Matching on Noise: Bias in the Synthetic Controls Estimator

Joseph Cummins, Douglas L. Miller, David Simon and Brock Smith

August 19, 2019

Abstract

We show that the synthetic control algorithm can have a systematic bias in a set of panel data settings commonly employed in empirical research. The bias comes from matching to idiosyncratic error terms (noise) in the treated unit and the donor units' pre-treatment outcome values. This in turn leads to a biased counterfactual for the post-treatment periods. We use a Monte Carlo analysis to illustrate the determinants of the bias in terms of error term variance, sample characteristics and DGP complexity. We present two procedures to reduce the bias: one based on a specification for the matching variables that includes estimates of unit-level polynomial trend parameters, and the other a direct computational bias-correction procedure based on re-sampling from a pilot model. Both of these can reduce the bias in empirically feasible implementations. However, both corrections increase the sampling variance and mean-squared error in our simulations relative to a baseline of matching on all pre-period outcome values.

Keywords: Synthetic Control, Over-fitting

JEL Codes: C23; C52

1 Introduction

The Synthetic Control method (Abadie and Gardeazabal, 2003) is a data-driven approach used to construct a counterfactual for a treatment unit from a pool of candidate untreated donor units. Athey and Imbens (2017) describe this method as "arguably the most important innovation in the policy evaluation literature in the last 15 years." The logic used to justify synthetic controls is simple and compelling: if the synthetic control group outcomes match those of the treated unit across the entire pre-intervention period, the group should project a valid counterfactual for the treated unit into the post-intervention period. Using a typical panel data generating process (DGP), we show that that the synthetic control group can systematically fail in this goal, leading to biased treatment effect estimates. This problem is most severe when the treated unit's unobserved structural parameters (e.g., level and trend) are away from the center of the distribution of the donor units' parameters. This bias can occur even when the synthetic control and treatment group outcomes appear to be well matched in the pre-treatment period.

The econometric work proving the consistency of the synthetic control estimator (Abadie et al., 2010) relies on a long pre-intervention period, a large number of potential donor groups, and a specific factor-based model for the unobserved heterogeneity in potential outcomes. These together guarantee that the matches generated by the algorithm will provide a good fit for the structural parameters of the "treatment group" and thus project a valid counterfactual into the post-period. In applied work, each of these assumptions may be suspect.

We focus on data environments with finite pre-intervention time periods and find that the synthetic control algorithm can regularly generate biased estimates. The bias we describe arises because the synthetic control algorithm systematically fails to match to donor units based on the underlying unobserved *structural components* of the treatment group, and instead matches in part on *idiosyncratic error terms* in the pre-period. This over-fitting to pre-period observed outcomes, combined with mean reversion in error terms in the post-period, leads to a systematic distortion in the post-period counterfactual. We evaluate the resulting bias using a series of Monte Carlo simulations, illustrate the determinants of the magnitude of bias by altering simulation parameters each in turn and explore two avenues for corrections aimed at reducing the bias. The bias arises whenever the treatment unit's unobserved parameters are away from the median relative to the distribution of control unit parameters and there is idiosyncratic noise in the DGP. The bias is decreasing in the number of control units and the length of pre-intervention period and increasing in the variance of error terms and in the extremity of the treatment unit's unobserved parameters relative to the distribution of control unit parameters.

The bias that we identify in this paper is a result of a type of "over-fitting" whereby donor unit weights are selected in part by matching on the error terms (noise) of both treatment and control units. This occurs at the cost of matching on the underlying structural parameters, and in a systematic way such that the resulting synthetic control tends to have structural parameters that are closer to the middle of the distribution than the treated unit's. This insight motivates two avenues for potential correction procedures. The first approach is to simplify the set of variables to match on. To do so we model the functional form of the underlying DGP, estimate the unit-level parameters, and match on the resulting parameter estimates. The second approach is to perform a direct bias correction. Using a parametric bootstrap, we simulate the magnitude of bias that results from matching on noise given the observable features of the data and characteristics of the treatment group relative to donor units. We then subtract this estimated bias from the main estimate. These methods perform reasonably well at reducing bias. This comes at a cost of increased variance, and in the simulations we consider this typically dominates the lowered bias, leading to higher mean squared error (MSE).

While our analysis is limited to a specific family of polynomial trend function DGPs, we believe it has relevance for both practitioners and econometricians. For practitioners, the analysis of the determinants of bias can provide guidance on when synthetic controls are more or less likely to generate unbiased estimates, given their empirical setting. Additionally, our correction procedures can provide grounds for alternative specifications and robustness checks when synthetic controls estimates are central to an empirical analysis. For econometricians, we hope that the imperfect correction procedures we outline here can provide a foundation for improved bias-correction methods or improvements in the algorithm itself.

