Synthetic learner: Model-Free Inference on Treatments over Time

Davide Viviano[1*] Jelena Bradic[2 †‡]

Stanford GSB[1] UC San Diego[1,2]

First version: April, 2019
This version: July, 2022

Abstract

Understanding the effect of a particular treatment or a policy pertains to many areas of interest, ranging from political economics, marketing to healthcare. In this paper, we develop a non-parametric algorithm for detecting the effects of treatment over time in the context of Synthetic Controls. The method builds on counterfactual predictions from many algorithms without necessarily assuming that the algorithms correctly capture the model. We introduce an inferential procedure for detecting treatment effects and show that the testing procedure is asymptotically valid for stationary, beta mixing processes without imposing any restriction on the set of base algorithms under consideration. We discuss consistency guarantees for average treatment effect estimates and derive regret bounds for the proposed methodology. The class of algorithms may include Random Forest, Lasso, or any other machine-learning estimator. Numerical studies and an application illustrate the advantages of the method.

Keywords: Synthetic Control, Difference In Differences, Causal Inference, Random Forests.
JEL Code: C10, C14, C20, C30.

*Present address: Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305. Email: dviviano@stanford.edu. This work was mostly conducted while at the Department of Economics, University of California at San Diego, La Jolla, CA, 92093.

†Department of Mathematics and Halıcıoğlu Data Science Institute, University of California at San Diego, La Jolla, CA, 92093. Email: jbradic@ucsd.edu.

‡Jelena Bradic gratefully acknowledges the support of the grant NSF-DMS #1712481.

1 Introduction

This paper discusses estimation and inference on the effect of a policy intervention on a single unit observed over multiple periods and exposed to treatment from one point in time onwards. We consider an aggregate time-series set-up where researchers observe the outcome of the unit of interest T_0 periods before treatment and $T - T_0$ after the treatment. Researchers' main goal is to conduct inference on the effect of a trajectory of treatment effects over the post-treatment period. Namely, by denoting $\tau_t = Y_{0t}^1 - Y_{0t}^0$ the difference between the two potential outcomes at time t, researchers want to test whether $\{\tau_t\}_{t>T_0} = \tau^o$, for some null trajectory of interest τ^o. Their second goal is to precisely estimate the average effect over time.

In the same spirit of synthetic controls (Abadie and Gardeazabal, 2003; Abadie et al., 2010), and aggregate panel data models (Hsiao et al., 2012), we construct counterfactuals using information from n (possibly finitely many) control units and covariates observed over T periods. We exploit information over the time series for conducting asymptotic inference. The applications of interest are those where we observe individuals at a relatively frequent level, e.g., quarterly or monthly, over multiple years. Examples are studying the effects of taxation (Bai et al., 2014), changes in welfare programs (Maclean et al., 2019) or (geo-localized) marketing experiments (Brodersen et al., 2015; Varian, 2016). We discuss additional applications at the end of Section 1.1.

The first contribution of this paper is to derive an inferential procedure that is valid for general parametric and non-parametric (machine-learning) estimators. We propose a resampling mechanism to guarantee exact asymptotic size control without imposing restrictions on the set of estimators. The key idea for inference is to combine the sample splitting procedure with the block-bootstrap (Politis and Romano, 1992) and exploit the differentiability properties of the proposed procedure. Our approach does not require correct model specifications.

As a second contribution, we introduce an ensemble procedure that combines many such estimators, e.g., Synthetic controls, factor models estimators, and fully non-parametric estimators such as Random Forest or Kernel Smoothing, into a single prediction. We penalize the methods with the worst out-of-sample performance over the pre-treatment period. The method's goal is to increase precision. Also, it permits replacing heuristic model selection criteria whose properties are unknown with time-dependent observations with a theoretically grounded procedure. This has important implications for applied economists, who often face the difficult model selection choice from a dictionary that includes many methods (e.g., factor models, synthetic control, difference-in-differences).[1] The ensemble method that we propose builds on the literature on exponential aggregation (Rigollet et al., 2012; Cesa-Bianchi and Lugosi, 2006), time-

[1] See also the discussion in Athey et al. (2019).

series and forecasts' combinations (Timmermann, 2006; Elliott and Timmermann, 2004), which we study here for causal inference. We show that the procedure consistently recovers the average treatment effect through a bias adjustment, and it inherits strong oracle properties for its prediction performance.

Throughout the text, we assume exogeneity of the treatment time and stationarity for inference. These conditions are common when conducting inference with Synthetic Controls in the presence of a long time series[2], since they permit to conduct inference without relying on symmetry assumptions of placebo testing (Firpo and Possebom, 2018; Ben-Michael et al., 2021). They imply that a pre-post treatment comparison of the means returns a consistent (but possibly inefficient) estimate of the average effect on the treated. In separate sections, we relax stationarity in two directions: (i) allowing for time-varying fixed effects, encompassing popular two-way fixed-effect models; (ii) deriving prediction guarantees for arbitrary non-stationary settings.[3]

We conclude our discussion with a simulation study and an empirical application. We show that our procedure leads to larger power than existing methods while controlling the size of the test. In an application, we study the effect of Tennessee's health-insurance dis-enrollment program on health-related outcomes. Using survey data from Behavioral Risk Factor Surveillance System Data, we show that the program decreased health insurance coverage and the likelihood of visiting a doctor.

The paper is organized as follows. In Section 2 we introduce the set-up, and identification strategy. In Section 3, we introduce the method for estimating counterfactuals and inference on sharp nulls. In Section 4, we discuss estimation of the average effect on the treated. In Section 5 we discuss prediction guarantees under non-stationarity. Section 6 discusses numerical experiments. Section 7 discusses applications in health economics. Finally, in Section 8 we provide extensions in the presence of carry-over effects.

1.1 Related Work

While non-parametric estimators have found widespread use in microeconomic applications (Athey and Imbens, 2019), their analysis (and consequently their applicability) in the presence of aggregate data has received much less attention. However, with aggregate data, such estimators can improve precision and better disentangle the effect of the policy from idiosyncratic shocks. This paper proposes a method that enables counterfactual prediction and hypothesis testing in the context of Synthetic Controls (SC) using predictions arising from many parametric and non-parametric estimators.

[2]For inference via Synthetic Control Chernozhukov et al. (2018) impose stationarity of the residuals under correct model specification and similar stationarity assumptions as to the one discussed above under misspecification to show the validity of permutation tests. Stationarity, strong mixing conditions on the joint distribution of the residual errors and covariates, are also imposed for valid inference in synthetic control settings in Carvalho et al. (2018), while covariance stationarity conditions are imposed in Li and Bell (2017). Further discussion is in Section 2.

[3]These are in Section 4.1 and Section 5 respectively.

Recent literature has proposed a wide variety of methods for predicting counterfactuals in the setting under consideration, including factor and panel data models (Bai, 2009; Hsiao et al., 2012), synthetic controls (Abadie et al., 2010; Xu, 2017; Doudchenko and Imbens, 2016; Arkhangelsky et al., 2021; Ferman and Pinto, 2016), two-ways fixed effects models (Imai and Kim, 2021), ridge regression methods (Ben-Michael et al., 2021), kernel balancing (Hazlett and Xu, 2018), functional methods (Gunsilius, 2020) among others. Additional references include Athey et al. (2021) who proposes a matrix completion methods for SC; Athey et al. (2019) shows in a simulation exercise that ensemble methods outperform individual SC predictions on many economic data sets; Amjad et al. (2018) proposes singular value thresholding, whereas matching has been discussed by Imai et al. (2018). Difference-in-difference methods were recently discussed in Athey and Imbens (2022) and Arkhangelsky et al. (2021) in the context of staggered adoption.

However, selection among these methods remains an open research question (Hsiao and Zhou, 2019). Combining different predictions offers a simple data-adaptive procedure to exploit information from all these models while improving the prediction performance (Timmermann, 2006). As a result, our contribution can be viewed as complementary to this literature. To the best of our knowledge, we provide the first set of conditions under which predictions made by any or many machine learning methods, including Random Forests, can be used to develop valid tests for SC.

A closely related method to our inference procedure is the permutation-based inference method, discussed in Chernozhukov et al. (2021). Their method accommodates a single (linear) model specification only and imposes stability conditions of the estimator. Also, while the permutation-based methods estimate counterfactuals using the entire sample (pre and post-treatment periods) as a training set, imposing the sharp null hypothesis of no treatment effects, here we estimate counterfactuals using the pre-treatment period only. This approach permits that estimation of the counterfactual does not depend on the outcomes observed over the post-treatment period. Our approach is particularly suited whenever the post-treatment period is proportional to the pre-treatment period as in the context of our empirical application.[4]

Our paper relates more broadly to inference using penalized methods, including Chernozhukov et al. (2018) and Carvalho et al. (2018), which propose penalized linear regressors for asymptotic inference on treatment effects in an SC setup. However, Carvalho et al. (2018) requires a consistent estimation of treatment effects, which is not required in our setting and focuses on a penalized model only. Differently, Chernozhukov et al. (2018) proposes a bias-adjustment procedure for asymptotic inference, which in our framework is not required for hypothesis testing but used for average treatment effect on the treated (ATT) estimation. Our resampling mechanism allows for more generality than the tests in Chernozhukov

[4]In our empirical application, the post-treatment period is approximately twenty, and the pre-treatment period is approximately sixty. The reader may refer to Example 3.1 for an illustration of the benefits.

et al. (2018) since we do not require the use of penalized linear regressor and the related conditions for estimation of counterfactuals, but we allow for general non-parametric estimators. Related literature also includes Li and Bell (2017), and Hsiao et al. (2012), which discuss properties of constrained least-squares methods under stationarity and correct model specification only. Li (2020) discusses a sub-sampling procedure for inference with constrained least-squares estimators only.

Finally, we relate to Künzel et al. (2019) who discuss model averaging with *i.i.d.* data and classifies method into three classes, denoted as S, T, and X learners. This paper pioneers the idea of T-learning in Synthetic control setting, offering an alternative and simple weighting scheme which is inspired by the literature on boosting (Schapire and Freund, 2012) and online learning (Cesa-Bianchi et al., 1997, 1999; Cesa-Bianchi and Lugosi, 2006).

To conclude, we can list several applications of interest. The first set of relevant applications includes changes in policy at a regional or state level. For instance, studying the effects of (i) exogenous variations such as environmental disasters on policy-changes (Potrafke and Wuthrich, 2020), or economic outcomes (Cavallo et al., 2013); (ii) of Medicaid expenditure expansions or contractions (Tello-Trillo, 2021); (iii) of tax-reform on prices (Bai et al., 2014); or (iv) of policy reforms on economic growth (Billmeier and Nannicini, 2013). The second avenue of interesting applications includes experiments on online platforms, for which synthetic controls are becoming increasingly popular, especially in the presence of geo-localized experiments (Varian, 2016; Li, 2017). The third set of applications includes studying the effect of new algorithms or financial instruments (Xie and Huang, 2014; Bojinov and Shephard, 2019).

2 Setup and Identification

Throughout this article for each unit i we observe outcome variable Y_{it}. We denote with Y_{0t} the unit treated if $D_t = 1$ and under control otherwise; the remaining $i = 1, \cdots, n$ units, $Y_{1t}, \ldots, Y_{nt}$ are units always observed in the control state. Additional covariate information for each unit are denoted in compact form as Z_{it}. Z_{it} may also contain past covariates and past outcomes.

2.1 Estimands and Null Hypothesis

Following the literature on panel data and synthetic control models (e.g., Hsiao and Zhou, 2019; Abadie et al., 2010), we define the treatment assignment and the outcome of interest, respectively as

$$D_t = 1\{t > T_0\}, \quad Y_{0t} = D_t Y_{0t}^1 + (1 - D_t) Y_{0t}^0, \quad Y_{jt} = Y_{jt}^0, \quad j > 0 \tag{1}$$

where Y_{0t}^1, Y_{0t}^0 denote the potential outcomes under treatment and control for the unit of interest $i = 0$, and Y_{jt}^0 denotes the potential outcome under control of unit j. Our definition of potential outcomes implicitely imposes that SUTVA holds (Rubin, 1990) and no carry-over effects (Imai et al., 2013). Extensions in the presence of carry-overs are discussed in Section 8 .

In applications, researchers may want to test whether the difference between two potential outcomes $Y_{0t}^1 - Y_{0t}^0$ equals zero for all post-treatment periods. Namely, they may be interested in conducting inference on the time-specific treatment effects defined below.

Definition 2.1 (Time-specific treatment effect)**.** The time-specific treatment effect is defined as follows: $\tau_t = Y_{0t}^1 - Y_{0t}^0$.

Here, τ_t defines the difference in potential outcomes at time t. We begin our discussion by introducing the null hypothesis of interest.

Definition 2.2 (Sharp Null hypothesis)**.** Define the sharp null hypothesis as

$$H_0 : \tau_t = \alpha_t^o, \quad t \in \{T_0 + 1, \cdots, T\}, \tag{2}$$

for a known sequence $\{\alpha_t^o\}_{t>T_0}$.

Equation (2) imposes that the potential outcome over the post-treatment period equals the potential outcome under control plus a known (possibly time-varying) constant. For example, we may consider $a_t^o = 0$ or we may consider also testing a linear trend of the form $a_t^o = \delta(t - T_0)$ for an arbitrary $\delta \in \mathbb{R}$. Our results also extend when we test the average of τ_t.[5] Finally, note that more generally, we can incorporate a more general class of hypothesis $H_0 : Y_{0t}^1 = f(Y_{0t}^0, a_t^o),\ a_t^o \in \mathbb{R},\ t > T_0$, for a function f being invertible in its first argument, omitted for the sake of brevity. We define

$$Y_{0t}^o = \begin{cases} Y_{0t} - \alpha_t^o & t > T_0, \\ Y_{0t} & t \leq T_0. \end{cases}$$

the (observed) potential outcome under control under the null hypothesis H_0.

