A REVIEW OF SYNTHETIC DIFFERENCE-IN-DIFFERENCES

YUEZHOU QU

Abstract. In various fields within social and natural sciences, conducting completely randomized experiments and assuming strong ignorability are often unfeasible. In such cases, there are many available methods to set up an alternative quasi-experimental design and conduct causal inference. This project focuses on a recent advancement that generalizes the difference-in-differences method and the synthetic control method. Our research includes a comprehensive literature review that critically evaluates the robustness of this new estimator, specifically its conditions for consistency and asymptotic normality. Additionally, we reproduce key results from the original study using the author-provided source code. We further apply this methodology to other real world panel data, demonstrating its practical utility.

1. Introduction

In the field of causal inference, researchers often confront the challenge of evaluating policy effects in scenarios where randomized experiments are not feasible. One particular structure where this issue arises is the panel data analysis, where the dataset consists of repeated observations of units over time. In such settings, units are subject to policy changes in some time periods but not others, creating a complex dynamic for assessing causal effects. Indeed, plenty of literature have addressed this issue with classical methods such as difference-in-differences (DID), as documented by [5]. One of the most seminal studies employing this framework is done by Card and Krueger [4], where they estimated the causal effect of increased minimum wage on the employment outcomes, contrasting the fast-food restaurants in New Jersey with Pennsylvania. More recently, the synthetic control method (SC) has gained increasing attention as an strong alternative method for such comparative case studies, as introduced by Abadie et al [2] [1], where they estimated the effectiveness of a comprehensive tobacco control legislation passed in California. In this project, we will focus on a recent development of the field that combines the beneficial features of both DID and SC, and we will conduct a comprehensive literature review of the paper from Arkhangelsky et al [3]. The paper provides some important insights for the similarities between the underlying assumptions of SC and DID, and they further propose a generalized method called synthetic difference-in-differences (SDID). As the DID and SC methods in this quasi-experimental setting is slightly beyond the scope of the course, this project will first provide a brief introduction to these methods here based on the lecture notes of Li [7] and Wager [9].

Date: December 9, 2023.

1.1. **Introduction to Difference-in-Differences (DID).** The DID method, a staple in empirical research, particularly in applied economics, estimates causal effects under the assumption of parallel trends. This assumption implies that, in the absence of treatment, the difference between treated and control units would remain constant over time. The classical model assumption is to specify a two-way additive structure for $Y_{it}(0)$, such that

$$Y_{it} = \alpha_i + \beta_t + W_{it}\tau + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon \mid \alpha, \beta, W] = 0. \tag{1}$$

This model assumes that each unit and each time period have a distinctive offset (or fixed effect), and that any deviation from this two-way structure is due to noise. This model is very indeed idealistic and may not hold in generalized scenarios, and a vast amount of literature improved this setting with different direction. However, it still leads to perfectly reasonable point estimates for the causal effect τ in some applications, and we will just work with some simple situations to understand the intuition behind the methods for now. An illustrative example is when there are only two time periods ($T = 2$), with some units never receiving treatment ($W_{i.} = (0, 0)$) and others beginning treatment in the second period ($W_i = (0, 1)$). In this case, an Ordinary Least Squares (OLS) approach to equation (1) allows for an exact solution, expressed as:

$$\hat{\tau}^{DID} = \frac{1}{\mid \{i : W_{i2} = 1\} \mid} \sum_{\{i:W_{i2}=1\}} (Y_{i2} - Y_{i1}) - \frac{1}{|\{i : W_{i2} = 0\}|} \sum_{\{i:W_{i2}=0\}} (Y_{i2} - Y_{i1}) \, . \tag{2}$$

Intuitively, we assess the outcome variations between exposed and unexposed units by comparing the differences before and after the intervention, and we denote it by $\hat{\tau}^{DID} := \hat{\tau}_{tr} - \hat{\tau}_{co}$. This method effectively captures a significant causal effect when the exposure is assigned randomly. To see this, however, we need to assume that treatment group and the control group experience the same trends in the absence of treatment, which is also known as the "parallel trends" assumption.

$$\mathbb{E}\left[Y_{i,2}(0) - Y_{i,1}(0) \mid W_{i2} = 1\right] = \mathbb{E}\left[Y_{i,2}(0) - Y_{i,1}(0) \mid W_{i2} = 0\right]. \tag{3}$$

With stable unit treatment values assumption (SUTVA), then one can easily show that

$$\begin{aligned} \mathbb{E}\left(\hat{\tau}_{co}\right) &= \mathbb{E}\left[Y_{i,2}(0) - Y_{i,1}(0) \mid W_{i2} = 0\right] \quad \text{(SUTVA)} \\ &= \mathbb{E}\left[Y_{i,2}(0) - Y_{i,1}(0) \mid W_{i2} = 1\right] \quad \text{(parallel trends)} \\ &= \mathbb{E}\left[Y_{i,2}(0) \mid W_{i2} = 1\right] - \mathbb{E}\left[Y_{i,1}(0) \mid W_{i2} = 1\right] \quad \text{(SUTVA)} \end{aligned}$$

Similarly, we can compute $\mathbb{E}\left(\hat{\tau}_{tr}\right) = \mathbb{E}\left[Y_{i,2}(1) \mid W_{i2} = 1\right] - \mathbb{E}\left[Y_{i,1}(0) \mid W_{i2} = 1\right]$, and

$$\mathbb{E}\left(\hat{\tau}^{DID}\right) = \mathbb{E}\left(\hat{\tau}_{tr} - \hat{\tau}_{co}\right) = \mathbb{E}\left[Y_{i,2}(1) - Y_{i,2}(0) \mid W_{i2} = 1\right] = \tau.$$