2 Synthetic Controls: Background, Estimation and Graphical Representation

Synthetic control techniques were developed primarily as a tool for quantitative comparative case studies that exploit panel variation (e.g. state and year) to estimate the impacts of an external event (e.g. policy changes) that affected a single "treatment unit" (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015). The synthetic control method compares the outcome of interest for the unit affected by the intervention or policy to an algorithmically-determined weighted average of donor (control) units that were not affected. In this way it reduces researcher degrees of freedom over choice of control groups and allows the data itself to determine appropriate matches. As with matching methods in general, the approach works when the treated unit, absent the intervention, would have evolved in a similar way to the synthetic control group. This means that selecting the correct weights for the control units is of central importance.

2.1 Estimation

The estimation of treatment effects using synthetic control methods involves three steps. First, a weight (w_i) is chosen for each control unit. These weights are derived from an optimization procedure that minimizes the mean squared-distance between a weighted average of control group variables (X) and their corresponding values in the treated unit. Second, these weights are applied to control group observations to create a "synthetic control unit" with period-specific weighted average outcomes defined as $Y_{S,t} = \sum_{1}^{G} w_i * Y_{it}$. The "fit" of the model is generally assessed informally and visually by the ability of the synthetic control outcome time-series to match that of the treatment unit in the pre-treatment periods. Conditional on an adequate pre-treatment fit, the treatment effect is then calculated as the difference between the treated unit and the synthetic control unit in the post-treatment periods. Finally, inference is typically performed through a permutation-based procedure. These steps are outlined in detail below.

2.1.1 Model of Unit Level Time-Series

Abadie et al (2010) proposes the following model for outcome Y in period t to justify the use of synthetic controls.

$$Y_{it}^{N} = \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \tag{1}$$

Y_{it}^{N} is the value of outcome Y_{it} absent any treatment. θ_t is a vector of time-varying coefficients on observed variables Z_i, which are constant across time within a unit (i). μ_i are the unobserved (and time-invariant) factor loadings for unit i, and the vector λ_t contains period-specific common factors.

Next, let the value of Y_{it} conditional on having been exposed to the treatment be defined as:

$$Y_{i=treated,t} = Y_{it}^{N} + \delta_t D_{it} \tag{2}$$

Where D_{it} is an indicator equal to one if unit i has been exposed to the treatment in time t, and zero otherwise. We observe a balanced panel of data with G control units and a single treatment unit,[1] each observed for T periods. There are T_0 pre-treatment periods, and in period $T_0 + 1$ the treated unit is exposed to the treatment and immediately experiences the treatment effect.

[1] It is possible to run this procedure for more than one treated unit. One approach is to take averages of observables to combine all treated units into a single unit (e.g.Kreif et al., 2016).

2.1.2 Choosing Weights

At the core of the synthetic control method is an algorithm that chooses a set of weights for the control units. These weights are chosen to make the weighted average of control group variables X match their treated-unit counterparts as closely as possible. A critical step in this process is for the researcher to choose the predictor variables (X) for the synthetic control machinery to match on . Predictors are generally linear combinations of pre-period covariates and outcomes, and it is typical to choose outcomes in different parts of the pre-treatment period in order to capture secular time trends driven by unobservable factors.

Let W be a (G x 1) vector of non-negative weights that sum to one. If the data has the factor structure in equation (1) above, Abadie et al (2010) show that if there exists a W* such that:

$\sum_{i=1}^{G} w_i^* Y_{i,1} = Y_{i=treated,1}$, ..., $\sum_{i=1}^{G} w_i^* Y_{i,T_0} = Y_{i=treated,T_0}$, and $\sum_{i=1}^{G} w_i^* Z_i = Z_{i=treated}$

and if the number of pre-periods is large relative to the scale of the errors, then the following is an unbiased estimate of δ_t:

$$\hat{\delta}_t = Y_{i=treated,t>T_0} - \sum_{i=1}^{G} w_i^* Y_{i,t} \tag{3}$$

Control unit weights W are selected such that the average distance between the resulting synthetic control and the treated unit for each predictor is minimized. An optimization procedure selects weights that minimize the following function:

$(X_1 - X_0W)'V(X_1 - X_0W)$

Where X_1 is a (K x 1) vector of predictors (chosen from Z and $Y_{t \leq T_0}$) for the treatment unit, X_0 is a (K x G) matrix of the predictors for the control units, and V is a (K x K) diagonal matrix with the diagonal elements representing the importance of each predictor. W is a (G x 1) vector of time-invariant, non-negative weights to be estimated for each control unit and constrained to sum to one.[2]

[2] Doudchenko and Imbens (2016) discusses an extension to synthetic controls that relaxes the assumption of non-

2.1.3 Inference

The typical inference procedure for synthetic control estimation involves a permutation test based on the null hypothesis of no effect. The procedure uses the donor unit pool as a set of "placebo" treatment states and estimates a "placebo treatment effect" for each control unit by assigning that unit as the placebo-treatment group and the original treatment year as the placebo treatment year. The distribution of placebo estimates then acts as an estimate of the sampling distribution of the estimator under the null hypothesis of no effect. A pseudo-p-value can then be computed, based on the rank of the magnitude of the actual estimated treatment effect relative to the distribution of placebo estimates. Abadie et al. (2010) recommend selecting only those placebo estimates which have a good pre-treatment period fit.