Testing for treatment effects may not be satisfactory to researchers, who may also be interested in reporting estimates of average treatment effects. This is defined below.

[5]Namely, we may also test $\mathbb{E}[\tau_t] = \alpha_t^o$, hence allowing potential outcomes under treatment and control having different idiosyncratic shocks. This is omitted for the sake of brevity only and discussed in the Appendix (Algorithm G.1) and Remark 3.

Definition 2.3 (Average Treatment Effect on the Treated)**.** The average treatment effect on the treated is defined as

$$\tau = \frac{1}{T - T_0} \sum_{t > T_0} \mathbb{E}\Big[Y_{0t}^1 - Y_{0t}^0\Big]. \tag{3}$$

Here, τ denotes the average effect on the treated, averaged also over the post-treatment period. The expectation is taken over the idiosyncratic shocks.[6]

2.2 Identification Conditions

The first condition that we impose is stationarity. This is common in the literature on Synthetic Control, see Carvalho et al. (2018); Li and Bell (2017); Chernozhukov et al. (2018).[7] We relax it in Section 5, where we allow for non-stationary observations.

Assumption 1 (Stationarity)**.** Suppose that $(Y_{0t}^0, Y_{1n}, \cdots, Y_{nt}, Z_{0t}, \cdots, Z_{nt}) \sim \mathcal{D}_0$ is stationary.

Assumption 1 imposes stationarity. In Section 4.1 we show how our results for inference extend under non-stationarity, in the presence of an (unknown) time-varying fixed-effects.

The second condition we impose is an identification assumption.

Assumption 2 (Identification Condition)**.** Suppose that $T_0 \perp (Y_{0t}^0, Y_{0t}^1, Y_{1:n,t}, Z_{0:n,t})_{t=1}^T$.

Assumption 2 states that the timing of the treatment is exogenous. The same conditions can be found in previous literature on Synthetic Controls. For instance, recent literature on Synthetic Control (Abadie et al., 2010; Chernozhukov et al., 2018; Arkhangelsky et al., 2021; Chernozhukov et al., 2018; Li and Bell, 2017) treated T_0 as deterministic, in which case exogeneity of T_0 implicitely holds. See for example, the discussion in Ferman and Pinto (2016) and Bottmer et al. (2021). Carvalho et al. (2018) also explicitely imposes exogeneity similarly to the above condition.

In the context of our empirical application, where the treatment consists of a dis-enrollment from Medicaid occurring in the early 2000s in Tennessee, and the outcomes are health-related outcomes, as argued in Argys et al. (2020), if the policy can be attributed to budget deficit interpretable as an exogenous variation, the assumption directly holds. We warn the reader, however, that failure of the assumption may invalidate the inferential strategy.

Motivated by Assumption 2 we will implicitely condition on T_0 throughout the rest of our discussion unless otherwise specified, since, conditional on T_0 the distribution of observables and unobserved potential

[6] Note that whenever τ_t is non-random $\tau = \frac{1}{T-T_0} \sum_{t>T_0} \tau_t$.

[7] Stationarity and beta-mixing conditions as stated above cover a large class of ARMA processes (Pham and Tran, 1985), AR-ARCH processes (Lange et al., 2011), Markov Switching Processes (see for example Lee, 2005), GARCH (Carrasco and Chen, 2002), to cite some.

outcomes remains invariant. We return instead to non-stationary conditions in Section 5 where we relax Assumption 1 and 2.

Example 2.1 (Stationary factor models)**.** Suppose that

$$Y_{jt}^0 = \mu_j + \theta_t + \lambda_j F_t + \gamma(Z_{jt}) + u_{j,t}, \quad j \in \{0, \cdots, n\}, \tag{4}$$

with $u_{j,t}$ denoting stationary idiosyncratic errors. Let $\theta_t \sim \mathcal{N}(0,1)$ and exogenous, with, F_t denoting common stationary unobserved exogenous factors and λ_j the (exogenous) individual specific effect, and $\gamma(Z_{jt})$ a stationary components which depends on covariates. Then Assumption 1 holds. □

2.3 Testable Implications and Identification of Average Effects

We conclude this discussion with two lemmas. The first lemma provides a testable implication for the sharp null hypothesis in Definition 2.2. This is stated below.

Lemma 2.1 (Sharp Null: Testable Implication)**.** *Under the null hypothesis in Equation (2) and Assumptions 1, 2, then* $(Y_{0t}^o, Y_{1t}, \cdots, Y_{nt}, Z_{0t}, \cdots, Z_{nt})$ *is stationarity and independent of* T_0 *for all* $t \in \{1, \cdots, T_0, \cdots, T\}$.

The above condition implies that the distribution before and after the intervention period T_0 must remain invariant under the null hypothesis. This assumption is testable since we observe the empirical distribution of the above vector before and after the intervention time T_0. Lemma 2.1 is at the basis of our approach for testing the null hypothesis, which we discuss in Section 3.

The second goal is to estimate the average effect. Identification of τ is discussed in the following lemma.

Lemma 2.2 (Identification of τ)**.** *Let Assumptions 1, 2 hold. Then*

$$\tau = \frac{1}{T - T_0} \sum_{t > T_0} \mathbb{E}\Big[Y_{0t}^1 | T_0\Big] - \frac{1}{T_0} \sum_{s=1}^{T_0} \mathbb{E}\Big[Y_{0s}^0 | T_0\Big].$$

Lemma 2.2 is an identification result. It states that the post-treatment difference in expectation equals the target estimand τ. While Lemma 2.2 is not invoked for constructing our test, it is used for estimating the average effect discussed in Section 4.

The proofs of the lemmas are contained in Appendix A.

Remark 1 (Extensions with time fixed effects)**.** In Section 4.1 we show how our results directly extend to the case where

$$Y_{0t}^0 = \kappa_t^0 + \iota_j + \varepsilon_{j,t}^0, \quad \mathbb{E}[\varepsilon_{j,t}^0] = 0,$$

where κ_t^0 denotes a time-fixed effect. This extension, while simple, has important implications: it permits incorporating non-stationary unobserved components. For this case, stationary can fail as long as the control group is "representative" of the treated unit, i.e., time fixed effects are the same between the control and treated unit. This follows in the same spirit of two-way fixed-effect models (Imai and Kim, 2021) that are commonly encountered in applications see, for example, Garthwaite et al. (2014).[8] In the presence of time-fixed effects, identification is achieved by a difference-in-difference as opposed to a pre-post treatment comparison as discussed in Lemma 4.2. □

3 Counterfactuals Predictions and Hypothesis Testing

In this section, we discuss the problem of estimating the counterfactual prediction $\hat{Y}_{0t}^0, t > T_0$ and conducting inference on the sharp null hypothesis in Definition 2.2.

Throughout our discussion, for expositional convenience, we denote in the compact form

$$X_t = (Y_{1t}, \ldots, Y_{nt}, Z_{0t}^\top, Z_{1t}^\top, \ldots, Z_{nt}^\top) \in \mathcal{X}.$$

To test the null hypothesis of interest, we first estimate the true potential outcome $Y_{0t}^0, t > T_0$, unobserved over the post-treatment period. We define $\widehat{Y}_{0t}^0$ its estimate constructed as follows

$$\widehat{Y}_{0t}^0 = w(F_0)^\top g(X_t). \tag{5}$$

Here $w(F_0)$ denotes a generic functional of the empirical distribution F_0 of (Y_{0t}, X_t) over the pre-treatment period, where Y_{0t} serves as the main outcome of interest. The choice of $w(F_0)$ can be arbitrary, with the only condition required that $w(\cdot)$ is a Hadamard differentiable functional (see Assumption 4).[9] In Section 5 we provide explicit expressions for $w(\cdot)$.[10]

It is important to note that the functions $g(X_t)$ can be data-dependent. Such functions are estimated as

[8] For example, in the context of our application for studying the effect of Tennessee dis-enrollment health-insurance program, Obamacare between 2010 and 2014 may act as a time-varying confounder. Therefore, we use as the control group for the effect of dis-enrollment in Tennessee the outcome from the other Southern States that, similarly to Tennessee, did not expand Medicaid between 2010-2014 due to Obamacare.

[9] Definition of Hadamard differentiability is provided in Appendix B.1.

[10] Since the functions $g(\cdot)$ can contain an intercept, the component $w(F_0)^\top g(X_t)$ also estimates the (time-invariant) shift in mean after subctracting $\bar{Y}_t$. See, for instance, Example 3.1.

described in Algorithm 1. Namely, first, we divide the pre-treatment period into two blocks $t \in \{T_-, \cdots, 0\}$ and $t \in \{1, \cdots, T_0\}$. We define F_- the empirical distribution for $t < 1$ of (Y_{0t}, X_t). We construct the predictors as follows

$$\Big\{ x \mapsto g_j(x; F_-) \Big\} \tag{6}$$

with F_- denoting the training set for such predictors, and g_j denoting some pre-specified regressor. For example $g_1(X_t; F_-)$ can denote the prediction of a Random Forest, trained over the sample $t < 0$, with empirical distribution F_-. Whenever clear from the context, we will omit the second argument F_- from the function $g_j(\cdot)$. Throughout the rest of our discussion, we will fix the size of T_- and consider asymptotics as $T \to \infty, T_0 \propto T$.

We conclude our discussion with two cases of interest.

Case 1: Ensemble To gain further intuition on Equation (5), observe that we can interpret $w(F_0)$ as some data-dependent weights, while the functions $g(X_t)$ are interpreted as predictors or "experts" that, based on information X_t, predict the counterfactual outcome Y_{jt} (net of time-fixed effects) at time t. This follows in the same spirit of forecasts combinations (Timmermann, 2006), with g_i denoting some "experts" or learners. Section 5 discusses examples and properties of weighting schemes. The construction of the weights is based on out-of-sample performance: it uses information F_0 which has not been used for the training of the algorithms $g(\cdot)$ (that instead use information before time $t = 0$).

Case 2: Single Regressor Equation (5) is, however, more general than only ensemble methods, since it also allows for estimation with a single regressor $g_1(X_t)$ (e.g., the Synthetic Control).

Algorithm 1 Counterfactual Estimation with Sample Splitting

Require: Observations $\{Y_{0t}, X_t\}_{t=T_-}^{T}$, time of the treatment-$T_0$, learners $F \mapsto g_1(\cdot, F), \ldots, g_p(\cdot, F)$

1: Split the pre-treatment period into two parts: $t \in [T_-, 0]$ and $t \in [1, T_0]$
2: Form predictions $g_j(\cdot; F_-)$ with F_- being the empirical distribution of $\{Y_{0t}, X_t\}_{t<1}$, $j \in 1, \ldots, p$.
3: Use the second pre-treatment period, $\{Y_{0t}, X_t\}_{t=1}^{T_0}$, to estimate the weights of the learners, $w(F_0)$.
4: Compute the predicted counterfactual

$$\hat{Y}_{0t}^0 = \textstyle\sum_{j=1}^{p} w_j(F_0) g_j(X_t) \text{ for } t > T_0$$

return the predictions $(\hat{Y}_{T_0+1}^0, \cdots, \hat{Y}_T^0)$.

3.1 Testing the Sharp Null Hypothesis

Given $\widehat{Y}_{0t}$, we construct a test statistics and test with coverage $1 - \alpha$ having the following form:

$$\mathcal{T} = (T - T_0)^{-1/2} \sum_{t=T_0+1}^{T} \left(Y_{0t}^o - \hat{Y}_{0t}^0\right)^2, \quad \phi_\alpha(\mathcal{T}) = 1\Big\{\mathcal{T} \geq q_{1-\alpha}^*\Big\} \tag{7}$$

where $q_{1-\alpha}^*$ is the estimated $1-\alpha$ quantile of $\mathcal{T}$. The test statistic depends on the prediction error between the estimated counterfactual outcome and the potential outcome *under* the null hypothesis. Observe that under Equation (5), we can represent in compact form the test statistic as a functional of the *empirical* distributions

$$\mathcal{T}(F_0, F_1) = (T - T_0)^{1/2} \int \Big(y - w(F_0)g(x)\Big)^2 dF_1(x, y), \tag{8}$$

where F_1 denotes the *empirical* distribution over the post-treatment period of (Y_{0t}^o, X_t).

We obtain the critical quantile $q_{1-\alpha}^*$ using the block bootstrap (Politis and Romano, 1992, 1994). Formally, we resample the entire vector $(Y_{0t}, X_t)_{t=1}^T$, over the pre-treatment period, constructing first T_0 units serving as pre-treatment period's observations and $T - T_0$ units serving as post-treatment period's observations. We then construct the empirical measure of the bootstrapped sample (F_0^*, F_1^*) and compute $\mathcal{T}^* = \mathcal{T}(F_0^*, F_1^*)$. A formal description is included in Algorithm 2.

The main intuition behind the inferential procedure is the following. We use a portion of the data to train predictors, while the remaining observations are used for the bootstrap estimate of the critical value when conducting hypothesis testing. We only estimate learners once and not on each bootstrapped sample. Figure 1 provides a graphical illustration.

Algorithm 2 Testing Sharp Nulls: Basic Algorithm

Require: Observations $\{Y_{0t}, X_t\}_{t>1}$, predictors $g_1(\cdot), \cdots, g_p(\cdot)$ estimated as in Algorithm 1.

