NBER WORKING PAPER SERIES

SYNTHETIC DIFFERENCE IN DIFFERENCES

Dmitry Arkhangelsky
Susan Athey
David A. Hirshberg
Guido W. Imbens
Stefan Wager

Working Paper 25532
http://www.nber.org/papers/w25532

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2019

This research was generously supported by ONR grant N00014-17-1-2131 and the Sloan Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w25532.ack

Synthetic Difference In Differences
Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager
NBER Working Paper No. 25532
February 2019
JEL No. C01

ABSTRACT

We present a new perspective on the Synthetic Control (SC) method as a weighted least squares regression estimator with time fixed effects and unit weights. This perspective suggests a generalization with two way (both unit and time) fixed effects, and both unit and time weights, which can be interpreted as a unit and time weighted version of the standard Difference In Differences (DID) estimator. We find that this new Synthetic Difference In Differences (SDID) estimator has attractive properties compared to the SC and DID estimators. Formally we show that our approach has double robustness properties: the SDID estimator is consistent under a wide variety of weighting schemes given a well-specified fixed effects model, and SDID is consistent with appropriately penalized SC weights when the basic fixed effects model is misspecified and instead the true data generating process involves a more general low-rank structure (e.g., a latent factor model). We also present results that justify standard inference based on weighted DID regression. Further generalizations include unit and time weighted factor models.

Dmitry Arkhangelsky
CEMFI
5 Calle Casado del Alisal
Madrid 28014
Spain
darkhangel@cemfi.es

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

David A. Hirshberg
Department of Statistics
Stanford University
Stanford, CA 94305
davidahirshberg@stanford.edu

Guido W. Imbens
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
Imbens@stanford.edu

Stefan Wager
Graduate School of Business
Stanford University
Stanford, CA 94305
swager@stanford.edu

1 Introduction

Synthetic Control (SC) methods, introduced in a seminal series of papers by Abadie and coauthors [Abadie and Gardeazabal, 2003, Abadie, Diamond, and Hainmueller, 2010, 2015, Abadie and L'Hour, 2016], have quickly become one of the most popular methods for estimating treatment effects in panel settings. By using data-driven weights to balance pre-treatment outcomes for treated and control units, the SC method imputes post-treatment control outcomes for the treated unit(s) by constructing a synthetic version of the treated unit(s) equal to a convex combination of control units.

In the current paper, we build on these ideas to provide a different perspective on the SC approach and to propose a new estimator with improved bias properties. First, we show that the SC estimator can be viewed as a weighted least squares regression estimator with unit-specific weights, where the regression model includes time fixed effects. We then propose adding unit fixed effects to this regression representation of the standard SC set up to add flexibility, as well as time weights to ensure that the weighted periods resemble more closely the period(s) for which we are imputing the counterfactual. We show that this leads to a doubly weighted, or local, version of the standard Difference In Differences (DID) estimator [e.g., Bertrand, Duflo, and Mullainathan, 2004, Card, 1990]. We then establish that the resulting estimator, which we call the Synthetic Difference In Differences (SDID) estimator, has attractive bias properties compared to both the SC and DID estimators. In particular the estimator satisfies a form of double robustness that both the SC and DID estimators lack: the estimator is consistent if either the model is correctly specified, or if the weights are well choosen, but consistency does not require both those conditions.

Consider the simplest case of a balanced panel with N units and T time periods, where outcomes are denoted by Y_{it}, and exposure to the binary treatment is denoted by $W_{it} \in \{0, 1\}$. Initially suppose that $W_{it} = 0$ unless $(i, t) = (N, T)$, so that only unit N is treated, and only in period T. Suppose also that there are no covariates. In that case, the SC estimator for the causal effect is $\hat{\tau}^{\text{sc}} = Y_{NT} - \hat{Y}^{\text{sc}}_{NT}$ where $\hat{Y}^{\text{sc}}_{NT}$ is a weighted average of the period T outcomes for the control units, $\hat{Y}^{\text{sc}}_{NT} = \sum_{i=1}^{N-1} \hat{\omega}^{\text{sc}}_i Y_{iT}$, with the weights $\hat{\omega}^{\text{sc}}_i$ chosen to make the weighted average of the controls in the pre-treatment period approximate the corresponding value for the treated unit, $\sum_i \hat{\omega}^{\text{sc}}_i Y_{it} \approx Y_{Nt}$ for all $t = 1, \ldots, T-1$. In this paper, we introduce a novel

characterization of the SC estimator $\hat{\tau}^{\text{sc}}$ as a weighted least squares regression estimator:

$$(\hat{\mu}, \hat{\beta}, \hat{\tau}^{\text{sc}}) = \arg\min_{\mu,\beta,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i^{\text{sc}}. \tag{1.1}$$

The regression has time fixed effects and unit-specific weights. In comparison, the standard DID estimator $\hat{\tau}^{\text{did}}$ for the treatment effect is

$$(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\tau}^{\text{did}}) = \arg\min_{\alpha,\beta,\mu,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2. \tag{1.2}$$

Our characterizations (1.1) and (1.2) make clear that relative to the SC estimator, the DID estimator adds a unit fixed effect to the specification of the regression function, but it removes the (unit) weights in the estimation. Contrasting the SC and DID estimators in this way suggests a natural modification. Specifically, we propose the SDID estimator $\hat{\tau}^{\text{sdid}}$, formally defined as:

$$(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\tau}^{\text{sdid}}) = \arg\min_{\alpha,\beta,\mu,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i \hat{\lambda}_t. \tag{1.3}$$

The regression in (1.3) includes both unit and time fixed effects as well as weights, where the weights are the product of unit weights $\hat{\omega}_i$ and time weights $\hat{\lambda}_t$, with both sets of weights are derived from the data. In the spirit of the SC approach, these time weights $\hat{\lambda}_t$ could be chosen so that within a unit, the weighted average outcomes across periods approximate the target period, $\sum_{t=1}^{T} \hat{\lambda}_t Y_{it} \approx Y_{iT}$ for all $i = 1, \ldots, N-1$. Alternatively, one may wish to choose the time weights partly to put more emphasis on recent periods. Thus, the proposed SDID estimator differs from the DID estimator by allowing for both unit and time weights, and it differs from the SC estimator by including unit-fixed effects and allowing for time weights.

Many approaches to SC settings, including Abadie, Diamond, and Hainmueller [2010, 2015], Doudchenko and Imbens [2016], Xu [2017], Athey, Bayati, Doudchenko, Imbens, and Khosravi [2017], Carvalho, Masini, and Medeiros [2018], Li and Bell [2017], can be thought of as either focusing on constructing balancing weights, or focusing on modeling the conditional outcomes. Ben-Michael, Feller, and Rothstein [2018] is an interesting exception. Their Augmented Synthetic Control (ASC) estimator uses a model for the conditional expectation of the last period's outcome Y_{iT} in terms of the lagged outcomes, in combination with the SC balancing weights,

in the spirit of unconfoundedness type methods, and in particular residual balancing methods [Robins, Rotnitzky, and Zhao, 1994, Athey, Imbens, and Wager, 2018]. Their method also has double robustness properties, but it cannot be characterized as a weighted regression estimator. The importance of combining such outcome modeling and balancing/weighting and the associated double robustness are prominent features of the general program evaluation literature [e.g., Chernozhukov, Escanciano, Ichimura, Newey, and Robins, 2018b, Hirshberg and Wager, 2018, Imbens and Rubin, 2015, Newey, Hsieh, and Robins, 2004, Scharfstein, Rotnitzky, and Robins, 1999], and most of the currently recommended estimators in that literature combine them.

An attraction of our regression set up is that it generalizes naturally to the case with multiple treated units and multiple treated periods. We can in that case choose the unit weights for the control units to balance the average of the treated units during the pre-treatment period, and the time weights for the pre-treatment periods to balance the average post-treatment outcomes for the control units. The regression set up can also easily accomodate covariates that vary by unit and time by including them in the regression function. Unit-specific but time-invariant covariates, which cannot be accomodated in the standard DID set up can be accomodated here by adjusting the unit weights so the weights also balance these unit-specific covariates, and similarly for time-varying covariates common to all units.

In the second half of the paper, we establish asymptotic properties of the SDID estimator in a regime where both N and T are large. Establishing formal asymptotic properties for estimators has been a major challenge in the SC literature. Throughout, we take the perspective, common in panel data settings, that $\boldsymbol{Y}$ is a noisy estimate of an underlying signal matrix $\boldsymbol{L}$, i.e., $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$, where W_{it} denotes treatment assignment and ε is noise. The matrix $\boldsymbol{L}$ could have a simple two-way fixed effect form, or have more generic low-rank structure (e.g., interactive fixed effects, latent factor models) as in Athey, Bayati, Doudchenko, Imbens, and Khosravi [2017], Bai and Ng [2002], Bai [2009], Bonhomme and Manresa [2015], Li and Bell [2017], and Xu [2017]. We prove a variety of consistency results under different assumptions to highlight the double robustness properties of our proposed estimator. One consistency result makes weak assumptions on the weights but relatively strong assumptions on the outcome model $\boldsymbol{L}$, while another makes weak assumptions on the conditional outcome model but stronger assumptions on the weights.

Ideally, the weights $\hat{\omega}$ and $\hat{\lambda}$ would balance out the rows and columns of the underlying signal matrix $\boldsymbol{L}$ in a way that eliminates bias, and moreover the weights would not depend

on the noise ε. This is essentially what occurs in the analysis of balancing methods under unconfoundedness, where pre-treatment covariates are taken to be noiseless [Athey, Imbens, and Wager, 2018, Graham, de Xavier Pinto, and Egel, 2012, Hainmueller, 2012, Imai and Ratkovic, 2014, Zubizarreta, 2015]. Here, however, the weights $\hat{\omega}$ and $\hat{\lambda}$ are optimized to balance $\boldsymbol{Y}$, not $\boldsymbol{L}$, and have a rich dependence on the noise ε that cannot be eliminated via sample splitting. In Section 4.3, we use tools from modern empirical risk minimization theory to address both challenges and to show that, despite being optimized to balance the observed $\boldsymbol{Y}$, and despite the fact that balancing the $\boldsymbol{Y}$ is a major challenge because the number of units is of the same order as the number of time periods, the weights $\hat{\omega}$ and $\hat{\lambda}$ balance the unobserved $\boldsymbol{L}$ well enough to achieve consistency. In addition to proving consistency of the SDID estimator, our results also allow us to establish conditions under which the original SC estimator is consistent given a low-rank $\boldsymbol{L}$. The conditions on the weights for consistency of the SC estimator are stronger than those needed for consistency of SDID because the latter has a double bias removal property thanks to the time weights. Our asymptotic results require that we modify the original SC weights using penalization to ensure that the number of units with positive weights increases in large samples.

Finally, we present conditions that justify calculating the standard error for $\hat{\tau}^{\text{sdid}}$ using standard robust inference methods for DID regressions; we show that the standard robust standard errors are valid despite the fact that they take the weights as fixed, that is, they do not algorithmically account for dependence of the weights on the data.

2 Synthetic Panel Methods

Suppose we have a balanced panel with observations on an outcome Y_{it}, $i = 1, \ldots, N$, $t = 1, \ldots, T$, with some units treated in some periods, and the binary treatment indicator denoted by $W_{it} \in \{0, 1\}$. Throughout this paper, we assume that the outcomes Y_{it} are generated as $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$, where ε_{it} is a noise term (potentially with correlation over time, that is, within rows of the matrix $\boldsymbol{Y}$), and $\boldsymbol{L}$ is a baseline expected response matrix that may be correlated with W_{it}. The methods proposed here can also be generalized to allow for heterogeneity in τ, e.g., perhaps treatment intensity depends on the number of time periods for which a unit has been exposed to treatment. We refer to Athey and Imbens [2018] for a design-based interpretation of

these estimands using potential outcomes.

As motivation for our approach, consider models that parametrize the conditional expectation of the full set of unit and time period pairs: $\boldsymbol{L} = g(\theta)$, where $g : \Theta \mapsto \mathbb{R}^N \times \mathbb{R}^T$ models the conditional expectation of the control outcomes in terms of an unknown parameter θ. Examples of such panel data models include

$$
\begin{aligned}
&g(\theta)_{it} = \theta, && && \text{(constant)}\\
&g(\mu, \alpha, \beta)_{it} = \mu + \alpha_i + \beta_t, && \theta = (\mu, \alpha, \beta), && \text{(two-way fixed effect)}\\
&g(\boldsymbol{A}, \boldsymbol{B})_{it} = \sum_{r=1}^{R} A_{ir} B_{tr}, && \theta = (\boldsymbol{A}, \boldsymbol{B}), && \text{(factor model).}
\end{aligned}
$$

(In the two-way fixed effect and factor models we also need some normalizations, e.g., $\alpha_1 = \beta_1 = 0$.) Now, given such a $g(\cdot)$, a natural approach would be to fit θ and τ on all the data:

$$
\left(\hat{\theta},\, \hat{\tau}\right) = \arg\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(Y_{it} - g(\theta)_{it} - \tau W_{it}\Big)^2. \tag{2.1}
$$

In the case with $g(\theta)_{it} = \mu + \alpha_i + \beta_t$, this leads us to the basic difference in differences estimation strategy.

A challenge with this approach, however, is that if both N and T are large it may be difficult to find a simple specification of $g(\cdot)$ that fits well over the entire panel. But accurately estimating the entire matrix $\boldsymbol{L}$ is more difficult than the actual challenge: in order to estimate τ, we only need the expectation L_{it} locally, that is, for the control potential outcomes in the treated cell or cells. Here, we propose addressing potential misspecification of $g(\cdot)$ through weighting. We apply weights $\hat{\omega}_i$ to the units and $\hat{\lambda}_t$ to the time periods to form a "synthetic panel" on which the model $g(\cdot)_{it}$ is approximately unbiased for L_{it}, and then we build a model of the conditional expectation on this weighted panel:

$$
\left(\hat{\theta}^{\text{sdid}},\, \hat{\tau}^{\text{sdid}}\right) = \arg\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T} \Big(Y_{it} - g(\theta)_{it} - \tau W_{it}\Big)^2 \hat{\omega}_i \hat{\lambda}_t. \tag{2.2}
$$

We can qualitatively think of the SDID estimator as relaxing the parallel trends assumption in the DID estimator. Instead of assuming parallel trends for all units and all time periods,

the SDID estimator assumes that there exist unit and time weights such that the averaged treated unit and the weighted average of the control units satisfy a parallel trends assumption, and satisfy it not for all time periods but only for the averaged post-treatment period and the weighted average of the pre-treatment periods.

One of our main findings is that, if we use synthetic control weights $\hat{\omega}_i$ and $\hat{\lambda}_t$, then relying on a simple two-way fixed effect model for $g(\cdot)$ in (2.2) allows for consistent (in the large N and large T sense) estimation of τ, even if the two-way fixed effects model may be badly misspecified over the full panel. This finding is in line with a key insight from the program evaluation literature is that often methods that combine weighting/balancing the treated and control units with modeling the control outcome distribution outperform methods that only model the outcomes, as well as methods that only balance treated and control units. "Better" here includes both formal bias properties, as well as simulation evidence. A key formal property is that of double robustness, where misspecification of only the balancing weights or the conditional outcome model does not lead to inconsistency of $\hat{\tau}$ [Athey, Imbens, and Wager, 2018, Belloni, Chernozhukov, and Hansen, 2014, Chernozhukov, Escanciano, Ichimura, Newey, and Robins, 2018b, Hirshberg and Wager, 2018, Imbens and Rubin, 2015, Newey, Hsieh, and Robins, 2004, Scharfstein, Rotnitzky, and Robins, 1999].

Before discussing our choice of weights further below, we note that (2.2) not only allows for practical estimation of τ, but can also be used to build confidence intervals for τ. In the case of two-way fixed effects with covariates,

$$\left(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\tau}^{\text{sdid}}\right) = \arg\min_{\alpha,\beta,\mu,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - X_{it}\gamma - W_{it}\tau\right)^2 \hat{\omega}_i \hat{\lambda}_t, \tag{2.3}$$

there is a wide variety of standard error estimates that that have been studied in the literature; see Arellano [2003], Bertrand, Duflo, and Mullainathan [2004], Hansen [2007], Liang and Zeger [1986], and Stock and Watson [2008] for examples and discussions. We present both formal and simulation evidence that we can obtain valid confidence intervals for τ by applying these methods directly to (2.2), treating the weights $\hat{\omega}$ and $\hat{\lambda}$ as fixed, despite the fact that these weights depend on the Y_{it}; see Section 5 for details.

2.1 Weighting for Synthetic Panels

In this paper, we focus on weighting in a generalization of the basic synthetic control setting of Abadie, Diamond, and Hainmueller [2010]. We assume that exposed units $i > N_0$ get treated in time periods $t > T_0$, i.e $W_{it} = 1\{i > N_0 \text{ and } t > T_0\}$. The weighting component of our approach focuses on weights to balance the sample towards the treated unit / time period pairs. A critical feature of our approach, appropriate to our treatment pattern W, is that we impose a factor structure on the weights: $\gamma_{it} = \omega_i \lambda_t$. In addition to the factor structure we impose some restrictions on the unit and time period weights. The weights are non-negative, and weight groups $\omega_{1:N_0}$, $\omega_{(N_0+1):N}$, $\lambda_{1:T_0}$, $\lambda_{(T_0+1):T}$ all sum to one. Moreover, we give equal weight to all the exposed units and time periods. Formally, the set of weights we consider satisfy

$$\begin{aligned}
\mathbb{W} &= \left\{ \omega \in \mathbb{R}^N \,\middle|\, \omega_i \geq 0; \quad \sum_{i=1}^{N_0} \omega_i = 1; \quad \omega_{N_0+1},\, ...,\, \omega_N = \frac{1}{N - N_0} \right\}, \\
\mathbb{L} &= \left\{ \lambda \in \mathbb{R}^T \,\middle|\, \lambda_t \geq 0; \quad \sum_{t=1}^{T_0} \lambda_i = 1; \quad \lambda_{T_0+1},\, ...,\, \lambda_T = \frac{1}{T - T_0} \right\}.
\end{aligned} \tag{2.4}$$

One possible choice for the weights is the SC weights Abadie, Diamond, and Hainmueller [2010, 2015], which Doudchenko and Imbens [2016] show, for the case without covariates, can be written as (in the basic form we consider, synthetic control analyses only have one exposed unit, i.e., $N_0 = N - 1$ and only one exposed time period, $T_0 = T - 1$; however, the generalization is immediate)

$$\hat{\omega}^{\text{sc}} = \underset{\omega \in \mathbb{W}}{\arg\min} \sum_{t=1}^{T-1} \left(\sum_{i=1}^{N-1} \omega_i Y_{it} - Y_{Nt} \right)^2. \tag{2.5}$$

In the current paper we modify these weights slightly by putting an L_2 (ridge) penalty on the weights to ensure that in larger samples there will be many units with non-zero weights, which is important for the asymptotic properties of the estimator. Note that using an L_1 (lasso) penalty does not work because the weights are nonnegative and sum to one. We also consider the time equivalent of the SC weights, which do not appear to have been considered in this literature:

$$\hat{\lambda}^{\text{sc}} = \underset{\lambda \in \mathbb{L}}{\arg\min} \sum_{i=1}^{N-1} \left(\sum_{t=1}^{T-1} \lambda_t Y_{it} - Y_{iT} \right)^2. \tag{2.6}$$

The time weights play somewhat of a different role than the unit weights. In some cases one may wish to explicitly put more weights on recent periods than on distant periods, and not solely have these weights determined by the similarity, in terms of outcomes, to the current period. We may also wish to regularize the time weights to avoid putting most of the weight on a very small number of units or time periods.

In cases where the data exhibit substantial trends, the SC time weights would tend to concentrate on the most recent values. One modification in that case is to allow for an intercept in the regression, and solve

$$\hat{\lambda}^{\text{isc}} = \underset{\lambda_0 \in \mathbb{R},\, \lambda \in \mathbb{L}}{\arg\min} \sum_{i=1}^{N-1} \left(\lambda_0 + \sum_{t=1}^{T-1} \lambda_t Y_{it} - Y_{iT} \right)^2 . \tag{2.7}$$

The intercept $\hat{\lambda}_0$ is not needed for weighting, as the time dummies β_t will be able to absorb any time trends during the modeling stage. We refer to these weights as the intercept weights $\hat{\lambda}_t^{\text{isc}}$, and note that these weights are invariant to adding in any global time trend to our observations, $Y_{it} \leftarrow Y_{it} + f(t)$.