Here, we proved the consistency. Notice that the parallel trends assumption we proposed here has strong observable implications. Among non-treated units, all trends should be parallel—because units only differ by their initial offset α_i. The method should not be used if parallel trends are not seen to hold. The natural estimator in this case is to estimate the variance of the differences:

$$\widehat{\operatorname{Var}}[\hat{\tau}] = \frac{\widehat{\operatorname{Var}}\left[Y_{i2} - Y_{i1} \mid W_{i2} = 1\right]}{|\{i : W_{i2} = 1\}|} + \frac{\widehat{\operatorname{Var}}\left[Y_{i2} - Y_{i1} \mid W_{i2} = 0\right]}{|\{i : W_{i2} = 0\}|}$$

This approach aligns with robust inference methods accounting for ε_{it} being correlated within rows. More broadly, for inference regarding equation (1), one could consider techniques like bootstrap or jackknife that sample entire rows of the panel. Note that our DID estimator is invariant to additive unit-level shifts, but the trade-off is to assume the parallel trends that may not hold in many situations. Thus, we may consider other methods to compensate this weakness.

1.2. Introduction to Synthetic Control Method (SC).

In contrast, the SC method compensates for the lack of parallel trends by reweighting units to match their pre-exposure trends, making it suitable for cases with a single or a small number of treated units, such as comparative case study research, where we want to evaluate the causal effects of policy interventions. Here, we consider a scenario with $J+1$ regions, where only the first region experiences an intervention, leaving J regions as potential control units. For simplicity, it's assumed that the first region is consistently exposed to the intervention after an initial period. The time frame prior to the intervention is denoted as T_0, $1 \leq T_0 < T$. The model assumes that the intervention has no impact on outcomes before its implementation, so $Y_{it}(0) = Y_{it}(1)$ for $t \in \{1, \ldots, T_0\}$, $i \in \{1, \ldots, N\}$. In this model, we can allow unobserved confounding within time-varying effects. A latent factor model is proposed for the outcome $Y_{it}(0)$ which is represented as

$$Y_{it}(0) = \mu_i' \lambda_t + \delta_t + X_i' \beta + \epsilon_{it}, \quad t = 1, \ldots, T. \tag{4}$$

Here, μ_i is a vector of unobserved confounders, λ_t is the corresponding time-varying coefficients, and X_i is the p vector of observed covariates. This model extends the traditional fixed-effects model used for DID by replacing α_i with $\mu_i' \lambda_t$ with α_i. Using, the model, the potential outcome is

$$Y_{1t}(1) = Y_{1t}(0) + \tau_{1t} W_{1t} \tag{5}$$

The goal here is to estimate the set of average treatment effects on the treated (ATT), $\{\tau_{1\ T_0+1}, \ldots, \tau_{1\ T}\}$ defined for post-intervention periods $T_0 + 1$ through T,

$$\tau_{1t} = Y_{1t}(1) - Y_{1t}(0) = Y_{1t} - Y_{1t}(0) \quad \text{for} \quad t = T_0 + 1, \ldots, T. \tag{6}$$

Note that here our $Y_{1t}(0)$ is counterfactual. The central task of the SC method is to estimate the unobserved potential outcome $Y_{1t}(0)$ by constructing a "synthetic control", as the name suggested. Consider a vector $W = (w_2, \ldots, w_{J+1})'$ with non-negative elements that sum to one, where each choice of W represents a different potential synthetic control. The objective is to select an optimal weight vector W^* that allows for the pre-treatment covariates and outcomes of the treated unit to be replicated by a convex combination of the control units. This is expressed as:

$$\sum_{j=2}^{J+1} w_j^* X_j = X_1, \ \sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \ \ldots, \ \sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}. \tag{7}$$

Under the latent factor model (4) and certain standard conditions (see [1] for detail), we could show that $Y_{1t}(0) - \sum_{j=2}^{J+1} w_j^* Y_{jt} \approx 0$, if the number of pre-treatment periods is sufficiently large compared to the residual variance. This assumption allows for an approximately unbiased estimator of the treatment effect τ_{1t}, given by:

$$\hat{\tau}_{1t}^{SC} = Y_{it} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, \qquad \text{for } t = T_0 + 1, \ldots, T. \tag{8}$$

The pre-treatment vector for the treated unit, $Z_1 = (X_1', Y_{11}, \ldots, Y_{1T_0})'$, and the matrix for the control units, Z_0, with each column mirroring the control units' variables, are used to calculate the weights W^*. These weights are determined by minimizing a discrepancy metric $\|Z_1 - Z_0 W\|$, under the constraint that the weights sum to unity. A typical discrepancy metric is defined as:

$$\|Z_1 - Z_0 W\|_V = \sqrt{(Z_1 - Z_0 W)' \, V \, (Z_1 - Z_0 W)} \tag{9}$$

where V is a symmetric and positive semidefinite matrix that typically is chosen to be diagonal to indicate the relative importance of each variable. An iterative algorithm is then implemented to find the optimal V^* and W^* that minimizes the mean square prediction error (MSPE) of the pre-treatment outcomes. The essence of the synthetic control method is to seek the balance of the covariates and outcomes of the treated unit against a weighted combination of control units to estimate the causal effect of an intervention.

