

Synthetic Difference In Differences

Dmitry Arkhangelsky* Susan Athey† David A. Hirshberg‡

Guido W. Imbens§ Stefan Wager¶

Current version December 2018

Abstract

We present a new perspective on the Synthetic Control (SC) method as a weighted regression estimator with time fixed effects. This perspective suggests a generalization with two way (both unit and time) fixed effects, which can be interpreted as a weighted version of the standard Difference In Differences (DID) estimator. We refer to this new estimator as the Synthetic Difference In Differences (SDID) estimator. We validate our approach formally, in simulations, and in an application, finding that this new SDID estimator has attractive properties compared to the SC and DID estimators. In particular, we find that our approach has doubly robust 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.

Keywords: Synthetic Controls, Causal Effects, Panel Data, Difference In Differences, Low-Rank Confounders

*Assistant Professor, CEMFI, Madrid, darkhangel@cemfi.es.

†Professor of Economics, Graduate School of Business, Stanford University, and NBER, athey@stanford.edu.

‡Postdoctoral Fellow, Department of Statistics and SIEPR, Stanford University, davidahirshberg@stanford.edu.

§Professor of Economics, Graduate School of Business, and Department of Economics, Stanford University, SIEPR, and NBER, imbens@stanford.edu.

¶Assistant Professor of Operations, Information and Technology, Graduate School of Business, and of Statistics (by courtesy), Stanford University, 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 fixed effect estimator, where the regression model allows for time fixed effects but not for unit fixed effects. We propose adding unit fixed effects to this representation of the standard SC set up, and we show that this leads to a weighted version of the standard Difference In Differences (DID) estimator. In addition to adding unit fixed effects to the implicit SC specification, we also allow for the inclusion of weights based on the time period. These additional weights ensure that the weighted periods resemble more closely the period for which we are imputing the counterfactual, for example by weighting more recent periods more heavily if that is warranted. We show 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: the estimator is consistent if *either* the model is correctly specified, *or* if the weights are well chosen, but not necessarily both.

Consider the simplest case of a panel with N units and T time periods, where outcomes are denoted Y_{it} , and where the only unit exposed to the treatment is N in period T . Suppose that there are no covariates. In that case, the SC estimator for the counterfactual control value Y_{NT} is a weighted average of the period T outcomes for the control units, $\hat{Y}_{NT}^{\text{sc}} = \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{iT}$, with the weights $\hat{\omega}_i^{\text{sc}}$ chosen to make the weighted average of the controls in the pre-treatment period approximate the corresponding value for the treated unit, so that $\sum_i \hat{\omega}_i^{\text{sc}} Y_{it} \approx Y_{Nt}$ for all $t = 1, \dots, T-1$. In this paper, we introduce an alternative characterization of the SC estimator

as a weighted fixed effect estimator:

$$\hat{Y}_{NT}^{\text{sc}} = \hat{\mu} + \hat{\beta}_T, \quad \text{where } (\hat{\mu}, \hat{\beta}) = \arg \min_{\mu, \beta} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \beta_t)^2 \hat{\omega}_i^{\text{sc}}, \quad (1.1)$$

where \mathcal{O} is the set of all pairs of indices (i, t) for which we observe the control outcome, that is, all pairs (i, t) other than (N, T) . This representation is helpful because it clarifies the connection to the DID literature. The standard DID estimator for $Y_{NT}(0)$ is in this setting equal to

$$\hat{Y}_{NT}^{\text{did}} = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T, \quad \text{where } (\hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta, \mu} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \alpha_i - \beta_t)^2. \quad (1.2)$$

Our characterization (1.1) makes clear that relative to the SC estimator, the DID estimator adds a unit fixed effect to the specification, but it removes the weights. Contrasting the SC and DID estimators in this way suggests a natural modification. Specifically, we propose the SDID estimator for $Y_{NT}(0)$, formally defined as:

$$\hat{Y}_{NT}^{\text{sdid}} = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T, \quad \text{where } (\hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta, \mu} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \alpha_i - \beta_t)^2 \hat{\omega}_i \hat{\lambda}_t. \quad (1.3)$$

Crucially, the regression in (1.3) includes both unit fixed effects and weights, where the weights are the product of unit weights $\hat{\omega}_i$ and time weights $\hat{\lambda}_t$, where 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 $i = 1, \dots, N - 1$. 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, while it differs from the SC estimator by including unit-fixed effects and allowing for time weights.

Many of the earlier approaches to SC settings, including Abadie, Diamond, and Hainmueller [2010, 2015], Doudchenko and Imbens [2016], Xu [2017], Athey, Bayati, Doudchenko, Imbens, and Khosravi [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 out-

comes, 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]. 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., 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], and most of the currently recommended estimators in that literature combine them.

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. Throughout, we take the perspective, common in panel data settings, that \mathbf{Y} is a noisy estimate of an underlying signal matrix \mathbf{L} , i.e., $Y_{it} = L_{it} + \varepsilon_{it}$ where ε is noise. The matrix \mathbf{L} , meanwhile, could be taken to have a simple fixed effect specification, or to 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], or Xu [2017]. We formalize double robustness properties by presenting two consistency results, one which makes weak assumptions on the weights but relatively strong assumptions on the outcome model \mathbf{L} , and one which makes weak assumptions on the conditional outcome model but stronger assumptions on the weights.

A major challenge in the SC literature has been the characterization of the asymptotic behavior of the estimators. Ideally, the weights $\hat{\omega}$ and $\hat{\lambda}$ would balance out the rows and columns of the underlying signal matrix \mathbf{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 (usually called \mathbf{X}) 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 \mathbf{Y} , not \mathbf{L} , and have a rich dependence on the noise ε that cannot be eliminated via sample splitting. In Section 5.3, we use tools from modern empirical process theory to address both challenges and to show that, despite being optimized to balance the observed \mathbf{Y} , the weights $\hat{\omega}$ and $\hat{\lambda}$ also balance the unobserved \mathbf{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 \mathbf{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. For both consistency and for asymptotic normality results for SDID and for consistency of SC, it is important that we penalize the original SC weights 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 $\hat{Y}_{NT}^{\text{sdid}}$ 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 Set Up

Suppose we have a balanced panel with observations on an outcome Y_{it} , $i = 1, \dots, N$, $t = 1, \dots, T$, with some units treated in some periods, and the binary treatment indicator denoted by $W_{it} \in \{0, 1\}$. Initially we focus on the case with just a single unit treated in a single period, unit N in period T , although the ideas are more general. Specifically, the methods developed here extend directly to the block treatment assignment, and to staggered adoption [Athey and Imbens, 2018], but not necessarily to general assignment settings where units switch in and out of treatment, [e.g., Arkhangelsky and Imbens, 2018, de Chaisemartin and D’Haultfoeulle, 2018]. In this special case the treatment is

$$W_{it} = \begin{cases} 1 & \text{if } i = N, t = T, \\ 0 & \text{otherwise.} \end{cases}$$

For units and in time periods with $W_{it} = 0$ we observe $Y_{it}(0)$ and for units in time periods with $W_{it} = 1$ we observe $Y_{it}(1)$, so that

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } W_{it} = 0, \\ Y_{it}(1) & \text{otherwise.} \end{cases}$$

We estimate the treatment effect for unit N in period T , $\tau = Y_{NT}(1) - Y_{NT}(0)$, as

$$\hat{\tau} = Y_{NT}(1) - \hat{Y}_{NT},$$

where \hat{Y}_{NT} is the imputed value for $Y_{NT}(0)$ based on the other $NT - 1$ control observations Y_{it} , for $(i, t) \in \mathcal{O}$. Here \mathcal{O} is the set of pairs of unit time indices for which we observe the control outcome, $\mathcal{O} = \left\{ (i, t) : i \in \{1, \dots, N\}, t \in \{1, \dots, T\}, W_{it} = 0 \right\}$. Define also the set of indices for units who are never treated, during periods where no-one is treated: $\mathcal{C} = \left\{ (i, t) : i \in \{1, \dots, N - 1\}, t \in \{1, \dots, T - 1\} \right\}$.

Partition the $N \times T$ matrix of observed outcomes \mathbf{Y} , and other conformable matrices, by treatment group and pre/post treatment period:

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

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

$$\bar{Y}^{\text{c,pre}} = \frac{1}{(N - 1)(T - 1)} \sum_{(i,t) \in \mathcal{C}} Y_{it}, \quad \bar{Y}^{\text{c,post}} = \frac{1}{N - 1} \sum_{i=1}^{N-1} Y_{iT},$$

and

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

Our general approach to imputing the missing $Y_{NT}(0)$ has two components. First, we weight the units and time periods to balance them with respect to the unit and time period that we are trying to impute, and second, we build a model for the conditional expectation of the full set of control outcomes. We now consider the two components of our strategy separately.

Starting with the latter, we separate \mathbf{Y} into two terms, $\mathbf{Y} = \mathbf{L} + \boldsymbol{\varepsilon}$, and consider models that parametrize the conditional expectation of the full set of unit and time period pairs:

$$\mathbf{L} = g(\theta),$$

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

$$g(\theta)_{it} = \theta, \quad (\text{constant})$$

$$g(\mu, \alpha, \beta)_{it} = \mu + \alpha_i + \beta_t, \quad \theta = (\mu, \alpha, \beta), \quad (\text{two-way fixed effect})$$

$$g(\mathbf{A}, \mathbf{B})_{it} = \sum_{r=1}^R A_{ir} B_{tr}, \quad \theta = (\mathbf{A}, \mathbf{B}), \quad (\text{factor model}).$$

In the two-way fixed effect and factor models we need some normalizations, e.g., $\alpha_1 = \beta_1 = 0$. For any of these models, we can estimate the unknown parameters by least squares

$$\hat{\theta} = \arg \min_{\theta} \sum_{(i,t) \in \mathcal{O}} \left(Y_{it} - g(\theta)_{it} \right)^2. \quad (2.1)$$

We then impute the missing value as $\hat{Y}_{NT} = g(\hat{\theta})_{NT}$.