1: **for** $b = 1, \ldots, B$ **do**

2: Sample observations with replacement $\{Y_{0t}^o, X_t\}_{t>1}$, and obtain bootstrap sample $\{Y_{0t}^{o*}, X_t^*\}_{t>1}$ by performing circular block bootstrap on $\{Y_{0t}^o, X_t\}$ for $t \in \{1, \ldots, T\}$;

3: Construct the empirical measure from the bootstrap sample F_0^*, F_1^*, and the test statistic $\mathcal{T}^* = \mathcal{T}(F_0^*, F_1^*)$;

4: **end for**

5: Compute $q_{1-\alpha}^*$ as $(1-\alpha)$-th quantile of the sample

return Reject the null hypothesis if $\mathcal{T} > q_{1-\alpha}^*$.

Below, we formalize the validity of the bootstrap.[11] We impose the following condition.

Assumption 3. Assume that $\{Y_{0t}^0, X_t\}_{t\geq 1}$ is β-mixing with mixing coefficients $\sum_{k=1}^{\infty}(k+1)^2\beta(k) < \infty$. In addition, $g(\cdot), Y_{0t}^0$ are uniformly bounded almost surely.

[11] Lemma 2.1 provides the main intuition. The lemma implies that the distribution before and after the treatment must remain the same under the null hypothesis. Therefore, intuitively, we may expect that (F_0^*, F_1^*) centered around the true empirical distribution converges to the same empirical process (after appropriate rescaling) of the limiting process of $(F_0 - \mathcal{D}_0, F_1 - \mathcal{D}_0)$, under the null hypothesis. We can then invoke Hadamard differentiability properties to show the bootstrap's validity. We use such properties in the derivation of the validity of the bootstrap.

Assumption 4. Suppose that $w(\cdot)$ is Hadamard differentiable at $\mathcal{D}_0$ and uniformly bounded.

In Appendix B.5 we show that Assumption 4 holds for exponential weights considered in the following sections. We now introduce the first theorem.

Theorem 3.1. *Let Assumptions 1-4 hold. Let* $\limsup_{T\to\infty} b(T)/\sqrt{T} < \infty$ *and* $\lim_{T\to\infty} b(T) \to \infty$. *Then, under the null hypothesis, whenever* $p < \infty$

$$\sup_x \left| \mathbb{P}(\mathcal{T}^* - \mathcal{T} \leq x | Y_{1:T}, X_{1:T}, T_0, H_0) - \mathbb{P}(\mathcal{T} - \mathbb{E}[\mathcal{T}] \leq x | T_0, H_0) \right| = o_p(1), \quad \text{for} \quad T_0 \propto T \to \infty.$$

Corollary (Size control)**.** *Let the conditions in Theorem 3.1 hold. Then* $\lim_{T\to\infty} P\Big(\phi_\alpha(\mathcal{T}) = 1 | H_0\Big) = \alpha$.

The above corollary follows from the validity of the bootstrap (see for example, Fang and Santos, 2018) and it guarantees exact asymptotic coverage.

The proof of Theorem 3.1 is contained in Appendix B. Theorem 3.1 does not impose any restriction on the predictor other than Hadamard differentiability. Conditions on Hadamard differentiability are often imposed in the literature (Belloni et al., 2017), and require that weights are smooth functionals of the data. These are satisfied under mild conditions, and simple examples of weights that satisfy such conditions include least squares (Lunde and Shalizi, 2017) or exponential weights we discuss in Section 5 (see Appendix B.5 for a formal discussion). Note that the theorem is valid under misspecification.

It is interesting to observe that if we were doing classical statistical inference, we would need to account for the estimation error generated by each individual prediction. Instead, the sample splitting procedure in Algorithm 1 combined with the resampling method in Algorithm 2 guarantees valid inference, and it overcomes the complicated generated regressor problem. The key intuition is that the empirical distributions corresponding to the group of observations used for the training and those for the resampling are asymptotically independent under standard mixing conditions. As a result, *sample splitting* guarantees that we can condition on the initial training period F_-, without affecting the asymptotic properties of the bootstrap. In this sense, our analysis extends the standard sample-splitting procedures employed in the *i.i.d.* setting (Rinaldo et al., 2019) to dependent observations.

Example 3.1 (Inference with the sample mean)**.** For an illustrative example, consider the following simple model:

$$Y_{jt} = \mu + \alpha 1\{t > T_0, j = 0\} + \varepsilon_{jt}, \quad \mathbb{E}[\varepsilon_t | T_0] = 0 \quad \forall t,$$

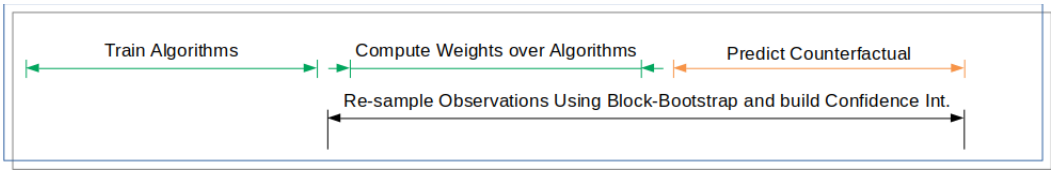

Figure 1: The algorithm, in the presence of multiple post-treatment periods, works as follows: train learners on an initial sample, compute the weights on a consecutive block of observations and then predict the counterfactual. The green color denotes the pre-treatment period while the orange color denotes the post-treatment period; bootstrap observations after imposing the null hypothesis.

and as estimator the difference-in-difference estimator

$$\hat{Y}_{0t} = \frac{1}{|T_-|}\sum_{s<0} Y_{0s},$$

where, for illustrative purposes, we use sample splitting for its construction.[12] Our corresponding test statistic takes the following form:

$$\mathcal{T} = \frac{1}{\sqrt{T-T_0}}\sum_{t>T_0}\left|\varepsilon_{0t} - \frac{1}{|T_-|}\sum_{s<0}\varepsilon_{0s} + \alpha\right|^2.$$

It is interesting to compare to the test statistic obtained from the permutation based method. This constructs the estimated counterfactual using the sample mean over the time-window $t \in \{1, \cdots, T_0, \cdots, T\}$, imposing a sharp null hypothesis (Chernozhukov et al., 2018). Its corresponding test statics takes the following form

$$\mathcal{T}^c = \frac{1}{\sqrt{T-T_0}}\sum_{t>T_0}\left|\varepsilon_{0t} - \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{0t} + (1-\lambda)\alpha\right|^2, \quad \lambda = \frac{T-T_0}{T}. \tag{9}$$

To gain further intuition, let $\lambda \approx 1$ (i.e., the post-treatment period is larger than the pre-treatment period). Then, the distribution of $\mathcal{T}^c$, corresponding to the permutation-based approach, does not depend on the treatment effect α, and therefore it cannot detect treatment effects. On the other hand, the dependence of the test statistic $\mathcal{T}$ with α remains invariant as λ changes. □

Remark 2 (Choice of the test statistics)**.** Alternative test statistics can be constructed by taking the k-norm of the residual error may also be considered (Chernozhukov et al., 2018),

$$\left((T-T_0)^{-1/2}\sum_{t>T_0}\left(Y^o_{0t} - \hat{Y}^0_{0t}\right)^k\right)^{1/k},$$

[12] Note that we could also have taken $\frac{1}{T_0}\sum_{s=1}^{T_0} Y_{0s}$, since the difference in means is Hadamard differentiable, with $w(F_0)g(X_t) = \int y dF_0(y)$ with $g(X_t) = 1$.

for a general k of the form. In this paper, the choice of $\mathcal{T}$ is motivated by two reasons: (i) the test statistics well captures permanent effects (i.e., effects exhibited over each post-treatment period) compared to test statistics having $k > 2$, which instead are better suited for the different case of large but temporary treatment effects, as in the case of $k = \infty$;[13] (ii) the test statistic presents desirable differentiable properties compared to the case of $k = 1$, which instead would not exhibit differentiability properties necessary for the validity of the resampling mechanism (see, for example Fang and Santos, 2018). □

Remark 3 (Testing the weak null hypothesis)**.** Our framework for inference discussed in Section 3 also extends to null hypothesis of the form

$$H_0^{avg} : \mathbb{E}\left[Y_{0t}^1 - Y_{0t}^0\right] = a^o, \quad a^o \in \mathbb{R}, t > T_0.$$

While we omit details for brevity, we note that, in this secenario, the test statistics takes the form $\mathcal{T}_A = \left((T - T_0)^{-1/2} \sum_{t>T_0} \left(Y_{0t}^o - \hat{Y}_{0t}^0\right)\right)^2$. The test statistics is attractive in the presence of additive treatment effects and small deviations,[14] while the square (instead of the absolute value) guarantees differentiability. □

4 Average Treatment Effects, Bias Adjustment and Time-Varying Fixed Effects

In applications, we may also be interested in estimating the average treatment effect consistently in Definition 2.3. A direct corollary of Lemma 2.2 is that a simple treated and controls difference, under standard mixing conditions, provides a consistent estimate of the treatment effect. However, such an estimate can present significant variance since we do not control the variation captured by covariates X_t, which can be non-vanishing with high-dimensional controls.

An alternative estimator for the average treatment effect takes the following form instead:

$$\frac{1}{T - T_0} \sum_{t>T_0} \left(Y_{0t} - \widehat{Y}_{0t}^0\right).$$

Unfortunately, such an estimator may be inconsistent for the true τ, due to misspecification of the functions $g(x)$. We consider a bias adjustment instead.

The bias adjustment works as follows. Define $F_{0,-1/2}$ the empirical distribution of (Y_t, X_t) for $t \in$

[13] This is of interest also in our empirical applications, where the effects of a decrease in expenditures in Medicaid is expected to have long term effects on health-care outcomes, instead of large and temporary effects.

[14] See for example the discussion at Page 65 in Imbens and Rubin (2015).

$\{1, \cdots, T_0/2\}$, and $F_{0,1/2}$ the empirical distribution of (Y_t, X_t) for $t \in \{T_0/2+1, \cdots, T_0\}$. Then we construct an estimator of the form

$$\widehat{\tau} = \frac{1}{T - T_0} \sum_{t > T_0} \left(Y_{0t} - \widehat{Y}^0_{0t} \right) - \frac{1}{T_0/2} \sum_{t=T_0/2+1}^{T_0} \left(Y_{0t} - \widehat{Y}^0_{0t,-1} \right), \quad \widehat{Y}^0_{0t,-1} = w(F_{0,-1/2}) g(X_t) \tag{10}$$

where $\widehat{Y}^0_{0t,-1}$ is the estimator whose weights are constructed using only the first half of the treatment period, and g is constructed as in Algorithm 1. The first difference is taken over the post-treatment period, with weights computed over the entire pre-treatment period. The second difference is taken over the second half of the pre-treatment period, but with weights computed over the first half. The second difference defines the bias adjustment.

To gain further intuition on the bias adjustment, note that we can write

$$\hat{\tau} = \tau + o_p(1) - \underbrace{\int \Big(w(F_0) g(x) - w(F_{0,-1/2}) g(x) \Big) d(F_1(y,x) - F_{0,1/2}(y,x))}_{= o_p(1)}$$

where F_1 denotes the empirical distribution over the treated period, $F_{0,-1/2}$ denotes the empirical distribution over the first half of the pre-treatment period, and $F_{0,1/2}$ denotes the empirical distribution over the second half. The above estimator is a difference in difference, where the first component converges to τ while the second component is of order $o_p(1)$ as shown in the following theorem.

We first impose the following condition.

Assumption 5 (Potential outcome Y^1_{0t})**.** Assume that

$$Y^1_{0t} = \mu^1_{0t} + \varepsilon^1_{0t}, \quad \mathbb{E}[\varepsilon^1_{0t}] = 0,$$

where $\mu^1_{0t} < \infty$ denotes some (possibly non-stationary) expectation and ε^1_{0t} is stationary.

Assumption 5 states that the potential outcome under control can be decomposed into two components: a (possibly non-stationary) expectation and an additive stationary idiosyncratic shock. We can now introduce the following theorem.

Theorem 4.1. *Let Assumption 1, 2, 4, 5 holds. Assume that $(\varepsilon^1_{0t}, Y^0_{0t}, X_t)$ is β-mixing with mixing coefficients $\sum_{k=1}^{\infty} (k+1)^2 \beta(k) < \infty$, with $g(\cdot; F_-)$ uniformly bounded. Then $\hat{\tau} - \tau \to_p 0$.*

Theorem 4.1 shows consistency of the estimated effect. It imposes standard mixing conditions. Its proof is in Appendix B.

4.1 Extension to Time-Varying Fixed Effects

In this section, we turn to the case where time-fixed effects occur and illustrate how our results extend to this case without necessitating estimating the fixed effect.

We consider the following data generating process.

Assumption 6 (Fixed effects and Stationarity)**.** Suppose that

$$Y_{jt}^0 = \kappa_t^0 + \iota_j^0 + \varepsilon_{jt}^0, \quad \mathbb{E}[\varepsilon_{jt}^0] = 0 \quad j \in \{0, \cdots, n\}, t \leq T,$$

where $\iota_j^0 < \infty, \kappa_t^0 < \infty$ denotes individual and time-fixed effect and $(\varepsilon_{0t}^0, \varepsilon_{1t}^0, \cdots, \varepsilon_{nt}^0, Z_{0t}, \cdots, Z_{nt}) \sim \mathcal{D}$, define arbitrary unobservables and observables which follow a stationary process.

The above assumption allows for the failure of stationary as long as the time fixed effects are the same between the control and treated unit. The following lemma illustrates identification for this case.