1.3. **Introduction to Synthetic Difference-in-Differences (SDID).** In this paper, the author provides an important insight that the fundamental assumptions of the two methods are closely related. Thus, they propose the new method SDID that combines the virtue of both methods. SDID not only reweights and matches pre-exposure trends to weaken the reliance on parallel trend type assumptions, but also is invariant to additive unit-level shifts and allows for valid large-panel inference. Take the first N_{co} units as control and remaining N_{tr} as treated, exposed after time T_{pre} (still assuming block treatment assignment here). Instead of just considering the unit weights in SC, it employs weights of both units and time periods in a two-way fixed effects regression. This emphasizes units and time periods with similar historical patterns to the treated units and periods. The SDID estimator in the regression setting can be formulated as the least square optilization problem:

$$\left(\hat{\tau}^{SDID}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i^{SDID} \hat{\lambda}_t^{SDID} \right\} \tag{10}$$

where $\hat{\omega}_i^{SDID}$ and $\hat{\lambda}_t^{SDID}$ are the weights for units and time periods. The paper made an interesting observation to relate this estimator with the previous two methods. To see this, we express the DID estimator in the form:

$$\left(\hat{\tau}^{DID}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\alpha,\beta,\mu,\tau}{\arg\min} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \right\}. \tag{11}$$

For the SC estimator:

$$\left(\hat{\tau}^{SC}, \hat{\mu}, \hat{\beta}\right) = \arg\min_{\mu,\beta,\tau} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \beta_t - W_i \tau\right)^2 \hat{\omega}_i^{SC} \right\}. \tag{12}$$

Note that the SC estimator formulated here may look slightly different to the one we introduced earlier, but it could be shown that they are indeed equivalent under certain assumptions. From this perspective, it is not hard to see that SDID is indeed a generalization of DID and SC. Note that DID does not consider the unit weights $\hat{\omega}_i$ employed in SC, and SC does not consider fixed effects across units α_i employed in DID; on the other hand, SDID incorporates both features in its model and includes additional weights for time periods as well. The intuition is that the weights employed here makes the two-way fixed effect regression "local", in the sense that they put similar units together. The rationale for including time weights in the SDID estimator is to enhance the estimator's flexibility and robustness, which is supported by empirical data showing that these weights can account for much of the variation in outcomes. Now we will delve into the intricacies of the SDID method, providing a comprehensive analysis of its theoretical underpinnings and practical applications.

2. Robustness Analysis

2.1. **Estimator Construction.** Here we introduce the core insight provided by the paper. For both DID and SC methods, the estimators can be rewritten as a weighted average difference in adjusted outcomes $\hat{\delta}_i$ for appropriate sample weights $\hat{\omega}_i$:

$$\hat{\tau} = \hat{\delta}_{tr} - \sum_{i=1}^{N_{co}} \hat{\omega}_i \hat{\delta}_i \quad \text{where} \quad \hat{\delta}_{tr} = \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} \hat{\delta}_i. \tag{13}$$

For the adjusted outcomes $\hat{\delta}_i$, SC uses unweighted treatment period averages, DID uses unweighted differences between average treatment period and pretreatment outcomes, and SDID uses weighted differences of the same:

$$\begin{aligned} \hat{\delta}_i^{SC} &= \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it}, \\ \hat{\delta}_i^{DID} &= \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it} - \frac{1}{T_{\text{pre}}} \sum_{t=1}^{T_{\text{pre}}} Y_{it}, \\ \hat{\delta}_i^{SDID} &= \frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it} - \sum_{t=1}^{T_{\text{pre}}} \hat{\lambda}_t^{SDID} Y_{it}. \end{aligned} \tag{14}$$

In this way, SDID generalizes the two methods. Also, the sample weights are $\hat{\omega}_i^{DID} = 1/N_{co}$ for DID. For the SC method, the weights are defined by

$$\hat{\omega}^{SC} = \arg\min_{\omega\in\Omega} \left\{ \sum_{t=1}^{T_{pre}} \left(\sum_{i=1}^{N_{co}} \omega_i Y_{it} - \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} Y_{it} \right)^2 \right\}, \tag{15}$$

subject to the constraint that

$$\Omega=\left\{\omega \in \mathbb{R}_{+}^{N}: \sum_{i=1}^{N_{co}} \omega_i=1, \omega_i=N_{tr}^{-1} \text{ for all } i=N_{co}+1, \ldots, N\right\} .$$

On the other hand, for the SDID estimator, the weights are defined by

(16)

$$\left(\hat{\omega}_0, \hat{\omega}^{SDID}\right)=\arg \min _{\omega_0 \in \mathbb{R}_{+}, \omega \in \Omega}\left\{\sum_{t=1}^{T_{pre}}\left(\omega_0+\sum_{i=1}^{N_{co}} \omega_i Y_{it}-\frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} Y_{it}\right)^2+\zeta^2 T_{pre}\|\omega\|_2^2\right\},$$

subject to the constraint that

$$\Omega=\left\{\omega \in \mathbb{R}_{+}^{N}: \sum_{i=1}^{N_{co}} \omega_i=1, \omega_i=N_{tr}^{-1} \text{ for all } i=N_{co}+1, \ldots, N\right\} .$$