An alternative, for both the unit and time weights, is to use kernel weights. For example, if there is only one exposed unit / time period, we could use

$$\hat{\omega}_i^{\text{kernel}} \propto K\left(\frac{\boldsymbol{Y}_{i(1:T-1)} - \boldsymbol{Y}_{N(1:T-1)}}{h_\omega} \right), \qquad \hat{\lambda}_t^{\text{kernel}} \propto K\left(\frac{\boldsymbol{Y}_{(1:N-1)t} - \boldsymbol{Y}_{(1:N-1)T}}{h_\lambda} \right), \tag{2.8}$$

for some kernel function $K(\cdot)$, e.g., $K(a) = \exp(-a^\top a)$. We allow the tuning parameter to be different for the unit and time dimension. We also consider nearest neigbor weights, where we given constant weights to the K_ω nearest units and the K_λ neast time periods [e.g., Abadie and Imbens, 2006].

Using nearest neighbor methods to construct weights stresses the challenges in obtaining formal large sample properties for the resulting estimators, and this explains partly the limited nature of large sample results in the SC literature. If N is large, it is impossible to obtain a "close" match for $\boldsymbol{Y}_{(1:N-1)T}$ because we are matching on $N-1$ variables with only $T-1$ potential matches (e.g., Abadie and Imbens [2006]). Similarly, if T is large it is impossible to obtain close matches for $\boldsymbol{Y}_{N(1:T-1)}$ because there are only $N-1$ potential matches and $T-1$ variables to match on. With both N and T large it is impossible to obtain close matches in either direction.

Nevertheless, under some assumptions on the outcome model, the closest matches may be good enough, in the sense that they match closely on the relevant underlying variables. For example, if the data are generated by a two-way fixed effect model for $\boldsymbol{L}$, matching on all the lagged outcomes will not give a close match in terms of all the lagged outcomes. But, in large N and large T such matching methods will lead to matches that are close in terms of the unit-fixed effects, which is what matters. Our formal results show that this holds in general factor models for $\boldsymbol{L}$.

3 Panel Estimators as Bias Reduction Methods

In the previous section, we introduced SDID as a flexible approach to estimating causal effects in panels where both N and T are moderately large. To gain further intuition about the method, we focus here on the behavior of the SDID estimator in the case where we fit a two-way fixed effects model without any additional covariates, and only unit N in time period T gets treated. In this case, the SDID estimator allows for a simple, closed form solution that allows for a transparent comparison with the SC estimator.

We use the following notation. Partition the $N \times T$ matrix of observed outcomes $\boldsymbol{Y}$, and other comformable matrices, by treatment group and pre/post treatment period:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_{::} & \boldsymbol{Y}_{:T} \\ \boldsymbol{Y}_{N:} & \boldsymbol{Y}_{NT} \end{pmatrix},$$

where $\boldsymbol{Y}_{::}$, $\boldsymbol{Y}_{:T}$, $\boldsymbol{Y}_{N:}$, and $\boldsymbol{Y}_{NT}$ are $(N-1)\times(T-1)$, $(N-1)\times 1$, $1\times(T-1)$, and 1×1 matrices respectively. Also define $\boldsymbol{Y}_{i:}$ to be a $T-1$ dimensional row vector and define $\boldsymbol{Y}_{:t}$ to be a $N-1$ dimensonal column vector, each with typical element $\boldsymbol{Y}_{it}$. Define the averages for the three sets of control outcomes,

$$\overline{Y}^{\text{c,pre}} = \frac{1}{(N-1)(T-1)} \sum_{i=1}^{N-1} \sum_{t=1}^{T-1} Y_{it}, \qquad \overline{Y}^{\text{c,post}} = \frac{1}{N-1} \sum_{i=1}^{N-1} Y_{iT},$$

and

$$\overline{Y}^{\text{t,pre}} = \frac{1}{T-1}\sum_{t=1}^{T-1} Y_{Nt}.$$

In this setting, the basic difference in difference estimator (1.2) can be written as

$$\begin{aligned} &\hat{\tau}^{\text{did}} = Y_{NT} - \hat{Y}_{NT}^{\text{did}}(0), \\ &\hat{Y}_{NT}^{\text{did}}(0) = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T = \overline{Y}^{\text{c,pre}} + \left(\overline{Y}^{\text{t,pre}} - \overline{Y}^{\text{c,pre}}\right) + \left(\overline{Y}^{\text{c,post}} - \overline{Y}^{\text{c,pre}}\right). \end{aligned} \tag{3.1}$$

We can see the DID estimator as doubly bias-adjusting the simple average $\overline{Y}^{\text{c,pre}}$, with the first bias adjustment, $\overline{Y}^{\text{t,pre}} - \overline{Y}^{\text{c,pre}}$, taking into account stable differences between the treated unit and the control units and the second bias adjustment, $\overline{Y}^{\text{c,post}} - \overline{Y}^{\text{c,pre}}$, taking into account stable differences over time for the control group. The main weakness of this basic DID estimator, however, is of course that it is only valid under a well-specified two-way fixed effects model, which is a very strong assumption when both N and T are moderately large.

3.1 Synthetic Control as a Single Bias Reduction Method

The main idea of the SC approach is to re-weight the control rows $i = 1, ..., N-1$ of the matrix Y with weights $\hat{\omega}_i^{\text{sc}}$ such as to make the time trends among the weighted controls and the treated unit track each other. In the spirit of (3.1), we can write the SC estimator (1.1) as a weighted bias-reduced estimator:

$$\hat{\tau}^{\text{sc}} = Y_{NT} - \hat{Y}_{NT}^{\text{sc}}(0), \quad \hat{Y}_{NT}^{\text{sc}}(0) = \frac{1}{T-1}\sum_{i=1}^{N-1}\sum_{t=1}^{T-1} \hat{\omega}_i^{\text{sc}} Y_{it} + \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} \left(Y_{iT} - \frac{1}{T-1}\sum_{t=1}^{T-1} Y_{it} \right). \tag{3.2}$$

The bias adjustment uses a weighted average of the post-treatment control outcomes, with weights $\hat{\omega}_i^{\text{sc}}$ minus a doubly weighted average of the pre-treatment control outcomes.

The SC estimator presents an obvious improvement over the DID estimator in its use of weights to address potential misspecification of the basic two-way fixed effects model. However, unlike (3.1), the estimator (3.2) appears to be "missing" a second bias correction term of the form $1/(T-1)\sum_{t=1}^{T-1}(Y_{Nt} - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{it})$ that seeks to correct for a potential systematic failure of the weights $\hat{\omega}_i^{\text{sc}}$ to achieve balance in the pre-treatment periods. It is interesting to note that

if the weights $\hat{\omega}_i^{\text{sc}}$ were to balance the pre-treatment periods perfectly, so that

$$Y_{Nt} - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{it} = 0, \qquad \text{for all } t = 1, \ldots, T-1, \tag{3.3}$$

then the second bias correction term would be numerically zero, and so SC could implicitly be seen as a double-bias reduction method.[1] Typically, however, perfect balance as in (3.3) does not hold, and so the lack of this second bias-correction term may affect the properties of the SC estimator.

3.2 Synthetic Difference In Differences as a Double Bias Reduction Method

The SDID estimator addresses the case where the synthetic control adjustment does not completely balance the underlying signal in the pre-treatment periods. In the special case with only unit N treated in period T, the SDID estimator (1.3) can be thought of as bias-adjusting the SC estimator based on the pre-treatment discrepancies, weighted by $\hat{\lambda}_t^{\text{sc}}$:

$$\hat{\tau}^{\text{sdid}} = Y_{NT} - \hat{Y}_{NT}^{\text{sdid}}(0), \quad \hat{Y}_{NT}^{\text{sdid}}(0) = \hat{Y}_{NT}^{\text{sc}}(0) + \sum_{t=1}^{T-1} \hat{\lambda}_t^{\text{sc}} \left(Y_{Nt} - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{it} \right).$$

We can also write the SDID estimator as a symmetric version of (3.2), with

$$\hat{Y}_{NT}^{\text{sdid}}(0) = \sum_{i=1}^{N-1} \sum_{t=1}^{T-1} \hat{\omega}_i^{\text{sc}} \hat{\lambda}_t^{\text{sc}} Y_{it} + \sum_{t=1}^{T-1} \hat{\lambda}_t^{\text{sc}} \left(Y_{Nt} - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{it} \right) + \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} \left(Y_{iT} - \sum_{t=1}^{T-1} \hat{\lambda}_t^{\text{sc}} Y_{it} \right). \tag{3.4}$$

That is, compared to the simple weighting estimator $\hat{Y}_{NT}^{\text{weight}}$ there are two (weighted) bias adjustments,

$$\sum_{t=1}^{T-1} \hat{\lambda}_t^{\text{sc}} \left(Y_{Nt} - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{it} \right) \qquad \text{and} \quad \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} \left(Y_{iT} - \sum_{t=1}^{T-1} \hat{\lambda}_t^{\text{sc}} Y_{it} \right),$$

[1] An equivalent statement of this fact is that if (3.3) were to hold, then adding row fixed effects to to the synthetic control estimator (1.1) would not change the point estimate $\hat{\tau}$.

whereas the SC estimator has only one bias adjustment (the second one), similar to the way the DID estimator has two bias adjustments in the unweighted case.

The problem of turning synthetic controls into a double-bias removal style estimator has also been recently considered by Ben-Michael, Feller, and Rothstein [2018]. Their main proposal, the augmented synthetic control (ASC) estimator, involves fitting a model for the conditinoal expectation $m(\cdot)$ for Y_{iT} in terms of the lagged outcomes $\boldsymbol{Y}_{i(1:(T-1))}$, and then using this fitted model to "augment" the basic synthetic control estimator

$$\begin{aligned}\hat{Y}^{\text{asc}}_{NT}(0) &= \sum_{i=1}^{N-1} \hat{\omega}^{\text{sc}}_i Y_{iT} + \left(\hat{m}(\boldsymbol{Y}_{N:}) - \sum_{i=1}^{N-1} \hat{\omega}^{\text{sc}}_i \hat{m}(\boldsymbol{Y}_{i(1:(T-1))}) \right) \\ &= \hat{m}(\boldsymbol{Y}_{N:}) + \left(\sum_{i=1}^{N-1} \hat{\omega}^{\text{sc}}_i \Big(Y_{it} - \hat{m}(\boldsymbol{Y}_{i(1:(T-1))}) \Big) \right).\end{aligned} \tag{3.5}$$

The first representation of the ASC estimator emphasizes its interpretation as a modification of the SC estimator. It uses a cross-section model for the last period's outcome to remove biases from the standard SC estimator. The second representation stesses the connections to the unconfoundedness literature. The starting point is a model for the potential outcomes in the last period as a function of lagged outcomes. On its own this would suggest the estimator $\hat{m}(\boldsymbol{Y}_{N:})$; the ASC estimator then robustifies this using a weighted average of the residuals. This construction is related to the residual balancing estimators in the original double robust literature [Robins, Rotnitzky, and Zhao, 1994] or in high-dimensional settings [Athey, Imbens, and Wager, 2018], where now the SC weights can be interpreted as a type of covariate-balancing inverse-propensity weights. Such adjustments make the estimator doubly robust under appropriate conditions. A more recent paper, Chernozhukov, Wuthrich, and Zhu [2018c], takes a similar approach to Ben-Michael, Feller, and Rothstein [2018]; however, they swap the role of units and time periods and present formal results under strong time-homogeneity assumptions that, in particular, rule out the low-rank model $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$.

The second representation of the ASC estimator makes clear that its formal justification would be standard under a unconfoundedness assumption with the lagged outcomes playing the role of the pre-treatment variables, given a fixed number of pretreatment periods and large N. By the same token it would make the justification more difficult under general factor structures and with large T. This representation also highlights the feature of this estimator that it includes

the lagged outcomes in exactly the same way that pre-treatment variables would be included in unconfoundedness-type analyses.[2] This is in contrast to many panel data models such as fixed effect and factor models that incorporate lagged outcomes in the model in a way that is similar to the way the last period outcomes are treated, namely as noisy measures of the underlying unobserved components that are critical for prediction.

Despite their different motivations, ASC and SDID share an interesting connection: In the special case with a single treated unit / period and with a linear $\hat{m}(\cdot)$ model, the SDID estimator in (3.4) and the ASC estimator in (3.5) are very similar. In fact, they would be equivalent if we impose the additional restriction on the ASC estimator that the slope coefficients are nonnegative, positive and to sum to one. This connection suggests that weighted double bias-removal methods are a natural way of working with panels where we do not believe the basic DID approach to be appropriate. This being said we emphasize that, once we move past the most basic model, e.g., we have covariates or multiple treated units and periods, or we use more flexible specifications for $m(\cdot)$, then the connection between the two methods breaks down. Moreover, as Ben-Michael, Feller, and Rothstein [2018] motivate their estimator using unconfoundedness type arguments, they do not provide consistency results for the type of factor models considered here. In addition the ASC estimator does not have a weighted least squares interpretation, which is helpful in accommodating covariates.

4 Formal Results

In this section, we develop the properties of the SDID estimator. First, we consider properties that hold when the DID model is correctly specified; second, we discuss robustness properties provided by weighting. For simplicity, in this section we focus on the single exposed unit / time period setting. In this case, we can write the SDID treatment effect estimate $\hat{\tau}$ as

$$\hat{\tau} = Y_{NT} - \widehat{L}_{NT}, \quad \widehat{L}_{NT} = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T, \tag{4.1}$$

[2]Ben-Michael, Feller, and Rothstein [2018] also propose a suite of methods that can be used when both unit-specific covariates and lagged outcomes are available. For example, they propose first projecting out the component of the outcomes that can be explained using the unit-specific covariates, and then running ASC on the residuals.

where the parameters $\hat{\mu}$, $\hat{\alpha}$ and $\hat{\beta}$ are as defined in (1.3). Here, we provide several results establishing convergence of $\widehat{L}_{NT}$ to L_{NT}, implying that the error of $\hat{\tau}$ is asymptotically fully determined by the intrinsic noise in Y_{NT}. In the following section, we then build on these results to provide methods for inference about τ in settings with more than one treated unit. Recall that we assume that Y_{it} is generated as below, and we follow the convention that ":" always indexes over unexposed units or time periods.

Assumption 1. *We have $N, T \to \infty$, and there is a deterministic matrix $\boldsymbol{L}$ such that $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$ with $W_{it} = 1\{i = N, t = T\}$ and $\varepsilon_{it} \sim \mathcal{N}(0,\ \sigma^2)$, independently for each cell $(i,\ t)$.*[3]

We consider two distinct sets of conditions: First, we examine the case where the two-way fixed effects model is well specified, i.e., $L_{it} = \mu + \alpha_i + \beta_t$, and show that SDID is consistent under very flexible conditions. The main point here is that using data-adaptive weights $\hat{\omega}$ and $\hat{\lambda}$ does not break DID when the outcome model is well specified. Second, we consider the generalized fixed effects model, i.e., where we make only weak assumptions on $\boldsymbol{L}$. Here, basic DID is inconsistent; however, we show that SDID with penalized SC weights is consistent whenever $\boldsymbol{L}$ is well-approximated by a matrix of rank $r \ll \min(N, T)$ and can be consistent at the rate $\sqrt{\log(T)/\min(N, T)}$ if it is well-approximated by a matrix of fixed rank.

4.1 Properties in the Well-Specified Two-Way Fixed Effects Model

Our first result shows that SDID is consistent and asymptotically normal in the well-specified model, that is, where $L_{it} = \mu + \alpha_i + \beta_t$. Here we consider the following kernel weights:

$$\hat{\omega}_i = \frac{1\left(\left\{\frac{1}{T-1}\left\|\boldsymbol{Y}_{i:} - \boldsymbol{Y}_{N:}\right\|_2^2 \le c_\omega\right\}\right)}{\sum_{i \ne N} 1\left(\left\{\frac{1}{T-1}\left\|\boldsymbol{Y}_{i:} - \boldsymbol{Y}_{N:}\right\|_2^2 \le c_\omega\right\}\right)}, \quad \hat{\lambda}_t = \frac{1\left(\left\{\frac{1}{N-1}\left\|\boldsymbol{Y}_{:t} - \boldsymbol{Y}_{:T}\right\|_2^2 \le c_\lambda\right\}\right)}{\sum_{t \ne T} 1\left(\left\{\frac{1}{N-1}\left\|\boldsymbol{Y}_{:t} - \boldsymbol{Y}_{:T}\right\|_2^2 \le c_\lambda\right\}\right)}, \tag{4.2}$$

for $i = 1, ..., N-1$ and $t = 1, ..., T-1$, where c_ω and c_λ are tuning parameters. For our result, we also make generative assumptions that let us characterize the behavior of nearest neighbor matching with noisy data; see Bonhomme, Lamadon, and Manresa [2017] for related results on the behavior of clustering panel data.

[3] We make this assumption for simplicity of exposition only. In Section 8.3 of the appendix we state and prove results for heteroskedastic and auto-correlated subgaussian errors, and also for choices of weights $\hat{\omega}, \hat{\lambda}$ that we do not consider here.

Assumption 2. *$L_{it} = \mu + \alpha_i + \beta_t$; $\delta_{\alpha,i} := |\alpha_i - \alpha_N|$ and $\delta_{\beta,t} := |\beta_T - \beta_t|$ are i.i.d. random variables such that corresponding densities f_{δ_α} and f_{δ_β} are bounded at zero.*

Theorem 1. *Suppose Assumptions 1 and 2 hold and $\lim N/T = \rho \in (0, 1)$; then for the weights $\hat{\omega}_i$ and $\hat{\lambda}_t$ defined above we have the following: $\widehat{L}_{NT}$ is consistent, that is,*

$$\widehat{L}_{NT} - L_{NT} \to_p 0,$$

and

$$\frac{1}{\sqrt{\|\hat{\omega}_{:}\|_2 + \|\hat{\lambda}_{:}\|_2}} \left(\widehat{L}_{NT} - L_{NT}\right) \to \mathcal{N}(0, \sigma^2) \tag{4.3}$$

provided $c_\omega = \sigma^2 + a_{N,T}\frac{\log(N)}{\sqrt{T}}$, $c_\omega = \sigma^2 + o(1)$, $a_{N,T} \to \infty$, and $c_\lambda = \sigma^2 + b_{N,T}\frac{\log(T)}{\sqrt{N}}$, $c_\lambda = \sigma^2 + o(1)$, $b_{N,T} \to \infty$

Note that matching discrepancies c_λ and c_ω in (4.2) do not go to zero, instead they go to σ^2, because with N and T large all rows and columns of Y will have distances that concentrate at σ^2 away. We also note that the weighting function considered here is approximately equivalent to k-nearest neighbors weighting, where $k_\omega = T\sqrt{c_\omega - \sigma^2}$ is approximately the number of units that we average over and $k_\lambda = T\sqrt{c_\lambda - \sigma^2}$ is the approximate number of used time periods.

4.2 Double Robustness Part I: The Fixed Effects Model with General Weights

Next we show that, if fixed-effects model is correct, then our method is consistent essentially regardless of the weights we use. Instead of requiring a specific functional form for the weights, we only ask that we not use the T-th time period when picking the row weights $\hat{\omega}$, and we not use the N-th row when picking $\hat{\lambda}$, and that the weights are not too concentrated on a few units or time periods. All algorithms considered in this paper, ranging from synthetic control weighting to nearest neighbor matching, satisfy this condition.

Assumption 3. *We choose weights such that $\hat{\omega}_{:} \perp\!\!\!\perp \boldsymbol{Y}_{:T}$ and $\hat{\lambda}_{:} \perp\!\!\!\perp \boldsymbol{Y}_{N:}$.*

Theorem 2. *Under Assumption 1, suppose moreover that $L_{it} = \mu + \alpha_i + \beta_t$. Then, provided we use weights $\hat{\omega}$ and $\hat{\lambda}$ satisfying Assumption 3 such that*

$$\|\hat{\omega}_{:}\|_2,\ \|\hat{\lambda}_{:}\|_2 \to_p 0, \qquad \sqrt{\max\{N,\, T\}}\|\hat{\omega}_{:}\|_2\|\hat{\lambda}_{:}\|_2 \to_p 0, \tag{4.4}$$

we have $\widehat{L}_{NT} - L_{NT} \to_p 0$.