2.2 The Development of Synthetic Control Methods

In a series of important papers, Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015), introduced social scientists to the synthetic control method. The method was first proposed in Abadie and Gardeazabal (2003) where it was applied to study the effects of political conflict in the Basque region on economic development. Abadie et al. (2010), which focused on the effects of a California anti-smoking measure on cigarette sales, further developed the algorithm and econometric grounding and introduced the permutation-based inference procedure described above. Following these papers, there has been a swell of interest in using the method on a wide range of topics in economics and other social sciences, and we briefly discuss the scope of this work in the following sub-sections.

2.2.1 Recent Empirical Work Using Synthetic Control Methods

Synthetic controls have most commonly been employed to evaluate state or region-level policy changes affecting aggregate production, health and labor market outcomes. Courtemanche and

negative weights, and we briefly discuss similar innovations in Section 2.2. Here we are discussing the standard synthetic control method proposed in Abadie et al. (2010), which is what we employ throughout the paper.

Zapata (2012) evaluate self-reported health outcomes after the implementation of universal health care in Massachusetts; Bohn et al. (2014) study the immigrant labor market using the 2006 Legal Arizona Workers Act in Arizona; Eren et al. (2011) study the passage of right-to-work laws in Idaho and Oklahoma and their effect on union membership, FDI and manufacturing employment; and Jones and Marinescu (2018) investigate the labor market impacts of universal cash income from the Alaska Permanent Fund.

Recently, researchers have expanded the domain of analyses to include the effects of a broad range of policies, such as the impact of educational reforms on pre-school (Fitzpatrick, 2008) and college (Klasik, 2013) performance, as well as criminal justice issues ranging from prostitution (Cunningham and Shah, 2017) to gun control (Donohue et al., 2019) A number of papers have alternatively used synthetic controls to study country-level events such as the effect of natural disasters (Cavallo et al., 2013), resource discovery (Smith, 2015), political systems (Nannicini and Ricciuti, 2010) and monetary policy on economic growth (Lee, 2010). Other analyses examine alternative group level units, such as universities (Hinrichs, 2014), financial firms (Acemoglu et al., 2016), grocery stores (Kiesel and Villas-Boas, 2013), and neighborhoods (Gautier et al., 2009).

2.2.2 Proposed Refinements and Methodological Considerations

Along with the growing number of synthetic control applications, a concurrent methodological literature has arisen that attempts to explore the properties of the synthetic control estimator and improve upon the technique. A number of papers attempt to relax assumptions in the original algorithm. For example, Doudchenko and Imbens (2016) propose a method that allows synthetic weights to be negative and to not sum to 1, while Kreif et al. (2016) propose a version allowing for multiple treated units. Other papers contextualize synthetic control within a broader framework of matching and difference estimators, including Xu (2017), which frames synthetic control in a fixed-effects framework, and Athey and Imbens (2017), which places the synthetic control estimator within the broader potential outcomes framework and treatment-effects estimation literature.

Closer to our own work, a few papers have attempted to evaluate specification issues directly using simulations similar in spirit to those employed here. Peri and Yasenov (2015) show that focusing on sub samples of control groups can cause inflated type 1 error rates due to measurement error; while Kaul et al. (2015) argue both theoretically and through a replication analysis that matching on all pre-period outcome variable values can be problematic.[3] Our results apply to a broader set of analyses and model specifications. Unlike Peri and Yasenov (2015), our documented bias applies even in cases where there are a relatively large number of control groups instead of a small sub-sample; and unlike Kaul et al. (2015), our bias does not depend solely on matching on all pre-period outcome values.

Our work most closely relates to recent work by Bruno Ferman and Cristine Pinto. In a series of thoughtful papers, they provide evidence from a linear factor model that weights on donor units might not converge to the treated unit weights in settings with a finite number control units (Ferman and Pinto, 2016), that visual displays of inference by placebo can be misleading (Ferman and Pinto, 2017), and that in applications with few pre-periods there is substantial room for specification hunting and cherry picking of results (Ferman et al., 2017). The closest of these papers to ours, Ferman and Pinto (2016), models bias as arising from time-invariant unobserved heterogeneity in the control units, which can be corrected for by demeaning the pre-treatment average of the outcome for each unit. In our analyses, which is based on DGPs that include special cases of the time-invariant unobserved heterogeneity described in Ferman and Pinto (2016), de-meaning is insufficient because the mismatch persists in the trend parameters. Our framework of relating the bias to idiosyncratic noise and asymmetry in weighting effects also motivates two unique paths for bias correction that do not necessarily follow from the frameworks employed in the previous literature.

