuest, Peichl, and Siegloch, 2018).

The traditional linear DiD estimator depends on a parallel-trend assumption that in the absence of treatment, the difference in outcomes between treated and untreated groups remains constant over time. In many situations, however, this assumption may not hold because there are other individual characteristics that may be associated with the variations of the outcomes. The treatment may be taken as exogenous only after controlling for these characteristics. However, as noted by Meyer, Viscusi, and Durbin (1995), including control variables in the linear specification of the traditional DiD estimator imposes strong constraints on the heterogeneous effect of treatment. To address this problem, Abadie (2005) proposed the semiparametric DiD estimator. Compared

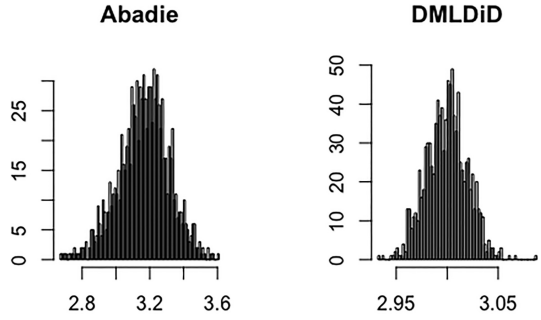


Figure 1. Comparison of Abadie's DiD and DMLDiD with the first-step Logit Lasso estimation. The true value is $\theta_0 = 3$. The results for other ML methods can be found in Section 4.

to the traditional linear DiD estimator, the advantages of Abadie's estimator are threefold. First, the characteristics are treated nonparametrically so that any estimation error caused by functional specification is avoided. Second, the effect of treatment is allowed to vary among individuals, whereas the traditional linear DiD estimator does not allow this heterogeneity. Third, the estimation framework proposed in Abadie (2005) will enable researchers to estimate how the effect of treatment varies with changes in the characteristics.

The present paper provides an orthogonal extension of Abadie's semiparametric DiD estimator (DMLDiD hereafter).[1] Abadie's semiparametric DiD estimator behaves well when researchers use conventional nonparametric methods, such as kernel and sieve estimators, to estimate the propensity score in the first-step estimation. As shown in the classical semiparametric estimation literature, Abadie's DiD estimator is $\sqrt{N}$-consistent and asymptotically normal when kernel or sieve is used in the first-step estimation. However, according to the general theory of inference developed in Chernozhukov et al. (2018), these desirable properties may fail if researchers use a rich set of newly developed nonparametric estimation methods, the so-called machine learning (ML) methods, such as Lasso, Logit Lasso, random forests, boosting, neural network, and their hybrids, in the first-step estimation. This is especially a problem when researchers confront a high-dimensional data set in which the number of potential control variables is more than the sample size, and thus the conventional nonparametric estimation methods do not apply.

In this paper, I propose DMLDiD for three different data structures: repeated outcomes, repeated cross sections, and multilevel treatment, which are all based on the original paper by Abadie (2005), as well as on the papers on the general inference theory of ML methods by Chernozhukov et al. (2018) and Chernozhukov et al. (2019). DMLDiD will allow researchers to apply a broad set of ML methods and still obtain valid inferences. The key difference is that DMLDiD, in contrast to Abadie's original DiD estimator, is constructed on the basis of a score function that satisfies the so-called Neyman orthogonality (Chernozhukov et al. 2018), which is an important property for obtaining valid inferences when applying ML methods. With this property, DMLDiD can overcome the bias caused by the first-step ML estimation and achieve $\sqrt{N}$-consistency and asymptotic normality as long as the ML estimator converges to its true value at a rate faster than $N^{-1/4}$. Figure 1 shows the Monte Carlo simulation that illustrates the negative effect of directly combining ML methods on Abadie's estimator and the benefit of using

[1] The R codes can be found on my Github: https://github.com/NengChiehChang/Diff-in-Diff

DMLDiD. The histogram in the left panel shows that the simulated distribution of Abadie's estimator is biased, whereas the simulated distribution of DMLDiD in the right panel is centred at the true value.

As an empirical example, I study the effect of tariff reduction on corruption by using the trade data between South Africa and Mozambique during 2006 and 2014. The source of exogenous variation is the large tariff reduction on certain commodities occurring in 2008. This natural experiment was studied previously by Sequeira (2016) using the traditional linear DiD estimator. On the basis of Sequeira's linear specification, I include the interaction terms between the treatment and a vector of control variables. After controlling for the interaction terms, I find that the traditional linear DiD estimate becomes insignificantly different from zero. This suggests the existence of heterogeneous treatment effects, and Sequeira's result can be interpreted as a weighted average of these heterogeneous effects. As pointed out by Abadie (2005), it is ideal to treat the control variables nonparametrically when there exists heterogeneity in treatment effects, to avoid any inconsistency caused by functional form misspecification. I apply both Abadie's semiparametric DiD and DMLDiD on the same data set (Table 9 of Sequeira, 2016). In comparison with Sequeira's result, though with the same sign, Abadie's estimator is at least twice as large as previously reported by Sequeira (2016). This large effect, however, may be due to the lack of robustness of this estimation method and the finite-sample bias in the first-step nonparametric estimation. DMLDiD removes the first-order bias and suggests a smaller effect that is closer to Sequeira's estimate. The value becomes only 60% higher than Sequeira's result. This extra effect can be explained by the misspecification of the traditional linear DiD estimator. Therefore, I obtain the same conclusion as Sequeira (2016) that tariff reduction decreases corruption, but my estimate suggests an even larger magnitude.

The DMLDiD proposed in the present paper relies heavily on the recent high-dimensional and ML literature: Belloni et al. (2012), Belloni, Chernozhukov, and Hansen (2014), Chernozhukov et al. (2015), Belloni et al. (2017), and Chernozhukov et al. (2018). The present paper is also closely related to the robustness of average treatment effect estimation discussed in Robins and Rotnitzky (1995) and the general discussion in Chernozhukov et al. (2019). The asymptotic properties of the robust estimators discussed in those papers remain unaffected if only one of the first-step estimation with classical nonparametric method is inconsistent. In independent and contemporaneous works, Zimmert (2019), Sant'Anna and Zhao (2019), Li (2019), and Lu, Nie, and Wager (2019) also considered the orthogonal property of Abadie's DiD estimator. Zimmert (2019) further discussed its efficiency, whereas Sant'Anna and Zhao (2019) and Li (2019) focused on classical first-step estimation. Lu, Nie, and Wager (2019) discussed the situation in which control variables are correlated with time.