4.3 Double Robustness Part II: The Approximate Factor Model with SC Weights

In this section we relax the modeling assumptions from the above section, and simply require that $\boldsymbol{L}$ be approximable by a low-rank matrix. This type of model was used to motivate the SC approach by Abadie, Diamond, and Hainmueller [2010], and has also been studied in other contexts by, e.g., Athey, Bayati, Doudchenko, Imbens, and Khosravi [2017] and Bai [2009]. Our goal is to show that, with well chosen weights, SDID remains consistent. Here, we focus on a form of penalized synthetic control weights:

$$\begin{aligned}
\hat{\omega}^{\text{sc}}(a_\omega) &= \arg\min_{\omega \in \mathbb{W}} \left\{ \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{i=1}^{N-1} \omega_i Y_{it} \right)^2 :\ \|\omega_{:}\|_2 \le a_\omega \right\}, \\
\hat{\lambda}^{\text{sc}}(a_\lambda) &= \arg\min_{\lambda \in \mathbb{L}} \left\{ \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2 :\ \|\lambda_{:}\|_2 \le a_\lambda \right\},
\end{aligned} \tag{4.5}$$

where a_λ and a_ω are tuning parameters. The penalization is important to ensure that in large samples there will be many units and time periods with positive weights.

The key difficulty in showing that these SC weights $\hat{\omega}$ and $\hat{\lambda}$ were chosen to balance rows and columns of Y; however, what we really need for useful inference with an approximately low-rank $\boldsymbol{L}$ is for $\hat{\omega}$ and $\hat{\lambda}$ to balance the the rows and columns of $\boldsymbol{L}$. Furthermore, the weights defined in (4.5) have a complicated dependence on the noise $\varepsilon = \boldsymbol{Y} - \boldsymbol{L}$, and the panel structure means that we cannot address this challenge via sample splitting as in, e.g., Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins [2018a]. Here, we establish conditions under which our SDID estimator with data-dependent weights (4.5) is consistent in the approximately low-rank model. As an additional benefit, we also prove that the basic SC estimator is consistent in

the motivating model from Section 2.2 of Abadie, Diamond, and Hainmueller [2010].

In order to spell out our result, we first define infeasible SC weights that balance the underlying effect matrix $\boldsymbol{L}$ rather than the observed matrix $\boldsymbol{Y}$:

$$\begin{aligned}\omega^\star(a_\omega) &= \arg\min_{\omega\in\mathbb{W}} \left\{ \sum_{t=1}^{T-1} \left(L_{Nt} - \sum_{i=1}^{N-1} \omega_i L_{it} \right)^2 : \ \|\omega_:\|_2 \leq a_\omega \right\}, \\ \lambda^\star(a_\lambda) &= \arg\min_{\lambda\in\mathbb{L}} \left\{ \sum_{i=1}^{N-1} \left(L_{iT} - \sum_{t=1}^{T-1} \lambda_t L_{it} \right)^2 : \ \|\lambda_:\|_2 \leq a_\lambda \right\},\end{aligned} \tag{4.6}$$

We then the following identification assumption in terms of these weights. Specifically we ask that these population weights succeed in obtaining balance, i.e., the last row and column of the matrix can in fact be usefully represented via a convex combination of other rows. Given this assumption, SDID with penalized SC weights is consistent.

Theorem 3. *Given Assumption 1, and that we choose weights via* (4.5) *with* a_ω *and* a_λ *satisfying*

$$\begin{aligned}\delta_\omega &= \|L_{N:} - \omega_:^\star(a_\omega)' L_{::}\|; \\ \delta_\lambda &= \|L_{:T} - L_{::}\lambda_:^\star(a_\lambda)\|; \\ \delta_{sdid} &= \left|L_{NT} - (\omega_:^\star(a_\omega)'\boldsymbol{L}_{:T} + \boldsymbol{L}_{N:}\lambda_:^\star(a_\lambda)) \ - \omega_:^\star(a_\omega)'\boldsymbol{L}_{::}\lambda_:^\star(a_\lambda)\right|.\end{aligned}$$

Then for r_λ, r_ω *defined in Lemma 4,*

$$\begin{aligned}\widehat{L}_{NT} - L_{NT} = \mathcal{O}_P\Bigg(&\delta_{sdid} + a_\lambda \left[\delta_\omega + \sigma \min\{\sqrt{\log(N)}, a_\omega\sqrt{N}\}\right] \\ &+ a_\omega \left[\delta_\lambda + \sigma \min\{\sqrt{\log(T)}, a_\lambda\sqrt{T}\}\right] \\ &+ \min(a_\omega r_\lambda, a_\lambda r_\omega)\Bigg).\end{aligned}$$

In the case that $N/T \to \kappa \in (0, \infty)$, $\sigma = \mathcal{O}(1)$, L *has exact (rather than approximate) low rank, and we choose* a_λ, $a_\omega, = \mathcal{O}(1/\sqrt{N})$, *this bound simplifies to*

$$\widehat{L}_{NT} - L_{NT} = \mathcal{O}_P\left(\delta_{sdid} + \sqrt{\frac{\max\left\{\log(N),\ \text{rank}(L),\ \delta_\omega,\ \delta_\lambda\right\}}{N}}\right).$$

The key technical result underlying Theorem 3 is the following lemma, which establishes convergence of the feasible SC weights (4.5) that balance Y to the infeasible weights (4.6) that balance $\boldsymbol{L}$. We emphasize that our result does not rely on the weights $\hat{\omega}_{:}$ and $\hat{\lambda}_{:}$ converging to $\hat{\omega}_{:}^{\star}$ and $\hat{\lambda}_{:}^{\star}$ respectively at a particularly fast rate. In our analysis, we only use the trivial bounds $\|\hat{\omega}_{:} - \hat{\omega}_{:}^{*}\| \leq \|\hat{\omega}_{:}\| + \|\hat{\omega}_{:}^{\star}\|$, etc., and in fact the weights $\hat{\omega}_{:}$ and $\hat{\lambda}_{:}$ do not appear to be particularly stable empirically. Rather, our result only relies on the feasible and oracle weights having similar "balancing" properties, i.e., for $L'_{::}(\hat{\omega}_{:} - \omega_{:}^{\star})$ and $L_{::}(\hat{\lambda}_{:} - \lambda_{:}^{\star})$ to be small as established below.

Lemma 4. *Given Assumption 1, choose weights via* (4.5) *with* a_ω, a_λ. *Then in terms of* $\delta_\omega, \delta_\lambda$ *defined in Theorem 3,*

$$\|L'_{::}(\hat{\omega}_{:} - \omega_{:}^{\star}(a_\omega))\|_2 = \mathcal{O}_P(r_\omega) \quad \textit{and} \quad \|L_{::}(\hat{\lambda}_{:} - \lambda_{:}^{\star}(a_\lambda))\|_2 = \mathcal{O}_P(r_\lambda),$$

where

$$\begin{aligned} r_\omega = \max\Bigg(&\delta_\omega,\ \sigma\sqrt{\text{approx-rank}_\omega},\\ &\sigma\sqrt{a_\omega \max\left(\sqrt{N}, \sqrt[4]{NT}\right)\log(N)},\ \sigma\sqrt{T}a_\omega, \sigma\min\left\{\sqrt{\log(N)}, a_\omega\sqrt{N}\right\}\Bigg);\\ r_\lambda = \max\Bigg(&\delta_\lambda,\ \sigma\sqrt{\text{approx-rank}_\lambda},\\ &\sigma\sqrt{a_\lambda \max\left(\sqrt{T}, \sqrt[4]{NT}\right)\log(T)},\ \sigma\sqrt{N}a_\lambda, \sigma\min\left\{\sqrt{\log(T)}, a_\lambda\sqrt{T}\right\}\Bigg). \end{aligned}$$

Here approx-rank$_\lambda$ *and* approx-rank$_\omega$ *are lower bounds on the rank of* $L_{::}$ *that ignore small nonzero singular values,*

$$\begin{aligned} \text{approx-rank}_\omega &= \min_{r\in 1,2,\ldots}\left\{\ r \geq \sigma^{-1}\min\left(a_\omega\sqrt{\sum_{k>r} s_k^2},\ s_{r+1}\min\left\{\sqrt{\log(N)}, a_\omega\sqrt{N}\right\}\right)\right\},\\ \text{approx-rank}_\lambda &= \min_{r\in 1,2,\ldots}\left\{\ r \geq \sigma^{-1}\min\left(a_\lambda\sqrt{\sum_{k>r} s_k^2},\ s_{r+1}\min\left\{\sqrt{\log(T)}, a_\lambda\sqrt{T}\right\}\right)\right\}, \end{aligned}$$

where $s_1, s_2, \ldots$ *is the decreasing sequence of singular values of* $L_{::}$.

Finally, as discussed above, we can also use Lemma 4 to prove consistency of SC estimation

in the approximately low-rank model under the assumptions of Theorem 3. The main difference between Theorem 5 and Theorem 3 above is that the error depends on the performance of an oracle SC estimator rather than that of the the oracle SDID estimator.

Theorem 5. *Given Assumption 1, choose $\hat{\omega}$ via* (4.5) *with constant a_ω. Then in terms of $\delta_{sc} = |\omega_\cdot^*(a_\omega) \cdot \boldsymbol{L}_{:T} - L_{NT}|$ and r_ω defined in Lemma 4,*

$$|\hat{\omega}_\cdot(a_\omega) \cdot Y_{:T} - L_{NT}| = \mathcal{O}_p\left(\delta_{sc} + a_\omega + \min_{\tilde{\lambda} \in \mathbb{R}^{T-1}} \left[r_\omega \|\tilde{\lambda}\|_2 + a_\omega \|L_{:T} - L_{::}\tilde{\lambda}\|_2\right]\right). \tag{4.7}$$

4.4 Asymptotic Properties with Multiple Exposed Units and Time Periods

We will now consider the problem of inference in a simple setting with N_1 exposed units and T_1 exposed time periods, in which all exposed units start treatment at the same time $T_0 + 1$. Relative to the setting of the previous section, we also consider autocorrelated errors.

Assumption 4. *We have N_0, $T_0 \to \infty$, and there is a deterministic $N_0 + N_1 \times T_0 + T_1$ matrix $\boldsymbol{L}$ such that $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$ with $W_{it} = 1\{i > N_0, t > T_0\}$ and the rows of ε are independent and distributed according to the gaussian AR(1) process $\varepsilon_{i,t+1} = \rho\varepsilon_{i,t} + \xi_{i,t+1}$ with $\rho \in [0, 1)$ and iid shocks $\xi_{i,t} \sim N(0, \sigma_\xi^2)$.*

In this setting, our essential result is that our estimator $\hat{\tau}$ is asymptotically normal and unbiased when our total number $N_1 T_1$ of treated observations is small relative to the number of untreated observations and our number of treated units N_1 is small relative to our number of post-treatment periods T_1. The following theorem formalizes this. Its proof, as well as a more complete discussion of our estimator in this setting, appears in Section 8.5 of the appendix.

Theorem 6. *Under Assumption 4, let $\hat{\tau}$ be defined as in* (4.1) *with weights*

$$\hat{\omega} = \arg\min_{\omega \in \mathbb{W}} \left\{ \sum_{t=1}^{T_0} \left(\frac{1}{N_1} \sum_{i=N_0+1}^{N_1} Y_{it} - \sum_{i=1}^{N_0} \omega_i Y_{it} \right)^2 : \ \|\omega_\cdot\|_2 \leq a_\omega \right\},$$

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{L}} \left\{ \sum_{i=1}^{N-1} \left(\frac{1}{T_1} \sum_{i=T_0+1} Y_{it} - \sum_{t=1}^{T_0} \lambda_t Y_{it} \right)^2 : \ \|\lambda_\cdot\|_2 \leq a_\lambda \right\}.$$

In terms of the correponding oracle weights $\omega^\star, \lambda^\star$ defined by substituting $\boldsymbol{L}$ for $\boldsymbol{Y}$ in the definitions above, let

$$\begin{aligned}
\delta_\omega &= \|N_1^{-1} \sum_{i=N_0+1}^{N_0+N_1} L_{i:} - \omega_\star' L_{::}\|, \\
\delta_\lambda &= \|T_1^{-1} \sum_{t=T_0+1}^{T_0+T_1} L_{:t} - L_{::}\lambda_\star\|, \\
\delta_{sdid} &= \left| (N_1 T_1)^{-1} \sum_{i=N_0+1}^{N_0+N_1} \sum_{t=T_0+1}^{T_0+T_1} L_{it} - (\omega_\star' \boldsymbol{L}_{:T} + \boldsymbol{L}_{N:}\lambda_\star - \omega_\star' L_{::}\lambda_\star) \right|.
\end{aligned}$$

If we choose $a_\omega \lesssim N_0^{-1/2}$ and $a_\lambda \lesssim T_0^{-1/2}$, then

$$\sqrt{N_1 T_1}(\hat{\tau} - \tau) = \frac{1}{\sqrt{N_1 T_1}} \sum_{i=N_0+1}^{N_0+N_1} \sum_{t=T_0+1}^{T_0+T_1} \varepsilon_{it} + o_p(1), \tag{4.8}$$

provided the following conditions hold:

$$\begin{aligned}
N_1 T_1 &\ll \min\left\{ \frac{1}{\delta_{sdid}^2}, \ \frac{N_0}{\max\{\delta_\lambda^2, 1\}}, \ \frac{T_0}{\max\{\delta_\omega^2, 1\}} \right\}; \\
N_1 T_1^{1/2} &\ll \min\left\{ \frac{N_0}{T_0^{1/2}\log(T_0)}, \ \frac{N_0^{1/2}}{\log(T_0)} \right\}; \\
N_1 &\ll \min\left\{ T_1, \ \frac{N_0}{\text{approx-rank}\left(L_{::}, T_1^{-1/2}\right)} \right\}.
\end{aligned}$$

The asymptotic characterization (4.8) implies that under the stated conditions, our estimator is asymptotically normal with variance V_τ on the order of $(N_1 T_1)^{-1}$.

5 Large-Sample Inference of Treatment Effects

In the literature on synthetic controls, the dominant approach to uncertainty quantification is via placebo tests [Abadie, Diamond, and Hainmueller, 2010, 2015]. The main idea is to consider the behavior of synthetic control estimation when we replace the unit that was in fact exposed to

the treatment with different units that were not exposed. Such a placebo test is closely connected to permutation tests in randomization inference. However, in many applications of synthetic controls, the exposed unit was not chosen at random, in which case placebo tests do not have the formal properties of randomization tests [Firpo and Possebom, 2018, Hahn and Shi, 2016], and so may need to be interpreted via a more qualitative lens. Here, we take a different perspective, and consider inferential methods motivated by large sample asymptotics. Our proposal builds on methods for robust inference in large panels that were originally developed for the well-specified two-way fixed effects models [e.g., Arellano, 2003, Hansen, 2007, Liang and Zeger, 1986].

As shown in Theorem 6, the synthetic difference in differences estimator is asymptotically Gaussian under appropriate conditions,

$$(\hat{\tau} - \tau) \,/\, V_\tau^{1/2} \Rightarrow \mathcal{N}(0,\, 1)\,, \tag{5.1}$$

where the asymptotic variance V_τ is determined by the sampling errors $\{\varepsilon_{it} : i > N_0, t > T_0\}$ of the observations under treatment and does not depend on the noise in the synthetic control weights $\hat{\omega}$ or $\hat{\lambda}$. The uphot is that we can estimate V_τ and build confidence intervals for τ using standard methods for large-sample inference for weighted panels, while treating $\hat{\omega}$ and $\hat{\lambda}$ as fixed.

Here, we focus on estimating V_τ by applying the jackknife [Miller, 1974, Efron and Stein, 1981] to the weighted regression (2.3)—again, with $\hat{\omega}$ and $\hat{\lambda}$ treated as fixed. Following Bertrand, Duflo, and Mullainathan [2004], we seek robustness to errors that may be correlated within rows, and so we repeatedly run the regression (2.3) with one row i omitted at a time to get $\hat{\tau}^{(-i)}$, and use the variation of these $\hat{\tau}^{(-i)}$ to get an estimate the variance V_τ of the original treatment effect estimate $\hat{\tau}$. Although we do not do so here, one could also estimate V_τ via other methods for heteroskedasticity-consistent variance estimation [Efron, 1982, MacKinnon and White, 1985].

In terms of connections to the literature, we note that a qualitatively result was used by Bonhomme and Manresa [2015] to provide large-sample inference for panels with grouped fixed effects: They rely on clustering to discover groups, but then show that inference remains asymptotically valid while ignoring the effect of clustering. Meanwhile, Chernozhukov, Wuthrich, and Zhu [2017] propose a method for inference in synthetic control problems with a single treated unit that relies on the prediction residuals $Y_{it} - \hat{L}_{it}$ over the control units being representative of the counterfactual untreated residuals $L_{it} + \varepsilon_{it} - \hat{L}_{it}$ over the treated units.

Finally, we note that although the above discuss has focused on inference that is robust to

within-row correlations, our jackknife-based procedure can be flexibly adapted to reflect different sampling assumptions. If we believe that the ε_{it} were all independent, we could also apply a cell-wise jackknife (i.e., where the jackknife omits one cell rather than one row at a time), potentially allowing for power gains when there are few treated units. Meanwhile, if we believe the ε_{it} may be correlated within rows but that the noise process eventually mixes (i.e., there are no long-range correlations), we could consider and intermediate solution that divides each row into blocks and then applies a block-wise jackknife [Kunsch, 1989].

6 Empirical Evaluation

6.1 Placebo Evaluation: Predicting the Prevalence of Smoking

In one of the original studies on synthetic control methods, Abadie, Diamond, and Hainmueller [2010] focus on estimating the causal effect of anti-smoking legislation in California (Proposition 99). As discussed above, when only a single cell (N, T) is treated, synthetic control methods can be understood as producing a prediction $\hat{L}_{NT}$ of what the unit-N time-T outcome would have been without treatment, and then estimating $\hat{\tau} = Y_{NT} - \hat{L}_{NT}$. This suggests a simple placebo procedure for evaluating various synthetic control methods in a realistic environment: If we run synthetic control methods while singling out as "treated" a cell (n, t) that did not actually receive treatment, we should expect $\hat{L}_{nt}$ to be a good prediction of the realized outcome Y_{nt}. Here, we benchmark difference in differences, synthetic controls, and synthetic differences in differences against each other by running such a placebo analysis, and comparing the errors of each method in predicting Y_{nt}.

The original dataset of Abadie, Diamond, and Hainmueller [2010] had observations for 39 states (including California) from 1970 through 2000, where California is treated from 1989 onwards. We follow Abadie, Diamond, and Hainmueller [2010] in using per capita smoking as the outcome. Here, we focus only on time periods 1970-1988, in which none of the units were treated, and seek to predict the outcome of a focal cell using all data from earlier years as well as data from different states in the target year by running different methods with the focal cell considered as "treated". For example, when predicting the outcome for Arizona in 1985, we run the methods under comparison with the other 38 states used as the "control states" and the years 1970-1984 used as the "pre-treatment years" (and we do not used any data from 1986 or

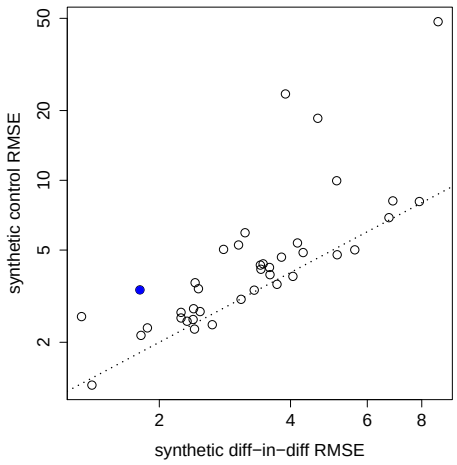


Figure 1: Comparison of the per-state root-mean squared error for SDID and SC. California is highlighted in blue.

later).