The regularization parameter ζ is set according to:

$$\zeta=\left(\frac{N_{tr}}{T_{\text{post}}}\right)^{1/4} \hat{\sigma} \quad \text{with} \quad \hat{\sigma}^2=\frac{1}{N_{co}\left(T_{pre}-1\right)} \sum_{i=1}^{N_{co}} \sum_{t=1}^{T_{pre}-1}\left(\Delta_{it}-\bar{\Delta}\right)^2,$$

where $\Delta_{it}=Y_{i(t+1)}-Y_{it}$ and $\bar{\Delta}$ is the average of these one-period changes for unexposed units. Although the weights for SDID are similar to SC here, it is important to notice that the paper includes an additional intercept term ω_0 for the SDID method. This means that the weight $\hat{\omega}^{SDID}$ no longer need to make the unexposed pre-trends of the synthetic control perfectly match the exposed treated unit. It is sufficient here for the weights to make the trends of synthetic control parallel to the treated unit, and the weighted differences in the adjusted outcomes will eliminate the bias similar to the DID estimator. Also, from the fixed effects α_i in the regression setting earlier will absorb any constant differences between different units. The paper includes a regularization penalty as well. This encourages the dispersion of the fitted weights, ensuring the uniqueness of the weights. For time weights λ^{SDID}, the optimization problem is:

$$(17) \qquad \left(\hat{\lambda}_0, \hat{\lambda}^{SDID}\right)=\underset{\lambda_0 \in \mathbb{R}, \lambda \in \Lambda}{\arg \min }\left\{\sum_{i=1}^{N_{co}}\left(\lambda_0+\sum_{t=1}^{T_{pre}} \lambda_t Y_{it}-\frac{1}{T_{\text{post}}} \sum_{t=T_{\text{pre}}+1}^{T} Y_{it}\right)^2\right\},$$

subject to the constraint that

$$\Lambda=\left\{\lambda \in \mathbb{R}_{+}^{T}: \sum_{t=1}^{T_{\text{pre}}} \lambda_t=1, \lambda_t=T_{\text{post}}^{-1} \text{ for all } t=T_{\text{pre}}+1, \ldots, T\right\} .$$

The main difference between the time weights here (17) and unit weights (16) is that we are not using a regularization this time. The paper argues that this is motivated by the model assumption that we only allow for correlated observations within time periods for the same units, but not across units within a time period, which is beyond what is captured by the systemic component of outcomes represented by a latent factor model in section 1 [3]. However, in real data, it is possible that we can have units also correlated with each other when time is fixed, and one popular example is the spatial-temporal data. Therefore, although the paper did an outstanding job generalizing two common methods DID and SC, the model could still be further generalized, and this could be a future direction for research. Finally, according to the paper, the complete pipeline of calculating the SDID estimator is summarized as the algorithm 1 here.

```
Algorithm 1 SDID
Data: $Y, W$
Result: Point estimate $\hat{\tau}^{SDID}$
    (1) Compute regularization parameter $\zeta$ using (5);
    (2) Compute unit weights $\hat{\omega}^{SDID}$ via (4);
    (3) Compute time weights $\hat{\lambda}^{SDID}$ via (6);
    (4) Compute the SDID estimator via the weighted DID regression
    $(\hat{\tau}^{SDID}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg\min_{\tau,\mu,\alpha,\beta} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \hat{\omega}_i^{SDID} \hat{\lambda}_t^{SDID} \right\};$
```

2.2. **Assumptions for Consistency and Asymptotic Normality.** In this section, we will outline a formal framework of the SDID estimator. First, the paper assumes the data generating process to follow a generalization of the latent factor model,

$$\mathbf{Y} = \mathbf{L} + \mathbf{W} \cdot \boldsymbol{\tau} + \mathbf{E}, \text{ where } (\mathbf{W} \cdot \boldsymbol{\tau})_{it} = \mathbf{W}_{it} \boldsymbol{\tau}_{it}. \tag{18}$$

Note that the paper only considers the situation that $\mathbf{W}_{it} = \mathbf{1}\left(\{i > N_{co}, t > T_{pre}\}\right)$ is the block assignment. Hence, this could be another direction to further generalize the results, where we may allow other treatment assignment patterns. Note that we can write a useful decomposition $\mathbf{L} = \boldsymbol{\Gamma}\boldsymbol{\Upsilon}^\top$, where $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{D}^{1/2}$ and $\boldsymbol{\Upsilon}^\top = \mathbf{D}^{1/2}\mathbf{V}^\top$ in terms of the singular value decomposition $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Then target estimand is the average treatment effect for the treated units (ATT) during the periods they were treated, which under block assignment is

$$\tau = \frac{1}{N_{tr} T_{\text{post}}} \sum_{i=N_{co}+1}^{N} \sum_{t=T_{pre}+1}^{T} \tau_{it}.$$

Note that here we are employing the finite population model as we are estimating the average of ATT, where the experimental units are fixed and the randomness comes solely from the treatment assignment. This might be the plausible scenario in certain situations. However, one may also try to move the setting of the problem to super population model, where the units are an independent sample from some hypothetical infinite population. It is hard to say which model may perform better. This could be a future direction of research.

The SDID estimator, like the SC estimator, seeks to recover τ by reweighting to remove the bias associated with L, using both unit weights and time weights, providing a double robustness property. The essence of the paper's argument is that not only $\hat{\omega}$ can be used to balance out Γ, but also the time weights $\hat{\lambda}$ can be used to balance out Υ. Also, the two-way fixed effects setting and the inclusion of the intercept contribute to SDID estimator's invariance to additive shocks in any row or column of L aligns it with DID, allowing it to remain unchanged under modifications of $\mathbf{L}_{it} \leftarrow \mathbf{L}_{it} + \alpha_i + \beta_t$, sharing an advantage with DID. The paper further shows that with certain assumptions, the bias introduced by $\mathbf{L}$ can be removed, leading to its major theorem regarding the asymptotic property.