Meanwhile, the weighting component of our approach focuses on weights to balance the sample towards the treated unit / time period pair. An critical feature of our approach 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 the unit and time weights sum (up to, but not including, the last period and the last unit) to one. In addition the weights for the last unit and time period are equal to one. Formally, the set of weights we consider satisfy

$$\mathbb{W} = \left\{ \omega \in \mathbb{R}^N \left| \omega_i \geq 0, \omega_N = 1, \sum_{i=1}^{N-1} \omega_i = 1 \right. \right\}, \quad (2.2)$$

and

$$\mathbb{L} = \left\{ \lambda \in \mathbb{R}^T \left| \lambda_t \geq 0, \lambda_T = 1, \sum_{t=1}^{T-1} \lambda_t = 1 \right. \right\}. \quad (2.3)$$

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

$$\hat{\omega}^{\text{sc}} = \arg \min_{\omega \in \mathbb{W}} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{i=1}^{N-1} \omega_i Y_{it} \right)^2. \quad (2.4)$$

Typically we modify these weights slightly by putting an L_2 penalty on the weights to ensure that in larger samples there will be many units with non-zero weights. We also consider the time equivalent of the SC weights, which we also give the SC label, although they do not appear to have been considered in this literature:

$$\hat{\lambda}^{\text{sc}} = \arg \min_{\lambda \in \mathbb{L}} \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2. \quad (2.5)$$

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. In both cases we may wish to regularize the 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

$$\hat{\lambda}^{\text{isc}} = \arg \min_{\lambda_0 \in \mathbb{R}, \lambda \in \mathbb{L}} \sum_{i=1}^{N-1} \left(Y_{iT} - \lambda_0 - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2. \quad (2.6)$$

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,

$$\hat{\omega}_i^{\text{kernel}} \propto K\left(\frac{(\mathbf{Y}_{i:} - \mathbf{Y}_{N:})}{h_\omega}\right), \quad \hat{\lambda}_t^{\text{kernel}} \propto K\left(\frac{\mathbf{Y}_{:t} - \mathbf{Y}_{:T}}{h_\lambda}\right), \quad (2.7)$$

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 neighbor weights, where we give constant weights to the K_ω nearest units and the K_λ nearest 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 $\mathbf{Y}_{:T}$ 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 \mathbf{Y}_N 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, 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 general weights ω and λ the weighted estimators will have the form

$$\hat{Y}_{NT} = g(\hat{\theta}^{\text{weight}})_{NT}, \quad \text{where } \hat{\theta}^{\text{weight}} = \arg \min_{\theta} \sum_{(i,t) \in \mathcal{O}} \left(Y_{it} - g(\theta)_{it}\right)^2 \omega_i \lambda_t. \quad (2.8)$$

In general we consider estimators with different models and different weighting strategies. 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, 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 introduce any bias [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].

Finally, although most of this paper is devoted to estimation of L_{NT} , our framework allows for several extensions. In particular, we can accommodate covariates in the conditional outcome model, and more general treatment assignment mechanisms with multiple units treated in multiple periods. Consider, for example, a case where we have covariates X_{it} , and a block of units are assigned to treatment: Units $1 : \dots, N_0$ are control units, and units $N_0 + 1, \dots, N_0 + N_1 = N$ are treated from period $T_0 + 1$ onwards. In that case we assign weight $1/N_1$ to the treated units, and find the control unit weights that balance the control and treated samples in periods 1 through T_0 . We also assign time weights $1/(T - T_0)$ to all the post-treatment periods, and find time weights for the periods $1, \dots, T_0$ to ensure balance between the pre and post treatment periods. Then, we could estimate a constant treatment effect model by weighted regression

$$\hat{\tau} = \arg \min \left\{ \sum_{i,t} (Y_{it} - \mu - \alpha_i - \beta_t - X_{it}\gamma - W_{it}\tau)^2 \hat{\omega}_i \hat{\lambda}_t \right\}. \quad (2.9)$$

We also consider conditions under which out-of-the-box robust standard errors for this weighted regression that take the weights as given can be used for inference about the corresponding treatment effect parameter τ .

3 The DID and SC estimators as Bias Reduction Methods

In this section we consider the DID and SC estimators. The point is to set up the new estimators we propose in the next section.

3.1 Difference In Differences as a Double Bias Reduction Method

The popular DID estimator for panel data settings has both unit and time fixed effects (e.g., Abadie [2003], Abadie and Cattaneo [2018], Card [1990], Meyer, Viscusi, and Durbin [1995]). In

the case with no covariates the estimator for the counterfactual Y_{NT} , can be characterized as:

$$\hat{Y}_{NT}^{\text{did}} = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T = \bar{Y}^{\text{c,pre}} + \left(\bar{Y}^{\text{t,pre}} - \bar{Y}^{\text{c,pre}} \right) + \left(\bar{Y}^{\text{c,post}} - \bar{Y}^{\text{c,pre}} \right),$$

where

$$(\hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\mu, \alpha, \beta} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \alpha_i - \beta_t)^2.$$

We can see this as doubly bias-adjusting the simple average $\bar{Y}^{\text{c,pre}}$, with the first bias adjustment, $\bar{Y}^{\text{t,pre}} - \bar{Y}^{\text{c,pre}}$, taking into account stable differences between the treated unit and the control units and the second bias adjustment, $\bar{Y}^{\text{c,post}} - \bar{Y}^{\text{c,pre}}$, taking into account stable differences over time for the control group.

3.2 Synthetic Control as a Single Bias Reduction Method

The original Abadie, Diamond, and Hainmueller [2010] SC estimator can be characterized as a weighted fixed effect estimator with only time fixed effects:

$$\hat{Y}_{NT}^{\text{sc}} = \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{iT} = \hat{\mu} + \hat{\beta}_T, \quad \text{where } (\hat{\mu}, \hat{\beta}) = \arg \min_{\beta, \mu} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \beta_t)^2 \hat{\omega}_i^{\text{sc}}. \quad (3.1)$$

First note that we can also write the objective function here as the sum of $(Y_{it} - \mu - \beta_t)^2 \hat{\omega}_i^{\text{sc}} \hat{\lambda}_t$ with any time weights. Next, we can also write the SC estimator as a bias-adjusted estimator:

$$\hat{Y}_{NT}^{\text{sc}} = \hat{Y}_{NT}^{\text{weight}} + \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} \left(Y_{iT} - \sum_{t=1}^{T-1} \hat{\lambda}_t^{\text{sc}} Y_{it} \right), \quad \text{where } \hat{Y}_{NT}^{\text{weight}} = \sum_{(i,t) \in \mathcal{C}} \hat{\omega}_i^{\text{sc}} \hat{\lambda}_t^{\text{sc}} Y_{it}.$$

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.

Another observation is that in the case where the SC weights balance the pre-treatment

periods perfectly, so that

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

adding the unit fixed effects to the model would not change anything, and the SC estimator would in that perfect balance case correspond to

$$\hat{Y}_{NT} = \hat{\mu} + \hat{\beta}_T + \hat{\alpha}_N, \quad \text{where } (\hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\mu, \alpha, \beta} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \alpha_i - \beta_t)^2 \hat{\omega}_i^{\text{sc}}. \quad (3.2)$$

In general, however, without perfect balance, adding the unit fixed effects would change the estimates. Even if there is perfect balance, this equivalence requires using only unit weights.

3.3 The Augmented Synthetic Control Method

Ben-Michael, Feller, and Rothstein [2018] augment the SC method by combining it with a model for the conditional outcome in the last period. Denoting the conditional expectation of Y_{iT} given the lagged outcomes Y_{i1}, \dots, Y_{iT-1} by $m(\mathbf{Y}_{i:})$, with estimator $\hat{m}(\mathbf{Y}_{i:})$, their proposed Augmented Synthetic Control (ASC) estimator is

$$\begin{aligned} \hat{Y}_{NT}^{\text{asc}} &= \sum_{i=1}^{N-1} \omega_i^{\text{sc}} Y_{iT} + \left(\hat{m}(\mathbf{Y}_{N:}) - \sum_{i=1}^{N-1} \omega_i \hat{m}(\mathbf{Y}_{i:}) \right) \\ &= \hat{m}(\mathbf{Y}_{N:}) + \left(\sum_{i=1}^{N-1} \omega_i (Y_{iT} - \hat{m}(\mathbf{Y}_{i:})) \right). \end{aligned}$$

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 stresses 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 $\hat{m}(\mathbf{Y}_{N:})$. The ASC estimator then robustifies this using a weighted average of the residuals, in essence similar to the residuals balancing estimator in the original double robust literature (e.g., Robins, Rotnitzky, and Zhao [1994]), or in high-dimensional settings in Athey, Imbens, and

Wager [2018]. Such adjustments make the estimator doubly robust under appropriate conditions. 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, although by the same token it would make the justification more difficult under factor structures. 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. 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.

4 Synthetic Difference In Differences

The SDID estimator addresses the case where the SC adjustment does not completely balance the underlying signal in the pre-treatment periods. Formally, we use the DID (two-way additive fixed effect) model for the potential outcome, in combination with time and units weights:

$$\hat{Y}_{NT}^{\text{sdid}} = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T, \quad (4.1)$$

where

$$(\hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta, \mu} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \alpha_i - \beta_t)^2 \hat{\omega}_i^{\text{sc}} \hat{\lambda}_t^{\text{sc}}.$$

The estimator can be thought of as bias-adjusting the SC estimator based on the pre-treatment discrepancies, weighted by $\hat{\lambda}_t^{\text{sc}}$:

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

Even without the time weights, that is, with $\hat{\lambda}_t = 1/(T - 1)$, the inclusion of the unit-fixed effects modifies the basic SC estimator by adding the average of the pre-treatment imbalances:

$$\frac{1}{T-1} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{i=1}^{N-1} \hat{\omega}_i^{\text{sc}} Y_{it} \right).$$

We can also write the SDID estimator as

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

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

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

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.

We can also think of the SID estimator 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 the treated unit and a particular weighted average of the control units satisfy a parallel trends assumption, but only for a particular weighted average of the periods. If there is only one control unit, and only one pre-treatment period, this reduces to the standard DID set up.