Given this setup, we make predictions for all states in years 1980-1988 (i.e., we run DID, SC and SDID separately for each focal state-year pair), and calculate the square root of the average squared error:

$$\text{RMSE}_i = \sqrt{\frac{1}{9}\sum_{t=1980}^{1988}\left(Y_{i,t} - \hat{L}_{i,t}\right)^2},$$

for all 39 states. In this example, we use L_2-penalized SC weights

$$\hat{\omega}^{\text{sc}} = \underset{\omega\in\mathbb{W}}{\arg\min}\left\{\frac{1}{T-1}\sum_{t=1}^{T-1}\left(Y_{Nt} - \sum_{i=1}^{N-1}\omega_i Y_{it}\right)^2 + \zeta\left\|\omega\right\|_2^2\right\}, \tag{6.1}$$

where we set ζ to be the average of $(Y_{i,t+1} - Y_{i,t})^2$ over the pre-treatment data. For time weights $\hat{\lambda}$, we use an analogously penalized version of the intercept weights λ_t^{isc} to deal with the trends in smoking rates.

We report the results on the RMSE in Figure 1 for each state and the average over all 39 states; Table 3 in the Appendix has detailed results. We find that the SDID method does substantially better than the SC and DID method in terms of predictive accuracy, with the SC outperforming the DID method: in Figure 1 almost all the pairs of RMSE lie above the 45

degree line, showing that the average RMSE based on the SDID estimator is lower than that based on the SC estimator, for almost every state. The median improvement of the state-wise root-mean-squared error of SDID over the state-wise root-mean-squared error of SC is 15% (and the corresponding improvement over DID is 50%).

We can gain further insight into the behavior of different methods by comparing the one-step-ahead predicted trajectories $\hat{L}_{i,t}$ to the true ones $Y_{i,t}$. We see that SC struggle when a state doesn't fit neatly within the convex hull of other states (e.g., in the case of Utah), whereas difference-in-differences does poorly when the temporal pattern of a state doesn't match the average temporal pattern (e.g., in Alabama). Of course, it is unlikely that practitioners would use SC to study a state that does not fit within the convex hull of other states, as is the case of Utah here, as standard goodness of fit checks would flag SC as an inappropriate method to use here. However, we find that SDID out-performs SC in states where the latter are appropriate (such as California), and remain robust in cases where the latter are not (such as Utah).

6.2 Simulation Results: Point Estimation

In this section we assess the properties of the proposed SDID estimator relative to the DID and SC estimators in finite samples using a simulation study. As in the above placebo study, we consider a setting with no treatment effect, and run all methods as though a single unit in cell (N, T) had been treated; then, we measure the accuracy of $\hat{L}_{NT}$ as an estimate for L_{NT}. In all our examples, the data is drawn as $Y_{it} \sim \mathcal{N}(L_{it}, \sigma^2)$, independently for each (i, t) pair. Meanwhile, the $N \times T$ signal matrix $\boldsymbol{L}$ is low rank, $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{N \times R}$ and $\boldsymbol{V} \in \mathbb{R}^{T \times R}$ for a rank parameter R.

The key choice is in how we generate this low-rank matrix $\boldsymbol{L}$. First, we consider a simulation where the N-th row and the N-th column of $\boldsymbol{L}$ are "typical"; formally, we generate $\boldsymbol{L}$ via an exchangeable process, such that $U_{il} \sim \text{Exp}(1)$ and $V_{tl} \sim \text{Exp}(1)$ independently for each (i, l) and (t, l). Second, we consider a case where the focal row and column are not "typical", and in particular the rows and columns are not exchangeable. Here, we draw $U_{il} \sim \text{Pois}(\sqrt{i/N})$ for each (i, l), and $V_{tl} \sim \text{Pois}(\sqrt{t/T})$ for each (t, l). Note that the N-th row and T-th column will on average have relatively large observations.

In all our simulations, we use consider penalized SC weights as in (6.1), with ζ set to the sample variance of the Y_{it}. Below, we first generate a random $\boldsymbol{L}$, and then simulate $\boldsymbol{Y}$ 20 times

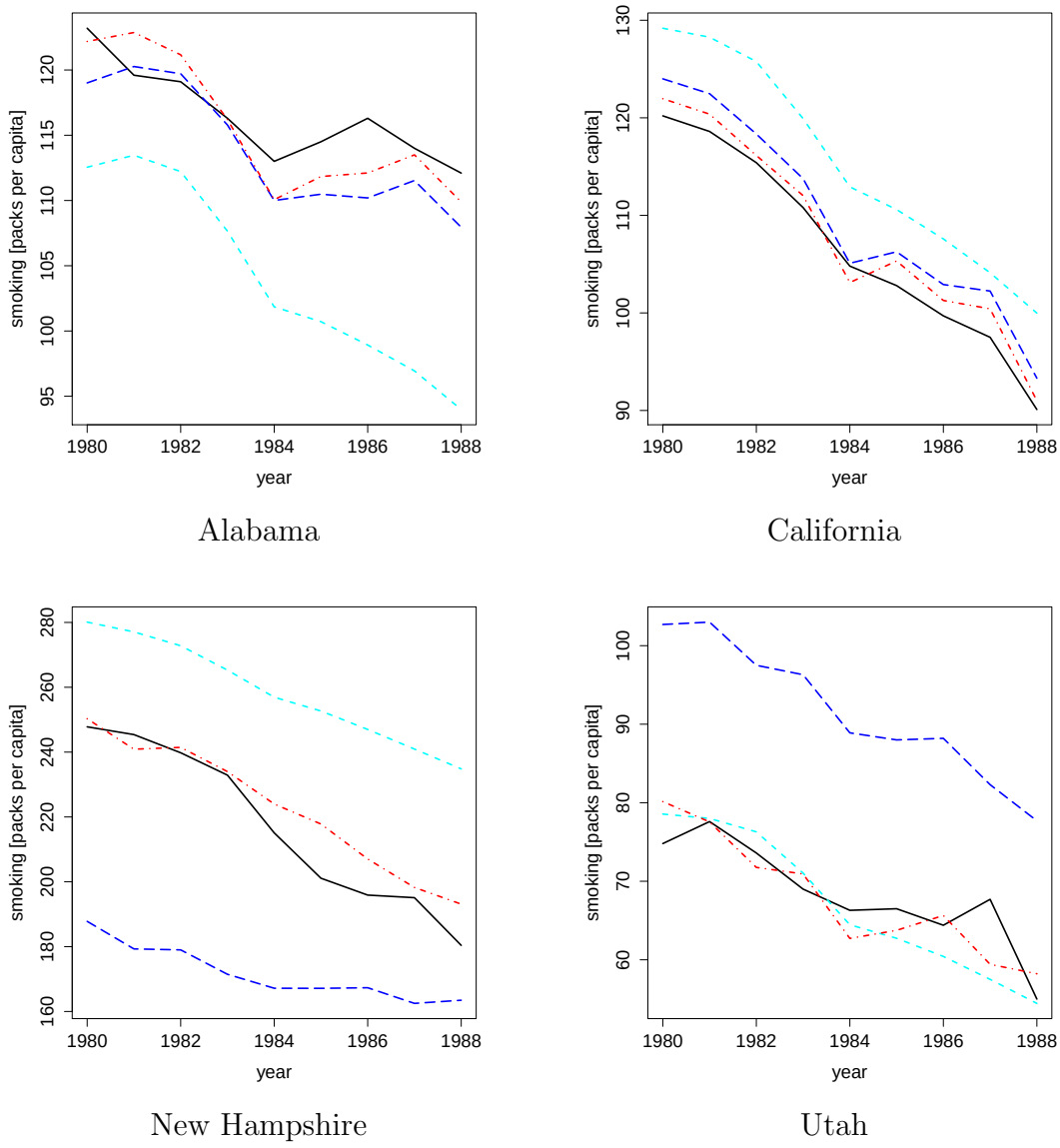


Figure 2: Predictions for per capita smoking rates for selected states, using as training data all years prior to the year indicated on the x-axis. The true yearly per-capita smoking $Y_{i,t}$ is in black. SDID estimates are in red. SC estimates are in blue. DID estimates are in teal.

given this $\boldsymbol{L}$. This lets us separate the contributions of bias and variance to the error. We report for the two designs, for different values of σ^2 and the rank R, and for different pairs of (N, T), the root-mean-squared-error and mean-absolute-bias for the three estimators, DID, SC, and SDID. We report results in Tables 1 and 2. In the Appendix, we also show results for unpenalized SC ($\zeta = 0$), in Tables 4 and 5. We find that in all cases the SDID estimator has substantially better bias properties than the DID and SC estimators, and in most cases also has better root-mean-squared-error.

6.3 Simulation Results: Confidence Intervals

Finally, we study the properties of confidence intervals derived via the weighted regression perspective of SDID. We work in the same data-generating distribution as for the "non-exchangeable" example in Section 6.2, except now with multiple treated units. Units $1 : \ldots, N_0$ are control units, and units $N_0 + 1, \ldots, N_0 + N_1 = N$ are treated from period $T_0 + 1$ onwards. Writing W_{it} for the treatment indicator, we draw data as

$$Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}\left(0,\, \sigma^2\right).$$

We use weights $\hat{\omega}_i = 1/N_1$ for $i = N_0 + 1$, .., N, and

$$\hat{\omega}^{\text{sc}}_{1:N_0} = \arg\min \left\{ \frac{1}{T_0} \sum_{t=1}^{T_0} \left(\frac{1}{N_1} \sum_{j=N_0+1}^{N} Y_{jt} - \sum_{i=1}^{N_0} \omega_i Y_{it} \right)^2 + \frac{\zeta}{N_1} \|\omega\|_2^2 : \omega_i \geq 0,\, \sum_{i=1}^{N_0} w_i = 1 \right\}, \quad (6.2)$$

and pick $\hat{\lambda}$ analogously. As above, we set ζ to the sample variance of the Y_{it} The, given these weights, we estimate

$$\hat{\tau} = \arg\min \left\{ \sum_{i,\,t} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i \hat{\lambda}_t \right\}. \quad (6.3)$$

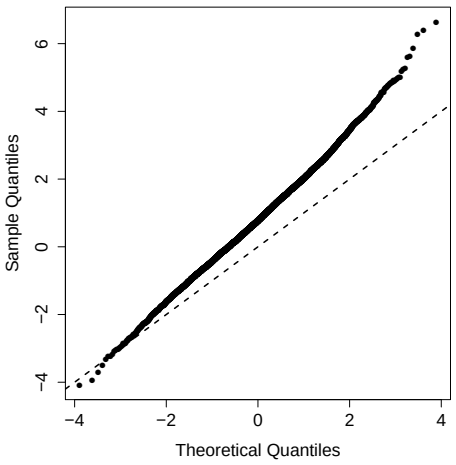
We perform inference via heteroskedasticity-consistent standard error as provided in the `R` package `sandwich` [Zeileis, 2004]. We estimate variance via the jackknife [Miller, 1974], which corresponds to `HC3` standard errors of MacKinnon and White [1985]. We run the weighted regression as though $\hat{\omega}_i$ and $\hat{\lambda}_t$ were deterministic and did not depend on the data.

				root-mean sq. error			mean absolute bias		
N	T	σ	rank	DID	SC	SDID	DID	SC	SDID
50	50	0.5	2	1.56	0.47	0.24	0.86	0.21	0.07
50	50	0.5	5	1.96	1.10	0.57	1.40	0.70	0.33
50	50	2	2	1.45	1.04	0.89	0.87	0.44	0.24
50	50	2	5	2.20	1.54	1.15	1.52	0.92	0.52
50	200	0.5	2	1.22	0.39	0.17	0.79	0.11	0.04
50	200	0.5	5	2.09	0.65	0.44	1.46	0.41	0.22
50	200	2	2	1.22	0.64	0.68	0.75	0.26	0.16
50	200	2	5	2.40	1.17	1.04	1.52	0.66	0.42
200	200	0.5	2	1.38	0.29	0.11	0.87	0.11	0.02
200	200	0.5	5	2.19	0.77	0.30	1.56	0.44	0.13
200	200	2	2	1.27	0.51	0.53	0.81	0.21	0.11
200	200	2	5	2.38	1.12	0.72	1.64	0.58	0.24

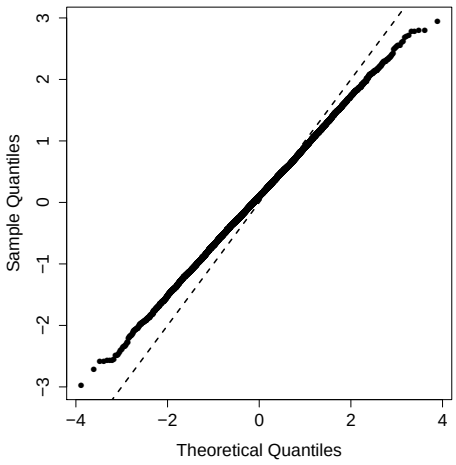
Table 1: Simulation study with an **exchangeable** distribution for $\boldsymbol{L}$ and penalized SC weights. Results are aggregated over 400 draws of the low-rank $\boldsymbol{L}$ matrix and 25 draws of $\boldsymbol{Y}$ for each $\boldsymbol{L}$ (for a total of 10,000 simulation replications).

				root-mean sq. error			mean absolute bias		
N	T	σ	rank	DID	SC	SDID	DID	SC	SDID
50	50	0.5	2	1.70	0.58	0.31	1.08	0.29	0.09
50	50	0.5	3	1.99	0.95	0.46	1.35	0.51	0.20
50	50	2	2	1.55	1.07	1.00	1.03	0.58	0.33
50	50	2	3	1.76	1.30	1.14	1.22	0.75	0.44
50	200	0.5	2	1.40	0.30	0.19	1.01	0.14	0.05
50	200	0.5	3	2.08	0.65	0.37	1.37	0.29	0.12
50	200	2	2	1.49	0.83	0.83	0.98	0.40	0.25
50	200	2	3	2.02	1.07	0.96	1.42	0.62	0.37
200	200	0.5	2	1.86	0.67	0.18	1.14	0.17	0.03
200	200	0.5	3	2.07	0.53	0.21	1.35	0.24	0.06
200	200	2	2	1.63	0.70	0.64	1.09	0.35	0.16
200	200	2	3	1.89	0.88	0.68	1.31	0.49	0.21

Table 2: Simulation study with an **non-exchangeable** distribution for $\boldsymbol{L}$ and penalized SC weights. Results are aggregated over 400 draws of the low-rank $\boldsymbol{L}$ matrix and 25 draws of $\boldsymbol{Y}$ for each $\boldsymbol{L}$ (for a total of 10,000 simulation replications).



standardized DID errors



standardized SDID errors

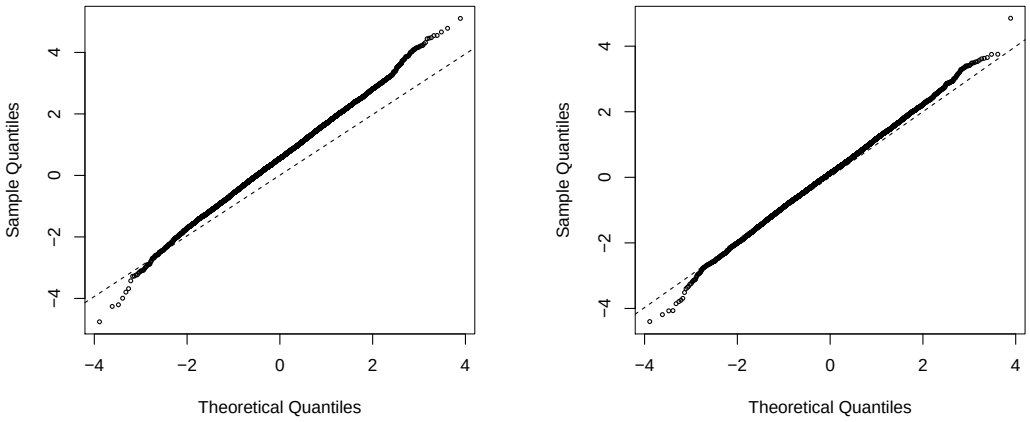
Figure 3: Standard Gaussian QQ-plot of the standardized errors $(\hat{\tau} - \tau)/\widehat{\mathrm{Var}}\,[\hat{\tau}]^{1/2}$, for the independent design for both DID and SDID, aggregated across 10,000 simulation replications. Points along the diagonal (dashed) would indicate perfectly calibrated Gaussian standard errors. Points along a centered line with a slope shallower than 45 degrees indicate that confidence intervals are conservative.

We generated data as in the non-exchangeable case above, with $N = 100$, $N_1 = 20$, $T = 120$, $T_1 = 5$, $\sigma = 2$, $\tau = 1$, and rank set to 2. It appears that SDID confidence intervals were well calibrated albeit slightly conservative: Nominal 95% confidence intervals achieved 98% coverage. The slight conservativeness may be due to the well-known mild upward bias of jackknife variance estimates [Efron and Stein, 1981]. In contrast, a basic difference-in-differences regression (6.3) but without weights $\hat{\omega}$ and $\hat{\lambda}$ did poorly: Nominal 95% confidence intervals achieved 82% coverage. Figure 3 shows a Gaussian QQ-plot of the standardized errors of both DID and SDID, mirroring the observation that SDID confidence intervals are well calibrated where DID ones are not.

To address a well-known critique of Bertrand et al. [2004] we also consider a design with dependent errors. The data is generated in the same way as above, but now the errors are correlated:

$$\mathbb{E}[\varepsilon_{it}\varepsilon_{il}] = \rho^{|t-l|}$$

We set $\rho = 0.7$ leaving all other parameters the same. To deal with the correlation in errors, we



standardized DID errors standardized SDID errors

Figure 4: Standard Gaussian QQ-plot of the standardized errors $(\hat{\tau} - \tau)/\widehat{\text{Var}}\,[\hat{\tau}]^{1/2}$, for the correlated design for both DID and SDID, aggregated across 10,000 simulation replications. Points along the diagonal (dashed) would indicate perfectly calibrated Gaussian standard errors.

estimate the variance using row-based jackknife. We run the weighted regression as though $\hat{\omega}_i$ and $\hat{\lambda}_t$ were deterministic and did not depend on the data.

Nominal 95% confidence intervals based on SDID now achieve 93% coverage, while those based on simple DID estimator achieve 88%. Figure 4 shows a Gaussian QQ-plot of the standardized errors of both DID and SDID, indicating that again SDID is better calibrated than DID.

7 Conclusion

We present a new estimator in a Synthetic Control setting, he synthetic difference in differences (SDID) estimator, which can be interpreted as a doubly weighted DID estimator. We find that the new estimator has attractive double robustness properties compared to the SC and DID estimators, both in simulations, in an application, and based on formal large N and large T symptotic results. By putting the new estimator as well as the original SC estimator in a weighted regression framework it allows us to connect the SC methodology to regression methods, which suggests alternative ways for accomodating time-invariant as well as time-varying covariates, as well as generalizations from two-way fixed effect models to factor models.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. American Economic Review, 93(-):113–132, 2003.

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. Econometrica, 74(1):235–267, 2006.

Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data, 2016.

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.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. American Journal of Political Science, pages 495–510, 2015.

Manuel Arellano. Panel data econometrics. Oxford university press, 2003.

Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research, 2018.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. arXiv preprint arXiv:1710.10251, 2017.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4):597–623, 2018.

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

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. Econometrica, 70(1):191–221, 2002.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. The Journal of Economic Perspectives, 28(2): 29–50, 2014.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. New perspectives on the synthetic control method. Technical report, UC Berkeley, 2018.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? The Quarterly journal of economics, 119(1):249–275, 2004.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. Econometrica, 83(3):1147–1184, 2015.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. Technical report, IFS Working Papers, 2017.

David Card. The impact of the mariel boatlift on the miami labor market. Industrial and Labor Relation, 43(2):245–257, 1990.

Carlos Carvalho, Ricardo Masini, and Marcelo C Medeiros. Arco: an artificial counterfactual approach for high-dimensional panel time-series data. Journal of Econometrics, 207(2):352–380, 2018.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. arXiv preprint arXiv:1712.09089, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and M Robins. Locally robust semiparametric estimation. arXiv preprint arXiv:1608.00033, 2018b.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. Inference on average treatment effects in aggregate panel data settings. arXiv preprint arXiv:1812.10820, 2018c.

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

Bradley Efron. The jackknife, the bootstrap, and other resampling plans, volume 38. Siam, 1982.

Bradley Efron and Charles Stein. The jackknife estimate of variance. The Annals of Statistics, pages 586–596, 1981.

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

Bryan S Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. The Review of Economic Studies, 79(3): 1053–1079, 2012.

Jinyong Hahn and Ruoyao Shi. Synthetic control and inference. Available at UCLA, 2016.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1):25–46, 2012.