2.2.1. *Asymptotic Properties.* Now we will discuss the asymptotic setting specified for this problem, starting with the assumptions listed by the paper. Although 4 assumptions are not numerous, we may want to evaluate how strong they are.

ASSUMPTION 1 (Properties of Errors)**.** *The rows $\mathbf{E}_i$ of the noise matrix are independent and identically distributed Gaussian vectors and the eigenvalues of its covariance matrix Σ are bounded and bounded away from zero.*

The first assumption assumes the error matrix $\mathbf{E}$ to satisfy some regularity properties about the dependence structure. As we commented earlier, the independence across units could be a strong assumption in some cases. There are plenty of examples such as spatial-temporal data that don't satisfy this condition. One may also suspect that the example provided in the paper about the proposition 99 in California can have correlated units causing dependence for $\mathbf{E_i}$'s, as the regulation proposed by California might effect the people living near the border of the adjacent states, or it can effect people's action in states with similar or opposing political affiliations. Of course, assuming dependence across units could be an extremely difficult generalization to realize, yet it still could be a future direction to investigate. Also, the statement about the eigenvalues of the covariance matrix actually requires that there are no strong long-range dependence in time. We think this is an intuitive and common assumption in time-series modeling.

ASSUMPTION 2 (Sample Sizes)**.** *We consider a sequence of populations where*

(i) the product $N_{tr}T_{post}$ goes to infinity, and both N_{co} and T_{pre} go to infinity,
(ii) the ratio $\frac{T_{pre}}{N_{co}}$ is bounded and bounded away from zero,
(iii) $\frac{N_{co}}{(N_{tr}T_{post}\max(N_{tr},T_{post})\log^2(N_{co}))} \to \infty$.

The second assumption focuses on the sample size, where the paper assumes that the panel size should be large, and the number of treated cells should grow to infinity but slower than the total panel size. In particular, we cannot let T_{post} and N_{tr} both grow to infinity. This is an intuitive assumption as we indeed want to use a synthetic control to approximate the counterfactual of the treated units. Examining the behavior at the limiting case might be an interesting experiment to carry out.

ASSUMPTION 3 (Properties of $\mathbf{L}$). *Letting $\sigma_1(\mathbf{\Gamma}), \sigma_2(\mathbf{\Gamma}), \ldots$ denote the singular values of the matrix $\mathbf{\Gamma}$ in decreasing order and R the largest integer less than $\sqrt{\min(T_{pre}, N_{co})}$,*

$$\sigma_R\left(\mathbf{L}_{co,pre}\right)/R = o\left(\min\left\{N_{tr}^{-1/2}\log^{-1/2}\left(N_{co}\right), T_{post}^{-1/2}\log^{-1/2}\left(T_{pre}\right)\right\}\right). \tag{19}$$

The third assumption states that $\mathbf{L}$ should not have a large number of significant singular values, although it can have many small ones. Specifically, the rank of L should be less than the square root of the minimum of T_{pre} and N_{co}. There is no lower bound assumed for the nonzero singular values of $\mathbf{L}$, which means that it can accommodate many nonzero but very small singular values. This flexibility allows the model to handle scenarios with many small signal coefficients, which is typical in highdimensional inference problems. The requirement is that the $\sqrt{\min\left(T_{pre}, N_{co}\right)}$th singular value of $L_{co,pre}$ must be sufficiently small for the model to be appropriate. This is an very interesting assumption on the data generating mechanism. Our intuition here is that this requires no strong confounding effects in our data, but more simulation to test the limit might be interesting as the condition is sufficient but not necessary, and one may speculate that it could be further improved.

ASSUMPTION 4 (Properties of Weights and $\mathbf{L}$). *The oracle unit weights $\tilde{\omega}$ satisfy*

$$\left\|\tilde{\omega}_{co}\right\|_2 = o\left(\left[\left(N_{tr}T_{post}\right)\log\left(N_{co}\right)\right]^{-1/2}\right) \tag{20}$$

and

$$\left\|\tilde{\omega}_0 + \tilde{\omega}_{co}^{\top}\mathbf{L}_{co,pre} - \tilde{\omega}_{tr}^{\top}\mathbf{L}_{tr,pre}\right\|_2 = o\left(N_{co}^{1/4}\left(N_{tr}T_{post}\max\left(N_{co}, T_{post}\right)\right)^{-1/4}\log^{-1/2}\left(N_{co}\right)\right), \tag{21}$$

the oracle time weights $\tilde{\lambda}$ satisfy

$$\left\|\tilde{\lambda}_{pre} - \psi\right\|_2 = o\left(\left[\left(N_{tr}T_{post}\right)\log\left(N_{co}\right)\right]^{-1/2}\right), \tag{22}$$

and the oracle weights jointly satisfy

$$\tilde{\omega}_{tr}^{\top}\mathbf{L}_{tr,post}\tilde{\lambda}_{post} - \tilde{\omega}_{co}^{\top}\mathbf{L}_{co,post}\tilde{\lambda}_{post} - \tilde{\omega}_{tr}^{\top}\mathbf{L}_{tr,pre}\tilde{\lambda}_{pre} + \tilde{\omega}_{co}^{\top}\mathbf{L}_{co,pre}\tilde{\lambda}_{pre} = o\left(\left(N_{tr}T_{post}\right)^{-1/2}\right). \tag{23}$$