5 Formal Results

In this section, we develop the properties of the SDID estimator. First, we consider properties that hold when the model is correctly specified; second, we discuss double robustness properties. For generic weights $\hat{\omega}$ and $\hat{\lambda}$ consider the corresponding SDID estimator:

$$\hat{Y}_{NT}^{\text{sdid}}(\hat{\omega}, \hat{\lambda}) = \hat{\mu} + \hat{\alpha}_N + \hat{\beta}_T,$$

where

$$(\hat{\mu}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta, \mu} \sum_{(i,t) \in \mathcal{O}} (Y_{it} - \mu - \alpha_i - \beta_t)^2 \hat{\omega}_i \hat{\lambda}_t.$$

This estimator can also be written as

$$\begin{aligned} \hat{Y}_{NT}^{\text{sdid}}(\hat{\omega}, \hat{\lambda}) &= \sum_{i=1}^{N-1} \hat{\omega}_i Y_{iT} + \sum_{t=1}^{T-1} \hat{\lambda}_t Y_{Nt} - \sum_{i=1}^{N-1} \sum_{t=1}^{T-1} \hat{\omega}_i \hat{\lambda}_t Y_{it} \\ &= Y_{:T} \cdot \hat{\omega}_{\cdot} + Y_{N\cdot} \cdot \hat{\lambda}_{\cdot} + \hat{\omega}'_{\cdot} Y_{\cdot\cdot} \hat{\lambda}_{\cdot}, \end{aligned} \quad (5.1)$$

where we follow the convention that “ \cdot ” always indexes over unexposed units or time periods. Throughout our analysis, we assume that $Y_{it} = L_{it} + \varepsilon_{it}$ as below, and study the properties of $\hat{Y}_{NT}^{\text{sdid}}$ as an estimator for L_{NT} .

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

We consider two distinct sets of conditions: First, we examine the case where the fixed effects model is well specified, i.e., $L_{it} = \mu + \alpha_i + \beta_t$, and show that synthetic difference in differences is consistent under very flexible conditions. The main point here is that using data-adaptive weights $\hat{\omega}$ and $\hat{\lambda}$ does not break difference in differences when the outcome model is well specified. Second, we consider the generalized fixed effects model, i.e., where \mathbf{L} is only assumed to have rank $r \ll \min\{N, T\}$. Here, basic difference in differences is inconsistent; however, we show that synthetic difference in differences with penalized synthetic control weights is consistent.

5.1 Properties in the Well-Specified Fixed Effects Model

Our first result shows that synthetic difference-in-differences is consistent and asymptotically normal in the well-specified model using the following kernel weights:

$$\hat{\omega}_i = \frac{1 \left(\left\{ \frac{1}{T-1} \|\mathbf{Y}_{i\cdot} - \mathbf{Y}_{N\cdot}\|_2^2 \leq c_\omega \right\} \right)}{\sum_{i \neq N} 1 \left(\left\{ \frac{1}{T-1} \|\mathbf{Y}_{i\cdot} - \mathbf{Y}_{N\cdot}\|_2^2 \leq c_\omega \right\} \right)}, \quad \hat{\lambda}_t = \frac{1 \left(\left\{ \frac{1}{N-1} \|\mathbf{Y}_{:t} - \mathbf{Y}_{:T}\|_2^2 \leq c_\lambda \right\} \right)}{\sum_{t \neq T} 1 \left(\left\{ \frac{1}{N-1} \|\mathbf{Y}_{:t} - \mathbf{Y}_{:T}\|_2^2 \leq c_\lambda \right\} \right)}, \quad (5.2)$$

¹Gaussianity of errors does not play an important role in our analysis, and it is plausible that our results also hold for heteroskedastic and auto-correlated sub-Gaussian errors.

for $i = 1, \dots, N-1$ and $t = 1, \dots, T-1$, where c_ω and c_λ are tuning parameters. In contrast most of the earlier results are based on properties of tests under randomization distributions, e.g., Abadie, Diamond, and Hainmueller [2010], Hahn and Shi [2016]. 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.

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. RVs 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: the estimator (5.1) is consistent, i.e., $\hat{Y}_{NT}^{\text{sdid}} - L_{NT} \rightarrow_p 0$. and

$$\frac{1}{\sqrt{\|\hat{\omega}_\cdot\|_2 + \|\hat{\lambda}_\cdot\|_2}} \left(\hat{Y}_{NT}^{\text{sdid}} - L_{NT} \right) \rightarrow \mathcal{N}(0, \sigma^2) \quad (5.3)$$

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

Note that c_λ and c_ω 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.

5.2 Double Robustness Part I: Consistency with a Correct Model

Next we show consistency under much weaker conditions on the weights, still assuming the fixed-effects model is correct. 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 $\hat{\omega}_\cdot \perp \mathbf{Y}_{\cdot T}$ and $\hat{\lambda}_\cdot \perp \mathbf{Y}_{N\cdot}$.

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

$$\|\hat{\omega}_\cdot\|_2, \|\hat{\lambda}_\cdot\|_2 \rightarrow_p 0, \quad \sqrt{\max\{N, T\}}\|\hat{\omega}_\cdot\|_2\|\hat{\lambda}_\cdot\|_2 \rightarrow_p 0, \quad (5.4)$$

the estimator (5.1) is consistent, i.e., $\hat{Y}_{NT}^{\text{sdid}} - L_{NT} \rightarrow_p 0$.

5.3 Double Robustness Part II: Consistency with the Factor Model

In this section we relax the modeling assumptions from the above section, and simply require that \mathbf{L} be a low-rank matrix, i.e.,

$$Y_{it} \sim \mathcal{N}(L_{it}, \sigma^2), \quad \text{rank}\{L_{it}\} \ll \min\{N, T\}. \quad (5.5)$$

This type of model was used to motivate the SC approach by Abadie, Diamond, and Hainmueller [2010], and has also been studied in other context 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:

$$\begin{aligned} \hat{\omega}^{\text{sc}}(a_\omega) &= \arg \min_{\omega \in \mathbb{W}} \left\{ \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{i=1}^{N-1} \omega_i Y_{it} \right)^2 : \|\omega_\cdot\|_2 \leq a_\omega \right\}, \\ \hat{\lambda}^{\text{sc}}(a_\lambda) &= \arg \min_{\lambda \in \mathbb{L}} \left\{ \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2 : \|\lambda_\cdot\|_2 \leq a_\lambda \right\}, \end{aligned} \quad (5.6)$$

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 synthetic control weights $\hat{\omega}$ and $\hat{\lambda}$ were chosen to balance rows and columns of \mathbf{Y} ; however, what we really need for useful inference in the model (5.5) is for $\hat{\omega}$ and $\hat{\lambda}$ to balance the the rows and columns of \mathbf{L} . Furthermore, the weights defined in (5.6) have a complicated dependence on the noise $\varepsilon = \mathbf{Y} - \mathbf{L}$, 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 use modern empirical process arguments following Mendelson [2014] to establish conditions under which our synthetic difference in differences estimator with data-dependent weights (5.6) is consistent in the low-rank model (5.5). As an additional benefit, we also prove that the basic synthetic control 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 synthetic control weights that balance the underlying effect matrix \mathbf{L} rather than the observations \mathbf{Y} :

$$\begin{aligned}\omega^*(a_\omega) &= \arg \min_{\omega \in \mathbb{W}} \left\{ \sum_{t=1}^{T-1} \left(L_{Nt} - \sum_{i=1}^{N-1} \omega_i L_{it} \right)^2 : \|\omega\|_2 \leq a_\omega \right\}, \\ \lambda^*(a_\lambda) &= \arg \min_{\lambda \in \mathbb{L}} \left\{ \sum_{i=1}^{N-1} \left(L_{iT} - \sum_{t=1}^{T-1} \lambda_t L_{it} \right)^2 : \|\lambda\|_2 \leq a_\lambda \right\},\end{aligned}\tag{5.7}$$

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, synthetic difference in differences with penalized synthetic control weights is consistent.

Assumption 4. *For our chosen a_ω and a_λ , the weights (5.7) satisfy:*

$$\begin{aligned}L_{NT} - (\omega^*(a_\omega) \cdot \mathbf{L}_{:T} + \lambda^*(a_\lambda) \cdot \mathbf{L}_{N:} - \omega^*(a_\omega)' \mathbf{L}_{::} \lambda^*(a_\lambda)) &\rightarrow 0, \\ \|\mathbf{L}_{::}' \omega^*(a_\omega) - \mathbf{L}_{N:}\|_2^2 &\leq \frac{N}{\|\mathbf{L}_{::}\|_{op}}, \quad \|\mathbf{L}_{::} \lambda^*(a_\lambda) - \mathbf{L}_{:T}\|_2^2 \leq \frac{T}{\|\mathbf{L}_{::}\|_{op}}.\end{aligned}$$

Theorem 3. *Given Assumption 1, suppose moreover that $N/T \rightarrow \rho$, and that we choose weights via (5.6) with a_ω and a_λ satisfying Assumption 4 and the conditions $\log(T)^{1/2} a_\omega \rightarrow 0$ and $\log(N)^{1/2} a_\lambda \rightarrow 0$. Writing $\mathbf{L}_{::}^{-1}$ for the Moore-Penrose pseudo-inverse of \mathbf{L} , we assume that $\sqrt{N+T} \|\mathbf{L}_{::}^{-1}\|_{op} \rightarrow 0$, and that $\|\mathbf{L}_{::}\|_{op} \|\mathbf{L}_{::}^{-1}\|_{op} = \mathcal{O}(1)$. Then,*

$$\hat{Y}_{NT}^{\text{sdid}} - L_{NT} \rightarrow_p 0.$$

The key technical result underlying Theorem 3 is the following lemma, which establishes convergence of the feasible synthetic control weights (5.6) that balance Y to the infeasible weights (5.7) that balance \mathbf{L} . In particular we show both that $\hat{\lambda}$ and λ^* are close in L_2 , and that they balance $\mathbf{L}_{\cdot:}$ in the same way. An analogous result holds for $\hat{\omega}$; then, the proof of Theorem 3 uses both bounds in a “product of rates” type of argument.

Lemma 4. *Let Λ be a convex subset of \mathbb{L} , and let λ^* be the oracle minimizer over \mathbb{L} as in (5.7). Moreover, assume that $N/T \rightarrow \rho > 0$, that $\mathbf{L} \neq 0$, that $\|\mathbf{L}_{\cdot:}\lambda^* - \mathbf{L}_{:T}\|_2^2 \lesssim \|\mathbf{L}_{\cdot:}\|_{op}\|\lambda^*\|_2$, and that $\sqrt{N+T}\|\mathbf{L}_{\cdot:}^{-1}\|_{op} \rightarrow 0$. Then the constrained least squares estimator $\hat{\lambda} = \min_{\lambda \in \Lambda} \|Y_{\cdot:}\lambda - Y_{:T}\|^2$ satisfies the bounds*

$$\begin{aligned} \|\mathbf{L}_{\cdot:}(\hat{\lambda} - \lambda^*)\|_2 &= \mathcal{O}_P(r_\star), \quad \|\hat{\lambda} - \lambda^*\|_2 = \mathcal{O}_P(N^{-1/2}r_\star), \quad \text{where} \\ r_\star &= \mathcal{O}\left(\sqrt{\log(T)\|\mathbf{L}_{\cdot:}\|_{op}\|\lambda^*\|_2}\right). \end{aligned} \tag{5.8}$$

Finally, as discussed above, we can also use Lemma 4 to prove consistency of synthetic control estimation in the low-rank model (5.5) under a strengthening of the assumptions in Theorem 3. The main difference between the assumptions of Theorem 3 and Theorem 5 is first that we need to assume consistency of the oracle synthetic control estimator (5.9) rather than consistency of the oracle difference-in-differences estimator as in Assumption 4; and second that, in (5.10), we assume that there exists a way of weighting of columns that can express $\mathbf{L}_{:T}$ well while placing roughly weight on all the columns. These stronger assumptions reflect the fact that because plain synthetic controls are not a double-debiasing method, we do not benefit from product-of-rates bounds and so have larger bias terms.