Christian B Hansen. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. Journal of Econometrics, 141(2):597–620, 2007.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. arXiv preprint arXiv:1712.00038, 2018.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):243–263, 2014.

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Hans R Kunsch. The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, pages 1217–1241, 1989.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. arXiv preprint arXiv:1305.4825, 2013.

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

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. Biometrika, 73(1):13–22, 1986.

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In Geometric aspects of functional analysis, pages 277–299. Springer, 2017.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of econometrics, 29(3):305–325, 1985.

Shahar Mendelson. Learning without concentration. In Conference on Learning Theory, pages 25–39, 2014.

Rupert G Miller. The jackknife-a review. Biometrika, 61(1):1–15, 1974.

Whitney K Newey, Fushing Hsieh, and James M Robins. Twicing kernels and a small bias property of semiparametric estimators. Econometrica, 72(3):947–962, 2004.

Juan Peypouquet. Convex optimization in normed spaces: theory, methods and examples. Springer, 2015.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866, 1994.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448):1096–1120, 1999.

James H Stock and Mark W Watson. Heteroskedasticity-robust standard errors for fixed effects panel data regression. Econometrica, 76(1):155–174, 2008.

William F Trench. Asymptotic distribution of the spectra of a class of generalized kac–murdock–szegö matrices. Linear algebra and its applications, 294(1-3):181–192, 1999.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge University Press, 2018.

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

Achim Zeileis. Econometric computing with hc and hac covariance matrix estimators. Journal of Statistical Software, Articles, 11(10):1–17, 2004. doi: 10.18637/jss.v011.i10. URL `https://www.jstatsoft.org/v011/i10`.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511):910–922, 2015.

8 Appendix

Notation. Throughout the proofs section, we omit the ":" subscript from ω and λ when there is no risk of ambiguity. In addition, we write $\omega_\star$ and $\lambda_\star$ with the same meaning as $\omega^\star$ and $\lambda^\star$ where conventient. We will use c to denote a universal constant, which may differ in value in each instance. Many of our bounds are phrased in terms of the gaussian width of a set $S \subseteq \mathbb{R}^n$, $\mathrm{w}(S) = \mathrm{E}\sup_{s\in S}\langle g, s\rangle$ where $g \in \mathbb{R}^n$ is a vector of iid standard gaussians, as well as the radius $\mathrm{rad}(S) = \sup_{s\in S}\|t\|$ and diameter $\mathrm{diam}(S) = \sup_{s,s'\in\S}\|s - s'\|$. We will use results which, in our references, may be phrased in terms of variants of gaussian width, $\gamma(S) = \mathrm{E}\sup_{s\in S}|\langle g, s\rangle|$ and $h(S) = \sqrt{\mathrm{E}\sup_{s\in S}\langle g, s\rangle^2}$, and will express them here without comment in terms of either $\mathrm{w}(S)$ or $\mathrm{w}(S - S)$ for $S - S := \{s - s' : s, s' \in S\}$ using equivalences discussed in Vershynin [2018, Section 7.6]. Unless otherwise specified, $\|v\|$ will mean the euclidean norm $\|v\|_2$ for a vector v and $\|A\|$ will mean the operator norm $\|A\|_{op} = \sup_{\|v\|_2\leq 1}\|Ax\|_2$ for a matrix A. The L_2 norm, subgaussian norm, and subexponential norms for a scalar random variable Z will be written $\|Z\|_{L_2}$, $\|Z\|_{\psi_2}$ and $\|Z\|_{\psi_1}$, and we extend them to random vectors Z by defining $\|Z\|_{L_2} = \sup_{\|x\|=1}\|\angle Z, x\|_{L_2}$ and the others analogously as in Vershynin [2018].

8.1 Proof of Theorem 2

When the fixed effects model is correctly specified, we can check that synthetic difference in differences perfectly captures the signal for any set of weights, and the error depends only on the noise ε:

$$\widehat{L}_{NT} - L_{NT} = \hat{\omega} \cdot \varepsilon_{:T} + \hat{\lambda} \cdot \varepsilon_{N:} - \hat{\omega}'\varepsilon_{::}\hat{\lambda}.$$

Next, by Assumption 3, we know that

$$\hat{\omega} \cdot \varepsilon_{:T} \,\big|\, \hat{\omega} \sim \mathcal{N}\left(0,\, \sigma^2\|\hat{\omega}\|_2^2\right),$$

and so by the first part of (4.4) the term $\hat{\omega} \cdot \varepsilon_{:T}$ converges in probability to 0; the same argument also applies to $\hat{\lambda} \cdot \varepsilon_{n:}$. Finally, for the last term, we invoke Cauchy-Schwarz to check that

$$\hat{\omega}'\varepsilon_{::}\hat{\lambda} \leq \|\hat{\omega}\|_2\|\varepsilon_{::}\|_{op}\|\hat{\lambda}\|_2 = \mathcal{O}_P(\|\hat{\omega}\|_2\|\hat{\lambda}\|_2\sqrt{\max\{N, T\}}),$$

	DID	SC	SDID
Alabama	12.95	3.41	2.46
Arkansas	16.24	5.03	2.81
California	8.79	3.37	1.81
Colorado	7.18	4.66	3.81
Connecticut	6.25	2.79	2.40
Delaware	3.89	5.26	3.04
Georgia	12.68	3.61	2.42
Idaho	7.60	2.55	2.24
Illinois	2.40	3.07	3.08
Indiana	6.31	4.36	3.46
Iowa	4.45	4.77	5.12
Kansas	6.29	3.92	3.59
Kentucky	9.24	18.52	4.62
Louisiana	5.42	2.71	2.48
Maine	4.25	5.01	5.62
Minnesota	6.43	3.56	3.72
Mississippi	8.09	2.31	1.88
Missouri	5.98	2.14	1.82
Montana	6.98	4.31	3.41
Nebraska	2.84	1.31	1.40
Nevada	27.34	8.10	7.90
New Hampshire	42.52	48.37	8.72
New Mexico	1.75	2.38	2.65
North Carolina	30.35	9.96	5.10
North Dakota	6.98	5.37	4.15
Ohio	9.59	2.58	1.33
Oklahoma	8.11	4.88	4.27
Pennsylvania	8.55	2.47	2.32
Rhode Island	6.58	6.90	6.73
South Carolina	8.74	2.69	2.24
South Dakota	3.44	2.28	2.41
Tennessee	17.22	5.94	3.15
Texas	7.93	4.21	3.58
Utah	4.26	23.59	3.89
Vermont	6.49	3.85	4.05
Virginia	2.18	2.51	2.39
West Virginia	4.34	4.13	3.42
Wisconsin	5.57	3.36	3.30
Wyoming	12.27	8.15	6.87

Table 3: Root-mean squared error for one-step-ahead predictions made by difference in differences regression, SCs, and SDID. Results are averaged over the time period 1980-1988.

				root-mean sq. error			mean absolute bias		
n	T	σ	rank	DID	SC	SDID	DID	SC	SDID
50	50	0.5	2	1.56	0.32	0.33	0.86	0.09	0.05
50	50	0.5	5	1.96	0.82	0.47	1.40	0.38	0.14
50	50	2	2	1.45	1.15	1.20	0.87	0.39	0.25
50	50	2	5	2.20	1.51	1.42	1.52	0.73	0.43
50	200	0.5	2	1.22	0.38	0.27	0.79	0.07	0.04
50	200	0.5	5	2.09	0.50	0.42	1.46	0.19	0.09
50	200	2	2	1.22	0.84	0.98	0.75	0.25	0.18
50	200	2	5	2.40	1.22	1.24	1.52	0.54	0.35
200	200	0.5	2	1.38	0.27	0.21	0.87	0.06	0.04
200	200	0.5	5	2.19	0.58	0.30	1.56	0.19	0.05
200	200	2	2	1.27	0.63	0.79	0.81	0.18	0.14
200	200	2	5	2.38	1.10	0.96	1.64	0.45	0.21

Table 4: Simulation study with an **exchangeable** distribution for $\boldsymbol{L}$ and **unpenalized** SC weights. Results are aggregated over 400 draws of the low-rank $\boldsymbol{L}$ matrix and 25 draws of Y for each $\boldsymbol{L}$ (for a total of 10,000 simulation replications).

				root-mean sq. error			mean absolute bias		
n	T	σ	rank	DID	SC	SDID	DID	SC	SDID
50	50	0.5	2	1.70	0.47	0.36	1.08	0.16	0.07
50	50	0.5	3	1.99	0.80	0.46	1.35	0.32	0.12
50	50	2	2	1.55	1.15	1.26	1.03	0.51	0.32
50	50	2	3	1.76	1.36	1.38	1.22	0.65	0.41
50	200	0.5	2	1.40	0.30	0.28	1.01	0.09	0.05
50	200	0.5	3	2.08	0.57	0.40	1.37	0.17	0.08
50	200	2	2	1.49	0.97	1.06	0.98	0.37	0.25
50	200	2	3	2.02	1.16	1.17	1.42	0.56	0.34
200	200	0.5	2	1.86	0.61	0.24	1.14	0.12	0.04
200	200	0.5	3	2.07	0.42	0.27	1.35	0.13	0.05
200	200	2	2	1.63	0.75	0.84	1.09	0.30	0.17
200	200	2	3	1.89	0.89	0.88	1.31	0.41	0.20

Table 5: Simulation study with an **non-exchangeable** distribution for $\boldsymbol{L}$ and **unpenalized** SC weights. Results are aggregated over 400 draws of the low-rank $\boldsymbol{L}$ matrix and 25 draws of Y for each $\boldsymbol{L}$ (for a total of 10,000 simulation replications).

recalling that, under Assumption 1, it is known that $\mathbb{E}\left[\|\varepsilon_{::}\|_{op}^2\right] = \mathcal{O}(\max\{N, T\})$.

8.2 Proof of Theorem 1

We start with the following high-level lemma.

Lemma 7. *Suppose that Assumption 1 is satisfied, further assume that the following conditions hold:*

$$\begin{aligned}
&\|\omega_\star\|_2 = o(1)\\
&\|\lambda_\star\|_2 = o(1)\\
&\|\hat{\lambda} - \lambda_\star\|_2 = o_p(\|\lambda_\star\|_2)\\
&\|\hat{\omega} - \omega_\star\|_2 = o_p(\|\omega_\star\|_2)\\
&\|\hat{\omega} - \omega_\star\|_1 = o_p(1/\sqrt{\log(T)})\\
&\|\hat{\lambda} - \lambda_\star\|_1 = o_p(1/\sqrt{\log(N)})
\end{aligned} \tag{8.1}$$

Then we have the following result:

$$\hat{\omega}'\varepsilon_{::}\hat{\lambda} = o_p(\max\{\|\omega_\star\|_2, \|\lambda_\star\|_2\}) \tag{8.2}$$

Proof. We decompose $\hat{\omega}^T\Sigma_{::}\hat{\lambda}$ into a sum of four terms $\xi_1 + \xi_2 + \xi_3 + \xi_4$ and bound each term:

$$\hat{\omega}^T\varepsilon_{::}\hat{\lambda} = \omega_\star'\varepsilon_{::}\lambda_\star + \omega_\star'\varepsilon_{::}(\hat{\lambda} - \lambda_\star) + (\hat{\omega} - \omega_\star)'\varepsilon_{::}\lambda + (\hat{\omega} - \omega_\star)'\varepsilon_{::}(\hat{\lambda} - \lambda_\star) \tag{8.3}$$

Our goal is to show that these terms are negligible:

$$\xi_k = o_p(\max\{\|\omega_\star\|_2, \|\lambda_\star\|_2\}) \tag{8.4}$$

For the first term we get the following:

$$\xi_1 \sim \mathcal{N}(0, \|\omega_\star\|_2^2\|\lambda_\star\|_2^2) \Rightarrow \xi_1 = O_p(\|\omega_\star\|_2\|\lambda_\star\|_2) = o_p\left(\max\{\|\omega_\star\|_2, \|\lambda_\star\|_2\}\right) \tag{8.5}$$

The second term ξ_2 is bounded, via Hölder's inequality, by $\|\omega_\star'\varepsilon_{::}\|_\infty\|\hat{\lambda} - \lambda_\star\|_1$. The first factor

is the maximum of $T-1$ independent mean-zero gaussians with variance $\sigma^2\|\omega\|_2^2$, which is $O_p(\|\omega_\star\|_2\sqrt{\log(T)})$, and the second is $o_p(1/\sqrt{\log(T)})$ by assumption, so the product is $o_p(\|\omega_\star\|_2)$. The third term ξ_3 is analogously $o_p(\|\lambda_\star\|_2)$.

To bound the fourth term ξ_4, we use Chevet's inequality [Vershynin, 2018, Theorem 8.7.1],

$$\mathrm{E} \sup_{x\in X, y\in Y} x'\varepsilon y \le \mathrm{rad}(X)\,\mathrm{w}(Y) + \mathrm{w}(X)\,\mathrm{rad}(Y).$$

In essence, this is a uniform version of the same Hölder's inequality bound, allowing us to bound the simultaneous supremum over X and Y as if either $x \in X$ were a constant vector of length $\mathrm{rad}(X)$ or $y \in Y$ were a constant vector of length $\mathrm{rad}(Y)$. Here we can take X to be a set of the form $\{x : \|x\|_1 \le a/\sqrt{\log(T)}, \|x\|_2 \le b\|\omega_\star\|_2\}$, for $a \to 0$, which will contain $\hat{\omega} - \omega_\star$ with high probability under our assumptions, and define Y analogously in terms of $\hat{\lambda}$ and $\lambda_\star$. Then $\mathrm{w}(X) \lesssim a/\sqrt{\log(T)} \cdot \sqrt{\log(T)} \to 0$ and $\mathrm{rad}(X) \lesssim \|\omega_\star\|$ and analogously $\mathrm{w}(Y) \to 0$ and $\mathrm{rad}(Y) \lesssim \|\lambda_\star\|$, which shows that $\xi_4 = o_p(\max\{\|\omega_\star\|_2, \|\lambda_\star\|_2\})$. □

We now move to prove the claimed result. Define deterministic weights:

$$\omega_i^\star = \frac{1\left(\{(\alpha_i - \alpha_N)^2 \le \tilde{c}_\omega\}\right)}{\sum_{i\neq N} 1\left(\{(\alpha_i - \alpha_N)^2 \le \tilde{c}_\omega\}\right)}, \quad \lambda_t^\star = \frac{1\left(\{(\beta_t - \beta_T)^2 \le \tilde{c}_\lambda\}\right)}{\sum_{t\neq T} 1\left(\{(\beta_t - \beta_T)^2 \le \tilde{c}_\lambda\}\right)}, \tag{8.6}$$

where $\tilde{c}_\omega = c_\omega - \sigma^2$ and $\tilde{c}_\lambda = c_\lambda - \sigma^2$.

First we verify that conditions for Lemma 7 hold for $\hat{\omega}$ and $\omega^\star$. Results for $\hat{\lambda}$ and $\lambda^\star$ follow in the same way. Define the following random variables:

$$\begin{aligned} K &= \sum_{i\neq N} 1\left(\left\{(\alpha_i - \alpha_N)^2 \le \tilde{c}_\omega\right\}\right) \\ \hat{K} &= \sum_{i\neq N} 1\left(\left\{\frac{1}{T-1}\|Y_{i:} - Y_{N:}\|_2^2 \le c_\omega\right\}\right) \end{aligned} \tag{8.7}$$

By definition we have the following:

$$\begin{aligned} \omega_j^* &= \frac{1\{\delta_j^2 \le \tilde{c}_\omega)\}}{K} \\ \hat{\omega}_j &= \frac{1\{\hat{\delta}_j^2 \le c_\omega\}}{\hat{K}} \end{aligned} \tag{8.8}$$

where $\hat{\delta}_j^2 = \frac{1}{T-1}\|Y_{i:} - Y_{N:}\|_2^2 = \delta_j^2 + \sigma^2 + \xi_j$, where ξ_j is a mean-zero random variable. Define the following random variable:

$$l = \|\omega^\star - \hat{\omega}\|_0 \tag{8.9}$$

By definition l is the sum of $n-1$ i.i.d. binary terms thus:

$$\begin{aligned} &l = \mathcal{O}_p\left(\mu_N\right) \\ &\mu_N := (N-1)\mathbb{E}\left[1\{1\{\delta_j^2 \leq \tilde{c}_\omega)\} \neq 1\{\hat{\delta}_j^2 \leq c_\omega\}\}\right] \end{aligned} \tag{8.10}$$

Since $\hat{\delta}_j^2 = \delta_j^2 + \sigma^2 + \xi_j$, where $\xi_j = \mathcal{O}_p\left(\frac{1}{\sqrt{T}}\right)$ we get the following:

$$\mu_N = O\left((N-1)\frac{f_{\delta^2}(\tilde{c}_\omega)}{\sqrt{T}}\right) \tag{8.11}$$

Since $f_{\delta^2}(x) = \frac{f_\delta(\sqrt{x})}{\sqrt{x}}$, and using the fact that $\tilde{c}_\omega = o(1)$ we get:

$$l = \mathcal{O}_p\left(\frac{N-1}{\sqrt{\tilde{c}_\omega T}}\right) \tag{8.12}$$

By construction we have the following:

$$K = \mathcal{O}_p\left((N-1)F_{\delta_j^2}(\tilde{c}_\omega)\right) \tag{8.13}$$

and since $\tilde{c}_\omega = o(1)$ and $\tilde{c}_\omega = a_{N,T}\frac{\log(N)}{\sqrt{T}}$ we have $K = \mathcal{O}_p((N-1)\sqrt{\tilde{c}_\omega}) \to \infty$. This implies the following:

$$\|\omega^\star\|_2 = \frac{1}{\sqrt{K}} = o_p(1) \tag{8.14}$$

We have the following relationship:

$$\begin{aligned} &|K - \hat{K}| \leq l \\ &\frac{l}{K} = \mathcal{O}_p\left(\frac{1}{\sqrt{T}\tilde{c}_\omega}\right) = o_p(1) \end{aligned} \tag{8.15}$$

that implies

$$\frac{\hat{K}}{K} = \mathcal{O}_p\left(1 + \frac{l}{K}\right) = \mathcal{O}_p\left(1 + o_p(1)\right) = \mathcal{O}_p(1) \tag{8.16}$$

As a result, we get that $\|\hat{\omega}\|_2 = \mathcal{O}_p(\|\omega^\star\|_2)$. Define the following weights (different normalization):

$$\tilde{\omega}_j = \frac{\{\hat{\delta}_j^2 \leq c_\omega\}}{K} \tag{8.17}$$

We have bounds on the squared norms:

$$\begin{aligned} \|\tilde{\omega} - \omega^\star\|_2^2 &= \frac{1}{K}\left(\frac{l}{K}\right) = o_p(\|\omega^\star\|_2^2) \\ \|\tilde{\omega} - \hat{\omega}\|_2^2 &= \hat{K}\left(\frac{1}{K} - \frac{1}{\hat{K}}\right)^2 \leq \frac{1}{\hat{K}}\left(\frac{l}{K}\right)^2 = o_p(\|\tilde{\omega} - \omega^\star\|_2^2) \end{aligned} \tag{8.18}$$

Finally, we have the following:

$$\|\hat{\omega} - \omega^\star\|_2 \sqrt{\|\hat{\omega} - \omega^\star\|_0 \log(N)} = \mathcal{O}_p\left(\frac{l}{K}\log(N)\right) = \mathcal{O}_p\left(\frac{\log(N)}{\tilde{c}_\omega\sqrt{T}}\right) = o_p(1) \tag{8.19}$$

As $\|\cdot\|_1 \leq \|\cdot\|_2\sqrt{\|\cdot\|_0}$, this implies that the conditions of Lemma 1 are satisfied. Thus, our estimator has the following decomposition:

$$\begin{aligned} \widehat{L}_{NT} - L_{NT} &= \hat{\omega} \cdot \varepsilon_{:T} + \hat{\lambda} \cdot \varepsilon_{N:} - \hat{\omega}' \varepsilon_{::} \hat{\lambda} \\ &= \omega^\star \cdot \varepsilon_{:T} + \lambda^\star \cdot \varepsilon_{N:} + (\hat{\omega} - \omega^\star) \cdot \varepsilon_{:T} + (\hat{\lambda} - \lambda^\star) \cdot \varepsilon_{N:} + o_p(\max\{\|\omega^\star\|_2, \|\lambda^\star\|_2\}) \\ &= \omega^\star \cdot \varepsilon_{:T} + \lambda \cdot \varepsilon_{N:} + o_p(\max\{\|\omega^\star\|_2, \|\lambda^\star\|_2\}) \end{aligned}$$

where the last equality uses the fact that $\|\omega^\star - \hat{\omega}\|_2 = o_p(\|\omega^\star\|_2)$, $\|\lambda^\star - \hat{\lambda}\|_2 = o_p(\|\lambda^\star\|_2)$ and the fact that $\hat{\omega}$ is independent of $\varepsilon_{:T}$ and $\hat{\lambda}$ is independent of $\varepsilon_{N:}$. This proves the result.