Lemma 4.2 (Identification of τ)**.** *Let Assumptions 2, 6 hold. Then*

$$\tau = \frac{1}{T - T_0} \sum_{t > T_0} \left\{ \mathbb{E}\Big[Y_{0t}^1 | T_0\Big] - \frac{1}{n} \sum_{j>0} \mathbb{E}\Big[Y_{jt}^0 \Big| T_0\Big] \right\} - \frac{1}{T_0} \sum_{s=1}^{T_0} \left\{ \mathbb{E}\Big[Y_{0s}^0 | T_0\Big] - \frac{1}{n} \sum_{j>0} \mathbb{E}\Big[Y_{js}^0 \Big| T_0\Big] \right\}.$$

The above lemma shows that we can identify the treatment effect by taking a difference in difference. The following lemma discusses testable implications for the sharp null hypothesis.

Lemma 4.3 (Sharp Null: Testable Implication)**.** *Under the null hypothesis in Equation (2) and Assumptions 2, 6 then* $(Y_{0t}^o - \bar{Y}_t, Y_{1t} - \bar{Y}_t, \cdots, Y_{nt} - \bar{Y}_t, Z_{0t}, \cdots, Z_{nt})$ *is stationarity and independent of* T_0 *for all* $t \in \{1, \cdots, T_0, \cdots, T\}$*, with* $\bar{Y}_t = \frac{1}{n} \sum_{j=1}^n Y_{jt}$*.*

The proofs of the above lemmas are in Appendix A. The counterfactuals' estimation follows similarly as before, with a minor modification: we subtract the controls' average outcome from the outcome of the treated unit and from the control unit to guarantee stationarity. This is illustrated below.

$$\hat{Y}_{0t}^0 - \bar{Y}_t = w(\tilde{F}_0) g(\tilde{X}_t), \quad \bar{Y}_t = \frac{1}{n} \sum_{j=1}^n Y_{jt}, \quad \tilde{X}_t = \Big(Y_{1t} - \bar{Y}_t, \cdots, Y_{nt} - \bar{Y}_t, Z_{0t}, \cdots, Z_{nt}\Big),$$

where $\tilde{F}_0$ denotes the empirical distribution of $(Y_{0t} - \bar{Y}_t, \tilde{X}_t)$ over the pre-treatment period. Here, $\bar{Y}_t$ denotes the time-specific average of the *control* units, but not of the treated. Theorem 3.1 and Theorem 4.1 directly follows once we subctract from the outcomes $(Y_{0t}, \cdots, Y_{nt})$ the controls' average $\bar{Y}_t$ to guarantee stationarity.

5 Ensamble Weights and Prediction Guarantees under Non-stationarity

This section discusses a particular weighting scheme, exponential weights, and derives its properties in terms of prediction guarantees.

5.1 Ensamble: Exponential weights

Although least-squares have been considered (see, e.g., Künzel et al., 2019; Polley and Van Der Laan, 2010; Elliott and Timmermann, 2004), these can perform poorly when the number of learners is large compared to the sample size.[15] Equal weighting, on the other hand, can perform poorly in the presence of many uninformative learners. To equip the method to have a better performance in the presence of a large number of algorithms, some of which may potentially be ineffective, we discuss an alternative weighting scheme.

With a slight abuse of notation, we index weights by the time period as $w(t)$, with $t \in \{1, \cdots, T_0\}$. For example $w(\tilde{t})$ denotes the weight estimated using the empirical distribution from time $t = 1$ to time $t = \tilde{t}$.

The weigthing scheme we focus are exponential weights of the following form: we write

$$w^{(j)}(T_0) = \frac{\exp\left\{-\eta \sum_{s=1}^{T_0} \left(Y_{0s} - g_j(X_s)\right)^2\right\}}{\sum_{i=1}^{p} \exp\left\{-\eta \sum_{s=1}^{T_0} \left(Y_{0s} - g_i(X_s)\right)^2\right\}}. \tag{11}$$

Such weights have been widely discussed in literature on exponential aggregation, see e.g., Cesa-Bianchi et al. (1999), Rigollet et al. (2012) among others.

By choosing the tuning parameter $\eta \propto 1/T_0$ exponential weights can be written as differentiable functionals of F_0. This is discussed in Appendix B.5.

We can motivate Equation (11) as the solution to a penalized surrogate risk minimization, discussed below.

Example 5.1 (Model average as a surrogate risk minimization problem)**.** Consider the following optimization problem (e.g., Rigollet et al., 2012)

$$\min_{w \in \mathcal{W}} \{\sum_{t=1}^{T_0} w_j (Y_{0t} - \sum_{j=1}^{p} g_j(X_t))^2 + \text{pen}(w)\}, \quad \mathcal{W} = \{w \in \mathcal{R}^p : w_j \geq 0; \sum_{j=1}^{p} w_j = 1\},$$

[15] Observe that weights computed via least-squares are Hadamard differentiable (Lunde and Shalizi, 2017), hence satisfying the conditions in Theorem 3.1.

where pen(w) denotes a penalty on the weights. Under convexity of the loss function, the above optimization problem is interpreted as minimizing a surrogate loss function of the penalized empirical risk.[16] By letting pen$(w) = \frac{1}{\eta}\sum_{j=1}^{p} w_j \log(w_j)$, the solution to the problem reads as

$$w^{(j)}(T_0) = \frac{\exp(-\eta \sum_{t=1}^{T_0}(Y_{0t} - g_j(X_t))^2)}{\sum_{i=1}^{p}\exp(-\eta \sum_{t=1}^{T_0}(Y_{0t} - g_i(X_t))^2)}. \tag{12}$$

Intuitively, the method assigns larger weights to those predictors that have the lowest out-of-sample loss function. These weights inherit oracle guarantees discussed in Section 5.2. □

5.2 Prediction Guarantees

We now derive prediction guarantees of the exponential weights without imposing stationarity conditions. Our result illustrates oracle guarantees of the exponential aggregation method (Cesa-Bianchi and Lugosi, 2006), here applied to the different contexts of counterfactual prediction.

We study the behavior of our algorithm trained only on $t-1$ observation and evaluated at the th observation, i.e., we are willing to provide theoretical guarantees on the following cumulative loss.

$$T_0^{-1}\sum_{t=1}^{T_0}(\hat{Y}_{0t}^0(\mathcal{F}_{t-1}) - Y_{0t}^0)^2, \tag{13}$$

where $\hat{Y}_t^0(\mathcal{F}_{t-1}) = w(t-1)g(X_t)$. Here, $\hat{Y}_{0t}^0(\mathcal{F}_{t-1})$ denotes the prediction at time t using only information at time $t-1$. Since $\hat{Y}_{0t}^0(\mathcal{F}_{t-1})$ is estimated only on the previous data and evaluated at X_t, this notion of performance is rooted in out-of-sample performance metric.

We first study the cumulative loss in (13) compared to the smallest cumulative loss incurred by any of the algorithms under consideration, defined as

$$\mathcal{R} = T_0^{-1}\sum_{t=1}^{T_0}(\hat{Y}_t^0(\mathcal{F}_{t-1}) - Y_{0t}^0)^2 - \min_{i\in\{1,\dots,p\}} T_0^{-1}\sum_{t=1}^{T_0}(g_i(X_t) - Y_{0t}^0)^2.$$

In the following theorem, we consider the case where $T_0 = \lambda T$ where $\lambda \in (\gamma, 1-\gamma)$ is potentially a random variable for some constant $\gamma > 0$.

Theorem 5.1. *Suppose that* $(Y_{0t}, g(X_t)) \in [-M, M]^{p+1}$, *for some* $M < \infty$. *Consider an exponential weighting scheme as in* (11) *with* $\eta \propto \sqrt{\log(p)/T_0}$. *Then with probability at least* $1-2\delta$, $\mathcal{R} \leq C_0\sqrt{\frac{\log(p/\delta)}{\gamma T}}$ *for* $C_0 < \infty$ *being a constant independent of* T_0 *or* p.

[16] Observe that we can write the risk as $\{\sum_{t=1}^{T_0} l(Y_{0t}, \sum_{j=1}^{p} w_j f_{j,t}) + \text{pen}(w)\}$ where, by convexity, $\sum_{j=1}^{p} w_j l(Y_{0t}, \sum_{j=1}^{p} f_{j,t}) \geq l(Y_{0t}, \sum_{j=1}^{p} w_j f_{j,t})$. Surrogacy is often considered in decision problem for its computational appeals.

The proof is presented in Appendix C.

Theorem 5.2 provides an error bound for the empirical one step ahead prediction error. Remarkably, it does not require any stationarity assumption. The bound scales logarithmically with the number of learners, and it scales at square-root T with the length of the sequence.

In the following lines, we provide stronger guarantees with respect to the *conditional* expectation of Y_{0t}^0. For the next theorem to hold, we need to introduce an additional condition, which replaces Assumptions 1 and 2 that we imposed in previous sections.

Assumption 7. (Additive Error Model and Sequential Ignorability) Let the following hold

$$Y_{0t}^0 = \mu_t(X_t) + \varepsilon_{0t}. \tag{14}$$

where $\mathbb{E}[\varepsilon_{0t}|X_t, \mathcal{F}_{t-1}] = 0$. Assume in addition that $\varepsilon_{0t} \perp T_0|X_t, \mathcal{F}_{t-1}$, where $\mathcal{F}_{t-1}$ denotes the filtration at time $t-1$.

Assumption 7 states that the potential outcome can be decomposed into two main components, a conditional expectation function $\mu_t(\cdot)$ and idiosyncratic shocks. Note that such a condition is only required to derive guarantees with respect to $\mu_t(\cdot)$. It is natural to compare the cumulative loss in (13) with the smallest cumulative loss incurred by any of the algorithms under consideration. We define such a metric of comparison as the Conditional Mean Proxy Regret (CMPR).

$$\mathcal{R}^{\mu} = T_0^{-1} \sum_{t=1}^{T_0} (\hat{Y}_t^0(\mathcal{F}_{t-1}) - \mu_t(X_t))^2 - \min_{i \in \{1,\dots,p\}} T_0^{-1} \sum_{t=1}^{T_0} (g_i(X_t) - \mu_t(X_t))^2.$$

The above definition incorporates notions of performance with respect to the conditional mean function (as opposed to the outcome itself). Our definitions above combine definitions in the literature on prediction of individual sequences (Cesa-Bianchi et al., 1999) with the literature on causal inference. The main difference with standard notions of regret is that CMPR is based on the unobserved deviation of the predicted counterfactual from the conditional mean evaluated at X_t, and not just on the cumulative loss of the predictor

Theorem 5.2. *Let Assumption 7 hold and let* $(Y_{0t}, \mu_t(X_t), g(X_t)) \in [-M, M]^{p+2}$, *for some* $M < \infty$. *Consider an exponential weighting scheme as in* (11) *with* $\eta \propto \sqrt{\log(p)/T_0}$. *Then with probability at least* $1 - 2\delta$,

$$\mathcal{R}^{\mu} \leq C_0 \sqrt{\frac{\log(p/\delta)}{\gamma T}},$$

for C_0 *being a constant independent of* T_0 *or* p.

The proof is presented in Appendix C.

Theorem 5.2 provides an error bound for the empirical one step ahead prediction error with respect to the *conditional mean*.

If we are willing to assume more, in that our class of algorithms contains one learner that consistently estimates the unknown model, then previous results imply that our synthetic learner will preserve that consistency regardless of the number of learners in the entire class. We consider below asymptotics for $T \to \infty$ and $T_0 = \lambda T$ where $\lambda \in (0, 1)$ is potentially a random variable.

Corollary. *Suppose that the number of learners is such that* $\log(p)/T^{1/2} = o(1)$ *and conditions in Theorem 5.2 hold. Assume also that the following holds* $\min_{i \in \{1,\dots,p\}} T_0^{-1} \sum_{t=1}^{T_0} |\mu_t(X_t) - g_i(X_t)|^2 = o_p(1)$. *Then,*

$$T_0^{-1} \sum_{t=1}^{T_0} (\hat{Y}_t^0(\mathcal{F}_{t-1}) - \mu_t(X_t))^2 = o_p(1).$$

for $T \to \infty$, $T_0 = \lambda T$.

6 Numerical Experiments

In this section, we study the performance of the method in the presence of linear and non-linear outcome models, allowing for the presence of many non-informative methods. We compare the methodology to existing testing procedures, including permutation tests of Synthetic Control as well as the Difference-in-Difference method, and showcase a significant improvement.