Theorem 5. *Given Assumption 1, suppose that $N/T \rightarrow \rho$, that we choose $\hat{\omega}$ via (5.6) with a_ω satisfying $\log(T)^{1/2}a_\omega \rightarrow 0$ and*

$$\|\mathbf{L}'_{\cdot:}\omega^*(a_\omega) - \mathbf{L}_{:T}\|_2^2 \leq \frac{N}{\|\mathbf{L}_{\cdot:}\|_{op}},$$

and that the oracle synthetic control estimator is consistent

$$\omega^*(a_\omega) \cdot \mathbf{L}_{:T} - L_{NT} \rightarrow 0. \tag{5.9}$$

Suppose, moreover, that there exists a weight vector $\tilde{\lambda} \in \mathbb{R}^{N-1}$ such

$$\left\| \mathbf{L}_{::} \tilde{\lambda} - \mathbf{L}_{:T} \right\|_2^2 \leq \frac{T}{\|\mathbf{L}_{::}\|_{op}}, \quad \|\mathbf{L}_{::}\|_{op} \left\| \tilde{\lambda} \right\|_2^2 = \mathcal{O}(1). \quad (5.10)$$

Writing $\mathbf{L}_{::}^{-1}$ for the Moore-Penrose pseudo-inverse of \mathbf{L} , we assume that $\sqrt{N+T} \|\mathbf{L}_{::}^{-1}\|_{op} \rightarrow 0$, and that $\|\mathbf{L}_{::}\|_{op} \|\mathbf{L}_{::}^{-1}\|_{op} = \mathcal{O}(1)$. Then

$$|\hat{\omega}_{\cdot}(a_{\omega}) \cdot Y_{:T} - L_{NT}| \rightarrow_P 0. \quad (5.11)$$

6 An Application

In one of the seminal studies on SC methods, Abadie, Diamond, and Hainmueller [2010] focus on estimating the causal effect of anti-smoking legislation in California (Proposition 99). We reanalyze these data to compare the DID, SC, and SDID estimators in a realistic setting. We follow Abadie, Diamond, and Hainmueller [2010] in using per capita smoking as the outcome. We use 39 states and the 17 pre-legislation years so that we can compare estimates to true values. Taking one state at a time, we use the years 1980 through 1988 as the treated years. For example, when using Arizona at the treated state and 1985 as the treated year, we use the other 38 states as the control states and the years 1970-1984 as the pre-treatment years. We then use the DID, SC, and SDID methods to construct $\hat{Y}_{AZ,1985}$, and compare that to the actual value $Y_{AZ,1985}$. We then calculate the square of the average squared error:

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

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

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

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

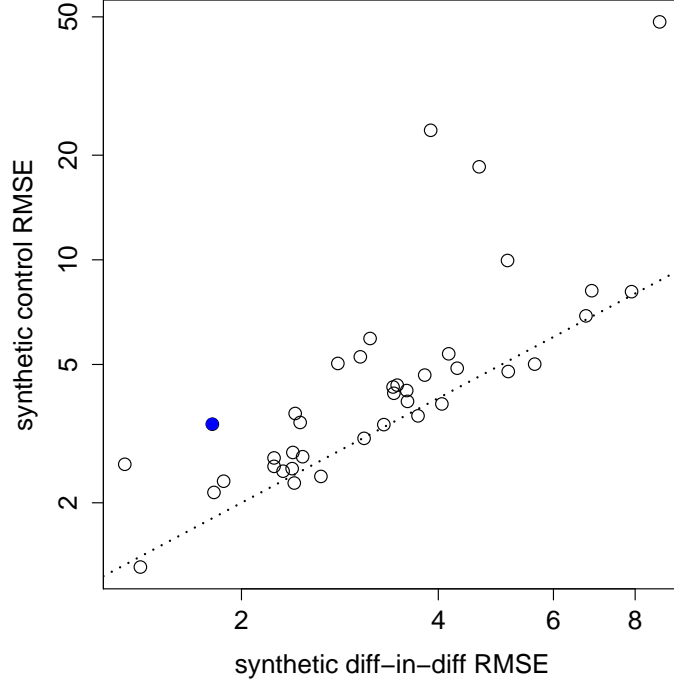
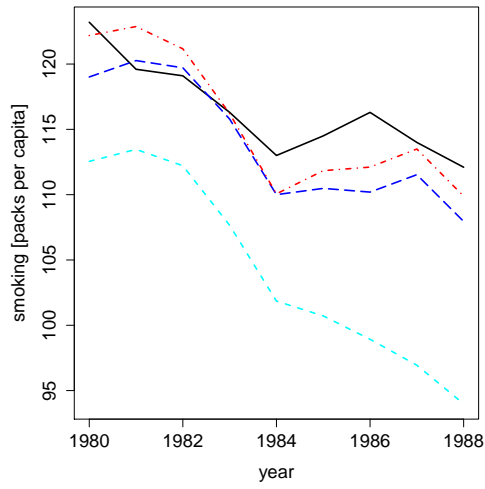


Figure 1: Comparison of the per-state root-mean squared error for synthetic difference in differences and synthetic controls. California is highlighted in blue.

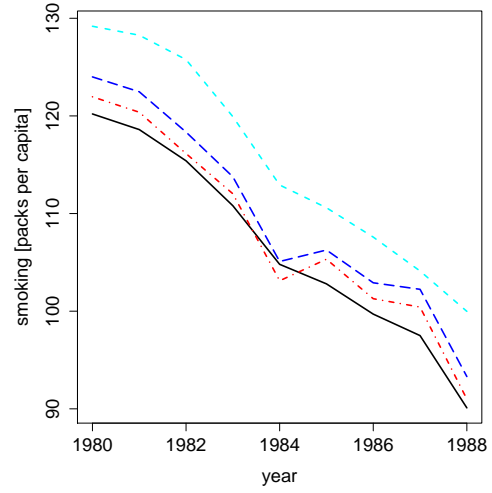
in smoking rates.

We report the results 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.

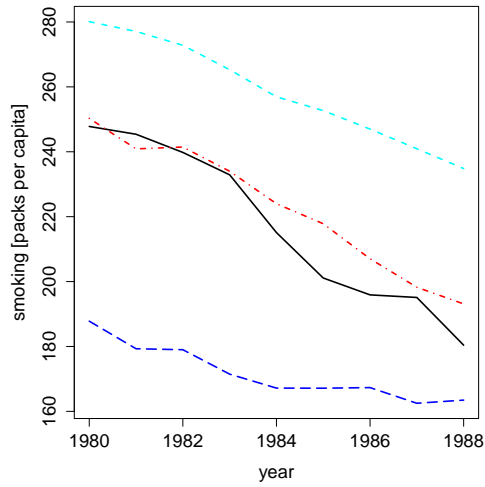
We can gain further insight into the behavior of different methods by comparing the one-step-ahead predicted trajectories $\hat{Y}_{i,t}$ to the true ones $Y_{i,t}$. We see that synthetic controls struggle when a state doesn't fit neatly within the convex hull of other states (e.g., in the case of Utah), 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 synthetic controls 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 synthetic controls as an inappropriate method to use here. However, we find that synthetic difference in differences out-performs synthetic controls in states where the latter are appropriate (such as California),



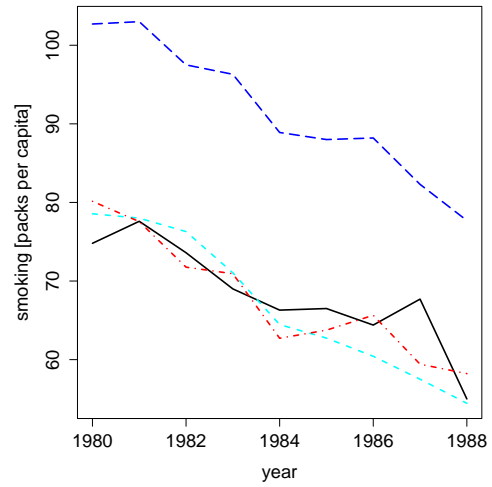
Alabama



California



New Hampshire



Utah

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.

and remain robust in cases where the latter are not (such as Utah).

7 Simulation Results

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. 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 \mathbf{L} is low rank, $\mathbf{L} = \mathbf{UV}^\top$, where $\mathbf{U} \in \mathbb{R}^{N \times R}$ and $\mathbf{V} \in \mathbb{R}^{T \times R}$ for a rank parameter R .

The key choice is in how we generate this low-rank matrix \mathbf{L} . First, we consider a simulation where the N -th row and the N -th column of \mathbf{L} are “typical”; formally, we generate \mathbf{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 synthetic control weights as in (6.1), with ζ set to the sample variance of the Y_{it} . Below, we first generate a random \mathbf{L} , and then simulate \mathbf{Y} 20 times given this \mathbf{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 synthetic controls ($\zeta = 0$), 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 better root-mean-squared-error.

8 Confidence Intervals via Weighted Regression

Finally, we study the properties of confidence intervals derived via the weighted regression perspective of synthetic difference in differences. We do not (yet) have formal results to motivate inference in this setting; however, simulation results appear promising. We work in the same data-generating distribution as for the “non-exchangeable” example in Section 7, except now

N	T	σ	rank	root-mean sq. error			mean absolute bias		
				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 \mathbf{L} and penalized synthetic control weights. Results are aggregated over 400 draws of the low-rank \mathbf{L} matrix and 25 draws of \mathbf{Y} for each \mathbf{L} (for a total of 10,000 simulation replications).