1.1. Plan of the paper

Section 2 reviews both the traditional linear DiD estimator and Abadie's semiparametric DiD estimator and discusses their limitations. Section 3 presents DMLDiD and discusses its theoretical properties. Section 4 conducts the Monte Carlo simulation to shed some light on the finite-sample performance of the proposed DiD estimator. Section 5 provides the empirical application, and Section 6 concludes the paper.

2. THE SEMIPARAMETRIC DID ESTIMATOR

In this section, I review the traditional linear DiD estimator and Abadie's semiparametric DiD estimator. Let $Y_i(t)$ be the outcome of interest for individual i at time t and $D_i(t) \in \{0, 1\}$ the treatment status. The population is observed in a pre-treatment period, $t = 0$, and in a post-treatment period, $t = 1$. With potential outcome notations (Rubin, 1974), we have $Y_i(t) = Y_i^0(t) + \left(Y_i^1(t) - Y_i^0(t)\right) D_i(t)$, where $Y_i^0(t)$ is the outcome that individual i would attain at time t in the absence of the treatment, and $Y_i^1(t)$ represents the outcome that individual i would attain at time t if exposed to the treatment. Because individuals are exposed to treatment only at $t = 1$, we have $D_i(0) = 0$ for all i. To reduce notation, I define $D_i \equiv D_i(1)$. Then, the specification for the traditional linear DiD without control variables is

$$Y_i(t) = \mu + \tau \cdot D_i + \delta \cdot t + \alpha \cdot D_i(t) + \varepsilon_i(t),$$

where α is the parameter of interest, $\varepsilon_i(t)$ is an exogenous shock that has mean zero, and (μ, τ, δ) are constant parameters. If the common trend assumption holds unconditionally, then the parameter α captures the effect of treatment. When the treated and untreated groups are thought to be unbalanced with some characteristics, researchers often include a vector of control variables, $X_i \in \mathbb{R}^d$, in the above linear specification:

$$Y_i(t) = \mu + X_i'\pi(t) + \tau \cdot D_i + \delta \cdot t + \alpha \cdot D_i(t) + \varepsilon_i(t).$$

As noted by Meyer, Viscusi, and Durbin (1995), including control variables in this linear specification may not be appropriate if the treatment has different effects for different groups in the population. One may also need to include the interaction terms between X_i and $D_i(t)$ to capture the heterogeneous effect of treatment. Hence, it is ideal to treat the control variables nonparametrically, as suggested by Abadie (2005). In the following, I review Abadie's semiparametric DiD estimator.

Let the parameter of interest be the average treatment effect on the treated (ATT):

$$\theta_0 \equiv E\left[Y_i^1(1) - Y_i^0(1) \mid D_i = 1\right].$$

Abadie (2005) discussed three data types: repeated outcomes, repeated cross sections, and multilevel treatment. To avoid repetition, I focus only on the first two cases. The discussion for multilevel treatments is provided in the appendix.

2.1. Case 1 (repeated outcomes)

Suppose that researchers observe both pre-treatment and post-treatment outcomes for the individual of interest. That is, researchers observe $\{Y_i(0), Y_i(1), D_i, X_i\}_{i=1}^N$. In this case, we can identify the ATT under the following assumptions (Abadie, 2005):

ASSUMPTION 2.1. $E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, D_i = 1\right] = E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, D_i = 0\right]$.

ASSUMPTION 2.2. *$P(D_i = 1) > 0$ and $P(D_i = 1|X_i) < 1$, with probability one.*

Assumption 2.1 is the conditional parallel-trend assumption. It states that conditional on the individual's characteristics, X_i, the average outcomes for treated and untreated groups would have followed parallel paths in the absence of treatment. Assumption 2.2 states that the support of the propensity score of the treated group is a subset of the support for the untreated. With these two

assumptions, Abadie (2005) identified the ATT:

$$\theta_0 = E\left[\frac{Y_i(1) - Y_i(0)}{P(D_i = 1)} \frac{D_i - P(D_i = 1 \mid X_i)}{1 - P(D_i = 1 \mid X_i)}\right]. \tag{2.1}$$

2.2. Case 2 (repeated cross sections)

Suppose what researchers observe is repeated cross-sectional data. That is, researchers observe $\{Y_i, D_i, T_i, X_i\}_{i=1}^N$, where $Y_i = Y_i(0) + T_i(Y_i(1) - Y_i(0))$, and T_i is a time indicator that takes value one if the observation belongs to the post-treatment sample.

ASSUMPTION 2.3. *Conditional on $T = 0$, the data are independent and identically distributed from the distribution of (Y(0), D, X), and conditional on $T = 1$, the data are independent and identically distributed from the distribution of (Y(1), D, X).*

Supposing Assumptions 2.1 through 2.3 hold, the ATT is identified (Abadie, 2005) as

$$\theta_0 = E\left[\frac{T_i - \lambda_0}{\lambda_0(1 - \lambda_0)} \frac{Y_i}{P(D_i = 1)} \frac{D_i - P(D_i = 1 \mid X_i)}{1 - P(D_i = 1 \mid X_i)}\right], \tag{2.2}$$

where $\lambda_0 \equiv P(T_i = 1)$.

Then, the semiparametric DiD estimator would be the sample analog of (2.1) and (2.2). For example, in Case 1, in which researchers confront repeated outcomes data, the sample analog of (2.1) is

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} \frac{Y_i(1) - Y_i(0)}{\hat{p}} \frac{D_i - \hat{g}(X_i)}{1 - \hat{g}(X_i)},$$

where $\hat{p}$ is the estimator of $p_0 \equiv P(D = 1)$, and $\hat{g}(X_i)$ is the estimator of the propensity score $g_0(X) \equiv P(D = 1|X)$. When $\hat{g}$ is estimated by using classical nonparametric methods, such as the kernel or series estimators, the estimator $\hat{\theta}$ is able to achieve $\sqrt{N}$-consistency and asymptotic normality under certain conditions, as shown in the semiparametric estimation literature (Newey, 1994; Newey and McFadden, 1994).