6.1 Experimental Setups

We describe our experiments in terms of the outcome model as well as the model of the design of the covariates and the error terms. In our first experiment, **DGP1**, we considered a simple Linear Outcome Model

$$Y_{0t} = X_t \beta + a_t D_t + \epsilon_t$$

and tested the ability of our method to detect changes in the treatment effect a_t. This example is intended to model a setting where classical Synthetic Control method is optimal. Here we set $\beta_j = 1/(1+j)^2$, $j = 1, \dots, p$, with the last beta chosen such that $\sum_j \beta_j = 1$, where we consider $p \in \{10, 50\}$. The parameter β will be kept as above for all our experiments. We considered a simple AR model for the errors ϵ_t with $\epsilon_t = 0.6\epsilon_{t-1} + v_t$ and $v_t \sim \mathcal{N}(0, 1 - 0.6^2)$. Control units are generated according to a factor model

as

$$X_{j,t} = \mu_j + \theta_t + \lambda_j F_t + u_t$$

with unit specific term $\lambda_j = \mu_j = (1+j)/j$ a time random effect $\theta_t \sim \mathcal{N}(0,1)$ and an unobserved factors $F_t \sim \mathcal{N}(0,1)$. Errors u_t follow an AR model $u_t = 0.6u_{t-1} + h_t$ with $h_t \sim \mathcal{N}(0, 1-0.6^2)$. In our second experiment, **DGP2**, we considered a Logistic-like Outcome Model

$$Y_{0t} = a_t D_t + \exp(X_t\beta + \epsilon_t)/(1+\exp(X_t\beta + \epsilon_t))$$

with $\epsilon_t = 0.5\epsilon_{t-1} + 0.3v_{t-1} + v_t$. This experiment has three settings: (a), (b) and (c). Setting (a) and (b) assume $v_t \sim \mathcal{N}(0, \sigma^2)$ with $(a)\sigma = 0.1$ and $(b)\sigma = 1$, respectively. Setting (c) assumes $\epsilon_t = 0.8\epsilon_{t-1} + v_t$, with $v_t = \sqrt{h_t}z_t, \quad h_t = 0.001 + 0.99v_{t-1}^2$ with $z_t \sim \mathcal{N}(0,1)$(AR-ARCH process). We report here (a) and (b). In addition we let $X_t = h_t + u_t$ with h_t being i.i.d over time with $\mathcal{N}(0,\Sigma)$ distribution with $\Sigma_{i,j} = 0.5^{|i-j|}$ and $u_t = 0.8u_{t-1} + k_t$ with $k_t \sim \mathcal{N}(0, 1-0.8^2)$. This setting is designed to test the ability of the proposed Synthetic Learner to adapt to nonlinear outcome model. We consider our third setting, **DGP3**, that follows a factor model

$$Y_{0t} = 0.5 + a_t D_t + \theta_t + 0.5F_t + \epsilon_t.$$

Error and design structures are the same as that of **DGP1**. **DGP4** considers an interaction outcome model that is polynomial in structure

$$Y_{0t} = a_t D_t + (X_{1,t} + X_{2,t} + \cdots + X_{10,t})^2 + \epsilon_t$$

with ϵ_t being the same as in **DGP2**(a) design X_t is the same as throughout **DGP2**; **DGP5** postulates a cosine, hence periodic, type of outcome model

$$Y_{0t} = \cos(X_t\beta + \epsilon_t) + a_t D_t.$$

Error and design setting have three components: (a), (b) and (c) that are following the setup of **DGP2** (a), (b) and (c), respectively. Finally, **DGP6** is a simple non-stationary model, which follows similarly to DGP3, but with $F_t \sim \mathcal{N}(\cos(t), 1)$, with $\cos(t)$ capturing a non-stationary component.[17] As discussed in the following subsections, we choose $p = 10$ for smaller $T(T \in \{60, 80, 100\})$ and $p = 50$ for larger $T \geq 300$.

[17]Note that the component cannot be removed from simple transformations such as differentiating.

6.2 Testing

We consider testing the following hypothesis H_0:

$$H_0 : Y_{0t}^1 - Y_{0t}^0 = 0, \quad t > T_0.$$

We consider the Synthetic Learner with experts, including a naive XGboost (which uses the default tuning parameter of the package XGboost in R), Support Vector Regression, and ARIMA(0,1,1) with external regressors together with 50 non-informative learners. Non-informative experts are randomly drawn from a multivariate gaussian with a full covariance matrix.

6.2.1 Power Study: Comparison with SC and DID

First, we compare Synthetic Learner's performance to existing procedures whose theoretical properties are well studied. In particular, we compare the Synthetic Control (SC) with weights being constrained to sum to one and an intercept according to Equation (7) and (8) in Chernozhukov et al. (2021), as well as the Difference in Difference (DiD) estimator, namely

$$\hat{Y}_t^{DiD} = \hat{\alpha} + (\hat{\beta} + \hat{\Delta})\mathbb{1}_{t>T_0}$$

with coefficient computed as in a standard DiD with the two periods corresponding to pre and post-treatment periods. We consider the test statistics for Synthetic Control

$$\frac{1}{\sqrt{T-T_0}} \sum_{t=T_0+1}^{T} |Y_{0t} - a_t^o - X_t \hat{w}_{SC}^0|^2 \tag{15}$$

where $\hat{w}_{SC}^0$ are computed via constrained Least Squares, with coefficients summing to one for Synthetic Control. Finally, we consider

$$\frac{1}{\sqrt{T-T_0}} \sum_{t=T_0+1}^{T} |Y_{0t} - a_t^o - \hat{Y}_t^{DiD}|^2 \tag{16}$$

for Difference-in-Differences. In Figures 2, 3 and 4 we compare the performance of our method to permutation tests where $\hat{w}_{SC}$ and $\hat{Y}_t^{DiD}$ must be computed on the entire sample, as described in Chernozhukov et al. (2021). We run the Synthetic Learner after training on the period running from 1 to $T_- = T_0/2, T \in \{60, 100\}, p = 10, T - T_0 = 10$, and we use the remaining observations for computing weights and bootstrap. We consider different treatment effects, denoted by α, and report on the x-axis the effect of the policy α divided by the (unconditional) standard deviation of the outcome.

We present power plots across different T in Figures 2, and 4 for $T = 60$ and $T = 100$ respectively, while Figure 3 collects results for the non-stationary DGP. Across all figures, we observe an improvement over permutation tests with both SC and DiD methods, with more significant improvements for the non-linear DGPs.

Improvements can be due to two factors: bootstrap outperforming permutation as well as Synthetic Learner's better performance in comparison to SC and DiD. Table 3 (see the discussion below) and Appendix E.3 provide suggestive evidence that improvements are due to both factors.

In particular, in Appendix E.3 we also consider the oracle case where the critical value is known. This case permits to compute $\hat{w}_{SC}$ and $\hat{Y}_t^{DiD}$ only using information until time T_0, as discussed in Doudchenko and Imbens (2016). We show improvements also in this setting. In Appendix E.2, we report a more extensive study with T_0 and $T - T_0$ vary and show the robustness of our results to different settings.

6.2.2 Size Control

Next, we study the size of our procedure for $T \in \{60, 80\}$ and $T^* = T - T_0 \in \{5, 10, 20\}$, as we vary the post-treatment period from small to longer post-treatment period. These are reported in Table 1. We observe that in a finite sample, for T relatively small, our test controls size across all DGP, except DGP3 and DGP6, where we observe a small size distortion of five percentage points for a short post-treatment period ($T^* = 5$). These results provide suggestive evidence of the correct size of the proposed test also in a finite sample.

Table 1: Size of the Synthetic Learner for $T \in \{60, 80\}$ and varying post-treatment period length T^*, for tests with size 5%. The first panel reports the size for $T = 60$ and the second panel for $T = 80$.

	$T = 60$			$T = 80$		
	$T^* = 5$	$T^* = 10$	$T^* = 20$	$T^* = 5$	$T^* = 10$	$T^* = 20$
DGP1	0.087	0.063	0.037	0.077	0.063	0.030
DGP2(a)	0.037	0.017	0.023	0.050	0.053	0.037
DGP2(b)	0.053	0.027	0.013	0.080	0.080	0.023
DGP2(c)	0.063	0.043	0.033	0.080	0.070	0.033
DGP3	0.090	0.087	0.050	0.097	0.060	0.040
DGP4(a)	0.060	0.040	0.027	0.040	0.063	0.033
DGP4(b)	0.067	0.047	0.040	0.090	0.053	0.043
DGP4(c)	0.080	0.057	0.037	0.057	0.067	0.020
DGP5(a)	0.073	0.047	0.020	0.060	0.077	0.037
DGP5(b)	0.050	0.063	0.027	0.067	0.043	0.030
DGP5(c)	0.103	0.040	0.020	0.060	0.057	0.040
DGP6	0.090	0.070	0.090	0.107	0.067	0.040

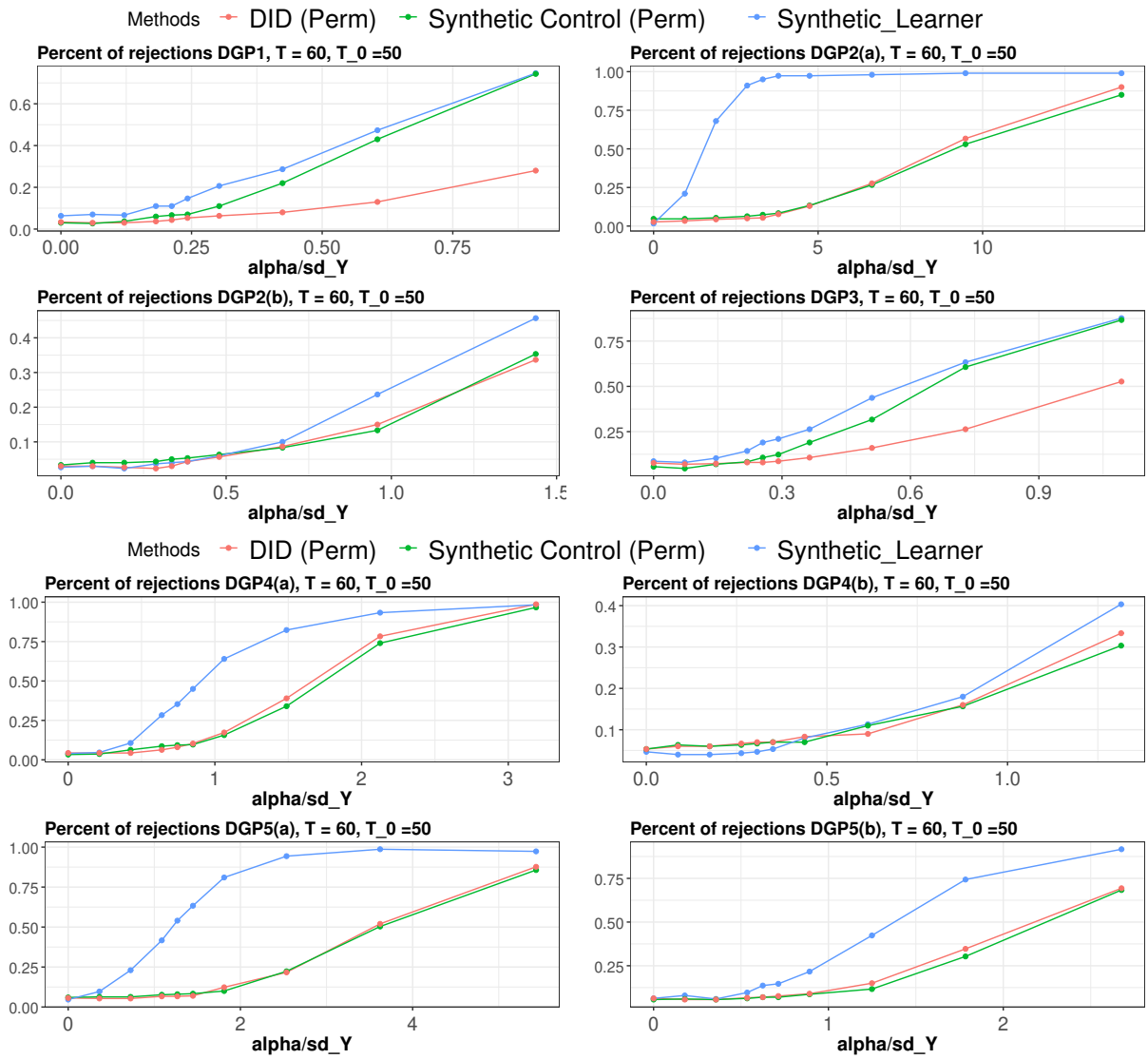

Figure 2: $T = 60, p = 10, T_0 = 50$. Percentage of rejections of the null hypothesis of no treatment effects over 300 repetitions. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. The blue line denotes the proposed method, the red line denotes the difference-in-differences, and the green line denotes the Synthetic Control.

6.2.3 Oracle Study: Learners' performance

In Table 2 we study the performance of the algorithm and each base algorithm for $T = 60, T - T_0 = 10$. Table 2 reports the power of each base algorithm and its corresponding weight assigned by the Synthetic Learner. We observe two striking facts: first, the largest weight is assigned to the best performing algorithm; second, the Synthetic Learner always performs approximately the same or *better* than any base algorithm under consideration. This result suggests the benefits of the ensemble procedure: the procedure combines predictions of different methods to maximize prediction (and ultimately power) optimally. For

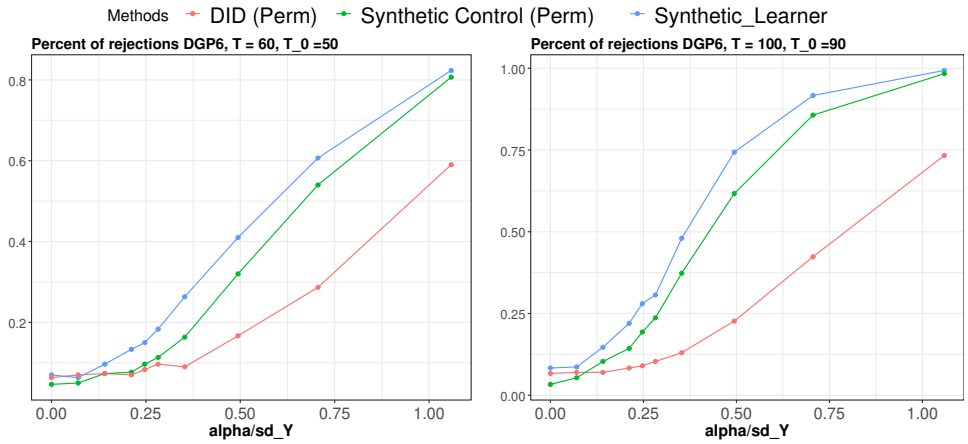

Figure 3: $DGP6, T \in \{60, 100\}, p = 10$. Percentage of rejections of the null hypothesis of no treatment effects over 300 repetitions. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. The blue line denotes the proposed method, the red line denotes the difference-in-differences, and green line denotes the Synthetic Control.