The last assumption concerns the relation between the factor structure $\mathbf{L}$ and the assignment mechanism $\mathbf{W}$. The intuition is that this plays a role of an identifying assumption, which guarantees that the oracle weights (weights that the optimization is achieved in (16) and (17)) are directly defined in terms of $\mathbf{L}$ are able to adequately cancel out $\mathbf{L}$ via the weighted double-differencing strategy. This requires that the fitted weights are reasonably spread out across units and time, and that the treated units and after periods not be too dissimilar from the control units and the before periods respectively similar to the identifying assumption. This might be a strong assumption in some cases. One possible scenario test out could be the case where we have a very long time series for treated units, and an interesting question could be how would the weights behave in those cases.

As the paper noted, the above assumptions are substantially weaker compared to those required for asymptotic normality in other comparative methods. Specifically, the SDID does not necessitate that double differencing alone should remove both individual and time effects as required by traditional Difference in Differences (DID) assumptions. It also does not hinge on the assumption that unit comparisons alone can remove biases between treated and control units, which is a requirement for Synthetic Control (SC) methods. Additionally, the SDID does not require a low rank factor model to be correctly specified, which is often a prerequisite in methods that explicitly estimate the matrix $\mathbf{L}$. The brilliant part of this SDID estimator is that it relies on a combination of three bias-reducing components: double differencing, unit weights, and time weights, to diminish the bias to an acceptably low level. Furthermore, the paper then provides a theorem to show that the SDID estimator is asymptotically normal with these assumptions, enabling us to conduct inference, hence making the method more convincing.

Theorem 2.1. *Under the model* (18) *with* $\mathbf{L}$ *and* $\mathbf{W}$ *taken as fixed, suppose that we run the SDID estimator with regularization parameter* ζ *satisfying* $(N_{tr}T_{post})^{1/2}\log(N_{co}) = o(\zeta^2)$. *Suppose moreover that Assumptions 1–4 hold. Then,*

$$\hat{\tau}^{SDID} - \tau = \frac{1}{N_{tr}} \sum_{i=N_{co}+1}^{N} \left(\frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} \varepsilon_{it} - \mathbf{E}_{i,pre}\psi \right) + o_p\left((N_{tr}T_{post})^{-1/2} \right), \tag{24}$$

and consequently

$$(\hat{\tau}^{sdid} - \tau)/\sqrt{V_\tau} \Rightarrow \mathcal{N}(0,1),$$

where

$$V_\tau = \frac{1}{N_{tr}} var \left(\frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} \varepsilon_{it} - \mathbf{E}_{i,pre}\psi \right).$$

Here V_τ *is on the order of* $1/(N_{tr}T_{post})$, *i.e.,* $N_{tr}T_{post}V_\tau$ *is bounded and bounded away from zero.*

Given these assumptions hold, and we can derive a consistent estimator for the asymptotic variance V_τ of the estimator can be obtained, the paper also proposes methods to construct conventional confidence intervals, given by the formula:

$$\tau \in \hat{\tau}^{sdid} \pm z_{\alpha/2}\sqrt{\hat{V}_\tau}. \tag{25}$$

Some given approaches are bootstrap, jackknife, and placebo variance estimation methods. Especially for the jackknife method, the paper shows that it is closely tailored to the SDID methodology. It involves running the full SDID algorithm just once to reduce the computational load significantly. This approach utilizes the weights from the SDID point estimate and applies the jackknife variance estimation method, which is very efficient. Here we provide the details in the following theorem from the paper.

Theorem 2.2. *Suppose that the elements of* $\mathbf{L}$ *are bounded. Then, under the conditions of Theorem 1, the jackknife variance estimator described in Algorithm 2 yields* *conservative confidence intervals, i.e., for any* $0 < \alpha < 1$,

$$\liminf \Pr\left[\tau \in \hat{\tau}^{sdid} \pm z_{\alpha/2}\sqrt{\widehat{V}_\tau^{jack}}\right] \geq 1 - \alpha. \tag{26}$$

Moreover, if the treatment effects $\tau_{it} = \tau$ *are constant and*

$$T_{post}\, N_{tr}^{-1} \left\| \hat{\lambda}_0 + \mathbf{L}_{tr,pre}\, \hat{\lambda}_{pre} - \mathbf{L}_{tr,post}\, \hat{\lambda}_{post} \right\|_2^2 \to_p 0, \tag{27}$$

that is, the time weights $\hat{\lambda}$ *are predictive enough on the exposed units, then the jackknife yields exact confidence intervals and* (26) *holds with equality.*

Algorithm 2 Jackknife Variance Estimation

```
Data: $\hat{\omega}, \hat{\lambda}, Y, W, \hat{\tau}$
Result: Variance estimator $\hat{V}_\tau$
  for $i \leftarrow 1$ to $N$ do
      Compute $\hat{\tau}^{-i} \leftarrow \arg\min_{\alpha,\beta} \sum_{j \neq i} (Y_{ij} - \alpha_j - \beta - \tau W_{ij})^2 \hat{\lambda}_j$
  end for
  Compute $\hat{V}_\tau^{jack} \leftarrow (N-1)N^{-1} \sum_{i=1}^{N} (\hat{\tau}^{-i} - \hat{\tau})^2$
```

Besides the analysis of methodology, the paper also has strong sections of simulation studies and applications real data. We will discuss them in the following.