When $\hat{g}$ is an ML estimator, however, the estimator $\hat{\theta}$ is not necessarily $\sqrt{N}$-consistent in general. According to the general theory of inference of ML methods developed in Chernozhukov et al. (2018), the reason is twofold. First, the score function based on (2.1), $\varphi(W, \theta_0, p_0, g_0) \equiv \frac{Y(1)-Y(0)}{P(D=1)} \frac{D-g_0(X)}{1-g_0(X)} - \theta_0$, has a non-zero directional (Gateaux) derivative with respect to the propensity score g_0:

$$\partial_g E[\varphi(W, \theta_0, p_0, g_0)][g - g_0] \neq 0,$$

where the directional (Gateaux) derivative is defined in Section 3. Second, ML estimators usually have a convergence rate slower than $N^{-1/2}$ because of regularisation bias. Similarly, the estimators obtained by directly plugging ML estimators into (2.2) will not be $\sqrt{N}$-consistent in general. The Monte Carlo simulation in Section 4 supports this theoretical insight and reveals significant bias on the estimators based on (2.1) and (2.2) when ML estimators are used in the first-step nonparametric estimation.

The next section proposes DMLDiD based on (2.1) and (2.2). A distinctive feature of DMLDiD is that the Gateaux derivatives of the score functions are zero with respect to their infinite-dimensional nuisance parameters. This property helps us remove the first-order bias of the first-step ML estimation.

3. THE DMLDID ESTIMATOR

In this section, I propose DMLDiD on the basis of Abadie's results (2.1) and (2.2). In Section 3.1, I present the new score functions derived from (2.1) and (2.2) and propose an algorithm to construct DMLDiD. In Section 3.2, I show the theoretical properties of the proposed estimator.

3.1. The Neyman-orthogonal score

Suppose Assumptions 2.1 through 2.3 hold, and consider the following new score functions.

Case 1 (repeated outcomes): The new score function for repeated outcomes is

$$\psi_1(W, \theta_0, p_0, \eta_{10}) = \frac{Y(1) - Y(0)}{P(D=1)} \frac{D - P(D=1 \mid X)}{1 - P(D=1 \mid X)} - \theta_0 \\ \underbrace{- \frac{D - P(D=1 \mid X)}{P(D=1)(1 - P(D=1 \mid X))} E[Y(1) - Y(0) \mid X, D=0]}_{c_1} \tag{3.1}$$

with the unknown constant $p_0 = P(D=1)$ and the infinite-dimensional nuisance parameter

$$\eta_{10} = (P(D=1 \mid X), E[Y(1) - Y(0) \mid X, D=0]) \equiv (g_0, \ell_{10}).$$

Case 2 (repeated cross sections): The new score function for repeated cross sections is

$$\psi_2(W, \theta_0, p_0, \lambda_0, \eta_{20}) = \frac{T - \lambda_0}{\lambda_0(1-\lambda_0)} \frac{Y}{P(D=1)} \frac{D - P(D=1 \mid X)}{1 - P(D=1 \mid X)} - \theta_0 - c_2, \tag{3.2}$$

where the adjustment term c_2 is

$$c_2 = \frac{D - P(D=1 \mid X)}{\lambda_0(1-\lambda_0) \cdot P(D=1) \cdot (1 - P(D=1 \mid X))} \times E[(T - \lambda_0) Y \mid X, D=0].$$

The nuisance parameters are the unknown constants $p_0 = P(D=1)$ and $\lambda_0 = P(T=1)$ and the unknown function

$$\eta_{20} = (P(D=1 \mid X), E[(T - \lambda) Y \mid X, D=0]) \equiv (g_0, \ell_{20}).$$

Notice that the above new functions are equal to the original score functions (2.1) and (2.2) plus the adjustment terms, (c_1, c_2), which have zero expectations. Thus, the new score functions (3.1) and (3.2) still identify the ATT in each case. The purpose of the adjustment terms is to make the Gateaux derivative of the new score functions zero with respect to infinite-dimensional nuisance parameters, which is the so-called Neyman-orthogonal property in Chernozhukov et al. (2018). I combine the new scores (3.1) and (3.2) with the cross-fitting algorithm of Chernozhukov et al. (2018) to propose DMLDiD.

DEFINITION 3.1. *(a) Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1, ..., N\}$. For simplicity, assume that each fold I_k has the same size $n = N/K$. For each $k \in [K] = \{1, ..., K\}$, define the auxiliary sample $I_k^c \equiv \{1, ..., N\} \setminus I_k$. (b) For each k, construct the intermediate ATT estimators,*

$$\tilde{\theta}_k = \frac{1}{n} \sum_{i \in I_k} \frac{D_i - \hat{g}_k(X_i)}{\hat{p}_k(1 - \hat{g}_k(X_i))} \times \left(Y_i(1) - Y_i(0) - \hat{\ell}_{1k}(X_i)\right) (rep - outcomes)$$

$$\tilde{\theta}_k = \frac{1}{n}\sum_{i\in I_k} \frac{D_i - \hat{g}_k(X_i)}{\hat{p}_k \hat{\lambda}_k \left(1-\hat{\lambda}_k\right)\left(1-\hat{g}_k(X_i)\right)} \times \left(\left(T_i - \hat{\lambda}_k\right) Y_i - \hat{\ell}_{2k}(X_i)\right) (rep-cross-sections)$$

,

where $\hat{p}_k = \frac{1}{n}\sum_{i\in I_k^c} D_i$, $\hat{\lambda}_k = \frac{1}{n}\sum_{i\in I_k^c} T_i$, and $\left(\hat{g}_k, \hat{\ell}_{1k}, \hat{\ell}_{2k}\right)$ *are the estimators of* $(g_0, \ell_{10}, \ell_{20})$ *constructed by using the auxiliary sample* I_k^c. *(c) Construct the final ATT estimator* $\tilde{\theta} = \frac{1}{K}\sum_{k=1}^{K} \tilde{\theta}_k$.

The estimators $\left(\hat{g}_k, \hat{\ell}_{1k}, \hat{\ell}_{2k}\right)$ can be constructed by using any ML methods or classical estimators, such as kernel or series estimators. For completeness, I present the Logit Lasso and Lasso estimators here.