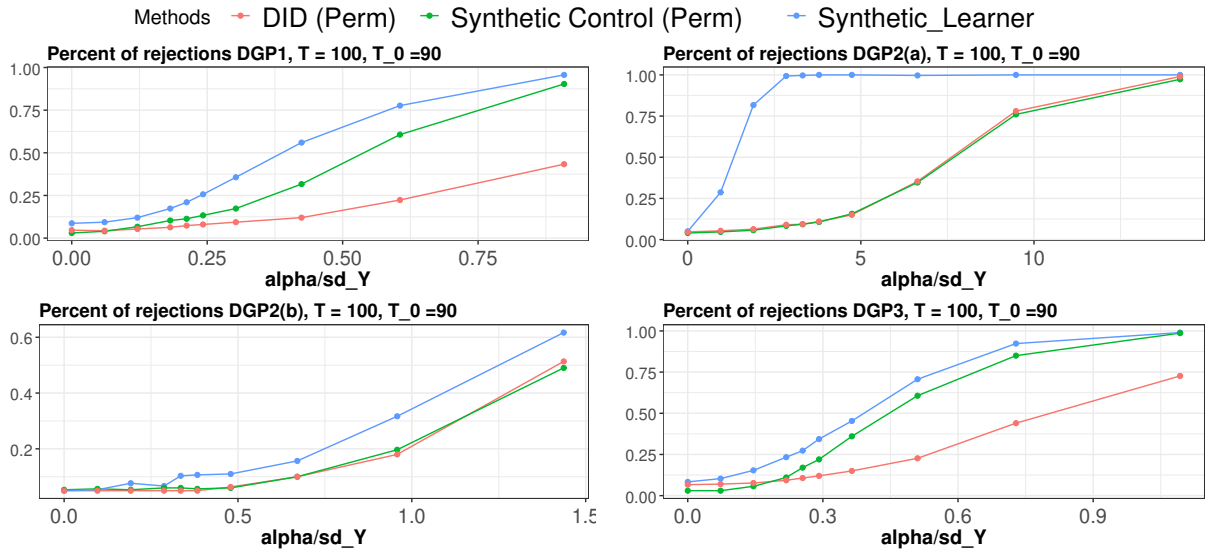

Figure 4: $T = 100, p = 10$ Percentage of rejections of the null hypothesis of no treatment effects over 300 repetitions. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. The blue line denotes the proposed method, the red line denotes the difference-in-differences, and the green line denotes the Synthetic Control.

the sake of brevity, we report results for two linear DGPs (DGP1 and DGP2) and two non-linear ones.

Table 2: $T = 60, p = 10$. Power of the Synthetic Learner and of each base algorithm. In each panel, the first two columns report the power for $\alpha \in \{0.1, 0.3\}$, and the third column reports the assigned weight of the Synthetic Learner.

	DGP1			DGP2(a)		
	0.1	0.3	Weights	0.1	0.3	Weights
Synthetic Learner	0.070	0.110		0.210	0.910	
XGBoost	0.030	0.047	0.153	0.157	0.823	0.345
SVM	0.040	0.037	0.044	0.213	0.910	0.375
Arima	0.060	0.103	0.803	0.163	0.693	0.280

	DGP2(b)			DGP3		
	0.1	0.3	Weights	0.1	0.3	Weights
Synthetic Learner	0.030	0.036		0.080	0.143	
XGBoost	0.020	0.023	0.132	0.053	0.056	0.217
SVM	0.030	0.033	0.831	0.057	0.050	0.091
Arima	0.020	0.026	0.037	0.057	0.106	0.693

6.2.4 Endogenous time of intervention

Next, we study the problem in the presence of an endogenous time of the treatment. In Figure 5 we report a representative set of results under the endogenous time of the treatment, where we simulate

$$T_0 = 280 + \min\{50, 1 + (\exp(1/\lambda) - 1/\lambda) \vee 1\}$$

where we choose $\lambda = |\sum_{j,t} X_{j,t}|$. The model follows a proportional hazard type model, similar to what is discussed in Shaikh and Toulis (2019), centered on $T_0 = 281$, with the time of treatment depending on other units' outcomes and constrained between 281 and 320. Figure 5 collects results with an endogenous time of treatment when critical quantiles are estimated via resampling, where the confidence intervals for the Synthetic Control and DiD are constructed using the permutation-based method in Chernozhukov et al. (2021). Results are consistent with the case of an exogenous treatment timing.

6.2.5 Variability in the quality of the learners

Next, we study the variability of the proposed method concerning the number and quality of learners included in learners' classes. We consider four different variations of the Synthetic Learner: Exponential and Least Squares weighting with 10 and 100 new non-informative learners. To guarantee the feasibility of the optimization problem given a large number of learners, we consider a large $T = 300$. Figure 6 contains the results. There we observe that many non-informative learners do little to nothing to the proposed

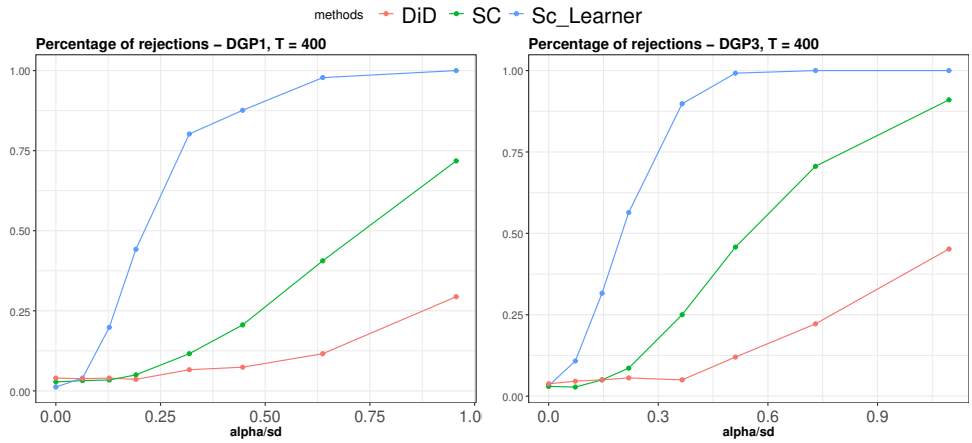

Figure 5: Percentage of rejections over 500 repetitions with $T = 400$, $T_0 = 280$, $p = 50$ and $T_- = 140$ when the critical quantile is estimated via resampling and the time of treatment is *endogenous*. The x-axis reports the policy's effect rescaled by the outcome's standard deviation.

Synthetic Learner. In sharp contrast, Least Squares' weighting suffers a substantial loss in power when the number of non-informative learners is increased.

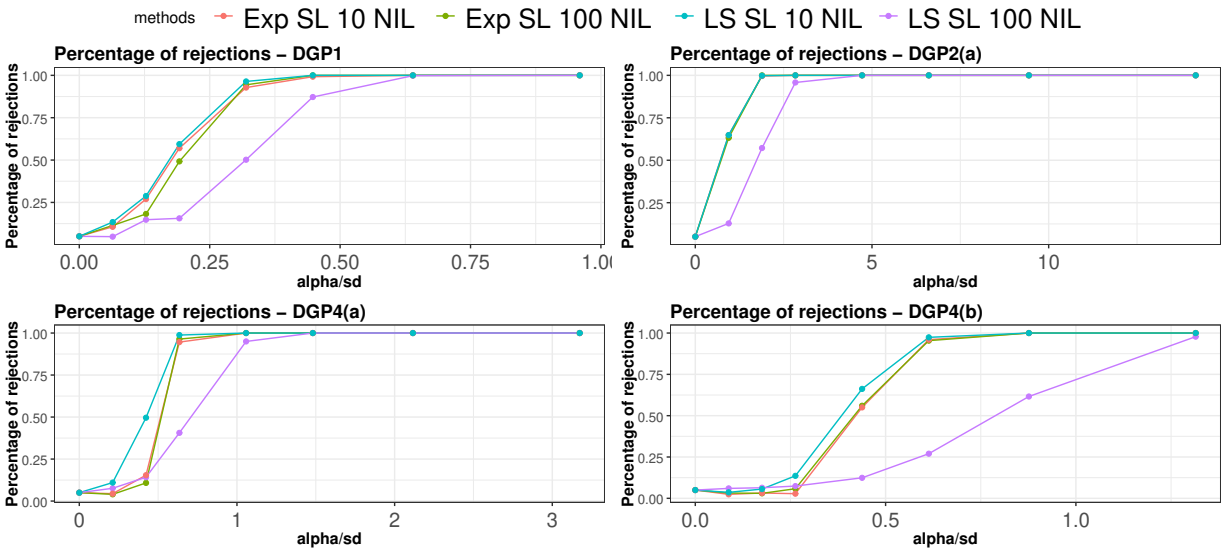

Figure 6: Percentage of rejections of null hypothesis H_0 over 500 repetitions with $T = 300$, $T_0 = 280$, $J = 50$ and $T_- = 140, p = 50$. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. As base learner we consider XGboost, Support Vector Regression, ARIMA(0,1,1) and either 100 or 10 additional non informative learners(NIL). We denote exp SL as the Synthetic Learner using exponential weights and LS SL as the Synthetic Learner using Least Squares weights.

6.2.6 Bootstrap vs permutations

We compare the performance of the circular bootstrap against permutations proposed in Chernozhukov et al. (2021). We consider only one learner: OLS. We compute the OLS coefficient for the bootstrap

method using only the first $T_0/2$ observations, and we bootstrap the remaining ones. We estimate the full sample's coefficient for the permutation method after imposing the null hypothesis of no effect. We consider the true effect is either $\alpha_t = 0.2$ or $\alpha_t = 0.3$. We vary $T \in \{60, 80\}$ and we consider $T - T_0 = 10$. Results are collected in Table 3. For the non-linear design, we mostly observe significant improvements in power, up to approximately fifty percentage points. Only for a few designs we observe comparable performances or slightly inferior by one to three percentage points.

Table 3: We compare the percentage of rejection of the sharp null hypothesis $\alpha_t^o = 0$ over 300 replications. We use the bootstrap method, while rejections via permutations are presented in the parenthesis. We predict the counterfactuals only using Least Squares. α denote the true policy effect. $T_0 = T - 10$.

	$T = 80$		$T = 60$	
	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.3$
DGP1	0.090(0.040)	0.107(0.067)	0.073(0.037)	0.057(0.060)
DGP2(a)	0.697(0.507)	0.940(0.793)	0.420(0.317)	0.747(0.667)
DGP2(b)	0.030(0.067)	0.040(0.087)	0.040(0.047)	0.030(0.070)
DGP2(c)	0.690(0.167)	0.833(0.303)	0.430(0.177)	0.567(0.253)
DGP3	0.033(0.050)	0.047(0.023)	0.030(0.027)	0.037(0.053)
DGP4(a)	0.143(0.087)	0.300(0.280)	0.113(0.107)	0.213(0.203)
DGP4(b)	0.070(0.040)	0.090(0.077)	0.043(0.053)	0.063(0.037)
DGP4(c)	0.180(0.063)	0.297(0.173)	0.110(0.080)	0.230(0.187)
DGP5	0.113(0.077)	0.167(0.133)	0.057(0.050)	0.127(0.057)
DGP6	0.027(0.060)	0.057(0.060)	0.033(0.027)	0.047(0.067)

7 The Effect of Public Health Insurance Ineligibility on Access to Medical Care

Understanding the effect of public health insurance coverage on health care access is a major concern in health economics (Kolstad and Kowalski, 2012; Long et al., 2009; Baicker et al., 2013; Anderson et al., 2012; Garthwaite et al., 2014).

The TennCare dis-enrollment program represents the largest reduction in public health insurance coverage ever experienced in the US. Between 2005 and 2006, approximately 170,000 individuals lost public health insurance coverage. Most of these individuals were childless adults who gained public health insurance coverage approximately ten years before, in 1994, during the Medicaid program expansion in Tennessee. In this section, we study the effect of the reform over childless adults on delayed medical care access due to medical costs. This population is of particular interest since most of the Affordable Care Act expansions target childless adults. Tello-Trillo (2021) estimates that the TennCare dis-enrollment

significantly decreased the likelihood of having health insurance between 2 and 5 percent. The author estimates an increase of between two and three percentage points in the probability of not going to a medical center when sick.[18] Our analysis provides supportive evidence for the claim, with positive effects ranging between one and five percentage points but with higher uncertainty in the absence of stationary. Details are discussed in the following lines.

7.1 Data

We use BFRSS data[19] to investigate the effect of the reform on the percentage of people who cannot afford healthcare expenses for medical costs. BFRSS is a national survey that has been continuously run over the years since 1984. The survey contains individual-specific information, including residence, state of health, access to health coverage, and others. The survey is run on a rolling basis, and the dataset can be organized as a long sequence of monthly and quarterly observations since we can cluster observations by the date of the interview. On average, we observe 150 childless adults between 18 and 64 years old in Tennessee per month from 2017 to 1993. To overcome survey variability, we aggregate data at a *quarter* level. The outcome variable is the percentage of childless adults who answered yes to the following survey question:"Was there a time in the past 12 months when you needed to see a doctor but could not because of the cost?".