3. Data Application

The author also provides a package in R with a detailed instruction to walk through the source code for the paper here [6]. Using the provided resource, we further evaluate the method with real data.

3.1. **Real Data Analysis.** We first consider the California proposition 99 data studied in the paper. As studied by Abadie et ell [1]. The goal of their analysis was to estimate the effect of increased cigarette taxes on smoking in California (based on the data from Orzechowski and Walker [8]). The data includes 39 states from 1970 to 2000. California (treated unit) passed Proposition 99 increasing cigarette taxes from 1989 onwards. Thus, we have $T_{pre} = 19$ pre treatment periods, $T_{post} = T - T_{pre} = 12$ post treatment periods, $N_{co} = 38$ unexposed states, and $N_{tr} = 1$ exposed state (California). We reproduced the results for the estimates and standard errors for the SDID, SC, and DID estimators in the paper in table 1. We also calculated a 95% confidence interval of $(-34.67, 3.46)$ using placebo variance estimation method. Therefore, we failed to reject the null hypothesis that there is no causal effect using the SDID method. We also reproduced the figure 1 in the paper in figure 1. Notice from the graph that the parallel trends does not seem to hold here, so DID probably overestimates the causal effect. SC method here matches the pre treatment period of the treated unit pretty well, so we may say that it satisfies the underlying model assumption. In this dataset, it is also noticeable that the SDID estimator gives the smallest estimate in all three methods. Here, we also print the unit weights $\hat{\omega}$ and time weights $\hat{\lambda}$ of three methods in

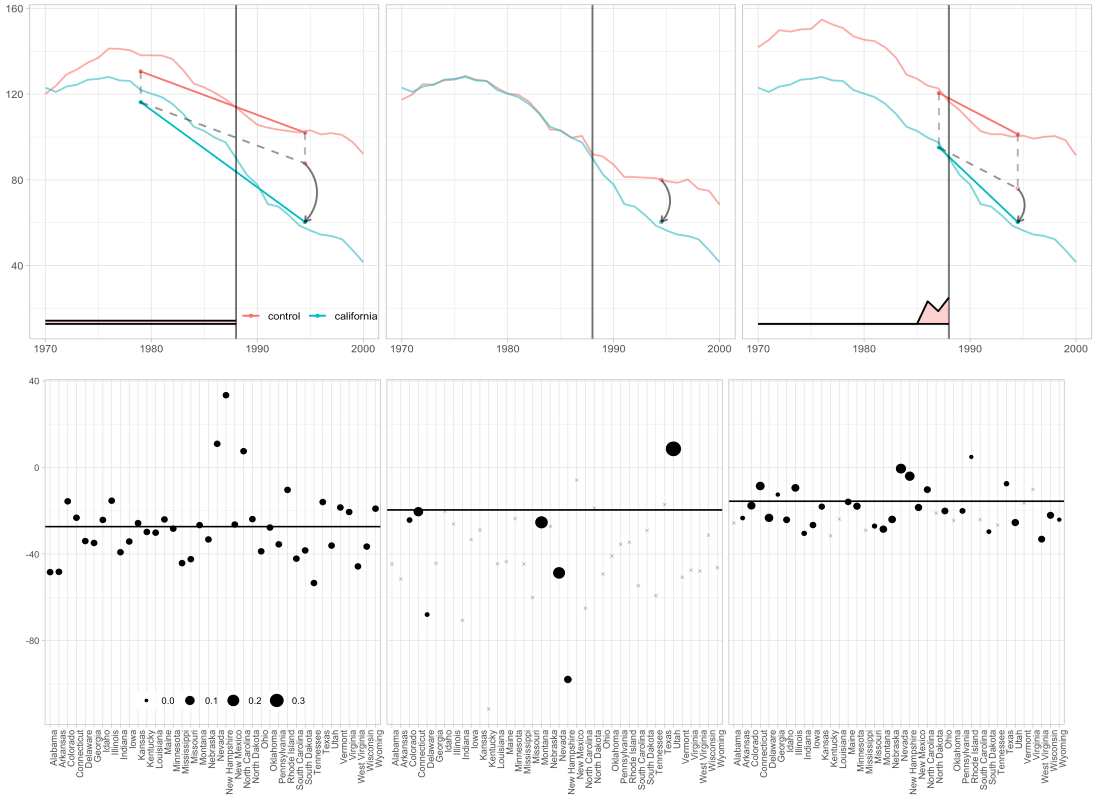

FIGURE 1. In the first row, we show trends in consumption over time for California and the relevant weighted average of control states, with the weights used to average pre-treatment time periods at the bottom of the graphs. The estimated effect is indicated by an arrow. In the second row, we show the state-by-state adjusted outcome difference $\hat{\hat{\delta}}_{tr} - \hat{\delta}_i$, with the weights $\hat{\omega}_i$ indicated by dot size and the weighted average of these differences: the estimated effect—indicated by a horizontal line. States are ordered alphabetically. Observations with zero weight are denoted by an × symbol.

table 2 and table 3. Notice that for our SDID method, the unit weights are reasonably dispersed, yet the time weights were concentrated on the most recent three years. This is expected as more recent time periods might be more reasonable than earlier ones. One may want to further test how does the distribution of weights interact with the model assumptions.