N	T	σ	rank	root-mean sq. error			mean absolute bias		
				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 \mathbf{L} and penalized synthetic control weights. Results are aggregated over 400 draws of the low-rank \mathbf{L} matrix and 25 draws of \mathbf{Y} for each \mathbf{L} (for a total of 10,000 simulation replications).

with multiple treated units. Units $1 : \dots, N_0$ are control units, and units $N_0+1, \dots, N_0+N_1 = N$ are treated from period $T_0 + 1$ onwards. Writing W_{it} for the treatment indicator, we draw data as

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

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

$$\hat{\omega}_{1:N_0}^{\text{sc}} = \arg \min \left\{ \frac{1}{T_0} \sum_{t=1}^{T_0} \left(\frac{1}{N_1} \sum_{j=N_0+1}^N Y_{jt} - \sum_{i=1}^{N_0} \omega_i Y_{it} \right)^2 + \frac{\zeta}{N_1} \|\omega\|_2^2 : \omega_i \geq 0, \sum_{i=1}^{N_0} \omega_i = 1 \right\}, \quad (8.1)$$

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

$$\hat{\tau} = \arg \min \left\{ \sum_{i,t} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \hat{\omega}_i \hat{\lambda}_t \right\}. \quad (8.2)$$

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.

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 (8.2) 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 SDID ones are not.

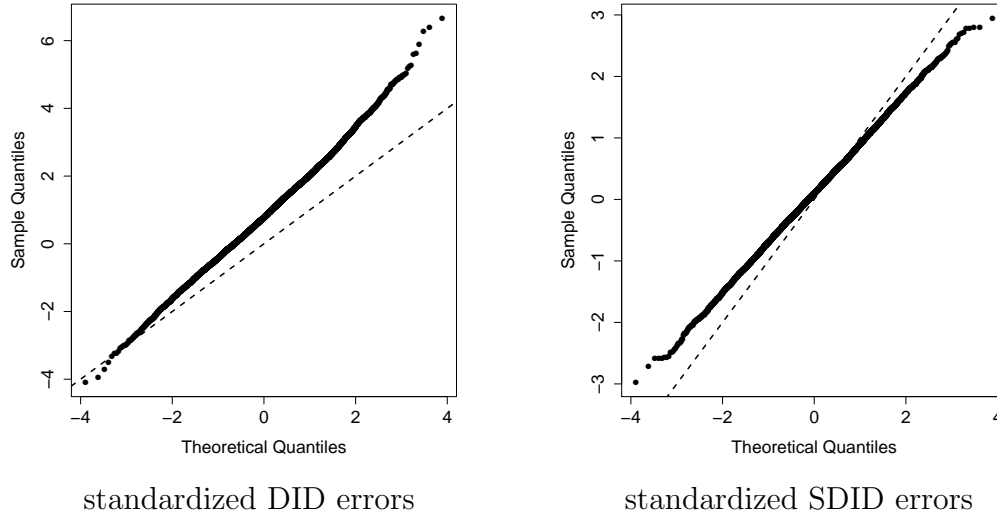


Figure 3: Standard Gaussian QQ-plot of the standardized errors $(\hat{\tau} - \tau)/\widehat{\text{Var}}[\hat{\tau}]^{1/2}$, for both DID and SDID, aggregated across 10,000 simulation replications. Points along the diagonal (dashed) would indicate perfectly calibrated Gaussian standard errors. Points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

9 Conclusion

We present a new estimator in a Synthetic Control setting which can be interpreted as a weighted Difference In Differences 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 results.

References

- Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.
- Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.
- 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.
- Dmitry Arkhangelsky and Guido Imbens. The role of the propensity score in fixed effect models. Technical report, National Bureau of Economic Research, 2018.
- 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.
- 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.
- 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.
- Clément de Chaisemartin and Xavier D’Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. 2018.
- 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 and Charles Stein. The jackknife estimate of variance. The Annals of Statistics, pages 586–596, 1981.
- 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.
- 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.
- 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.
- Bruce D Meyer, W Kip Viscusi, and David L Durbin. Workers’ compensation and injury duration: evidence from a natural experiment. The American Economic Review, pages 322–340, 1995.
- 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.
- 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.
- 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.

10 Appendix

Throughout the proofs section, we omit the “:” subscript from ω and λ when there is no risk of ambiguity.

10.1 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 :

$$\hat{Y}_{NT}^{\text{sdid}} - L_{NT} = \hat{\omega} \cdot \varepsilon_{:T} + \hat{\lambda} \cdot \varepsilon_{N:} - \hat{\omega}' \varepsilon_{::} \hat{\lambda}.$$

Next, by Assumption 3, we know that

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

and so by the first part of (5.4) 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

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

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

	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, synthetic controls, and synthetic difference in differences. Results are averaged over the time period 1980-1988.

n	T	σ	rank	root-mean sq. error			mean absolute bias		
				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 \mathbf{L} and **unpenalized** synthetic control weights. Results are aggregated over 400 draws of the low-rank \mathbf{L} matrix and 25 draws of Y for each \mathbf{L} (for a total of 10,000 simulation replications).

n	T	σ	rank	root-mean sq. error			mean absolute bias		
				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 \mathbf{L} and **unpenalized** synthetic control weights. Results are aggregated over 400 draws of the low-rank \mathbf{L} matrix and 25 draws of Y for each \mathbf{L} (for a total of 10,000 simulation replications).

10.2 Proof of Theorem 1

We start with the following high-level lemma.

Lemma 6. *Suppose that Assumption 1 is satisfied, further assume that the following conditions hold:*

$$\begin{aligned}
\|\omega^*\|_2 &= o(1) \\
\|\lambda^*\|_2 &= o(1) \\
\|\hat{\omega}\|_2 &= O_p(\|\omega^*\|_2) \\
\|\hat{\lambda}\|_2 &= O_p(\|\lambda^*\|_2) \\
\|\hat{\lambda} - \lambda^*\|_2 &= o_p(\|\lambda^*\|_2) \\
\|\hat{\omega} - \omega^*\|_2 &= o_p(\|\omega^*\|_2) \\
\|\hat{\omega} - \omega^*\|_2 \sqrt{\|\hat{\omega} - \omega^*\|_0 \log(N)} &= o_p(1) \\
\|\hat{\lambda} - \lambda^*\|_2 \sqrt{\|\hat{\lambda} - \lambda^*\|_0 \log(T)} &= o_p(1)
\end{aligned} \tag{10.1}$$

Then we have the following result:

$$\hat{\omega}' \varepsilon_{::} \hat{\lambda} = o_p(\max\{\|\omega^*\|_2, \|\lambda^*\|_2\}) \tag{10.2}$$

Proof. Define the following objects:

$$\begin{aligned}
\zeta &:= \hat{\omega}^T \Sigma_{::} \hat{\lambda} \\
v_1 &= \|\hat{\omega} - \omega^*\|_2 \\
v_2 &:= \|\hat{\lambda} - \lambda^*\|_2 \\
l_1 &:= \|\hat{\omega} - \omega^*\|_0 \\
l_2 &:= \|\hat{\lambda} - \lambda^*\|_0
\end{aligned} \tag{10.3}$$

We can decompose ζ :

$$\zeta = (\omega^*)^T \varepsilon_{::} \lambda^* + (\omega^*)^T \varepsilon_{::} (\hat{\lambda} - \lambda^*) + (\hat{\omega} - \omega^*)^T \varepsilon_{::} \lambda + (\hat{\omega} - \omega^*)^T \varepsilon_{::} (\hat{\lambda} - \lambda^*) =: \xi_1 + \xi_2 + \xi_3 + \xi_4 \quad (10.4)$$

Our goal is to show that these terms are negligible:

$$\xi_k = o_p(\max\{\|\omega\|_2, \|\lambda\|_2\}) \quad (10.5)$$

For the first term we get the following:

$$\xi_1 \sim \mathcal{N}(0, \|\omega^*\|_2 \|\lambda^*\|_2) \Rightarrow \xi_1 = O_p(\|\omega^*\|_2 \|\lambda^*\|_2) = o_p(\max\{\|\omega\|_2, \|\lambda\|_2\}) \quad (10.6)$$

For the second term we have the following:

$$\begin{aligned} \mathbb{E}[\{ |(\omega^*)^T \Sigma (\hat{\lambda} - \lambda^*)| \geq t \}] &\leq \mathbb{E}[\{ \sup_{x: \|x\|_2=v, \|x\|_0=l_2} |(\omega^*)^T \varepsilon_{::} x| \geq t \}] \leq \\ &\binom{T}{l_2} \mathbb{E}[\{ \sup_{x: \|x\|_2=v} |(\omega^*)^T \varepsilon_{:(l_2)} x_{(l_2)}| \geq t \}] \leq C_1 \binom{T}{l_2} \exp\left(C_2 l_2 - \frac{t^2}{C_3 \sigma^2 \|\omega\|_2^2 v^2}\right) \end{aligned} \quad (10.7)$$

where $\varepsilon_{:(l_2)}$ is the submatrix of $\varepsilon_{::}$ with l_2 columns and $x_{(l_2)}$ is the subvector of x with l_2 components. The last inequality follows from the fact that $\varepsilon_{::}$ is a gaussian matrix. This implies the order for ξ_2 :

$$\xi_2 = O_p(v_2 \|\omega^*\|_2 \sqrt{l_2 \log(T)}) = o_p(\|\omega^*\|_2) \quad (10.8)$$

The third term is analogous:

$$\xi_3 = O_p(v_1 \|\lambda^*\|_2 \sqrt{l_1 \log(N)}) = o_p(\|\lambda^*\|_2) \quad (10.9)$$

For the last term we use the same argument:

$$\begin{aligned}
& \mathbb{E} \left[\left\{ \sup_{x: \|x\|_2=v_1, \|y\|_2=v_2, \|x\|_0=l_1, \|y\|_0=l_2} |x^T \varepsilon_{::y}| \geq t \right\} \right] \leq \\
& \binom{T}{l_2} \binom{n}{l_1} \mathbb{E} \left[\left\{ \sup_{x: \|x\|_2=v_1, \|y\|_2=v_2} |x_{(l_1)}^T \varepsilon_{(l_1), (l_2)} y_{(l_2)}| \geq t \right\} \right] \leq \\
& \binom{T}{l_2} \binom{n}{l_1} \mathbb{E} \left[\left\{ \|\varepsilon_{(l_1), (l_2)}\| - \mathbb{E}[\|\varepsilon_{(l_1), (l_2)}\|] \geq \frac{t}{v_1 v_2} - \|\varepsilon_{(l_1), (l_2)}\|_{op} \right\} \right] \leq \\
& C_1 \binom{T}{l_2} \binom{n}{l_1} \exp \left(- \frac{\left(\frac{t}{v_1 v_2} - \|\varepsilon_{(l_1), (l_2)}\|_{op} \right)^2}{C_2 \sigma^2} \right) \quad (10.10)
\end{aligned}$$