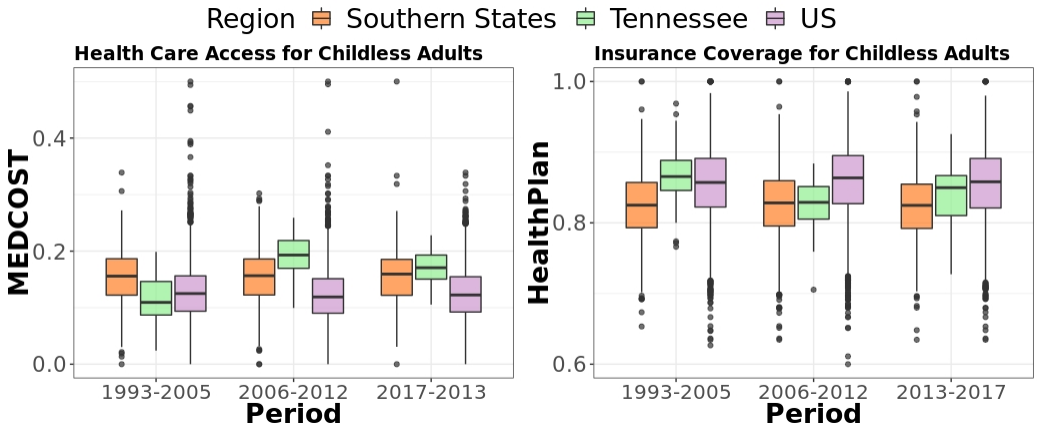

Figure 7: Sample distribution of childless adults between 18 and 64 years old in Tennessee, the Southern States, and the US who were not able to afford health care expenses (left panel) and who are covered by health insurance (right panel). BFRSS data.

In Figure 7, we report the distribution of respondents who were not able to afford medical costs in the past 12 months(left panel) and who are covered by health insurance[20](right panel), after clustering

[18]The reader might refer to Panel C, Table 5 in Tello-Trillo (2021).
[19]Behavioral Risk Factor Surveillance System Data: https://www.cdc.gov/brfss/annual_data/annual_data.htm.
[20]For the latter questions, we count the number of individuals who answer yes to the question: "Do you have any kind

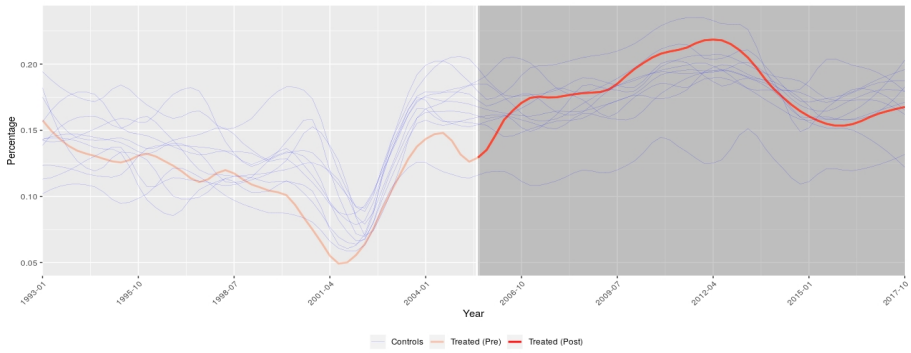

Figure 8: BRFFS data. Percentage of adults in Tennessee (treated unit, red) and Southern States (control, blue) who could not afford health care from January 1993 to December 2017.

over the period 1993-2005, 2006-2012 and 2013-2017 for Tennessee, other Southern States, and the United States. We observe a shift in the mean of Tennessee's outcome over these three periods, with a larger shift in the period just after the policy, between 2006 and 2012, while the variance remains approximately stable.

To check for stationarity of observed time-series, we test for unit roots at 95% confidence level. We reject the null hypothesis of a unit root in the time series of interest displayed in Figure 8.[21] However, we warn the reader that a lack of stationarity or confounders may invalidate the analysis. To accommodate failures of stationarity in the presence of time-varying fixed effects, we consider two alternative estimators with and without fixed effects adjustments, as discussed in the following subsection.

7.2 Estimation

As proposed in the Synthetic Control literature (Abadie et al., 2010), we impute the potential outcome under no dis-enrollment using a set of control variables in the other states. The starting date of the treatment corresponds to the second half of 2005.[22]

While the dis-enrollment program may be mostly attributable to an exogenous budget deficit (Argys et al., 2020), the series may still be affected by confounding sources over the period study, one of which is Obamacare's launch in 2014.[23] To control for potential confounders related to Obamacare, we consider as the post-treatment period the series until $t = 2014$, while we replicate the analysis also including periods

of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare or Indian Health Service?" We consider observations who answer "I do not know" as not having a plan.

[21] We use an Augmented Dickey-Fuller test, with constant and without time trend, and include one, two, or three lags. P-values are respectively < 0.01 for the first two tests and 0.05 for the latter.

[22] The overall dis-enrollment started in July 2005, and it lasted until June 2006. Most childless adults who dis-enrolled during this period were not able to requalify for Medicare (Garthwaite et al., 2014).

[23] The Affordable Health Care Act, also known as Obama Care, was officially approved in 2010, but the major change entered into force in 2014.

until 2017 in Appendix F. The selection into Obamacare from the states may reflect structural differences among different states. Motivated by this observation, we use a pool of control units only those Southern States that, similarly to Tennessee, did not expand Medicaid between 2010 and 2014, namely South and North Carolina, Mississippi, Alabama, Florida, and Georgia. In the Appendix, we replicate the analysis with all the states.

We construct the "Synthetic Control" using the Synthetic Learner described in the current paper. We consider the share of individuals in other countries who could not afford necessary health care expenses as control variables. To allow for time-varying fixed effects, we consider two variations of the Synthetic Learner, with and without fixed effects adjustments (see Section 4.1). We refer to the Synthetic Learner with fixed effect adjustment as "Demeaned Synthetic Learner" and SL otherwise.

We train Random Forest, Lasso, and a Factor model[24], and Diff-in-Diff mean proxy as discussed in Doudchenko and Imbens (2016) as base predictors.

Random Forest also contains additional covariates, such as the employment level in each state. Hyperparameters for the base predictors are chosen via cross-validation within the sample used to train such predictors. We construct weights with a tuning parameter $\eta = \frac{1}{\sqrt{T}\text{Var}(Y_t)}$ where the rescaling by the variance guarantees that the estimated weights are scale-invariant.[25] In Appendix F, we show the robustness of our results for several other choices of η. We consider two alternative sample splitting rules. First, observations between 1998 and 2006 are used to train the algorithms; observations between 1993 and 1997 are used to compute the weights and bootstrap. Second, the reverse is considered, where the training occurs over the earliest pre-treatment period. Observations from January 2006 onwards are used to compute the test statistic.

7.3 Results

In Table 4 we report the estimated test statistic for testing the null hypothesis of no effect, namely $H_0 : Y_{0t}^1 - Y_{0t}^0 = 0,\ t > T_0$, over the post-treatment period 2010-2014. The table reports the critical quantiles, the test statistic, and the estimated ATT when predictors are trained on the period closest to the post-treatment period (Period 1) or an earlier period (Period 2). The ATT oscillates between one and four percentage points, and its sign remains robust throughout all the setups considered. Significant effects are detected when training predictors on Period 1 for the test with a size of ten percent for Synthetic

[24] To guarantee the validity of the algorithm through the sample splitting procedure, the factor model consists of estimating the principal component over the training period, regressing the principal component on the control states over the training period and making counterfactual predictions using the predicted factor on the remaining periods.

[25] Since the loss is the squared loss, by rescaling by the variance we have losses of the form $(Y_t - \hat{Y}_t)^2/\text{Var}(Y_t) = (Y_t/\text{SD}(Y_t) - \hat{Y}_t/\text{SD}(Y_t))^2$ which are unit free in the outcome's unit. We rescale by $1/\sqrt{T}$ following the theoretical results of predictions. In the Appendix, we report results also after choosing different rescaling.

Learner. When estimating treatment effects by training the predictors more distant from the treatment timing (Period 2), results become non-significant, possibly reflecting higher uncertainty. Similarly, when we consider the adjustment for fixed effects, we observe p-values close to or larger than twenty percent, suggesting higher uncertainty for this case.

The Synthetic Learner predicts an effect larger than Lasso's one but smaller than a factor model. The reader may refer to the Appendix for further details. In Appendix F, we report results over the period 2010-2017, showing attenuated results over the time window 2010-2017.

Table 4: 90% and 80% critical values, t-statistic, and ATT using the southern states as controls. The effect estimated is over the time window 2010-2014 (first row), and consecutive windows 2011-2014 ($m = 1yr$), 2012-2014 ($m = 2yr$), 2013-2014 ($m = 3yr$). Period 1 collects results when learners are estimated using the window between 1998-2006, and weights are estimated over the period 1993-1997. Period 2 corresponds to the opposite scenario. Demeaned SL denotes the SL with time-varying fixed effects.

Period 1	SL				Demeaned SL			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.332	1.249	1.354	1.844	0.903	0.846	0.836	0.994
m = 1yr	1.261	1.176	1.281	1.998	0.794	0.745	0.729	1.060
m = 2yr	1.217	1.133	1.273	2.166	0.758	0.711	0.696	1.104
m = 3yr	1.160	1.070	1.229	2.253	0.714	0.664	0.651	1.096
Period 2	SL				Demeaned SL			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.264	1.173	0.691	5.223	0.519	0.471	0.243	3.137
m = 1yr	1.211	1.118	0.622	5.362	0.460	0.415	0.163	3.142
m = 2yr	1.189	1.100	0.625	5.516	0.453	0.407	0.154	3.202
m = 3yr	1.154	1.060	0.611	5.587	0.450	0.404	0.149	3.197

In Table 5 we collect the weights assigned to each base-algorithm. We observe that the Synthetic Learner assigns a larger weight to Lasso in the absence of fixed effects, and a larger weight to Random Forest in the presence of fixed effects.

Table 5: Weights estimated by the Synthetic Learner and by the Synthetic Learner after subtracting the control's mean to allow for fixed effects over Period 1.

	Synthetic Learner	Demeaned Synthetic Learner
Factor	0.242	0.202
CV Lasso	0.298	0.253
Random Forest	0.243	0.295
DID	0.217	0.251

In Figure 9, we report the test statistics and the acceptance region for Tennessee and for placebo

tests performed on the other Southern States that did not adopt Medicaid expansion. A placebo test consists of testing a policy's effect from 2006 to 2014 in a state different from Tennessee. Since none of the other Southern States had significant changes in the Medicaid system, we would expect no rejections for all Southern States except Tennessee. This is shown in Figure 9. We observe that we do not reject the null hypothesis when using only the simple DiD method, potentially due to the underpowered test. This result is consistent with what we observed in simulations, where the synthetic learner outperformed other methods in terms of power.

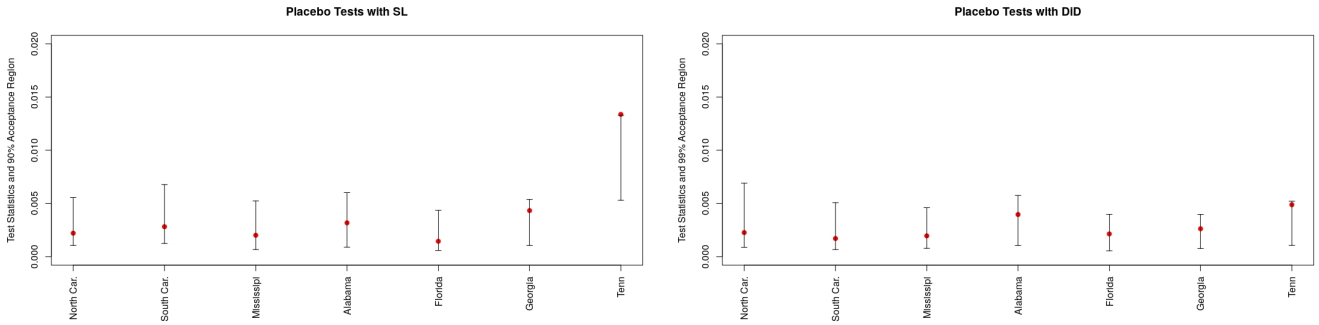

Figure 9: We test significant changes in the percentage of childless adults who are not able to afford medical expenses in those Southern states that did not adopt Medicaid Obamacare. We report the test statistic (red dot) and 90% confidence region for each of the states, including Tennessee, over Period 1. Left-panel reports the test when using Synthetic Learner. Right panel when only using Difference in Differences to predict the counterfactual.

7.4 Discussion: Assumptions and Possible Sources of Confounding

We conclude this section with a discussion on the assumptions and possible sources of confounding. The Tennessee dis-enrollment program was a result of a budget deficit. Whenever the budget deficit was due to an exogenous variation (Argys et al., 2020), the variation in state-level Medicaid expenses can be interpreted as exogenous to childless adults' financial status validating the exogeneity of the treatment timing. In this scenario, the ATT estimator is a consistent estimator of the underlying treatment effect. However, this condition may not necessarily hold. For instance, the budget deficit may be attributed to the increase in Medicaid expenses in previous years (Garthwaite et al., 2014), which may itself depend on individuals' past average income. In such a case, exogeneity may be replaced by conditional exogeneity given the past filtration without violating the prediction guarantees of the proposed algorithm.