3.2. **Future Works in Simulation.** Besides the real data application, the paper also provides interesting simulation using placebo method, and it could be enlightening to further investigate through that. Moreover, there are also multiple ways to generate data to test the strength of our assumption for the SDID method, as discussed in section 2. Unfortunately, as the project is already exceeding the planned length and

	SDID	SC	DID
Estimate	-15.6	-19.6	-27.3
Standard error	9.4	11.0	17.9

Table 1. Estimates for ATT of increased cigarette taxes on California per capita cigarette sales over 12 post-treatment years, based on SDID, SC, DID.

the limited amount of time, we may have to list them as our future directions of research.

4. Conclusion

In the preceding sections, we have meticulously examined the Synthetic Difference-in-Differences (SDID) approach as a substantial advancement in the field of comparative case studies and panel data analysis. Our comprehensive review has scrutinized the theoretical underpinnings of the SDID estimator, revealing its robustness in the face of the challenging conditions that often beset empirical research where randomized experiments are untenable. The SDID method, through its combination of the Difference-in-Differences (DID) and Synthetic Control (SC) methodologies, exhibits a remarkable resilience to biases that typically afflict observational studies, generalizing their assumptions and extending their applicability. The SDID estimator's double robustness, invariance to additive shifts, and capacity for large-panel inference are particularly commendable, reflecting a methodological improvement in estimating the causal effects of certain types of data.

The assumptions required for the consistency and asymptotic normality of the SDID estimator, while could be stringent in some cases, are less onerous than those required by comparative methods. The asymptotic properties and variance estimation strategies discussed, such as jackknife, further cement the SDID's utility as a tool for practical inference in large-sample scenarios. Besides the highlighted methodological triumphs of the SDID, we also offer a few avenues for future exploration. These include the potential to relax certain assumptions, extend the model to accommodate various treatment assignment patterns, and explore the implications of the method for spatial-temporal data, where unit correlations present additional complexities.

In conclusion, the SDID stands as a robust, versatile, and theoretically sound method that significantly advances the toolkit available for causal inference in quasi-experimental settings. With a readily available package available in R library [6], researchers could conduct causal inference on a broader range of data with higher prediction accuracy. The method could have a profound impact on policy evaluation and social science research for years to come.

References

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

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

[3] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

[4] David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, 1993.

[5] Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2020.

[6] David A. Hirshberg. synthdid library. https://synth-inference.github.io/synthdid/.

[7] Fan Li. Lecture notes for causal inference. https://www2.stat.duke.edu/ fl35/CausalInferenceClass.html.

[8] Walker Orzechowski and Robert C Walker. The tax burden on tobacco: historical compilation. *The Tax Burden on Tobacco. Arlington, VA*, 38, 2003.

[9] Stefan Wager. Stats 361 lecture notes for causal inference. https://web.stanford.edu/ swager/teaching.html.

5. Appendix: Table for Estimated Weights in Three Methods

	did	sc	sdid
Alabama	0.03	0.00	0.00
Arkansas	0.03	0.00	0.00
Colorado	0.03	0.01	0.06
Connecticut	0.03	0.10	0.08
Delaware	0.03	0.00	0.07
Georgia	0.03	0.00	0.00
Idaho	0.03	0.00	0.03
Illinois	0.03	0.00	0.05
Indiana	0.03	0.00	0.01
Iowa	0.03	0.00	0.03
Kansas	0.03	0.00	0.02
Kentucky	0.03	0.00	0.00
Louisiana	0.03	0.00	0.00
Maine	0.03	0.00	0.03
Minnesota	0.03	0.00	0.04
Mississippi	0.03	0.00	0.00
Missouri	0.03	0.00	0.01
Montana	0.03	0.23	0.04
Nebraska	0.03	0.00	0.05
Nevada	0.03	0.20	0.12
New Hampshire	0.03	0.04	0.10
New Mexico	0.03	0.00	0.04
North Carolina	0.03	0.00	0.03
North Dakota	0.03	0.00	0.00
Ohio	0.03	0.00	0.03
Oklahoma	0.03	0.00	0.00
Pennsylvania	0.03	0.00	0.01
Rhode Island	0.03	0.00	0.00
South Carolina	0.03	0.00	0.00
South Dakota	0.03	0.00	0.00
Tennessee	0.03	0.00	0.00
Texas	0.03	0.00	0.01
Utah	0.03	0.40	0.04
Vermont	0.03	0.00	0.00
Virginia	0.03	0.00	0.00
West Virginia	0.03	0.00	0.03
Wisconsin	0.03	0.00	0.04
Wyoming	0.03	0.00	0.00

Table 2. Table of unit weights $\hat{\omega}$ for all three methods.

	did	sc	sdid
1988	0.05	0.00	0.43
1987	0.05	0.00	0.21
1986	0.05	0.00	0.37
1985	0.05	0.00	0.00
1984	0.05	0.00	0.00
1983	0.05	0.00	0.00
1982	0.05	0.00	0.00
1981	0.05	0.00	0.00
1980	0.05	0.00	0.00
1979	0.05	0.00	0.00
1978	0.05	0.00	0.00
1977	0.05	0.00	0.00
1976	0.05	0.00	0.00
1975	0.05	0.00	0.00
1974	0.05	0.00	0.00
1973	0.05	0.00	0.00
1972	0.05	0.00	0.00
1971	0.05	0.00	0.00
1970	0.05	0.00	0.00

Table 3. Table of time weights $\hat{\lambda}$ for all three methods.