Consider a class of approximating functions of X_i,

$$q_i \equiv \left(q_1(X_i), ..., q_p(X_i)\right)'.$$

For example, q_i can be polynomials or B-splines. Let $\Lambda(u) \equiv 1/(1+\exp(-u))$ be the cumulative distribution function of the standard logistic distribution. Construct the estimator of the propensity score g_0 by

$$\hat{g}_k(x_i) \equiv \Lambda\left(q_i'\hat{\beta}_k\right), \tag{3.3}$$

where

$$\hat{\beta}_k \equiv \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{M}\sum_{i\in I_k^c} \left\{-D_i(q_i'\beta) + \log\left(1+\exp\left(q_i'\beta\right)\right)\right\} + \lambda_k \parallel \beta \parallel_1$$

is the Logit Lasso estimator, and $M = N - n$ is the sample size of the auxiliary sample I_k^c. Next, define M_k as the sample size of $I_k^c \cap \{i : D_i = 0\}$. Construct the estimators of ℓ_{10} and ℓ_{20} by

$$\hat{\ell}_{1k}(x_i) \equiv q_i'\hat{\beta}_{1k},$$

$$\hat{\ell}_{2k}(x_i) \equiv q_i'\hat{\beta}_{2k},$$

where

$$\hat{\beta}_{1k} \in \arg\min_{\beta\in\mathbb{R}^p} \left[\frac{1}{M_k}\sum_{i\in I_k^c} (1-D_i)\left(Y_i(1) - Y_i(0) - q_i'\beta\right)^2\right] + \frac{\lambda_{1k}}{M_k} \parallel \hat{\Upsilon}_{1k}\beta \parallel_1$$

and

$$\hat{\beta}_{2k} \in \arg\min_{\beta\in\mathbb{R}^p} \left[\frac{1}{M_k}\sum_{i\in I_k^c} (1-D_i)\left(\left(T_i - \hat{\lambda}_k\right) Y_i - q_i'\beta\right)^2\right] + \frac{\lambda_{2k}}{M_k} \parallel \hat{\Upsilon}_{2k}\beta \parallel_1$$

are the modified Lasso estimators proposed in Belloni et al. (2012). The choices of the penalty levels and loadings $\left(\lambda_{1k}, \lambda_{2k}, \hat{\Upsilon}_{1k}, \hat{\Upsilon}_{2k}\right)$ suggested by Belloni et al. (2012) are provided in the appendix.

3.2. Asymptotic properties

In this section, I show the theoretical properties of the DMLDiD estimator $\tilde{\theta}$. In particular, I will show that the estimator $\tilde{\theta}$ can achieve $\sqrt{N}$-consistency and asymptotic normality as long as the

first-step estimators converge at rates faster than $N^{-1/4}$. This rate of convergence can be achieved by many ML methods, including Lasso and Logit Lasso.

The critical difference between DMLDiD and Abadie's DiD estimator is the score functions on which they are based. The new score functions (3.1) and (3.2) have the directional (or the Gateaux) derivatives equal to zero with respect to their infinite-dimensional nuisance parameters, whereas the scores based on (2.1) and (2.2) do not have this property. This property is the so-called Neyman orthogonality in Chernozhukov et al. (2018).

The definition of the Neyman-orthogonal score provided here is slightly different from the definition in Chernozhukov et al. (2018). Instead of being orthogonal against all nuisance parameters, the Neyman-orthogonal score defined here is orthogonal against only those infinite-dimensional nuisance parameters. Formally, let $\theta_0 \in \Theta$ be the low-dimensional parameter of interest, ρ_0 be the true value of the finite-dimensional nuisance parameter ρ, and η_0 the true value of the infinite-dimensional nuisance parameter $\eta \in \mathcal{T}$. Suppose that W is a random element taking values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$, with probability measure P. Define the directional (or the Gateaux) derivative against the infinite-dimensional nuisance parameter $D_r : \tilde{\mathcal{T}} \to \mathbb{R}$, where $\tilde{\mathcal{T}} = \{\eta - \eta_0 : \eta \in \mathcal{T}\}$,

$$D_r\left[\eta - \eta_0\right] \equiv \partial_r \left\{E_P\left[\psi\left(W, \theta_0, \rho_0, \eta_0 + r\left(\eta - \eta_0\right)\right)\right]\right\}, \eta \in \mathcal{T},$$

for all $r \in [0, 1)$. For convenience, denote

$$\partial_\eta E_P \psi\left(W, \theta_0, \rho_0, \eta_0\right)\left[\eta - \eta_0\right] \equiv D_0\left[\eta - \eta_0\right], \eta \in \mathcal{T}.$$

In addition, let $\mathcal{T}_N \subset \mathcal{T}$ be a nuisance realisation set such that the estimator of η_0 takes values in this set with high probability.

DEFINITION 2. *The score ψ obeys the Neyman orthogonality condition at $(\theta_0, \rho_0, \eta_0)$ with respect to the nuisance parameter realisation set $\mathcal{T}_N \subset \mathcal{T}$ if the directional derivative map $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_N$ and vanishes at $r = 0$:*

$$\partial_\eta E_P \psi\left(W, \theta_0, \rho_0, \eta_0\right)\left[\eta - \eta_0\right] = 0, \text{ for all } \eta \in \mathcal{T}_N.$$

LEMMA 3.1. *The new score functions (3.1) and (3.2) obey the Neyman orthogonality.*

The proof of this lemma can be found in the online appendix. In fact, it is also possible to derive the Neyman-orthogonal scores with respect to both finite- and infinite-dimensional nuisance parameters. However, the functional forms are much more complicated than the score functions (3.1) and (3.2), and this will make the corresponding estimator not as neat as the estimators proposed here. Because they will enjoy the same asymptotic properties, here I focus only on the estimators based on (3.1) and (3.2).

In the following, I will discuss the theoretical properties of the new estimator $\tilde{\theta}$ when data belong to repeated outcomes and repeated cross sections. The results of multilevel treatment can be proved by using the same arguments. Let κ and C be strictly positive constants, $K \geq 2$ be a fixed integer, and ε_N be a sequence of positive constants approaching zero. Denote by $\|\cdot\|_{P,q}$ the L^q norm of some probability measure P: $\|f\|_{P,q} \equiv \left(\int |f(w)|^q dP(w)\right)^{1/q}$ and $\| f \|_{P,\infty} \equiv \sup_w | f(w) |$.

ASSUMPTION 3.1 (REGULARITY CONDITIONS FOR REPEATED OUTCOMES). *Let P be the probability law for (Y(0), Y(1), D, X). Let $D = g_0(X) + U$ and $Y(1) - Y(0) = \ell_{10}(X) + V_1$, with $E_P[U|X] = 0$ and $E_P[V_1|X, D = 0] = 0$. Define $G_{1p0} \equiv E_P[\partial_p \psi_1(W, \theta_0, p_0, \eta_{10})]$ and $\Sigma_{10} \equiv E_P[(\psi_1(W, \theta_0, p_0, \eta_{10}) + G_{1p0}(D - p_0))^2]$. For the above definition, the following conditions hold: (a) $Pr(\kappa \leq g_0(X) \leq 1 - \kappa) = 1$; (b) $\|UV_1\|_{P,4} \leq C$; (c) $E[U^2|X] \leq C$; (d) $E\left[V_1^2 \mid X\right] \leq C$; (e)*

$\Sigma_{10} > 0$; *and (f) given the auxiliary sample* I_k^c, *the estimator* $\hat{\eta}_{1k} = \left(\hat{g}_k, \hat{\ell}_{1k}\right)$ *obeys the following conditions. With probability* $1 - o(1)$, $\| \hat{\eta}_{1k} - \eta_{10} \|_{P,2} \leq \varepsilon_N$, $\| \hat{g}_k - 1/2 \|_{P,\infty} \leq 1/2 - \kappa$, *and* $\| \hat{g}_k - g_0 \|_{P,2}^2 + \| \hat{g}_k - g_0 \|_{P,2} \times \| \hat{\ell}_{1k} - \ell_{10} \|_{P,2} \leq (\varepsilon_N)^2$.