Using the fact that $\|\varepsilon_{(l_1), (l_2)}\| = O(\sqrt{\max\{l_1, l_2\}})$ we get the rate for the last term:

$$\begin{aligned}
\xi_4 &= O_p \left(v_1 v_2 \sqrt{l_1 \log(N) + l_2 \log(T)} \right) = \\
& o_p(\max\{\|\omega^*\|_2, \|\lambda^*\|_2\} \sqrt{\min\{v_1^2, v_2^2\} (l_1 \log(N) + l_2 \log(T))}) = o_p(\max\{\|\lambda\|_2, \|\omega\|_2\}) \quad (10.11)
\end{aligned}$$

This shows that $\zeta = o_p(\max\{\|\omega^*\|_2, \|\lambda^*\|_2\})$. □

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

$$\omega_i^* = \frac{1(\{(\alpha_i - \alpha_N)^2 \leq \tilde{c}_\omega\})}{\sum_{i \neq N} 1(\{(\alpha_i - \alpha_N)^2 \leq \tilde{c}_\omega\})}, \quad \lambda_t^* = \frac{1(\{(\beta_t - \beta_T)^2 \leq \tilde{c}_\lambda\})}{\sum_{t \neq T} 1(\{(\beta_t - \beta_T)^2 \leq \tilde{c}_\lambda\})}, \quad (10.12)$$

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

First we verify that conditions for Lemma 6 hold for $\hat{\omega}$ and ω^* . Results for $\hat{\lambda}$ and λ^* follow in the same way. Define the following random variables:

$$\begin{aligned}
K &= \sum_{i \neq N} 1(\{(\alpha_i - \alpha_N)^2 \leq \tilde{c}_\omega\}) \\
\hat{K} &= \sum_{i \neq N} 1 \left(\left\{ \frac{1}{T-1} \|Y_{i:} - Y_{N:}\|_2^2 \leq c_\omega \right\} \right) \quad (10.13)
\end{aligned}$$

By definition we have the following:

$$\begin{aligned}\omega_j^* &= \frac{1\{\delta_j^2 \leq \tilde{c}_\omega\}}{K} \\ \hat{\omega}_j &= \frac{1\{\hat{\delta}_j^2 \leq c_\omega\}}{\hat{K}}\end{aligned}\tag{10.14}$$

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

$$l = \|\omega^* - \hat{\omega}\|_0\tag{10.15}$$

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

$$\begin{aligned}l &= \mathcal{O}_p(\mu_N) \\ \mu_N &:= (N - 1)\mathbb{E}\left[1\{1\{\delta_j^2 \leq \tilde{c}_\omega\} \neq 1\{\hat{\delta}_j^2 \leq c_\omega\}\}\right]\end{aligned}\tag{10.16}$$

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:

$$\mu_N = O\left((N - 1)\frac{f_{\delta^2}(\tilde{c}_\omega)}{\sqrt{T}}\right)\tag{10.17}$$

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

$$l = \mathcal{O}_p\left(\frac{N - 1}{\sqrt{\tilde{c}_\omega T}}\right)\tag{10.18}$$

By construction we have the following:

$$K = \mathcal{O}_p\left((N - 1)F_{\delta^2}(\tilde{c}_\omega)\right)\tag{10.19}$$

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

$$\|\omega^*\|_2 = \frac{1}{\sqrt{K}} = o_p(1)\tag{10.20}$$

We have the following relationship:

$$\begin{aligned} |K - \hat{K}| &\leq l \\ \frac{l}{K} &= \mathcal{O}_p\left(\frac{1}{\sqrt{T}\tilde{c}_\omega}\right) = o_p(1) \end{aligned} \quad (10.21)$$

that implies

$$\frac{\hat{K}}{K} = \mathcal{O}_p\left(1 + \frac{l}{K}\right) = \mathcal{O}_p(1 + o_p(1)) = \mathcal{O}_p(1) \quad (10.22)$$

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

$$\tilde{\omega}_j = \frac{\{\hat{\delta}_j^2 \leq c_\omega\}}{K} \quad (10.23)$$

We have bounds on the squared norms:

$$\begin{aligned} \|\tilde{\omega} - \omega^*\|_2^2 &= \frac{1}{K} \left(\frac{l}{K}\right) = o_p(\|\omega^*\|_2^2) \\ \|\tilde{\omega} - \hat{\omega}\|_2^2 &= \hat{K} \left(\frac{1}{K} - \frac{1}{\hat{K}}\right)^2 \leq \frac{1}{\hat{K}} \left(\frac{l}{K}\right)^2 = o_p(\|\tilde{\omega} - \omega^*\|_2^2) \end{aligned} \quad (10.24)$$

Finally, we have the following:

$$\|\hat{\omega} - \omega^*\|_2 \sqrt{\|\hat{\omega} - \omega^*\|_0 \log(N)} = \mathcal{O}_p\left(\frac{l}{K} \log(N)\right) = \mathcal{O}_p\left(\frac{\log(N)}{\tilde{c}_\omega \sqrt{T}}\right) = o_p(1) \quad (10.25)$$

This implies that the conditions of Lemma 1 are satisfied and thus estimator has the following decomposition:

$$\begin{aligned} \hat{Y}_{NT}^{\text{sdid}} - L_{NT} &= \hat{\omega} \cdot \varepsilon_{:T} + \hat{\lambda} \cdot \varepsilon_{N:} - \hat{\omega}' \varepsilon_{:, \hat{\lambda}} = \\ &= \omega^* \cdot \varepsilon_{:T} + \lambda^* \cdot \varepsilon_{N:} + (\hat{\omega} - \omega^*) \cdot \varepsilon_{:T} + (\hat{\lambda} - \lambda^*) \cdot \varepsilon_{N:} + o_p(\max\{\|\omega^*\|_2, \|\lambda^*\|_2\}) = \\ &= \omega^* \cdot \varepsilon_{:T} + \lambda \cdot \varepsilon_{N:} + o_p(\max\{\|\omega^*\|_2, \|\lambda^*\|_2\}) \end{aligned} \quad (10.26)$$

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

10.3 Proof of Theorem 3

For any vectors $\omega \in \mathbb{R}^{N-1}$, $\lambda \in \mathbb{R}^{T-1}$, let

$$\delta = [L_{N:}\lambda^* + (\omega^*)'L_{:T} - (\omega^*)'L_{::}\lambda^*] - L_{NT}. \quad (10.27)$$

This is the error of an infeasible estimator essentially of the form we consider. Then our estimator's error is the difference between our estimator and this infeasible estimator, plus the infeasible estimator's error δ , i.e.

$$\begin{aligned} & \hat{Y}_{NT}^{\text{sdid}} - L_{NT} \\ &= \left[Y_{N:}\hat{\lambda} + \hat{\omega}'Y_{:T} - \hat{\omega}'Y_{::}\hat{\lambda} \right] - [(Y_{N:} - \varepsilon_{N:})\lambda^* + (\omega^*)'(Y_{:T} - \varepsilon_{:T}) - (\omega^*)'(Y_{::} - \varepsilon_{::})\lambda^*] + \delta \\ &= Y_{N:}(\hat{\lambda} - \lambda^*) + (\hat{\omega} - \omega^*)'Y_{:T} - \left[(\hat{\omega} - \omega^*)'Y_{::}(\hat{\lambda} - \lambda^*) + (\omega^*)'Y_{::}(\hat{\lambda} - \lambda^*) + (\hat{\omega} - \omega^*)'Y_{::}\lambda^* \right] \\ &\quad + \delta - \varepsilon'_{N:}\lambda^* - (\omega^*)'\varepsilon_{:T} + (\omega^*)'\varepsilon_{::}\lambda^* \\ &= (Y_{N:} - (\omega^*)'Y_{::})(\hat{\lambda} - \lambda^*) + (\hat{\omega} - \omega^*)'(Y_{:T} - Y_{::}\lambda^*) - (\hat{\omega} - \omega^*)'Y_{::}(\hat{\lambda} - \lambda^*) \\ &\quad + \delta - \varepsilon'_{N:}\lambda^* - (\omega^*)'\varepsilon_{:T} + (\omega^*)'\varepsilon_{::}\lambda^* \\ &= (L_{:N} - (\omega^*)'L_{::})(\hat{\lambda} - \lambda^*) + (\hat{\omega} - \omega^*)'(L_{:T} - L_{::}\lambda^*) + \delta \\ &\quad + (\varepsilon_{N:} - (\omega^*)'\varepsilon_{::})(\hat{\lambda} - \lambda^*) + (\hat{\omega} - \omega^*)'(\varepsilon_{:T} - \varepsilon_{::}\lambda^*) - (\hat{\omega} - \omega^*)'\varepsilon_{::}(\hat{\lambda} - \lambda^*) \\ &\quad - \varepsilon'_{N:}\lambda^* - (\omega^*)'\varepsilon_{:T} + (\omega^*)'\varepsilon_{::}\lambda^* \\ &\quad - (\hat{\omega} - \omega^*)'L_{::}(\hat{\lambda} - \lambda^*). \end{aligned}$$

Starting with the first term on the first line above, we can pair Cauchy-Schwarz with Assumption 4 and Lemma 4 to check that

$$\begin{aligned}
& \left| (L_{:,N} - (\omega^*)' L_{::})(\hat{\lambda} - \lambda^*) \right| \\
& \leq \| (L_{:,N} - (\omega^*)' L_{::}) \|_2 \left\| \hat{\lambda} - \lambda^* \right\|_2 \\
& \lesssim_P \sqrt{\frac{N}{\|L_{::}\|_{op}}} \sqrt{\frac{\log(T) \|L_{::}\|_{op} a_\lambda}{N}} \rightarrow 0.
\end{aligned}$$

The second term on the first line can be bounded analogously. The term δ is bounded by Assumption 4.