8.3 Generalizations of Theorem 3 and Lemma 4

In this section, we will replace Assumption 1 with the following generalization, which allows for heteroskedastic and autocorrelated errors. In this setting, we consider the behavior of our synthetic difference-in-difference estimator when we use least squares weights $\hat{\omega}, \hat{\lambda}$ subject to arbitrary constraints.

Assumption 5. *$Y = L + \varepsilon$ is an $N \times T$ matrix where L is deterministic and $\mathrm{E}\,\varepsilon = 0$; the rows of ε are independent and subgaussian; and $\mathrm{E}\,\varepsilon_{i:}'\varepsilon_{i:} = \Sigma$ for all $i \leq N-1$. Here subscripting by $:$ takes the rows or columns for $i < N, j < T$.*

Theorem 8. *Under Assumption 5, consider the least squares estimators*

$$\hat{\omega} = \underset{\omega \in \Omega}{\arg\min} \|\omega' Y_{::} - Y_{N:}\|_2^2 \quad \textit{and} \quad \hat{\lambda} = \underset{\lambda \in \Lambda}{\arg\min} \|Y_{::}\lambda - Y_{:T}\|_2^2$$

and the oracle estimators $\omega_\star, \lambda_\star$ defined analogously with L substituted for Y and define

$$\begin{aligned}
\delta_\omega &= \|L_{N:} - \omega_\star' L_{::}\|; \\
\delta_\lambda &= \|L_{:T} - L_{::}\lambda_\star\|; \\
\delta_{sdid} &= \left|L_{NT} - (\omega_\star' \boldsymbol{L}_{:T} + \boldsymbol{L}_{N:}\lambda_\star - \omega_\star' L_{::}\lambda_\star)\right|.
\end{aligned}$$

Then, for r_λ defined in Lemma 9,

$$\begin{aligned}
\left|\widehat{L}_{NT} - L_{NT}\right| \leq\, & \operatorname{diam}(\Lambda)\left[\delta_\omega + \mathcal{O}_p\left(\|\Sigma\|^{1/2} + \mathrm{w}(\Omega)\max_{i<N}\|\varepsilon_{i:}\|_{\psi_2}\right)\right] \\
& + \operatorname{diam}(\Omega)\left[\delta_\lambda + \mathcal{O}_p\left(\max_{i<N}\operatorname{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}]^{1/2} + \mathrm{w}(\Lambda)\max_{i<N}\|\varepsilon_{i:}\|_{\psi_2}\right)\right] \\
& + \delta_{sdid} + \mathcal{O}_p\left(\operatorname{diam}(\Omega) r_\lambda + \|\omega_\star'\varepsilon_{::}\|_{\psi_2}\,\mathrm{w}(\Lambda) + \left[\|\varepsilon_{::}\lambda_\star\|_{\psi_2} + \|\mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}]\|_{\psi_2}\right]\mathrm{w}(\Omega)\right) \\
& + \left|\varepsilon_{N:}'\lambda_\star + \omega_\star'\varepsilon_{:T} - \omega_\star'\varepsilon_{::}\lambda_\star\right|
\end{aligned}$$

If the elements of $\varepsilon_{::}$ are independent and identically distributed, we may substitute $\min\{\operatorname{diam}(\Omega) r_\lambda, \operatorname{diam}(\Lambda) r_\omega\}$ for $\operatorname{diam}(\Omega) r_\lambda$, where r_ω is defined analogously to r_λ, as the bound established by Lemma 9 on $\|(\hat{\omega} - \omega_\star)' L_{::}\|$.

Lemma 9. *Under Assumption 5, for any subset Λ of $\mathbb{R}^{T-1}$, the least squares estimator and oracle least squares estimator*

$$\hat{\lambda} = \min_{\lambda \in \Lambda} \|Y_{::}\lambda - Y_{:T}\|_2^2 \quad \textit{and} \quad \lambda^\star = \min_{\lambda \in \Lambda} \|L_{::}\lambda - L_{:T}\|_2^2$$

satisfy the bound $\|L_{::}(\hat{\lambda} - \lambda^\star)\|_2 = \mathcal{O}_P(r_\lambda)$ where

$$\begin{aligned} r_\lambda = \max \Bigg(& \|L_{::}\lambda^\star - L_{:T}\|, \\ & x\sqrt{\text{approx-rank}(L_{::}, x)} \quad \textit{for} \quad x = \|\varepsilon_{::}\lambda^\star\|_{\psi_2} + \|\varepsilon_{:T}\|_{\psi_2}, \\ & \sqrt{\sup_{\delta \in \Lambda} |\bar{\gamma}'\delta|}, \\ & \sqrt{\operatorname{diam}(\Lambda) \max\left(\sqrt{T}, \sqrt[4]{NT}\right) \log(T) \max_j \|\varepsilon_{iT}\|_{\psi_2} \|\varepsilon_{ij}\|_{\psi_2}}, \\ & \sqrt{N\|\Sigma\|} \operatorname{diam}(\Lambda) + \|\varepsilon_{i:}\|_{\psi_2}^2 \|\Sigma^{-1}\| \sqrt{\|\Sigma\|}\, \mathrm{w}(\Lambda) \Bigg). \end{aligned}$$

Here $\bar{\gamma} = (N-1)^{-1} \sum_{i=1}^{N-1} \mathrm{E}\, \varepsilon_{iT}\varepsilon_{i:}$; $\mathrm{w}(\Lambda)$ is the gaussian width of the set Λ; and $\text{approx-rank}(L_{::}, x)$ *is an approximation to the rank of $L_{::}$ that ignores small nonzero singular values, defined*

$$\text{approx-rank}(L_{::}, x) := \min \left\{ r \in 1, 2, \dots \;\middle|\; r \geq x^{-1} \min \left(\operatorname{diam}(\Lambda) \sqrt{\sum_{k>r} s_k^2},\ s_{r+1}\, \mathrm{w}(\Lambda) \right) \right\}$$

in terms of the decreasing sequence of singular values $s_1, s_2, \dots$ of $L_{::}$.

8.3.1 Proof of Theorem 8

Our estimator's error is the difference between our estimator and the corresponding infeasible estimator, plus the infeasible estimator's error δ_{sdid}, i.e.

$$
\begin{aligned}
&\widehat{L}_{NT} - L_{NT} \\
&= \left[Y_{N:}\hat{\lambda} + \hat{\omega}' Y_{:T} - \hat{\omega}' Y_{::}\hat{\lambda}\right] - \left[(Y_{N:} - \varepsilon_{N:})\lambda_\star + \omega_\star'(Y_{:T} - \varepsilon_{:T}) - \omega_\star'(Y_{::} - \varepsilon_{::})\lambda_\star\right] + \delta_{sdid} \\
&= Y_{N:}(\hat{\lambda} - \lambda_\star) + (\hat{\omega} - \omega_\star)' Y_{:T} - \left[(\hat{\omega} - \omega_\star)' Y_{::}(\hat{\lambda} - \lambda_\star) + \omega_\star' Y_{::}(\hat{\lambda} - \lambda_\star) + (\hat{\omega} - \omega_\star)' Y_{::}\lambda_\star\right] \\
&\qquad + \delta_{sdid} - \varepsilon_{N:}'\lambda_\star - \omega_\star'\varepsilon_{:T} + \omega_\star'\varepsilon_{::}\lambda_\star \\
&= (Y_{N:} - \omega_\star' Y_{::})(\hat{\lambda} - \lambda_\star) + (\hat{\omega} - \omega_\star)'(Y_{:T} - Y_{::}\lambda_\star) - (\hat{\omega} - \omega_\star)' Y_{::}(\hat{\lambda} - \lambda_\star) \\
&\qquad + \delta_{sdid} - \varepsilon_{N:}'\lambda_\star - \omega_\star'\varepsilon_{:T} + \omega_\star'\varepsilon_{::}\lambda_\star \\
&= (L_{:N} - \omega_\star' L_{::})(\hat{\lambda} - \lambda_\star) + (\hat{\omega} - \omega_\star)'(L_{:T} - L_{::}\lambda_\star) + \delta_{sdid} \\
&\qquad + (\varepsilon_{N:} - \omega_\star'\varepsilon_{::})(\hat{\lambda} - \lambda_\star) + (\hat{\omega} - \omega_\star)'(\varepsilon_{:T} - \varepsilon_{::}\lambda_\star) - (\hat{\omega} - \omega_\star)'\varepsilon_{::}(\hat{\lambda} - \lambda_\star) \\
&\qquad - \varepsilon_{N:}'\lambda_\star - \omega_\star'\varepsilon_{:T} + \omega_\star'\varepsilon_{::}\lambda_\star \\
&\qquad - (\hat{\omega} - \omega_\star)' L_{::}(\hat{\lambda} - \lambda_\star).
\end{aligned}
$$

We bound each line.

1. The first line is bounded by $\delta_\omega \operatorname{diam}(\Lambda) + \delta_\lambda \operatorname{diam}(\Omega) + \delta_{sdid}$, which follows by applying Cauchy-Schwarz to the first two terms.

2. The second line is

$$
\begin{aligned}
\mathcal{O}_p\Bigg(& \operatorname{diam}(\Lambda)\left[\|\Sigma\|^{1/2} + \mathrm{w}(\Omega)\max_{i<N}\|\varepsilon_{i:}\|_{\psi_2}\right] \\
& + \operatorname{diam}(\Omega)\left[\max_{i<N}\operatorname{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}]^{1/2} + \mathrm{w}(\Lambda)\max_{i<N}\|\varepsilon_{i:}\|_{\psi_2}\right] \\
& + \|\omega_\star'\varepsilon_{::}\|_{\psi_2}\,\mathrm{w}(\Lambda) + \left[\|\varepsilon_{::}\lambda_\star\|_{\psi_2} + \|\mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}]\|_{\psi_2}\right]\mathrm{w}(\Omega)
\end{aligned}
$$

   (a) The first term is the sum of two pieces, $\varepsilon_{N:}'(\hat{\lambda} - \lambda_\star)$ and $-\omega_\star'\varepsilon_{::}(\hat{\lambda} - \lambda_\star)$. The first piece is $\mathcal{O}_p(\|\Sigma\|^{1/2}\operatorname{diam}(\Lambda))$. Because the row $\varepsilon_{N:}$ is independent of the noise submatrices $\varepsilon_{::}, \varepsilon_{:T}$ that are used to define $\hat{\lambda}$, it has mean zero and variance bounded by

$\|\Sigma\| \operatorname{diam}(\Lambda)^2$ conditional on $\hat{\lambda}$. The second piece is $\mathcal{O}_p(\|\omega_\star' \varepsilon_{::}\|_{\psi_2} \operatorname{w}(\Lambda))$, as the vector $\omega_\star' \varepsilon_{::}$ is subgaussian, and by Talagrand's comparison inequality [Vershynin, 2018, Corollary 8.6.3],

$$\begin{aligned}
&\mathrm{E} \sup_{\delta \in \Lambda - \lambda_\star} |\omega_\star' \varepsilon_{::} \delta| \leq cK \operatorname{w}(\Lambda - \lambda_\star); \\
&K' = \sup_{x,y \in \Lambda - \lambda_\star} \|\omega_\star' \varepsilon_{::}(x-y)\|_{\psi_2} / \|x - y\| \leq \|\omega_\star' \varepsilon_{::}\|_{\psi_2}.
\end{aligned}$$

(b) The second term is the sum of two pieces, $(\hat{\omega} - \omega_\star)' \varepsilon_{:T}$ and $(\hat{\omega} - \omega_\star)' \varepsilon_{::} \lambda_\star$. The second piece, by Talagrand's comparison inequality as above, is $\mathcal{O}_p(\|\varepsilon_{::} \lambda_\star\|_{\psi_2} \operatorname{w}(\Omega))$. The first piece is

$$\mathcal{O}_p\left(\operatorname{diam}(\Omega) \max_{i<N} \operatorname{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}]^{1/2} + \operatorname{w}(\Omega) \|\mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}]\|_{\psi_2} \right)$$

with terms bounding those in the decomposition

$$(\hat{\omega} - \omega_\star)' \varepsilon_{:T} = (\hat{\omega} - \omega_\star)' (\varepsilon_{:T} - \mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}]) + (\hat{\omega} - \omega_\star) \mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}].$$

The bound on the second of these terms follows from Talagrand's comparison inequality as above, and the first of them has a conditional Chebyshev bound

$$\begin{aligned}
&\mathrm{P}\left(|(\hat{\omega} - \omega_\star)' (\varepsilon_{:T} - \mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}])| > t \mid \varepsilon_{::}, \varepsilon_{N:} \right) \\
&\leq t^{-2} \sum_{i=1}^{N-1} (\hat{\omega} - \omega_\star)_i^2 \operatorname{Var}[\varepsilon_{iT} \mid \varepsilon_{::}] \leq t^{-2} \operatorname{diam}(\Omega)^2 \max_{i<N} \operatorname{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}].
\end{aligned}$$

(c) The third term is $\mathcal{O}_p\left(\max_{i<N} \|\varepsilon_{i:}\|_{\psi_2} \left[\operatorname{diam}(\Omega) \operatorname{w}(\Lambda) + \operatorname{diam}(\Lambda) \operatorname{w}(\Omega)\right]\right)$. This follows from Chevet's inequality for random matrices with iid subgaussian rows,

$$\mathrm{E} \sup_{x \in X, y \in Y} x' \varepsilon y \leq c \max_{i<N} \|\varepsilon_{i:}\|_{\psi_2} \left[\operatorname{rad}(X) \operatorname{w}(Y) + \operatorname{w}(X) \operatorname{rad}(Y)\right].$$

The proof of Vershynin [2018, Theorem 8.7.1], which addresses the case of random matrices with iid subgaussian elements, can be adapted for iid subgaussian rows by applying Hoeffding's inequality (i.e. Vershynin [2018, Proposition 2.6.1]) to row sums

rather than elementwise when bounding the increments of this subgaussian process.

3. The third line is included in the bound.

4. The fourth line is $\mathcal{O}_p(\text{diam}(\Omega)r_\lambda)$. This follows from the Cauchy-Schwarz bound $\|\hat{\omega} - \omega_\star\|\|L_{::}(\hat{\lambda} - \lambda_\star)\|$ and Lemma 9. If the elements of $\varepsilon_{::}$ are iid, then Lemma 9 implies a bound $\|(\hat{\omega} - \omega_\star)'L_{::}\| \leq r_\omega$, and we can also bound this term by $\|(\hat{\omega} - \omega_\star)'L_{::}\|\|\hat{\lambda} - \lambda_\star\|$, so the fourth line will be $\mathcal{O}_p\left(\min\{\text{diam}(\Omega)r_\lambda, \text{diam}(\Lambda)r_\omega\}\right)$.

Collecting all of these results yields our claimed bound.

8.3.2 Proof of Lemma 9

To simplify our notation in this proof, we will write N and T for the dimensions of $Y_{::}$, which are called $N-1$ and $T-1$ in the lemma statement.

Our proof is based on the well-known isomorphic bounds argument in empirical risk minimization [see e.g. Lecué and Mendelson, 2013, Mendelson, 2014].

$$
\begin{aligned}
0 &\geq \|Y_{::}\hat{\lambda} - Y_{:T}\|_2^2 - \|Y_{::}\lambda_\star - Y_{:T}\|_2^2 \\
&= \|Y_{::}\hat{\lambda}\|_2^2 - \|Y_{::}\lambda_\star\|_2^2 - 2Y_{:T}'Y_{::}(\hat{\lambda} - \lambda_\star) \\
&\geq \|Y_{::}(\hat{\lambda} - \lambda_\star)\|_2^2 + 2(Y_{::}\lambda_\star - Y_{:T})'Y_{::}(\hat{\lambda} - \lambda_\star) + 1_{\{\Lambda = \text{conv}(\Lambda)\}} \cdot 2(L_{:T} - L_{::}\lambda_\star)'L_{::}(\hat{\lambda} - \lambda_\star) \\
&\geq \underbrace{\|Y_{::}(\hat{\lambda} - \lambda_\star)\|_2^2}_{Q\left(\hat{\lambda} - \lambda_\star\right)} + 2\underbrace{(Y_{::}\lambda_\star - Y_{:T})'Y_{::}(\hat{\lambda} - \lambda_\star) - 1_{\{\Lambda = \text{conv}(\Lambda)\}}(L_{::}\lambda_\star - L_{:T})'L_{::}(\hat{\lambda} - \lambda_\star)}_{M(\hat{\lambda} - \lambda_\star)}.
\end{aligned}
\tag{8.20}
$$

Here the addition of the term $1_{\{\Lambda = \text{conv}(\Lambda)\}} \cdot 2(L_{:T} - L_{::}\lambda_\star)'L_{::}(\hat{\lambda} - \lambda_\star)$ in the third line is justified by its negativity. This is a consequence of the convexity of the set $L_{::}\Lambda$ when the term is nonzero:

$L_{:T} - L_{::}\lambda_\star$ points 'outward' from the projection of $L_{:T}$ onto $L_{::}\Lambda$ to $L_{:T}$ itself, whereas $L_{::}(\hat{\lambda} - \lambda_\star)$ points 'inward' toward another element of $L_{::}\Lambda$ [see e.g. Peypouquet, 2015, Proposition 1.37].

Let $\Lambda - \lambda_\star$ be the set of possible deviations from $\lambda_\star$; $\hat{\delta} = \hat{\lambda} - \lambda_\star$ be the realized deviation; and $r^2(\delta) = \|L_{::}\delta\|^2$. We will show that, with probability tending to one, (8.20) cannot be satisfied if $r(\hat{\delta}) \geq r_\star$. This implies our claimed bounds.

To do this, we will show that with probability tending to one, we have a uniform quadratic lower bound on $Q(\delta)$ and a corresponding upper bound on $\left|M(\hat{\delta})\right|$,

$$\inf_{\substack{\delta\in\Lambda-\lambda_\star\\ r(\delta)\geq r_\star}} \frac{Q(\delta)}{r^2(\delta)} \geq 1-\eta, \tag{8.21}$$

$$\sup_{\substack{\delta\in\Lambda-\lambda_\star\\ r(\delta)\geq r_\star}} \frac{|M(\delta)|}{r^2(\delta)} < (1-\eta)/2. \tag{8.22}$$

These bounds are implied by simpler bounds of the form

$$\inf_{\delta\in\Lambda_\star} Q(\delta) \geq (1-\eta) r_\star^2, \tag{8.23}$$

$$\sup_{\delta\in\Lambda_\star} |M(\delta)| < (1-\eta)/2 \cdot r_\star^2. \tag{8.24}$$

where $\Lambda_\star = \{\delta \in [0,1](\Lambda - \lambda_\star) : r(\delta) = r_\star\}$. To see this, consider $\delta \in \Lambda - \lambda_\star$ with $r(\delta) \geq r$, and observe that because Q is quadratic and L is linear and $\delta r_\star / r(\delta) \in \Lambda_\star$, (8.23) and (8.24) imply that

$$\begin{aligned}
Q(\delta) &= (r(\delta)/r_\star)^2 Q(\delta r_\star / r(\delta)) \geq (r(\delta)/r_\star)^2 \cdot (1-\eta) r_\star^2 = (1-\eta) r(\delta)^2; \\
|M(\delta)| &= (r(\delta)/r_\star)\, |M(\delta r_\star / r(\delta))| \leq (r(\delta)/r_\star) \cdot (1-\eta/2) r_\star^2 \leq (1-\eta/2) r(\delta)^2.
\end{aligned}$$

We now move on to the core of our proof, which involves proving these bounds (8.23) and (8.24).