ASSUMPTION 3.2 (REGULARITY CONDITIONS FOR REPEATED CROSS SECTIONS). *Let P be the probability law for (Y, T, D, X). Let* $D = g_0(X) + U$ *and* $(T - \lambda_0)Y = \ell_{20}(X) + V_2$, *with* $E_p[U|X] = 0$ *and* $E_p[V_2|X, D = 0] = 0$. *Define* $G_{2p0} \equiv E_P[\partial_p \psi_2(W, \theta_0, p_0, \lambda_0, \eta_{20})]$, $G_{2\lambda 0} \equiv E_P[\partial_\lambda \psi_2(W, \theta_0, p_0, \lambda_0, \eta_{20})]$, *and* $\Sigma_{20} \equiv E_P[(\psi_1(W, \theta_0, p_0, \eta_{10}) + G_{2p0}(D - p_0) + G_{2\lambda 0}(T - \lambda_0))^2]$. *For the above definition, the following conditions hold: (a)* $Pr(\kappa \leq g_0(X) \leq 1 - \kappa) = 1$; *(b)* $\|UV_2\|_{P,4} \leq C$; *(c)* $E[U^2|X] \leq C$; *(d)* $E\left[V_2^2 \mid X\right] \leq C$; *(e)* $E_P[Y^2|X] \leq C$; *(f)* $|E_P[YU]| \leq C$; *(g)* $\Sigma_{20} > 0$; *and (h) given the auxiliary sample* I_k^c, *the estimator* $\hat{\eta}_{2k} = \left(\hat{g}_k, \hat{\ell}_{2k}\right)$ *obeys the following conditions. With probability* $1 - o(1)$, $\| \hat{\eta}_{2k} - \eta_{20} \|_{P,2} \leq \varepsilon_N$, $\| \hat{g}_k - 1/2 \|_{P,\infty} \leq 1/2 - \kappa$, *and* $\| \hat{g}_k - g_0 \|_{P,2}^2 + \| \hat{g}_k - g_0 \|_{P,2} \times \| \hat{\ell}_{2k} - \ell_{20} \|_{P,2} \leq (\varepsilon_N)^2$.

THEOREM 3.1. *For repeated outcomes, suppose Assumptions 2.1, 2.2, and 3.1 hold. For repeated cross sections, suppose Assumptions 2.1 through 2.3 and 3.2 hold. If* $\varepsilon_N = o(N^{-1/4})$, *the new ATT estimator* $\tilde{\theta}$ *obeys*

$$\sqrt{N}\left(\tilde{\theta} - \theta_0\right) \to N(0, \Sigma)$$

,*with* $\Sigma = \Sigma_{10}$ *for repeated outcomes and* $\Sigma = \Sigma_{20}$ *for repeated cross sections.*

THEOREM 3.2. *Construct the estimators of the asymptotic variances as*

$$\hat{\Sigma}_1 = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\left(\psi_1\left(W, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right) + \hat{G}_{1p}\left(D - \hat{p}_k\right)\right)^2\right] \qquad \text{(repeated outcomes)}$$

$$\hat{\Sigma}_2 = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\left(\psi_2\left(W, \tilde{\theta}, \hat{p}_k, \hat{\lambda}_k, \hat{\eta}_{2k}\right) + \hat{G}_{2p}\left(D - \hat{p}_k\right) + \hat{G}_{2\lambda}\left(T - \hat{\lambda}_k\right)\right)^2\right]$$

(repeated cross sections)

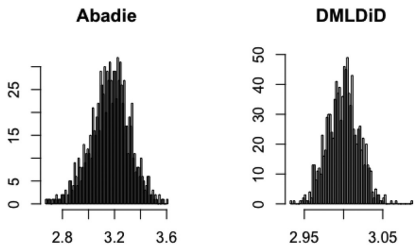
,

where $\mathbb{E}_{n,k}[f(W)] = n^{-1}\sum_{i \in I_k} f(W_i)$, $\hat{G}_{1p} = \hat{G}_{2p} = -\tilde{\theta}/\hat{p}_k$, *and* $\hat{G}_{2\lambda}$ *is a consistent estimator of* $G_{2\lambda 0}$. *If the assumptions of Theorem 1 hold,* $\hat{\Sigma}_1 = \Sigma_{10} + o_P(1)$ *and* $\hat{\Sigma}_2 = \Sigma_{20} + o_P(1)$.

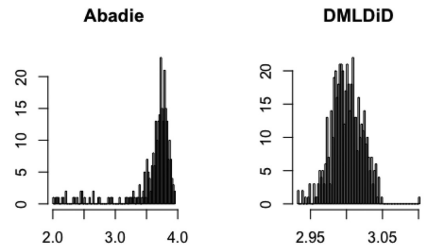
Theorem 3.1 shows that DMLDiD $\tilde{\theta}$ can achieve $\sqrt{N}$-consistency and asymptotic normality if the first-step estimators of the infinite-dimensional nuisance parameters converge at a rate faster than $N^{-1/4}$. This rate of convergence can be achieved by many ML methods. In particular, Van de Geer (2008) and Belloni et al. (2012) provided detailed conditions for Logit Lasso and the modified Lasso estimators to satisfy this rate of convergence. Theorem 3.2 provides consistent estimators for the asymptotic variance of $\tilde{\theta}$. The proofs of Theorem 3.1 and Theorem 3.2 can be found in the online appendix.