Next, the first and second terms on the second line are analogous to one another. Consider the two pieces of the first, $\varepsilon'_{N:}(\hat{\lambda} - \lambda^*)$ and $(\omega^*)'\varepsilon_{::}(\hat{\lambda} - \lambda^*)$. The first of these is gaussian with mean zero and variance $\|\hat{\lambda} - \lambda^*\|^2 \rightarrow 0$ conditionally on $\hat{\lambda}$. The second of these, as well as the last term on the line, can be handled by Chevet's inequality, $E \sup_{x \in X, y \in Y} x' \varepsilon y \lesssim \text{rad}(X)w(Y) + w(X)\text{rad}(Y)$ [Vershynin, 2018, Theorem 8.7.1]: $\hat{\lambda} - \lambda^*, \hat{\omega} - \omega^*$ are in a scaled $\|\cdot\|_1$ balls $2B_1$ with $w(2B_1) \lesssim \log(T)^{1/2}$, and also in $\|\cdot\|_2$ balls $a_\omega B_2, a_\lambda B_2$ of radius shrinking faster than $\log(T)^{1/2}$, so the products appearing in Chevet's bound go to zero. And the terms on the third line add up to a standard Gaussian with small variance. For the last term, we can use Cauchy-Schwarz to check that, under our assumption that $\|L^{-1}\|_{op} \ll 1/\sqrt{T}$,

$$\begin{aligned}
(\hat{\omega} - \omega^*)' L_{::}(\hat{\lambda} - \lambda^*) & \leq \|L^{-1}\|_{op} \|L'_{::}(\hat{\omega} - \omega^*)\|_2 \|L_{::}(\hat{\lambda} - \lambda^*)\|_2 \\
& = o_P \left(\log(T) \sqrt{\|\omega^*\|_2 \|\lambda^*\|_2} \|L^{-1}\|_{op} \|L\|_{op} \right),
\end{aligned}$$

where the last inequality follows from Lemma 4 and Assumption 4. Finally, by construction we know that $\|\omega^*\|_2 \|\lambda^*\|_2 \leq a_\omega a_\lambda \ll 1/\log(T)$, and so we are done.

10.4 Proof of Lemma 4

We will prove a more general result than the one we claimed, in which $\Lambda \subseteq \mathbb{R}^{T-1}$ may be any convex subset of a $\|\cdot\|_1$ ball. Our argument is based on the “learning without concentration” argument of Mendelson [2014].

For any $\lambda_\star \in \Lambda$, the minimizer $\hat{\lambda} \in \Lambda$ satisfies

$$\begin{aligned}
0 &\geq \|Y_{::}\hat{\lambda} - Y_{:T}\|_2^2 - \|Y_{::}\lambda_\star - Y_{:T}\|_2^2 \\
&= \|Y_{::}\hat{\lambda}\|_2^2 - \|Y_{::}\lambda_\star\|_2^2 - 2Y_{:T}'Y_{::}(\hat{\lambda} - \lambda_\star) \\
&= \|Y_{::}(\hat{\lambda} - \lambda_\star)\|_2^2 + 2(Y_{::}\lambda_\star - Y_{:T})'Y_{::}(\hat{\lambda} - \lambda_\star) \\
&\geq \underbrace{\|Y_{::}(\hat{\lambda} - \lambda_\star)\|_2^2}_{Q(\hat{\lambda} - \lambda_\star)} - 2 \underbrace{\left| (Y_{::}\lambda_\star - Y_{:T})'Y_{::}(\hat{\lambda} - \lambda_\star) + Z(\hat{\lambda} - \lambda_\star) \right|}_{M(\lambda - \lambda_\star)} - 2 \left| Z(\hat{\lambda} - \lambda_\star) \right|,
\end{aligned} \tag{10.28}$$

where $Z(\delta)$ can be any stochastic process. Define $\Lambda_\star = \Lambda - \lambda_\star$, the set of deviations from a deterministic oracle estimator λ_\star , and $r^2(\delta) = \mathbb{E}\|Y_{::}\delta\|^2 = \|L_{::}\delta\|^2 + (N-1)\kappa^2\|\delta\|^2$. Our main goal is to exhibit choices r_Q , r_M and η , as well as a stochastic process $Z(\delta)$, such that, with probability tending to 1,

$$\inf_{\substack{\delta \in \Lambda_\star \\ r(\delta) \geq r_Q}} \frac{Q(\delta)}{r^2(\delta)} \geq 3\eta, \tag{10.29}$$

$$\sup_{\substack{\delta \in \Lambda_\star \\ r(\delta) \geq r_M}} \frac{M(\delta)}{r^2(\delta)} < \eta, \tag{10.30}$$

$$\left| Z(\hat{\lambda} - \lambda_\star) \right| \leq (\eta/2) \max \{r_Q^2, r_M^2\}. \tag{10.31}$$

The first two bounds immediately imply that, with probability tending to 1,

$$Q(\delta) - 2M(\delta) > \eta \max \{r_Q^2, r_M^2\}$$

for all $r(\delta) \geq \max \{r_Q^2, r_M^2\}$; this, paired with (10.31), implies that $r(\hat{\lambda} - \lambda_\star) \leq \max \{r_Q, r_M\}$, again with probability tending to 1. And this implies that

$$\|L(\hat{\lambda} - \lambda_\star)\| = \mathcal{O}_P(\max \{r_Q, r_M\}), \text{ and } \|\hat{\lambda} - \lambda_\star\| = \mathcal{O}_P\left(\max \{r_Q, r_M\} / (\kappa\sqrt{N-1})\right).$$

We now move to the core part of the proof, which involves proving the bounds (10.29-10.31).

Notation. We will use c to denote every universal constant, which will not cause problems because what we will ultimately prove is a rate. Many of our bounds are phrased in terms of the

gaussian width, $w(S) = \mathbb{E} \sup_{s \in S} \langle g, s \rangle$ where g is a vector of iid standard gaussians, as well as the radius $rad(S) = \sup_{s \in S} \|s\|$. In some of our references, these bounds are instead phrased in terms of gaussian complexity, $\gamma(S) = \mathbb{E} \sup_{s \in S} |\langle g, s \rangle|$. Our use of $w(S)$ in its place is justified by the presence of an unspecified constant factor c and the bound $\gamma(S) \leq 2w(S)$ for sets S which, like Λ_\star , contain the origin [see e.g. Vershynin, 2018, Section 7.6.2]. We will denote by B_1 and B_2 the unit balls in norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively, and we will work with intersections $\Lambda_\star(r) = \Lambda_\star \cap rB_2$ of Λ_\star and a r -scaled unit $\|\cdot\|_2$ ball frequently. Unless otherwise specified, $\|v\|$ will mean the euclidean norm $\|v\|_2$ for a vector v and $\|A\|$ will mean the operator norm $\|A\|_{op} = \sup_{\|v\|_2 \leq 1} \|Av\|_2$.

The Z term. In what follows, we take $Z(\hat{\lambda} - \lambda_\star) = \varepsilon'_{:T} L_{::}(\hat{\lambda} - \lambda_\star)$. This process is easy to bound for our estimator $\hat{\lambda}$, as it is defined as a function of $Y_{::}$ and $Y_{N::}$, and is therefore independent of the noise $\varepsilon_{:T}$ in this process (i.e., our weights $\hat{\lambda}$ satisfy Assumption 3). To get a bound of the form (10.31), observe that the conditional distribution $Z(\hat{\lambda} - \lambda_\star) \mid Y_{::}, Y_{N::}$ is gaussian with standard deviation $s(\hat{\lambda} - \lambda_\star) = \sigma \|L_{::}(\hat{\lambda} - \lambda_\star)\|$, and $\mathbb{E} \left| Z(\hat{\lambda} - \lambda_\star) \right| = \mathbb{E} \left[s(\hat{\lambda} - \lambda_\star) \mathbb{E}[s(\hat{\lambda} - \lambda_\star)^{-1} Z(\hat{\lambda} - \lambda_\star) \mid Y_{::}, Y_{N::}] \right] = c \mathbb{E}[s(\hat{\lambda})]$ where $c = \mathbb{E} |g|$ for a unit gaussian. Because $\hat{\lambda} \in \Lambda$, $\mathbb{E} s(\hat{\lambda} - \lambda_\star) \leq c \|L_{::}\| rad(\Lambda_\star)$, and we can define a bound of the form (10.31) via Markov's inequality.

The lower bound (10.29). We begin with a three term expansion,

$$Q(\delta) = \|L_{::}\delta + \varepsilon_{::}\delta\|_2^2 = \|L_{::}\delta\|^2 + 2\delta' L'_{::} \varepsilon \delta + \|\varepsilon_{::}\delta\|^2,$$

and prove a uniform lower bound on $Q(\delta)/r(\delta)^2$ of the type

$$\inf_{\substack{\delta \in \Lambda_\star \\ r(\delta) \geq r_Q}} \frac{Q(\delta)}{r(\delta)^2} \geq \min \left\{ \frac{\|L_{::}\delta\|^2 + 2\delta' L'_{::} \varepsilon \delta}{2\|L_{::}\delta\|^2}, \inf_{\substack{\delta \in \Lambda_\star \\ r(\delta) \geq r_Q}} \frac{\|\varepsilon_{::}\delta\|^2}{2\sigma^2(N-1)\|\delta\|^2} \inf_{\delta \in \mathbb{R}^{T-1}} \right\} \quad (10.32)$$

The first term is a lower bound on $Q(\delta)/r(\delta)^2$ when $r(\delta) = \|L_{::}\delta\|$ when the first term of $r(\delta)^2$, $\|L_{::}\delta\|^2$, is at least half of $r(\delta)^2$ itself; the second term is a lower bound on $Q(\delta)/r(\delta)^2$ when it is not, i.e. when $\sigma^2(N-1)\|\delta\|^2$ is more than half of $r(\delta)^2$.

The second term above is shown in Mendelson [2014, Corollary 4.3] to have a lower bound $\eta >$

0 that holds with probability going to one for any r_Q satisfying the condition $(T-1)^{-1}w(\Lambda_\star(r)) \leq cr$.