The lower bound (8.21)**.** Much of our proof will rely on $\varepsilon_{::}\delta$ being small relative to $L_{::}\delta$. For $\varepsilon_{::}$ having independent rows with the same correlation structure, i.e. $\mathrm{E}\, \varepsilon_{i:}'\varepsilon_{i:} = \Sigma$ with $\|\varepsilon_{i:}\|_{\psi_2} \leq K$, we can establish this using the matrix deviation inequality of Liaw et al. [2017],

$$\mathrm{E} \sup_{\delta\in\Lambda_\star} \left| \|\varepsilon_{::}\delta\| - \sqrt{N}\|\sqrt{\Sigma}\delta\| \right| \leq c\|\Sigma^{-1}\| \sqrt{\|\Sigma\|} K^2\, \mathrm{w}(\Lambda_\star). \tag{8.25}$$

For simplicity, we use Markov's inequality rather than subgaussian concentration to derive a tail bound from this, letting it hold on an event $\mathcal{A}$.

By the triangle inequality, this controls the degree to which $r_\star = \|L_{::}\delta\|$ can exceed $\|Y_{::}\delta\|$,

as

$$\|L_{::}\delta\| - \|Y_{::}\delta\| \leq \|L_{::}\delta - Y_{::}\delta\| = \|\varepsilon_{::}\delta\|.$$

In particular, every $\delta \in \Lambda_\star$ satisfying

$$\sqrt{N}\|\sqrt{\Sigma}\delta\| + c(1-\mathrm{P}(\mathcal{A}))^{-1}K^2\|\Sigma^{-1}\|\sqrt{\|\Sigma\|}\,\mathrm{w}(\Lambda_\star) \leq \eta/2 \cdot r_\star$$

, will also satisfy the bound $\|L_{::}\delta\| - \|Y_{::}\delta\| \leq \eta/2 \cdot r_\star$. For

$$r_\star \geq 2\eta^{-1}\sqrt{N\|\Sigma\|}\,\mathrm{rad}(\Lambda_\star) + c(1-\mathrm{P}(\mathcal{A}))^{-1}\eta^{-1}K^2\|\Sigma^{-1}\|\sqrt{\|\Sigma\|}\,\mathrm{w}(\Lambda_\star), \tag{8.26}$$

this latter bound will be satisfied for all $\delta \in \Lambda_\star$. And rearranging it gives $(1-\eta/2)\|L_{::}\delta\| \leq \|Y_{::}\delta\|$, which implies (8.23). Thus, (8.21) holds on the event $\mathcal{A}$ for $r_\star$ above.

The upper bound (8.22)**.**

$$\begin{aligned}
M(\delta) &= (Y_{::}\lambda_\star - Y_{:T})'Y_{::}\delta - 1_{\{\Lambda=\mathrm{conv}(\Lambda)\}}(L_{::}\lambda_\star - L_{:T})'L_{::}\delta \\
&= 1_{\{\Lambda\neq\mathrm{conv}(\Lambda)\}}(L_{::}\lambda_\star - L_{:T})'L_{::}\delta && (8.27)\\
&+ (L_{::}\lambda_\star - L_{:T})'\varepsilon_{::}\delta && (8.28)\\
&+ \lambda_\star'\varepsilon_{::}'L_{::}\delta && (8.29)\\
&+ \lambda_\star'\varepsilon_{::}'\varepsilon_{::}\delta && (8.30)\\
&- \varepsilon_{:T}'L_{::}\delta. && (8.31)\\
&- \varepsilon_{:T}'\varepsilon_{::}\delta. && (8.32)
\end{aligned}$$

The first term (8.27) is deterministic, and has the Cauchy-Schwarz bound

$$1_{\{\Lambda\neq\mathrm{conv}(\Lambda)\}}\|L_{::}\lambda_\star - L_{:T}\|r_\star \text{ for } \delta \in \Lambda_\star.$$

The second term (8.28) has straightforward bound based Cauchy-Schwarz and (8.25). On $\mathcal{A}$,

$$|(L_{::}\lambda_\star - L_{:T})'\varepsilon_{::}\delta| \le \|L_{::}\lambda_\star - L_{:T}\| \cdot (\eta/2)r_\star \text{ for } \delta \in \Lambda_\star.$$

Similarly, Cauchy-Schwarz, (8.25), and the analogous bound

$$\mathrm{E}\left|\|\varepsilon\lambda_\star\| - \sqrt{N}\|\sqrt{\Sigma}\lambda_\star\|\right| \le c\|\Sigma^{-1}\|\sqrt{\|\Sigma\|}K^2\,\mathrm{w}(\{0,\lambda_\star\}) \le c\|\Sigma^{-1}\|\sqrt{\|\Sigma\|}K^2\|\lambda_\star\| \tag{8.33}$$

suffice to bound (8.30). They imply that for $\delta \in \Lambda_\star$, on the intersection of $\mathcal{A}$ and the analogous event on which the bound above holds with probability $\mathrm{P}(\mathcal{A})$,

$$\begin{aligned}
|\lambda_\star'\varepsilon_{::}'\varepsilon_{::}\delta| &\le \|\varepsilon_{::}\lambda_\star\|\|\varepsilon_{::}\delta\| \\
&\le \left[\sqrt{N}\|\Sigma^{1/2}\|\|\lambda_\star\| + (1-\mathrm{P}(\mathcal{A}))^{-1}c\|\Sigma^{-1}\|\sqrt{\|\Sigma\|}K^2\|\lambda_\star\|\right][(\eta/2)\cdot r_\star] \\
&\le (\eta/2)^2 r_\star^2\|\lambda_\star\|/\operatorname{rad}(\Lambda_\star).
\end{aligned}$$

The third term (8.29) is the supremum of the inner product of a subgaussian random vector $\varepsilon_{::}\lambda_\star$ and a vector $L_{::}\delta$ in the intersection of the image of $\Lambda_\star$ under $L_{::}$ and the $\|\cdot\|_2$ ball of radius $r_\star$. and via Talagrand's comparison inequality [Vershynin, 2018, Corollary 8.6.3],

$$\begin{aligned}
&\mathrm{E}\sup_{x\in L_{::}\Lambda_\star}|(\varepsilon_{::}\lambda_\star)'x| \le cK'\,\mathrm{w}(L_{::}\lambda_\star) \\
&K' = \sup_{x,y\in L_{::}\Lambda_\star}\|(\varepsilon_{::}\lambda_\star)'(x-y)\|_{\psi_2}/\|x-y\|_{L_2} \le \|\varepsilon_{::}\lambda_\star\|_{\psi_2}.
\end{aligned}$$

To bound the gaussian width $\mathrm{w}(L_{::}\Lambda_\star)$, we split $L_{::}$ into two pieces, a low rank approximation $\tilde{L}_R$ defined as the sum of the first R terms in the singular value decomposition $L_{::} = \sum_k \sigma_k u_k v_k'$, and the remainder. $\tilde{L}_R\delta$ is contained in a R-dimensional ball of radius $r_\star$, so in terms of a standard gaussian vector g

$$\mathrm{w}(L_{::}\lambda_\star) = \mathrm{E}\sup_{\delta\in\Lambda_\star} g'L_{::}\delta \le \mathrm{E}\sup_{\delta\in\Lambda_\star} g'\tilde{L}_R\delta + \mathrm{E}\sup_{\delta\in\Lambda_\star} g'(L_{::}-\tilde{L}_R)\delta \le c\sqrt{R}r_\star + \mathrm{w}((L_{::}-\tilde{L}_R)\Lambda_\star).$$

The fifth term (8.31) is analogous, with the difference that we substitute $\varepsilon_{:T}$ for $\varepsilon_{::}\lambda_\star$, so in the corresponding bound we have $K' = \|\varepsilon_{:T}\|_{\psi_2}$.

To bound the sixth term, $\|\varepsilon'_{:T}\varepsilon_{::}\delta\|$, we begin by characterizing the vector $\varepsilon'_{:T}\varepsilon_{::}$. Because the rows of ε are independent, $\varepsilon'_{:T}\varepsilon_{:j} = \sum_i \varepsilon_{iT}\varepsilon_{ij}$ is a sum of independent subexponential random variables with $\|\varepsilon_{iT}\varepsilon_{ij}\|_{\psi_1} \leq \|\varepsilon_{iT}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2}$ [Vershynin, 2018, Lemma 2.7.7]. In terms of the averaged autocorrelation $\bar{\gamma}_j = N^{-1}\sum_i E\varepsilon_{iT}\varepsilon_{ij}$, $\varepsilon'_{:T}\varepsilon_{::} = N\bar{\gamma} + Z$ where $Z_j = \varepsilon'_{:T}\varepsilon_{:j} - N\bar{\gamma}_j$ is a sum of independent subexponential random variables with mean zero and $\|Z_i\|_{\psi_1} \leq c\|\varepsilon_{iT}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2}$. By the Triangle inequality and Cauchy-Schwartz,

$$|\varepsilon'_{:T}\varepsilon_{::}\delta| \leq N\,|\bar{\gamma}'\delta| + |Z'\delta| \leq N\,|\bar{\gamma}'\delta| + 2\,\mathrm{rad}(\Lambda_\star)\|Z\|,$$

and by Bernstein's inequality [Vershynin, 2018, Theorem 2.8.1],

$$\mathrm{P}(|Z_j| \geq t_j) \leq 2\exp\left(-c\min\left(\frac{t_j^2}{\sum_i \|\varepsilon_{iT}\|_{\psi_2}^2\|\varepsilon_{ij}\|_{\psi_2}^2}, \frac{t_j}{\max_i \|\varepsilon_{iT}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2}}\right)\right),$$

so we have a bound of the form $2\,\mathrm{rad}(\Lambda_\star)\|Z\| < \xi r_\star^2$ by the union bound for $\sum_j t_j^2 \leq \xi^2 r_\star^4/(4\,\mathrm{rad}(\Lambda_\star)^2)$. Simply taking $t_j^2 = \xi^2 r_\star^4/(4\,\mathrm{rad}(\Lambda_\star)^2 T)$ yields

$$\begin{aligned}
&\mathrm{P}(2\,\mathrm{rad}(\Lambda_\star)\|Z\| \geq \xi r_\star^2) \\
&\leq 2\sum_j \exp\left(-c\min\left(\frac{\xi^2 r_\star^4}{\mathrm{rad}(\Lambda_\star)^2 T\sum_i \|\varepsilon_{iT}\|_{\psi_2}^2\|\varepsilon_{ij}\|_{\psi_2}^2}, \frac{\xi r_\star^2}{\mathrm{rad}(\Lambda_\star)\sqrt{T}\max_i \|\varepsilon_{iT}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2}}\right)\right) \\
&\leq 2T\exp\left(-c\min\left(\frac{\xi^2 r_\star^4}{\mathrm{rad}(\Lambda_\star)^2 TN\max_{ij} \|\varepsilon_{iT}\|_{\psi_2}^2\|\varepsilon_{ij}\|_{\psi_2}^2}, \frac{\xi r_\star^2}{\mathrm{rad}(\Lambda_\star)\sqrt{T}\max_{ij} \|\varepsilon_{iT}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2}}\right)\right).
\end{aligned}$$

Putting everything together, our claims on $M(\delta)$, as well as our claims on $Q(\delta)$ from the

previous section, hold with high probability when $\eta \le \min\{1, 4\,\mathrm{rad}(\Lambda_\star)/\|\lambda_\star\|\}$ and $r_\star$ satisfies

$$
\begin{aligned}
r_\star \ge c \max \Bigg(& [1_{\{\Lambda \neq \mathrm{conv}(\Lambda)\}} + \eta](1-\eta)^{-1} \|L_{::}\lambda_\star - L_{:T}\|, \\
& \sqrt{(1-\eta)^{-1}\left[\|\varepsilon_{::}\lambda_\star\|_{\psi_2} + \|\varepsilon_{:T}\|_{\psi_2}\right]\left[\sqrt{R} r_\star + \mathrm{w}((L_{::} - \tilde{L}_R)\Lambda)\right]}, \\
& \sqrt{(1-\eta)^{-1} N \sup_{\delta \in \Lambda_\star} |\bar{\gamma}'\delta|}, \\
& \sqrt{(1-\eta)^{-1}\,\mathrm{rad}(\Lambda_\star) \max\left(\sqrt{T}, \sqrt[4]{NT}\right) \max_{ij} \|\varepsilon_{iT}\|_{\psi_2} \|\varepsilon_{ij}\|_{\psi_2} \log(T)}, \\
& \eta^{-1}\sqrt{N}\|\Sigma\|\,\mathrm{rad}(\Lambda_\star) + \max_i \|\varepsilon_{i:}\|_{\psi_2}^2 \|\Sigma^{-1}\| \sqrt{\|\Sigma\|}\,\mathrm{w}(\Lambda_\star) \Bigg).
\end{aligned}
$$

Here the constraint on η comes from (8.30), the first line in $r_\star$ comes from (8.27) and (8.28), the second from (8.29) and (8.31), the third and fourth from (8.32), and the fifth incorporates (8.26) from the lower bound section.

Simplifications This is a fixed point condition, as $r_\star$ appears in the second line of the right side explicitly in the third line and implicitly throughout, as $\Lambda_\star$ is a function of $r_\star$. To eliminate the dependence of the right side on $r_\star$ through $\Lambda_\star$, we simply substitute for $\mathrm{w}(\Lambda_\star)$ and $\mathrm{rad}(\Lambda_\star)$ the upper bounds $\mathrm{w}(\Lambda)$ and $\mathrm{diam}(\Lambda)$. This also allows us to drop the constraint $\eta \le 4\,\mathrm{rad}(\Lambda_\star)/\|\lambda_\star\|$, as because $\lambda_\star \in \Lambda$, it becomes vacuous if we perform this substitution.[4]

This leaves us with an expression on the right that depends on $r_\star$ only through the second line, which will have the form $c\sqrt{x[\sqrt{R} r_\star + w(R)]}$ for $x = (1-\eta)^{-1}\left[\|\varepsilon_{::}\lambda_\star\|_{\psi_2} + \|\varepsilon_{:T}\|_{\psi_2}\right]$ and $w(R) = \mathrm{w}((L_{::} - \tilde{L}_R)\Lambda)$. To obtain the claimed result, we simplify this fixed point condition to a bound, observing that $r_\star^2 \ge cx[\sqrt{R} r_\star + w(R)]$ if $r_\star \ge 2cx\sqrt{\max(R, x^{-1}w(R))}$. In addition, we bound $w(R) = \mathrm{w}((L_{::} - \tilde{L}_R)\Lambda)$ by the minimum of $\|L_{::} - \tilde{L}_R\|\,\mathrm{w}(\Lambda)$ and the gaussian width of the ellipse with axes $\mathrm{diam}(\Lambda)\sigma_{R+1}, \mathrm{diam}(\Lambda)\sigma_{R+2}, \ldots$, which is proportional to $\mathrm{diam}(\Lambda)\sqrt{\sum_{k>R}\sigma_k^2}$ [see e.g Vershynin, 2018, Section 7.6.1], and we minimize this bound over R.

[4] To justify this, we can check that this constraint would not have arisen from our bound on (8.30) had we made this substitution earlier.

8.4 Specializing Theorem 8 and Lemma 9

In this section, we will prove more concrete versions of the results of the previous section. This includes Theorem 3 and Lemma 4 from Section 4, which make the assumption that $\varepsilon_{it} \sim N(0, \sigma^2)$ is independent for each cell (i, t), and allowing us to dramatically simplify our bounds. We also prove Theorem 5 from Section 4, a variant of Theorem 3 which characterizes the synthetic control estimator under the same assumptions.

Proof of Lemma 4. In the bound of Lemma 9, we simplify several expressions under Assumption 1: $\|\varepsilon_{::}\lambda_\star\|_{\psi_2} = \|\varepsilon_{i:}\lambda_\star\|_{\psi_2} = \sigma\|\lambda_\star\|$, $\|\varepsilon_{:T}\|_{\psi_2} = \|\varepsilon_{iT}\|_{\psi_2} = \sigma$, $\|\varepsilon_{iT}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2} = \sigma^2$, $\sqrt{\|\Sigma\|} = \sigma$, $\|\varepsilon_{i:}\|_{\psi_2}^2\|\Sigma^{-1}\|\sqrt{\|\Sigma\|} = \sigma^2\sigma^{-2}\sigma = \sigma$, and $\bar{\gamma} = 0$. With these simplifications, as well as the bounds $2a_\omega \geq \text{diam}(\Omega)$, $2a_\lambda \geq \text{diam}(\Lambda)$, and $\min\{\sqrt{\log(N)}, a_\omega\sqrt{N}\} \gtrsim \text{w}(\Omega), \min\{\sqrt{\log(T)}, a_\lambda\sqrt{T}\} \gtrsim \text{w}(\Lambda)$, the bound of Lemma 9 reduces to that of Lemma 4. □

Proof of Theorem 3. In the bound of Theorem 8 for iid noise, we substitute $\text{E}[\varepsilon_{:T} \mid \varepsilon_{::}] = 0$; $\tilde{\omega} = \hat{\omega}$; $\|\Sigma\|^{1/2} = \max_{i<N}\|\text{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}\|^{1/2} = \sigma$; $\|\varepsilon_{::}\lambda_\star\|_{\psi_2} = \sigma\|\lambda_\star\|$ and $\|\omega_\star'\varepsilon_{::}\|_{\psi_2} = \sigma\|\omega_\star\|$; $a_\omega \gtrsim \text{diam}(\Omega) + \|\omega_\star\|$ and $a_\lambda \gtrsim \text{diam}(\Lambda) + \|\lambda_\star\|$; $\min\{\sqrt{\log(N)}, a_\omega\sqrt{N}\} \gtrsim \text{w}(\Omega)$ and $\min\{\sqrt{\log(T)}, a_\lambda\sqrt{T}\} \gtrsim \text{w}(\Lambda)$; and $\varepsilon_{N:}'\lambda_\star + \omega_\star'\varepsilon_{:T} - \omega_\star'\varepsilon_{::}\lambda_\star = \mathcal{O}_p(\sigma\|\lambda_\star\| + \sigma\|\omega_\star\|)$. We do not have an $\mathcal{O}_p(\sigma\|\omega_\star\|\|\lambda_\star\|)$ term in our bound corresponding to the third term $\omega_\star'\varepsilon_{::}\lambda_\star$ in this last expression because the bounds for the other terms suffice: $\|\lambda_\star\| \leq 1$ for $\lambda_\star \in \Lambda \subseteq \mathbb{L}$, so $\sigma\|\omega_\star\|\|\lambda_\star\| \leq \sigma\|\omega_\star\|$. □

Proof of Theorem 5. We can express our estimator as

$$\hat{\omega} \cdot Y_{:T} - L_{NT} = (\omega_\star' L_{:T} - L_{NT}) + (\hat{\omega} - \omega_\star)' L_{:T} + \hat{\omega}'\varepsilon_{:T}.$$

The first term above is simply δ_{sc}, while the last is $O_p(a_\omega)$, as it is gaussian with standard deviation $\|\hat{\omega}\| \leq a_\omega$ conditional on $\varepsilon_{::}, \varepsilon_{N:}$. It remains to bound the middle term. To do so, note that for any $\tilde{\lambda}$,

$$\begin{aligned}
(\hat{\omega} - \omega_\star)' L_{:T} &= (\hat{\omega} - \omega_\star)' L_{::}\tilde{\lambda} + (\hat{\omega} - \omega_\star)' \left(L_{:T} - L_{::}\tilde{\lambda}\right) \\
&\leq \|L_{::}'(\hat{\omega} - \omega_\star)\|_2 \left\|\tilde{\lambda}\right\|_2 + \|\hat{\omega} - \omega_\star\|_2 \left\|L_{:T} - L_{::}\tilde{\lambda}\right\| \\
&= O_p\left(r_\omega \left\|\tilde{\lambda}\right\|_2 + a_\omega \left\|L_{:T} - L_{::}\tilde{\lambda}\right\|\right),
\end{aligned}$$

where the last line follows from Lemma 4. □

8.5 Proof of Theorem 6

We will now consider the problem of inference in a simple setting with N_1 treated units and T_1 treated time periods, in which all units with $i > N_0$ start treatment at the same time $T_0 + 1$. Our estimator (4.1) is, in this setting, of essentially the same form as those we've discussed for a single treated unit and time period. Having observed such an $N_0 + N_1 \times T_0 + T_1$ matrix $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{L}} + \tilde{\boldsymbol{\varepsilon}}$, we define a $N_0 + 1 \times T_0 + 1$ variant $\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{\varepsilon}$ in which treated units and rows are averaged,

$$\boldsymbol{Y} = \begin{pmatrix} \tilde{Y}_{1:N_0,1:T_0} & T_1^{-1} \sum_{t>T_0} \tilde{Y}_{1:N_0,t} \\ N_1^{-1} \sum_{i>N_0} \tilde{Y}_{i,1:T_0} & (N_1 T_1)^{-1} \sum_{n>N_0,t>T_0} \tilde{Y}_{it} \end{pmatrix}. \tag{8.34}$$

In these terms, letting $N = N_0 + 1$ and $T = T_0 + 1$, $\hat{\tau} = Y_{NT} - \hat{L}_{NT}$. Thus, the problem of estimating the averaged unobserved control potential outcome in this setting is essentially the same as the problem we've considered in the previous section, in which we had only one unobserved potential outcome. It differs only in that this averaging results in heteroskedasticity in the errors $\boldsymbol{\varepsilon}$, making the elements of $\varepsilon_{N:}$ and $\varepsilon_{:T}$ small relative to those of $\varepsilon_{::}$. Furthermore, because averaging drives ε_{NT} to zero, it is possible for our estimator $\hat{\tau}$ to be consistent in this setting, whereas in the previous section the constant order variance of Y_{NT} made this impossible. Our core result is that, under Assumption 4 and some restrictions on N_1 and T_1, $\hat{\tau}$ is approximately normal with bias negligible relative to variance $\text{Var}(\hat{\tau}) \approx \text{Var}(\varepsilon_{NT})$.