4. SIMULATION

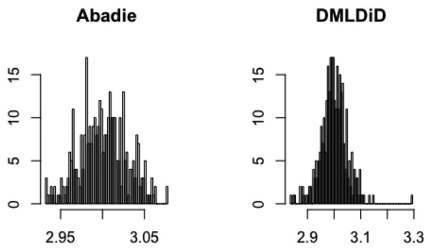
In the online appendix, I conduct Monte Carlo simulations to shed some light on the finite-sample properties of the DiD estimator of Abadie (2005) and the DMLDiD estimator $\tilde{\theta}$ in all three data structures: repeated outcomes, repeated cross sections, and multilevel treatment. For the first-step ML estimation, I generate high-dimensional data and estimate the propensity score



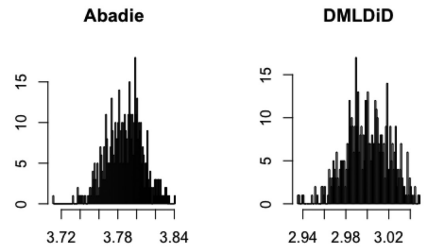
(a) Logit Lasso



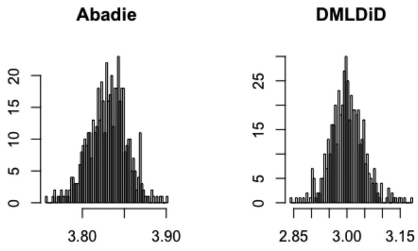
(b) SVM



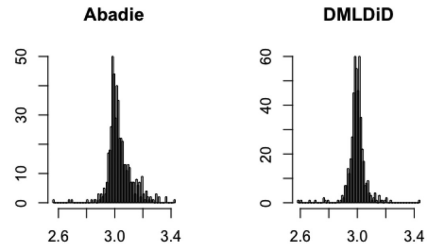
(c) Regression Tree



(d) Random Forests



(e) Boosting



(f) Neural Nets

Figure 2. The simulation for repeated outcomes with the true value $\theta_0 = 3$.

by Logit Lasso, Support vector machine (SVM), regression tree, random forests, boosting, and neural nets. I use random forests with 500 regression trees to estimate the remaining infinite-dimensional nuisance parameters. I find that although Abadie's DiD estimator suffers from the bias of a variety of ML methods, the DMLDiD estimator $\tilde{\theta}$ can successfully correct the bias and is centred at the true value. Figure 2 shows the Monte Carlo simulation and the data-generating process for *repeated outcomes*. Other cases and details are provided in the online appendix.

The data-generating process for repeated outcomes: Let $N = 200$ be the sample size and $p = 100$ the dimension of control variables, $X_i \sim N(0, I_{p \times p})$. Let $\gamma_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0) \in \mathbb{R}^p$, and D_i is generated by the propensity score $P(D = 1 \mid X) = \frac{1}{1+\exp(-X'\gamma_0)}$. Also, let the potential outcomes be $Y_i^0(0) = X_i'\beta_0 + \varepsilon_1, Y_i^0(1) = Y_i^0(0) + 1 + \varepsilon_2$, and $Y_i^1(1) = \theta_0 + Y_i^0(1) + \varepsilon_3$, where $\beta_0 = \gamma_0 + 0.5$ and $\theta_0 = 3$, and all error terms follow $N(0, 0.1)$. Researchers observe $\{Y_i(0), Y_i(1), D_i, X_i\}$ for $i = 1, ..., N$, where $Y_i(0) = Y_i^0(0)$ and $Y_i(1) = Y_i^0(1)(1 - D_i) + Y_i^1(1) D_i$.

5. EMPIRICAL EXAMPLE

In this example, I analyse the effect of tariff reduction on corruption behaviors by using the bribe payment data collected by Sequeira (2016) between South Africa and Mozambique. There have been theoretical and empirical debates on whether higher tariff rates increase incentives for corruption (Clotfelter, 1983; Sequeira and Djankov, 2014) or lower tariffs encourage agents to pay higher bribes through an income effect (Feinstein, 1991; Slemrod and Yitzhaki, 2002). The former argues that an increase in the tariff rate makes it more profitable to evade taxes on the margin, whereas the latter asserts that an increased tariff rate makes the taxpayers less wealthy, and this, under the decreasing risk aversion of being penalised, tends to reduce evasion (Allingham and Sandmo, 1972).

Sequeira (2016) collected primary data on the bribe payments between the ports in Mozambique and South Africa from 2007 to 2013. In exchange for tariff evasion, the cargo owners bribed the border officials who were in charge of validating clearance documentation and collecting all tariff payments. The exogenous variation used in Sequeira (2016) to study the effect of tariff reduction on corruption was the significant reduction in the average nominal tariff rate (of 5 percent) on certain products occurring in 2008. Because not all products were on the tariff reduction list, a credible control group of products is available. This credible control group allows for a DiD estimation. Sequeira (2016) pooled together the cross-section data between 2007 and 2013 and estimated the effect of treatment through the traditional linear DiD with many control variables. Table 9 of Sequeira (2016) presented the result of the following specification:

$$\begin{aligned} y_{it} = {} & \gamma_1 TariffChangeCategory_i \times POST \\ & + \mu POST + \beta_1 TariffChangeCategory_i \\ & + \beta_2 BaselineTariff i + \Gamma_i + p_i + w_t + \delta_i + \epsilon_{it}, \end{aligned} \tag{5.1}$$

where y_{it} is the natural log of the amount of bribe paid for shipment i in period t, conditional on paying a bribe. *TariffChangeCategory* $\in \{0, 1\}$ denotes the treatment status of commodities, *POST* $\in \{0, 1\}$ is an indicator for the years following 2008, and *BaselineTariff* is the tariff rate before the tariff reduction. The specification also includes a vector of characteristics Γ_i, and time and individual fixed effects p_i, w_t, and δ_i. The parameter γ_1 is the parameter of interest in Equation (5.1). Sequeira (2016) found that the amount of bribe paid dropped after

Table 1. Estimation results for (5.1) and (5.2).