The first term is lower bounded by $1/2 - \|\varepsilon_{::}L_{::}^{-1}\|$, as

$$|\delta' L_{::}' \varepsilon_{::} \delta| = |\delta' L_{::}' \varepsilon_{::} L_{::}^{-1} L_{::} \delta| \leq \|L_{::} \delta\|^2 \|\varepsilon_{::} L_{::}^{-1}\|.$$

It is well known that $\|\varepsilon_{::}\| = \mathcal{O}_p(\sqrt{N+T})$ [see e.g Vershynin, 2018, Theorem 4.4.5], so our condition that the pseudoinverse satisfies $\sqrt{N+T}\|L_{::}^{-1}\| \rightarrow 0$ implies that our lower bound $1/2 - \|\varepsilon_{::}\|\|L_{::}^{-1}\|$ converges to $1/2$. Thus, we may take r_Q to be the minimum of such r , i.e.,

$$r_Q = \max \left\{ \frac{1}{4}, \inf \{ r : (T-1)^{-1}w(\Lambda_\star(r)) \leq cr \} \right\}.$$

The upper bound (10.30). We will first show that this bound is implied by the bound

$$\sup_{\delta \in \Lambda_\star(r_M)} M(\delta) < \eta r_M^2. \tag{10.33}$$

We will use the properties that $M(\delta)$ is linear in δ and Λ_\star is star-shaped around zero. For any $\delta \in \Lambda_\star$ such that $r(\delta) \geq r_M$, (10.33) implies that $\tilde{\delta} = \delta \cdot r_M/r(\delta) \in \Lambda_\star(r)$ satisfies the bound $M(\tilde{\delta}) < \eta r_M^2$. Thus,

$$M(\delta) = M(\tilde{\delta}) \cdot r(\delta)/r_M < \eta r_M^2 \cdot r(\delta)/r_M = \eta r_M r(\delta) \leq \eta r^2(\delta).$$

To complete our proof, we will find r_M such that (10.33) is satisfied with high probability. To do this, we decompose $M(\delta)$ into five terms and bound each uniformly over $\Lambda_\star(r_M)$.

$$\begin{aligned} M(\delta) &= (Y_{::}\lambda_\star - Y_{:T})'Y_{::}\delta - Z(\delta) \\ &= (L_{::}\lambda_\star - L_{:T})'L_{::}\delta \end{aligned} \tag{10.34}$$

$$+ (L_{::}\lambda_\star - L_{:T})'\varepsilon_{::}\delta \tag{10.35}$$

$$+ \lambda'_\star \varepsilon'_{::} L_{::}\delta \tag{10.36}$$

$$+ \lambda'_\star \varepsilon'_{::} \varepsilon_{::}\delta \tag{10.37}$$

$$- \varepsilon'_{:T} \varepsilon_{::}\delta. \tag{10.38}$$

The first term here is deterministic, and we have the Cauchy-Schwarz bound

$$\sup_{\delta \in \Lambda_\star(r)} |(L_{::}\lambda_\star - L_{:T})'L_{::}\delta| \leq \|L_{::}\lambda_\star - L_{:T}\| \sup_{\delta \in \Lambda_\star(r)} \|L_{::}\delta\| \leq \|L_{::}\lambda_\star - L_{:T}\| r \tag{10.39}$$

In the second, the elements of the vector $(L_{::}\lambda_\star - L_{:T})'\varepsilon_{::}$ are independent and identically gaussian with mean zero and standard deviation $\sigma\|L_{::}\lambda_\star - L_{:T}\|$. Thus, if scaled by the inverse of this standard deviation, this term has the distribution of the inner product between a vector of iid unit gaussians g_i and δ . We call this the canonical gaussian process, and we define the gaussian width $w(S)$ of a set S as the expected supremum of this process over $s \in S$ [see e.g. Vershynin, 2018, Chapter 7]. Thus,

$$\mathbb{E} \sup_{\delta \in \Lambda_\star(r)} |(L_{::}\lambda_\star - L_{:T})'\varepsilon_{::}\delta| \leq \sigma\|L_{::}\lambda_\star - L_{:T}\|w(\Lambda_\star(r)). \tag{10.40}$$

The third is also a gaussian process, and analogously,

$$\mathbb{E} \sup_{\delta \in \Lambda_\star(r)} |\lambda'_\star \varepsilon'_{::} L_{::}\delta| \leq \sigma\|\lambda_\star\|w(L_{::}\Lambda_\star(r)) \tag{10.41}$$

Here $L_{::}\Lambda_\star(r)$ indicates the image of the set $\Lambda_\star(r)$ under $L_{::}$.

Now we will bound the fourth term above, (10.37), using Cauchy-Schwarz. Below, the first factor is the norm of a vector of independent gaussians with mean zero and standard deviation

$\sigma\|\lambda_\star\|$, and is bounded by $\sigma\|\lambda_\star\| \left(\sqrt{N-1} + c\sqrt{\log(1/\tau)} \right)$ on an event of probability $1 - \tau$ [see e.g. Vershynin, 2018, Theorem 3.1.1]. The second factor's deviation above its root-mean-square $\sigma\sqrt{N-1}\|\delta\|$ is bounded by $c\sqrt{\log(1/\tau)}\sigma^2w(S)$ with probability $1 - \tau$ by a uniform-over- S version of the same inequality [Liaw et al., 2017, Theorem 1.4]. Thus, with probability $1 - 2\tau$ by the union bound simultaneously for all $\delta \in \Lambda_\star(r)$, both of these bounds apply, and letting $u = \sqrt{\log(1/\tau)}$,

$$\begin{aligned}
|*| \lambda'_\star \varepsilon'_{:,} \varepsilon_{:,} \delta &\leq \sigma^2 \|\varepsilon_{:,} \lambda_\star\| \|\varepsilon_{:,} \delta\| \\
&\leq \sigma^2 \|\lambda_\star\| \left[\sqrt{N-1} + cu \right] \left[\sqrt{N-1} \|\delta\| + cuw(\Lambda_\star) \right] \\
&\leq \sigma^2 \|\lambda_\star\| \left[\sqrt{N-1} + cu \right] \left[\sigma^{-1}r + cuw(\Lambda_\star) \right]
\end{aligned} \tag{10.42}$$

Finally, we will bound the fifth term above, (10.38), using Hölder's inequality.

$$\mathbb{E} |\varepsilon'_{:T} \varepsilon_{:,} \delta| \leq \|\varepsilon'_{:T} \varepsilon_{:,}\|_\infty \|\delta\|_1. \tag{10.43}$$

In the length $T-1$ vector $\varepsilon'_{:T} \varepsilon_{:,}$, the elements are iid and distributed as σ^2 times the inner product between two isotropic gaussian vectors. This inner product is $O_p(N^{1/2})$ and subexponential [Vershynin, 2018, Lemma 6.2.2], so the maximum of $T-1$ of them is $O_p(N^{1/2} \log(T))$. Because $\|\delta\|_1$ is bounded, this is the rate for the term.

Invoking Markov's inequality to get a probability $1 - \tau$ bounds on the terms whose expectations we've calculated and then taking the union bound, we get a tail bound of the form $\sup_{\delta \in \Lambda_\star(r)} M(\delta) \leq A(r)r + B(r)$. Ignoring constant factors including σ and τ ,

$$\begin{aligned}
A(r) &\lesssim \|L_{:,} \lambda_\star - L_{:T}\| + N^{1/2} \|\lambda_\star\| \\
B(r) &\lesssim \|L_{:,} \lambda_\star - L_{:T}\| w(\Lambda_\star(r)) + \|\lambda_\star\| w(L_{:,} \Lambda_\star(r)) + N^{1/2} \|\lambda_\star\| w(\Lambda_\star(r)) + N^{1/2} \log(T)
\end{aligned} \tag{10.44}$$

(10.33) and therefore (10.30) hold on this event if r_M is any r satisfying $A(r)r + B(r) \leq (\eta/2)r^2$. To simplify our calculations, we work with the sufficient conditions $A(r) \leq (\eta/2)r$ and $B(r) \leq$

$(\eta/2)r^2$, which hold for some r_M satisfying

$$r_M \lesssim \max \left\{ \|L_{::}\lambda_\star - L_{:T}\|, N^{1/2}\|\lambda_\star\|, w(\Lambda_\star)^{1/2} \max \left\{ \|L_{::}\lambda_\star - L_{:T}\|^{1/2}, N^{1/4}\|\lambda_\star\|^{1/2} \right\}, \right. \\ \left. \|\lambda_\star\|^{1/2} w(L_{::}\Lambda_\star(r))^{1/2}, N^{1/4} \log(T)^{1/2} \right\}.$$

Pulling everything together, we have shown that

$$r_\star = \max \left\{ w^{1/2}(\Lambda), 1 \right\} \max \left\{ \|L_{::}\lambda_\star - L_{:T}\|_2, N^{1/2}\|\lambda_\star\|_2, \|L_{::}\|_{op}^{1/2}\|\lambda_\star\|_2^{1/2}, 1 \right\}. \quad (10.45)$$

Finally, the bound (5.8) then follow from a few simple observations: (i) modulo factors of $w(\Lambda_\star)^{1/2}$, the first line contains two expressions and their square roots, and the square roots will be smaller unless the expressions themselves less than one; (ii) for $N \lesssim T$, $N^{1/2}\|\lambda_\star\| \lesssim T^{1/2}\|\lambda_\star\| \leq \|\lambda_\star\|_1$, which is $O(1)$ by assumption; (iii) $w(L_{::}\Lambda_\star(r)) \leq \|L_{::}\|w(\Lambda_\star(r))$ and $w(\Lambda_\star(r)) \leq w(\Lambda_\star) = w(\Lambda)$; (iv) if we use that bound and the bound $w(\Lambda_\star) \lesssim \log(T)^{1/2}$, because $\|L_{::}\| \gtrsim \|L_{::}^{-1}\|^{-1} \gtrsim \sqrt{N+T}$ and $\|\lambda_\star\| \gtrsim N^{1/2}$, the resulting bound on the penultimate term is larger than the final term.

10.5 Proof of Theorem 5

We can express our estimator as

$$\hat{\omega} \cdot Y_{:T} = \omega^\star \cdot L_{:T} + (\hat{\omega} - \omega^\star) \cdot L_{:T} + \hat{\omega} \cdot \varepsilon_{:T}.$$

Now, the first term above is consistent for L_{NT} by (5.9), while the last one can be bounded as in the proof of Theorem 2. It remains to bound the middle term. To do so, note that

$$\begin{aligned} (\hat{\omega} - \omega^\star) \cdot L_{:T} &= (\hat{\omega} - \omega^\star)' L_{::} \tilde{\lambda} + (\hat{\omega} - \omega^\star) \cdot (L_{:T} - L_{::} \tilde{\lambda}) \\ &\leq \|L_{::}' (\hat{\omega} - \omega^\star)\|_2 \|\tilde{\lambda}\|_2 + \|\hat{\omega} - \omega^\star\|_2 \|L_{:T} - L_{::} \tilde{\lambda}\| \\ &\lesssim_P \sqrt{\log(N) \|L_{::}\|_{op} a_\omega} \|\tilde{\lambda}\|_2 + \sqrt{\frac{\log(N) \|L_{::}\|_{op} a_\omega}{T}} \|L_{:T} - L_{::} \tilde{\lambda}\|, \end{aligned}$$

where the last line follows from Lemma 4. Finally, by the assumptions made on $\tilde{\lambda}$ in (5.10), the right-hand-side of the bound above goes to 0.