Our first step is to establish a formal equivalence between estimation of $\hat{\tau}$ under Assumption 4 on $\tilde{\boldsymbol{Y}}$ and under a specialization of Assumption 5 on $\boldsymbol{Y}$. This is straightforward.

Proposition 10. *If $\tilde{\boldsymbol{Y}}$ satisfies Assumption 4, then $\boldsymbol{Y}$ defined in* (8.34) *satisfies Assumption 6 with $\sigma^2 = \sigma_\xi^2/(1-\rho^2)$ and the estimator $\hat{\tau}$* (4.1) *based on $\tilde{\boldsymbol{Y}}$ is the same as the one based on $\boldsymbol{Y}$.*

Assumption 6. *We have $N = N_0 + 1$, $T = T_0 + 1 \to \infty$, and there is a $N \times T$ deterministic matrix $\boldsymbol{L}$ such that $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$ with $W_{it} = 1\{i = N, t = T\}$, and the rows of ε are independent gaussian vectors with, for $i < N$,*

1. $\text{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma^2 \rho^{\ell - k}$ *for $j \leq k < T$,*

2. $\text{Cov}\left(\varepsilon_{ij}, \varepsilon_{iT}\right) = T_1^{-1}\sigma^2 \sum_{k=1}^{T_1} \rho^{T_0+k-j} = T_1^{-1}\sigma^2\rho^{T_0+1-j}(1-\rho^{T_1})/(1-\rho)$ *for* $j < T$,

3. $\text{Var}\left(\varepsilon_{iT}\right) = T_1^{-2} \sum_{k=1}^{T_1} \sum_{\ell=1}^{T_1} \sigma^2\rho^{|k-\ell|}$,

for $\rho \geq 0$ *and, for* $i = N$, *the corresponding terms are* N_1^{-1} *times these quantities.*

Thus, as our estimator's error is

$$\hat{\tau} - \tau = (Y_{NT} - L_{NT}) + (L_{NT} - \hat{L}_{NT}) = \varepsilon_{NT} + \mathcal{O}_p(L_{NT} - \hat{L}_{NT}),$$

a specialized version of Theorem 8 under Assumption 6 is sufficient to establish conditions under which the second term is negligible relative to the first, i.e. conditions under the latter term is $o_p((N_1T_1)^{-1/2})$. Our general result, of which Theorem 6 is a corollary, follows.

Lemma 11. *Under Assumption 6, consider the least squares estimators*

$$\hat{\omega} = \underset{\omega\in\Omega}{\arg\min}\|\omega' Y_{::} - Y_{N:}\|_2^2 \quad \textit{and} \quad \hat{\lambda} = \underset{\lambda\in\Lambda}{\arg\min}\|Y_{::}\lambda - Y_{:T}\|_2^2$$

and the oracle estimators $\omega_\star, \lambda_\star$ *defined analogously with* L *substituted for* Y *and define*

$$\begin{aligned}
\delta_\omega &= \|L_{N:} - \omega_\star' L_{::}\|; \\
\delta_\lambda &= \|L_{:T} - L_{::}\lambda_\star\|; \\
\delta_{sdid} &= \left|L_{NT} - (\omega_\star' \boldsymbol{L}_{:T} + \boldsymbol{L}_{N:}\lambda_\star - \omega_\star' L_{::}\lambda_\star)\right|.
\end{aligned}$$

If $\sigma = O(1)$ *and* $\limsup_{N,T\to\infty} \rho < 1$, *and we take* $\text{diam}(\Omega) = O(1)$, *then* $\left|\widehat{L}_{NT} - L_{NT}\right| =$

$o_p((N_1T_1)^{-1/2})$ *if*

$$\begin{aligned}
N_1T_1 \ll \min\Bigg\{ & \frac{1}{\delta_{sdid}^2}, \\
& \frac{1}{\operatorname{diam}(\Omega)^2 \max\left\{\delta_\lambda^2, \mathrm{w}^2(\Lambda)\right\}}, \\
& \frac{1}{\operatorname{diam}(\Lambda)^2 \max\left\{\delta_\omega^2, \mathrm{w}^2(\Omega)\right\}}, \\
& \frac{1}{\operatorname{diam}(\Omega)^2 \operatorname{diam}(\Lambda)^2 N_0} \Bigg\}; \\
N_1T_1^{1/2} \ll & \frac{1}{\operatorname{diam}(\Omega)^2 \operatorname{diam}(\Lambda) \max\left\{T_0, \sqrt{N_0T_0}\right\} \log(T_0)}; \\
N_1 \ll \min\Bigg\{ & \frac{T_1}{\mathrm{w}(\Omega)^2}, \frac{1}{\operatorname{diam}(\Omega)^2 \operatorname{approx\text{-}rank}\left(L_{::}, \max\left\{\|\lambda_\star\|, T_1^{-1/2}\right\}\right)} \Bigg\}.
\end{aligned}$$

This follows from Corollary 13 and Corollary 14 below by a straightforward calculation. These corollaries are specializations of Lemma 9 and Theorem 8 respectively, which follow from our more general results with a few calculations collect in Lemma 12 below.

Lemma 12. *Under Assumption 6 for $\rho > 0$, letting Σ be the covariance matrix $\mathrm{E}\,\varepsilon_{i:}\varepsilon_{i:}'$ and γ be the vector of autocovariances $\mathrm{E}\,\varepsilon_{:T}'\varepsilon_{::}$,*

1. $\|\Sigma\| \le \sigma^2 \frac{1+\rho}{1-\rho}$
2. $\|\Sigma^{-1}\| \le \sigma^{-2} \frac{1+\rho}{1-\rho}$
3. $\|\gamma\| \le T_1^{-1}\sigma^2\rho^2(1-\rho)^{-1}(1-\rho^2)^{-1/2}$
4. $\|\varepsilon_{i:}\|_{\psi_2} = \|\Sigma\|^{1/2} \le \sqrt{2}\sigma(1-\rho)^{-1/2}$.
5. $\|\varepsilon_{::}\lambda_\star\|_{\psi_2} = \|\Sigma\|^{1/2}\|\lambda_\star\| \le \sqrt{2}\sigma(1-\rho)^{-1/2}\|\lambda_\star\|$
6. $\|\omega_\star'\varepsilon_{::}\|_{\psi_2} = \sigma\|\omega_\star\|$
7. $\|\varepsilon_{:T}\|_{\psi_2} = \|\varepsilon_{iT}\|_{\psi_2} \le \sqrt{2}T_1^{-1/2}\sigma(1-\rho)^{-1/2}$
8. $\|\mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}]\|_{\psi_2} = \|\mathrm{E}[\varepsilon_{iT} \mid \varepsilon_{::}]\|_{\psi_2} \le T_1^{-1}\rho^{1/2}(1-\rho)^{-1/2}\sigma$

9. $\mathrm{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}]^{1/2} = \mathcal{O}_p\left(T_1^{-1/2}\sigma(1-\rho)^{-1/2}\right)$

10. $\varepsilon_{N:}\lambda_\star + \omega_\star'\varepsilon\colon T - \omega_\star'\varepsilon_{::}\lambda_\star = \mathcal{O}_p\left(\left[N_1^{-1/2}\|\lambda_\star\| + T_1^{-1/2}\|\omega_\star\| + \|\omega_\star\|\|\lambda_\star\|\right]\sigma(1-\rho)^{-1/2}\right)$.

Corollary 13. *Under Assumption 6, for any subset* Λ *of* $\mathbb{R}^{T-1}$*, the least squares estimator and oracle least squares estimator*

$$\hat{\lambda} = \min_{\lambda\in\Lambda}\|Y_{::}\lambda - Y_{:T}\|_2^2 \quad \textit{and} \quad \lambda^\star = \min_{\lambda\in\Lambda}\|L_{::}\lambda - L_{:T}\|_2^2$$

satisfy the bound $\|L_{::}(\hat{\lambda}-\lambda^\star)\|_2 = \mathcal{O}_P(r_\lambda)$ *where, for* approx-rank *defined in Lemma 9,*

$$\begin{aligned} r_\lambda = \max\Bigg(&\|L_{::}\lambda^\star - L_{:T}\|,\\ &x\sqrt{\text{approx-rank}(L_{::}, x)} \quad \textit{for} \quad x = \max\left\{\|\lambda_\star\|, T_1^{-1/2}\right\}\sigma(1-\rho)^{-1/2},\\ &T_1^{-1/2}\operatorname{diam}^{1/2}(\Lambda)\cdot\sigma\rho(1-\rho)^{-1/2}(1-\rho^2)^{-1/4},\\ &T_1^{-1/4}\sqrt{\operatorname{diam}(\Lambda)\max\left(\sqrt{T_0}, \sqrt[4]{N_0T_0}\right)\log(T_0)}\cdot\sigma(1-\rho)^{-1/4},\\ &N_0^{1/2}\operatorname{diam}(\Lambda)\cdot\sigma(1-\rho)^{-1/2},\\ &\mathrm{w}(\Lambda)\cdot\sigma\left(\frac{1+\rho}{1-\rho}\right)^{3/2}\Bigg). \end{aligned}$$

Corollary 14. *Under Assumption 6, consider the least squares estimators*

$$\hat{\omega} = \arg\min_{\omega\in\Omega}\|\omega' Y_{::} - Y_{N:}\|_2^2 \quad \textit{and} \quad \hat{\lambda} = \arg\min_{\lambda\in\Lambda}\|Y_{::}\lambda - Y_{:T}\|_2^2$$

and the oracle estimators $\omega_\star, \lambda_\star$ *defined analogously with* L *substituted for* Y *and define*

$$\begin{aligned} \delta_\omega &= \|L_{N:} - \omega_\star' L_{::}\|;\\ \delta_\lambda &= \|L_{:T} - L_{::}\lambda_\star\|;\\ \delta_{sdid} &= \left|L_{NT} - (\omega_\star'\boldsymbol{L}_{:T} + \boldsymbol{L}_{N:}\lambda_\star - \omega_\star' L_{::}\lambda_\star)\right|. \end{aligned}$$

Then, for r_λ defined in Corollary 13,

$$\begin{aligned}\left|\widehat{L}_{NT} - L_{NT}\right| \leq & \operatorname{diam}(\Lambda)\left[\delta_\omega + \mathcal{O}_p\left(\sigma(1-\rho)^{-1/2}\max\left\{1, \mathrm{w}(\Omega)\right\}\right)\right] \\ & + \operatorname{diam}(\Omega)\left[\delta_\lambda + \mathcal{O}_p\left(\sigma(1-\rho)^{-1/2}\max\left\{T_1^{-1}, \mathrm{w}(\Lambda)\right\}\right)\right] \\ & + \delta_{sdid} + \mathcal{O}_p\left(\operatorname{diam}(\Omega) r_\lambda + \sigma(1-\rho)^{-1/2}\rho^{1/2}T_1^{-1}\,\mathrm{w}(\Omega)\right).\end{aligned}$$

8.6 Proof of Lemma 12

Because all random variables involved are gaussian, the subgaussian norm $\|\cdot\|_{\psi_2}$ is equivalent to $\|\cdot\|_{L_2}$. Furthermore, for a random vector v of identically distributed elements v_i, $\|v\|_{L_2} = \sup_{\|x\|=1}\sqrt{\sum_i x_i^2 \operatorname{E} v_i^2} = \|v_i\|_{L_2}$.

1,2. Our bounds (i-ii) are determined by upper and lower bounds on the maximal and minimal eigenvalues of the correlation matrix Σ/σ^2 respectively, which for our $AR(1)$ process are no larger than $\frac{1+\rho}{1-\rho}$ and no smaller than $\frac{1-\rho}{1+\rho}$ respectively [see e.g. Trench, 1999, Section 1].

3.

$$\begin{aligned}\|\gamma\| &= T_1^{-1}\sigma\rho\frac{1-\rho^{T_1}}{1-\rho}\sqrt{\sum_{j=1}^{T_0}\rho^{2(T_0+1-j)}} \\ &= T_1^{-1}\sigma^2\rho\frac{1-\rho^{T_1}}{1-\rho}\sqrt{\rho^2\frac{1-\rho^{2T_0}}{1-\rho^2}} \\ &\leq T_1^{-1}\sigma^2\rho^2(1-\rho)^{-1}(1-\rho^2)^{-1/2}.\end{aligned}$$

4,5,6. For any vector v, $\|\varepsilon_{::}v\|_{\psi_2} = \|\varepsilon_{i:}\Sigma^{-1/2}\Sigma^{1/2}v\|_{L_2} = \|\Sigma^{1/2}v\| \leq \sqrt{2}\sigma(1-\rho)^{-1/2}\|v\|$, reducing our problem to one about an identically distributed vector and in the last step using our bound (i) and substituting $2 > 1+\rho$. Similarly, $\|u'\varepsilon_{::}\|_{\psi_2} = \|u'\varepsilon_{::}\Sigma^{-1/2}\Sigma\|_{L_2} \leq \|\Sigma\|\|u\| \leq \sqrt{2}\sigma(1-\rho)^{-1/2}\|u\|$.

7. $\|\varepsilon_{:T}\|_{\psi_2} = \|\varepsilon_{iT}\|_{L_2}$, as it's a vector of identically distributed gaussian elements, and

$$\begin{aligned}\|\varepsilon_{iT}\|_{L_2}^2 &= T_1^{-2} \sum_{k=1}^{T_1} \sum_{\ell=1}^{T_1} \sigma^2 \rho^{|k-\ell|} \\ &= T_1^{-2}\sigma^2 \left(\sum_{s=0}^{T_1-1} 2(T_1 - s)\rho^s - T_1 \right) \\ &\le T_1^{-2}\sigma^2 \cdot 2T_1 \sum_{s=0}^{T_1-1} \rho^s \le 2T_1^{-1}\sigma^2(1-\rho)^{-1}.\end{aligned}$$

8. For the same reason, $\|\mathrm{E}[\varepsilon_{:T} \mid \varepsilon_{::}]\|_{\psi_2} = \|\mathrm{E}[\varepsilon_{iT} \mid \varepsilon_{i:}]\|_{L_2}$, and $\|\mathrm{E}[\varepsilon_{iT} \mid \varepsilon_{i:}]\|_{L_2} = \|T_1^{-1} \sum_{k=1}^{T_1} \rho^k \varepsilon_{i,T_0}\|_{L_2} = T_1^{-1}\sqrt{\rho\frac{1-\rho^{T_1}}{1-\rho}}\sigma \le T_1^{-1}\rho^{1/2}(1-\rho)^{-1/2}\sigma$.

9. $\mathrm{E}\,\mathrm{Var}[\varepsilon_{iT} \mid \varepsilon_{i:}] \le \mathrm{E}\,\varepsilon_{iT}^2 \le 2T_1^{-1}\sigma^2(1-\rho)^{-1}$.

10.

$$\begin{aligned}\|\tilde{\varepsilon}'_{N:}\lambda_\star\|_{L_2} &= N_1^{-1/2}\|\Sigma^{1/2}\lambda_\star\| \le \sqrt{2}N_1^{-1/2}\|\lambda_\star\|\sigma(1-\rho)^{-1/2} \\ \|\omega'_\star\tilde{\varepsilon}_{:T}\|_{L_2} &= \|\omega_\star\|\|\tilde{\varepsilon}_{iT}\|_{L_2} \le \sqrt{2}T_1^{-1/2}\|\omega_\star\|\sigma(1-\rho)^{-1/2} \\ \|\omega'_\star\tilde{\varepsilon}_{::}\lambda_\star\|_{L_2} &= \|\omega_\star\|\|\Sigma^{1/2}\lambda_\star\| \le \sqrt{2}\|\omega_\star\|\|\lambda_\star\|\sigma(1-\rho)^{-1/2}.\end{aligned}$$

8.6.1 Proof of Lemma 11

Substituting the bound of Corollary 13 into that of Corollary 14,

$$\begin{aligned}\left|\widehat{L}_{NT} - L_{NT}\right| =& \delta_{sdid} + \\ & \mathcal{O}_p\Bigg(\quad \operatorname{diam}(\Lambda) \max\left\{\delta_\omega,\ \mathrm{w}(\Omega)\right\} \\ & \quad + \operatorname{diam}(\Omega) \max\left\{\delta_\lambda,\ \mathrm{w}(\Lambda)\right\} \\ & \quad + \operatorname{diam}(\Omega) \max\left\{\|\lambda_\star\|, T_1^{-1/2}\right\} \sqrt{\text{approx-rank}\left(L_{::}, \max\left\{\|\lambda_\star\|, T_1^{-1/2}\right\}\right)} \\ & \quad + \operatorname{diam}(\Omega) T_1^{-1/2} \operatorname{diam}^{1/2}(\Lambda) \\ & \quad + \operatorname{diam}(\Omega) T_1^{-1/4} \sqrt{\operatorname{diam}(\Lambda) \max\left(\sqrt{T_0}, \sqrt[4]{N_0 T_0}\right) \log(T_0)} \\ & \quad + \operatorname{diam}(\Omega) N_0^{1/2} \operatorname{diam}(\Lambda) \\ & \quad + T_1^{-1} \mathrm{w}(\Omega) \Bigg).\end{aligned}$$

Several of the terms in this bound can be ignored, as they are bounded by others:

$$\operatorname{diam}(\Omega)\|\lambda_\star\| \sqrt{\text{approx-rank}\left(L_{::}, \max\left\{\|\lambda_\star\|, T_1^{-1/2}\right\}\right)} \lesssim \operatorname{diam}(\Omega) N_0^{1/2} \operatorname{diam}(\Lambda);$$

$$\operatorname{diam}(\Omega) T_1^{-1/2} \operatorname{diam}^{1/2}(\Lambda) \lesssim \operatorname{diam}(\Omega) T_1^{-1/2} \sqrt{\text{approx-rank}\left(L_{::}, \max\left\{\|\lambda_\star\|, T_1^{-1/2}\right\}\right)}.$$

Under the stated conditions on N_1 and T_1, each of the remaining terms is $o((N_1 T_1)^{-1/2})$.

1. The condition on $N_1 T_1$ in the lemma statement arises from terms in the expression above with no factors of N_1, T_1. It is equivalent to the condition

$$\begin{aligned}(N_1 T_1)^{-1/2} \gg \max\Bigg\{ & \delta_{sdid}, \\ & \operatorname{diam}(\Lambda) \max\left\{\delta_\omega,\ \mathrm{w}(\Omega)\right\}, \\ & \operatorname{diam}(\Omega) \max\left\{\delta_\lambda,\ \mathrm{w}(\Lambda)\right\}, \\ & \operatorname{diam}(\Omega) \operatorname{diam}(\Lambda) N_0^{1/2} \Bigg\}.\end{aligned}$$

2. The condition on $N_1 T_1^{1/2}$ in the lemma statment arises from the term above with the leading factor of $T_1^{-1/4}$.

3. The condition on N_1 arises from the terms above with leading factors of $T_1^{-1/2}$ and T_1^{-1}.