	Equation (5.1)	Equation (5.2)
$\hat{\gamma}_1$	-2.928^{**}	0.934
	(0.944)	(2.690)
$TP \times diff$		-0.986
		(0.959)
$TP \times agri$		-1.170^{**}
		(0.580)
$TP \times lvalue$		-0.098
		(0.129)
$TP \times perishable$		0.859
		(1.213)
$TP \times largefirm$		-0.576
		(0.988)
$TP \times day\_arri$		-0.002
		(0.106)
$TP \times inspection$		-0.525
		(0.911)
$TP \times monitor$		-0.482
		(0.713)
$TP \times 2007tariff$		0.009
		(0.048)
$TP \times SouthAfrica$		-2.706^{***}
		(0.912)

the tariff reduction ($\hat{\gamma}_1 = -2.928^{**}$). However, as noted by Meyer et al. (1995), this result of Equation (5.1) excludes the heterogeneous treatment effects. The estimate might be different if we take into account the heterogeneity. To shed some light on the heterogeneous treatment effect, I incorporate the interaction terms between *TariffChangeCategory* × *POST* (*TP*) and the characteristics Γ_i into (5.1). The specification becomes

$$\begin{aligned} y_{it} &= \gamma_1 TariffChangeCategory_i \times POST + \gamma_2 TP_i \times \Gamma_i \\ &\quad + \mu POST + \beta_1 TariffChangeCategory_i \\ &\quad + \beta_2 BaselineTariff_i + \Gamma_i + p_i + w_t + \delta_i + \epsilon_{it}, \end{aligned} \tag{5.2}$$

where γ_2 is a 10×1 vector. Table 1 shows the comparison of the estimates of (5.1) and (5.2).

Column 2 of Table 1 shows that (a) after controlling for the interaction terms, the estimate for γ_1 becomes insignificantly different from zero, and (b) most of the coefficients of the interaction terms are negative. This suggests that there exists a large set of negative heterogeneous treatment effects and that Sequeira's estimate may be a weighted average of these heterogeneous treatment effects. The negative coefficients of the interaction terms justify the sign of Sequeira's estimate. However, it is ideal to treat the covariates nonparametrically when there exists heterogeneity in treatment effects, to avoid any potential inconsistency created by functional form misspecification (Abadie, 2005).

Table 2. The results of semiparametric DiD estimation.

	Sequeira (2016)	Abadie (kernel)	DMLDiD (kernel)	Abadie (Lasso)	DMLDiD (Lasso)
ATT	-2.928^{**}	-8.168^{**}	-6.998^{*}	-6.432^{**}	-5.222^{*}
	(0.944)	(3.072)	(3.752)	(2.737)	(2.647)

I estimate the ATT using both Abadie's DiD estimator and DMLDiD. Because the data are repeated cross sections, I construct the estimators on the basis of (2.2) and (3.2), respectively. The estimators with first-step kernel estimation contain one individual characteristic (the natural log of shipment value per ton), which is the only significant and continuous control variable in Table 9 of Sequeira (2016). The estimators with first-step Lasso estimation contain a list of the covariates included in Table 9 of Sequeira (2016), which consists of the characteristics of product, shipment, firm, and border officials. I choose both the bandwidth kernel and penalty level of Lasso by 10-fold cross validations. Table 2 shows the estimation result. First, we can observe that the estimates with first-step kernel are much larger than the estimates with first-step Lasso. The reason may be that more control variables are included in the latter estimates. Second, though with the same sign, Abadie's estimator (-8.168 or -6.432) is at least twice as large as previously reported by Sequeira (2016). This large effect, however, may be due to not only the robustness of semiparametric estimation on the functional form but also the finite-sample bias in the first-step nonparametric estimation. The DMLDiD estimator (-5.222) removes the first-order bias and suggests a smaller effect that is closer to Sequeira's estimate. Its value is only 60% higher than Sequeira's result. This extra effect can be explained by the misspecification of the traditional linear DiD estimator. Therefore, I obtain the same conclusion as Sequeira (2016) that tariff reduction decreases corruption, but my estimate suggests an even larger magnitude.

6. CONCLUSION

The DiD estimator survives as one of the most popular methods in the causal inference literature. A practical problem that empirical researchers face is the selection of important control variables when they confront a large number of candidate variables. Researchers may want to use ML methods to handle a rich set of control variables while taking the strength of the DiD estimator. I improve its original versions by proposing DMLDiD to allow researchers to use ML methods while still obtaining valid inferences. This additional benefit will make DiD more flexible for empirical researchers to explore a broader set of popular estimation methods and analyse more types of data sets.

REFERENCES

Abadie, A. (2005). Semiparametric difference-in-differences estimators, *Review of Economic Studies*, *72*, 1–19.

Akee, R., W. Copeland, E. J. Costello and E. Simeonova (2018). How does household income affect child personality traits and behaviors?, *American Economic Review*, *108*(3), 775–827.

Allingham, M. G. and A. Sandmo (1972). Income tax evasion: a theoretical analysis, *Journal of Public Economics*, *1*, 323–38.

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain, *Econometrica*, *80*, 2369–2429.

Belloni, A., V. Chernozhukov, I. Fernández-Val and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data, *Econometrica*, *85*, 233–98.

Belloni, A., V. Chernozhukov and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls, *Review of Economic Studies*, *81*, 608–50.

Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market, *ILR Review*, *43*, 245–57.

Card, D. and A. Krueger (1994). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review*, *84*(4), 772–93.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters, *Econometrics Journal*, *21*, C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura and W. K. Newey (2019). Locally robust semiparametric estimation, arXiv:1608.00033 [math.ST].

Chernozhukov, V., C. Hansen and M. Spindler (2015). Valid post-selection and post-regularization inference: an elementary, general approach, *Annual Review of Economics*, *7*, 649–88.

Clotfelter, C. T. (1983). Tax evasion and tax rates: an analysis of individual returns, *Review of Economics and Statistics*, *65*, 363–73.

Feinstein, J. S. (1991). An econometric analysis of income tax evasion and its detection, *RAND Journal of Economics*, *22*, 14–35.

Fuest, C., A. Peichl and S. Siegloch (2018). Do higher corporate taxes reduce wages? Micro evidence from Germany, *American Economic Review*, *108*(2), 393–418.

Li, F. (2019). Double-robust estimation in difference-in-differences with an application to traffic safety evaluation, arXiv:1901.02152 [stat.AP].

Lu, C., X. Nie and S. Wager (2019). Robust nonparametric difference-in-differences estimation, arXiv:1905.11622 [stat.ME].

Meyer, B. D., W. K. Viscusi and D. L. Durbin (1995). Workers' compensation and injury duration: evidence from a natural experiment, *American Economic Review*, *85*, 322–40.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators, *Econometrica: Journal of the Econometric Society*, *62*, 1349–82.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing, *Handbook of Econometrics*, *4*, 2111–2245.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association*, *90*, 122–29.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, *66*, 688.

Sant'Anna, P. H. and J. B. Zhao (2019). Doubly robust difference-in-differences estimators, doi:10.2139/ssrn.3293315.

Sequeira, S. (2016). Corruption, trade costs, and gains from tariff liberalization: evidence from southern Africa, *American Economic Review*, *106*(10), 3029–63.