The results of our testing procedure should instead be interpreted as conditional on the treatment assignment mechanism, similar to what is discussed in Ferman and Pinto (2016). Here, the assumption of stationarity may fail if, for example, spillover effects of the dis-enrollment program occur over adjacent

states, therefore changing the distribution of control units (see the discussion in Appendix D.1). While the study of Synthetic controls under spillovers goes beyond this paper's scope, we observe that in this scenario, it may be necessary to estimate counterfactuals also on the other states, which may be affected by the policy intervention.

Confounding may also result from other events, such as the 1996 Clinton welfare reform and the great depression. In the former case, although the reform in Tennessee was targeted at families with children (so-called Family First policies[26]), which are excluded from our analysis, it may act as a confounder in our analysis in the presence of general equilibrium effects. Time-fixed effects can (partially) accommodate for these cases *if* the confounder affects the mean only, and the effect is homogenous across the states considered in our analysis. These restrictions motivate studies that focus on sets of control units most similar to the treated, which, in our case, correspond to a subset of Southern States.

8 Extension: Carry-over Effects

In many applications, treatment effects may carry over in time (Imai and Kim, 2021). Here we extend the proposed framework to carry-over effects as follows. We consider binary treatment and denote the treatment path up to time t as a vector $\mathbf{d}_{1:t} \in \{0,1\}^t$. Following the potential outcomes framework, we then posit the existence of potential outcomes $\tilde{Y}_{0t}(\mathbf{d}_{1:t})$, corresponding respectively to the response the treated subject would have experienced at time t while being exposed to the treatment assignment contained in the treatment path $\mathbf{d}_{1:t}$. Formulating treatments and potential outcomes as paths were introduced initially by Robins (1986).

Notation implicitly assumes no lead effects (Athey and Imbens, 2022). Also, we require that the realizations of potential outcomes do not depend on past m lags or more. Using the same notation as in Rambachan and Shephard (2019), for $t \in \{T_-, \dots, 0, 1, \dots, T_0, \dots, T\}$ we assume throughout the rest of this paper that the following holds.

Assumption 8 (Finite carry-over)**.** For all $\mathbf{d}_{T_-:t}$, $\tilde{Y}_{0t}(\mathbf{d}_{T_-:t}) = Y_{0t}(\mathbf{d}_t, \cdots, \mathbf{d}_{t-m})$, for some function $Y_{0t}(\cdot)$

The assumption explicitely defines carry-over effects of size m. The no-anticipation assumption has been previously discussed in Abbring and Heckman (2007), Athey and Imbens (2022), while the restricted carryover effect is analogous to the identification assumption stated in Imai et al. (2018), Iavor Bojinov (2019), Blackwell and Glynn (2018) among others. The estimand of interest is now defined as follows

$$\mathbb{E}\Big[Y_{0t}(\mathbf{1}) - Y_{0t}(\mathbf{0})\Big], \quad t > T_0 + m$$

[26] The reader may refer to https://haslam.utk.edu/sites/default/files/ffoct00.pdf.

which denotes the (long-run) ATT, comparing two policies always and never implemented.

The key idea in this setting consists in estimating treatment effects after removing the m lag components. Formally, we construct an ATT estimator of the form

$$\widehat{ATT}_m = (T - T_0)^{-1} \sum_{t > T_0 + m} (Y_{0t}^1 - \hat{Y}_{0t}^0) - (T_0/2)^{-1} \sum_{t=T_0/2+1}^{T_0} (Y_{0t}^0 - \hat{Y}_{0t,-1}^0) \tag{17}$$

where the estimator averages after m periods that the policy has been implemented. Our testing procedure remains invariant (and valid) after removing the periods $t \in \{T_0, \cdots, T_0 + m\}$.

Example 8.1 (Why considering carry-overs?)**.** Wrongly assuming the absence of carry-over effects can lead to misspecified causal estimands and hence possibly biased estimates. For example, consider a simple case

$$Y_t(\mathbf{d}_{(t-m):t}) = Y_t(\mathbf{0}) + \sum_{s=0}^{m} \alpha_{s+1} d_{t-s} \quad \Rightarrow \quad Y_t(\mathbf{1}) - Y_t(\mathbf{0}) = \sum_{s=0}^{m} \alpha_{s+1} d_{t-s}, \tag{18}$$

for a sequence of constants $\alpha_s \in \mathbb{R}$. The naive ATE estimate, defined as a difference between pre- and post-treatment averages is possibly biased. In fact, it's mean equal $|T - T_0|^{-1} \sum_{s=0}^{m-1} (m - s)\alpha_{s+1} + \sum_{s=0}^{m} \alpha_s$, where $|T - T_0|^{-1} \sum_{s=0}^{m-1} (m - s)\alpha_{s+1}$ defines its bias. □

9 Discussion

In this paper, we have introduced a novel strategy for estimating treatment effects and testing the null hypothesis of interest in the presence of time-dependent observations. We developed a novel algorithm, denoted as Synthetic Learner, that predicts the counterfactual building on multiple regression methods. Our framework provides a starting point for performing estimation and inference, which is valid regardless of the class of models under consideration.

The presence of one single treated unit at a given time of adopting the policy brings substantial challenges from an identification perspective. We considered three scenarios of interest. First, (i) the adoption date is deterministic, T_0 and fixed treatment effects similarly to Chernozhukov et al. (2021), Chernozhukov et al. (2018), Arkhangelsky et al. (2021) among others. We show that, under stationarity and mixing conditions, our algorithm controls the nominal size regardless of the class of base algorithm under consideration, even in the presence of misspecification bias. Extending this result to non-deterministic T_0 is conceptually feasible in the presence of multiple units treated at different points in time. (ii) We consider a random and exogenous time of treatment, and under stationarity assumptions, we show that the estimator for the average treatment effect is consistent under weak assumptions, letting T_0 be non-deterministic. Finally,

(iii) we let the treatment time be *sequentially* exogenous, without assuming any stationarity condition. We provide bounds on the predictive performance under this complex scenario. Our paper also opens new questions on constructing valid machine learning methods for causal inference when units exhibit dependence. We leave it to future research its study for inference on conditional average treatment effects under heterogeneous effects and endogenous treatment time.

10 Acknowledgments

We thank the editor, associate editor, three anonymous referees, and Graham Elliott, Yinchu Zhu, and Kaspar Wüthrich for helpful comments.

References

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

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American economic review 93*(1), 113–132.

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of econometrics 6*, 5145–5303.

Amjad, M., D. Shah, and D. Shen (2018). Robust synthetic control. *The Journal of Machine Learning Research 19*(1), 802–852.

Anderson, M., C. Dobkin, and T. Gross (2012). The effect of health insurance coverage on the use of medical services. *American Economic Journal: Economic Policy 4*(1), 1–27.

Argys, L. M., A. I. Friedson, M. M. Pitts, and D. S. Tello-Trillo (2020). Losing public health insurance: Tenncare reform and personal financial distress. *Journal of Public Economics 187*, 104202.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–4118.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association 116*(536), 1716–1730.

Athey, S., M. Bayati, G. Imbens, and Z. Qu (2019). Ensemble methods for causal effects in panel data settings. In *AEA Papers and Proceedings*, Volume 109, pp. 65–70.

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*, 685–725.

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

Bai, C., Q. Li, and M. Ouyang (2014). Property taxes and home prices: A tale of two cities. *Journal of Econometrics 180*(1), 1–15.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Baicker, K., S. L. Taubman, H. L. Allen, M. Bernstein, J. H. Gruber, J. P. Newhouse, E. C. Schneider, B. J. Wright, A. M. Zaslavsky, and A. N. Finkelstein (2013). The oregon experiment—effects of medicaid on clinical outcomes. *New England Journal of Medicine 368*(18), 1713–1722.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Ben-Michael, E., A. Feller, and J. Rothstein (2021). The augmented synthetic control method. *Journal of the American Statistical Association 116*(536), 1789–1803.

Billmeier, A. and T. Nannicini (2013). Assessing economic liberalization episodes: A synthetic control approach. *Review of Economics and Statistics 95*(3), 983–1001.

Blackwell, M. and A. N. Glynn (2018). How to make causal inferences with time-series cross-sectional data under selection on observables. *American Political Science Review 112*(4), 1067–1082.

Bojinov, I. and N. Shephard (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association 114*(528), 1665–1682.

Bottmer, L., G. Imbens, J. Spiess, and M. Warnick (2021). A design-based perspective on synthetic control methods. *arXiv preprint arXiv:2101.09398*.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, S. L. Scott, et al. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics 9*(1), 247–274.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory 18*(1), 17–39.

Carvalho, C., R. Masini, and M. C. Medeiros (2018). Arco: an artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*.

Cavallo, E., S. Galiani, I. Noy, and J. Pantano (2013). Catastrophic natural disasters and economic growth. *Review of Economics and Statistics 95*(5), 1549–1561.

Cesa-Bianchi, N., Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth (1997). How to use expert advice. *Journal of the ACM (JACM) 44*(3), 427–485.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge university press.

Cesa-Bianchi, N., G. Lugosi, et al. (1999). On prediction of individual sequences. *The Annals of Statistics 27*(6), 1865–1895.

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. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., K. Wüthrich, and Z. Yinchu (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pp. 732–749. PMLR.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2018). A t-test for synthetic controls. *arXiv preprint arXiv:1812.10820*.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association 116*(536), 1849–1864.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Elliott, G. and A. Timmermann (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics 122*(1), 47–79.

Fang, Z. and A. Santos (2018). Inference on directionally differentiable functions. *The Review of Economic Studies 86*(1), 377–412.

Ferman, B. and C. Pinto (2016). Revisiting the synthetic control estimator.

Firpo, S. and V. Possebom (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference 6*(2).

Garthwaite, C., T. Gross, and M. J. Notowidigdo (2014). Public health insurance, labor supply, and employment lock. *The Quarterly Journal of Economics 129*(2), 653–696.

Gunsilius, F. (2020). Distributional synthetic controls. *arXiv preprint arXiv:2001.06118*.

Hazlett, C. and Y. Xu (2018). Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data. *https://ssrn.com/abstract=3214231*.

Hsiao, C., H. Steve Ching, and S. Ki Wan (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics 27*(5), 705–740.

Hsiao, C. and Q. Zhou (2019). Panel parametric, semiparametric, and nonparametric construction of counterfactuals. *Journal of Applied Econometrics 34*(4), 463–481.

Iavor Bojinov, N. S. (2019, Forthcoming). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*.

Imai, K. and I. S. Kim (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis 29*(3), 405–415.

Imai, K., I. S. Kim, and E. Wang (2018). Matching methods for causal inference with time-series cross-section data. *https://imai.fas.harvard.edu/research/files/tscs.pdf*.

Imai, K., M. Ratkovic, et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*(1), 443–470.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kolstad, J. T. and A. E. Kowalski (2012). The impact of health care reform on hospital and preventive care: evidence from massachusetts. *Journal of Public Economics 96*(11-12), 909–929.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*(10), 4156–4165.

Lange, T., A. Rahbek, and S. T. Jensen (2011). Estimation and asymptotic inference in the ar-arch model. *Econometric Reviews 30*(2), 129–153.

Lee, O. (2005). Probabilistic properties of a nonlinear arma process with markov switching. *Communications in Statistics-Theory and Methods 34*(1), 193–204.

Li, K. T. (2017). Estimating average treatment effects using a modified synthetic control method: Theory and applications. *The Wharton School, the University of Pennsylvania*.

Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association 115*(532), 2068–2083.

Li, K. T. and D. R. Bell (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of econometrics 197*(1), 65–75.

Long, S. K., K. Stockley, and A. Yemane (2009). Another look at the impacts of health reform in massachusetts: evidence using new data and a stronger model. *American Economic Review 99*(2), 508–11.

Lunde, R. and C. R. Shalizi (2017). Bootstrapping generalization error bounds for time series. *arXiv preprint arXiv:1711.02834*.

Maclean, J. C., S. Tello-Trillo, and D. Webber (2019). Losing insurance and behavioral health inpatient care: Evidence from a large-scale medicaid disenrollment. Technical report, National Bureau of Economic Research.

Pham, T. D. and L. T. Tran (1985). Some mixing properties of time series models. *Stochastic processes and their applications 19*(2), 297–303.

Politis, D. N. and J. P. Romano (1992). A circular block-resampling procedure for stationary data. *Exploring the limits of bootstrap 2635270*.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical association 89*(428), 1303–1313.

Polley, E. C. and M. J. Van Der Laan (2010). Super learner in prediction.

Potrafke, N. and K. Wuthrich (2020). Green governments. *arXiv preprint arXiv:2012.09906*.

Rambachan, A. and N. Shephard (2019). A nonparametric dynamic causal model for macroeconometrics. *arXiv preprint arXiv:1903.01637*.

Rigollet, P., A. B. Tsybakov, et al. (2012). Sparse estimation by exponential weighting. *Statistical Science 27*(4), 558–575.

Rinaldo, A., L. Wasserman, and M. G'Sell (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics 47*(6), 3438–3469.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling 7*(9-12), 1393–1512.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference 25*(3), 279–292.

Schapire, R. E. and Y. Freund (2012). *Boosting: Foundations and algorithms*. MIT press.

Shaikh, A. and P. Toulis (2019). Randomization tests in observational studies with staggered adoption of treatment. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-144).

Tello-Trillo, D. S. (2021). Effects of losing public health insurance on preventative care, health, and emergency department use: Evidence from the tenncare disenrollment. *Southern Economic Journal 88*(1), 322–366.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting 1*, 135–196.

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences 113*(27), 7310–7315.

Xie, S. and J. Huang (2014). The impact of index futures on spot market volatility in china. *Emerging Markets Finance and Trade 50*(sup1), 167–177.

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