Sequeira, S. and S. Djankov (2014). Corruption and firm behavior: evidence from African ports, *Journal of International Economics*, *94*, 277–294.

Slemrod, J. and S. Yitzhaki (2002). Tax avoidance, evasion, and administration, In *Handbook of Public Economics*, Alan J. and Martin , Vol. *3*, pp. 1423–70, Elsevier.

Van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso, *Annals of Statistics*, *36*, 614–45.

Zimmert, M. (2019). Efficient difference-in-differences estimation with high-dimensional common trend confounding, arXiv:1809.01643 [econ.EM].

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website.

Online Supplement
Replication package

Co-editor Victor Chernozhukov handled this manuscript.

APPENDIX A: MORE ON ESTIMATION

Multilevel treatments: Individuals can also be exposed to different levels of treatment. Let $W \in \{0, w_1, ..., w_J\}$ be the level of treatment, where $W = 0$ denotes the untreated individuals. Researchers observe $\{Y_i(0), Y_i(1), W_i, X_i\}_{i=1}^N$. For $w \in \{0, w_1, ..., w_J\}$ and $t \in \{0, 1\}$, let $Y^w(t)$ be the potential outcome for treatment level w at period t. Denote the ATT for each level of treatment w by

$$\theta_0^w \equiv E\left[Y^w(1) - Y^0(1) \mid W = w\right].$$

Suppose that Assumptions (2.1) and (2.2) hold for each $w \in \{w_1, ..., w_J\}$:

$$E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, W_i = w\right] = E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, W_i = 0\right],$$

$P(W_i = w) > 0$, and with probability one, $P(W_i = w|X_i) < 1$. Then, we have (Abadie, 2005),

$$\theta_0^w = E\left[\frac{Y(1) - Y(0)}{P(W = w)} \frac{I(W = w) \cdot P(W = 0 \mid X) - I(W = 0) \cdot P(W = w \mid X)}{P(W = 0 \mid X)}\right],$$

where $I(\cdot)$ is an indicator function. The Neyman-orthogonal score function for multilevel treatments is

$$\psi_w(W, \theta_{w0}, p_{w0}, \eta_{w0}) = \frac{Y(1) - Y(0)}{P(W = w)} \frac{I(W = w)P(W = 0 \mid X) - I(W = 0)P(W = w \mid X)}{P(W = 0 \mid X)} - \theta_{w0} - c_w.$$

The adjustment term c_w is

$$c_w = \left(\frac{I(W = w) \cdot P(W = 0 \mid X) - I(W = 0) \cdot P(W = w \mid X)}{P(W = w) \cdot P(W = 0 \mid X)}\right) \times E\left[Y(1) - Y(0) \mid X, I(W = 0) = 1\right].$$

The nuisance parameters are the unknown constant $p_{w0} \equiv P(W = w)$ and the infinite-dimensional parameter $\eta_{w0} = (g_{w0}, g_{z0}, \ell_{30})$, where $g_{w0} = P(W = w|X)$, $g_{z0} = P(W = 0|X)$, and $\ell_{30} = E[Y(1) - Y(0)|X, I(W = 0) = 1]$.

Multilevel treatments algorithm:

(1) *Take a K-fold random partition* $(I_k)_{k=1}^K$ *of observation indices* $[N] = \{1, ..., N\}$, *such that the size of each fold* I_k *is* $n = N/K$. *For each* $k \in [K] = \{1, ..., K\}$, *define the auxiliary sample* $I_k^c \equiv \{1, ..., N\} \setminus I_k$.

(2) *For each* $k \in [K]$, *construct the estimator of* p_0 *and* λ_0 *by* $\hat{p}_w = \frac{1}{n}\sum_{i \in I_k^c} D_i$. *Also, construct the estimators of* g_w, g_z, *and* ℓ_{30} *by using the auxiliary sample* I_k^c: $\hat{g}_{wk} = \hat{g}_w\left((W_i)_{i \in I_k^c}\right)$, $\hat{g}_{zk} = \hat{g}_z\left((W_i)_{i \in I_k^c}\right)$, *and* $\hat{\ell}_{3k} = \hat{\ell}_3\left((W_i)_{i \in I_k^c}\right)$.

(3) *For each k, construct the intermediate ATT estimators:*

$$\tilde{\theta}_{wk} = \frac{1}{n}\sum_{i \in I_k} \frac{I(W_i = w) \cdot \hat{g}_{zk}(X_i) - I(W_i = 0) \cdot \hat{g}_{wk}(X_i)}{\hat{p}_w \hat{g}_{zk}(X_i)} \times \left(Y(1) - Y(0) - \hat{\ell}_{3k}(X_i)\right).$$

(4) *Construct the final ATT estimators*: $\tilde{\theta} = \frac{1}{K}\sum_{k=1}^K \tilde{\theta}_k$.

Lasso penalty. The following is suggested by Belloni et al. (2012). Let y_i denote $Y_i(1) - Y_i(0)$ or $\left(T_i - \hat{\lambda}_k\right)$, λ_k denote λ_{1k} or λ_{2k}, and $\hat{\Upsilon}_k$ denote $\hat{\Upsilon}_{1k}$ or $\hat{\Upsilon}_{2k}$. For $k \in [K]$, the loading $\hat{\Upsilon}_k$ is a diagonal matrix with entries $\hat{\gamma}_{kj}, j = 1, ..., p$, constructed by the following steps:

$$\text{Initial } \hat{\gamma}_{kj} = \sqrt{\frac{1}{M_k}\sum_{i\in I_k^c}(1-D_i)\,q_{ij}^2\,(y_i-\bar{y}_k)^2},\ \lambda_k = 2c\sqrt{M_k}\,\Phi^{-1}(1-\gamma/2p),$$

$$\text{Refined } \hat{\gamma}_{kj} = \sqrt{\frac{1}{M_k}\sum_{i\in I_k^c}(1-D_i)\,q_{ij}^2\hat{\varepsilon}_i^2},\ \lambda_k = 2c\sqrt{M_k}\,\Phi^{-1}(1-\gamma/2p),$$

where $\bar{y}_k = M^{-1}\sum_{i\in I_k^c} y_i$, $c > 1$ and $\gamma \to 0$. The empirical residual $\hat{\varepsilon}_i$ is calculated by the modified Lasso estimator β_k^* in the previous step: $\hat{\varepsilon}_i = y_i - q_i'\beta_k^*$. Repeat the second step $B > 0$ times to obtain the final loading.