[3] Our baseline estimates come from this commonly employed specification of matching on all pre-period outcome variables. However, the bias we describe persists whether we control for all pre-period outcomes or only some periods; the bias we descrbe is not the result of matching on all pre-period outcome values.

2.3 Graphical Demonstration and Data Simulation

Much like regression discontinuity and event study analyses, synthetic controls methods derive rhetorical power from the elegance and transparency of their graphical presentation. In order to bridge the background discussion above with the simulation-based results to come, we generate and analyze one realization of our simulation model and display the results.

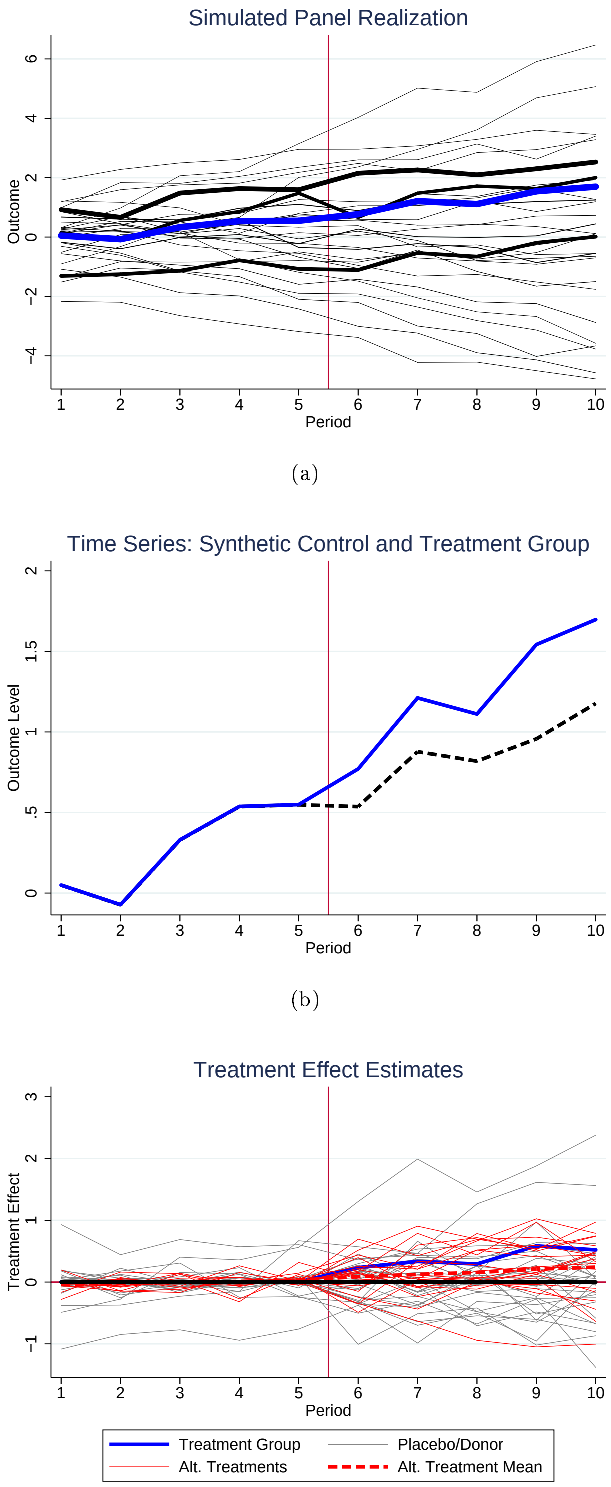
We generate time-series data from a random-intercepts and random-slopes DGP:

$$Y_{it} = \mu_{0,i} + \mu_{1,i} * t + \delta * Treated_{it} + \epsilon_{it} \tag{4}$$

We draw $\mu_0 \sim N(0,1)$ and $\mu_1 \sim N(0,0.25)$ independently for each of 30 units. There are 10 time periods ($T = 10$), 5 of which are pre-period ($T_0 = 5$). ϵ_{it} is drawn independently for all observations from a $N(0,0.2)$ distribution. The "treatment unit" has the 8th ranked intercept and the26th ranked trend from the realized sample of 31 values of μ_0. and μ_1. For this and all further simulations, we set the treatment effect to zero ($\delta = 0$).[4]

The raw data from a single realization of the DGP described above are presented as a panel of unit-level time-series graphs in Figure 1a. The thick blue line represents the treated unit. The gray lines represent the 30 donor units from which the procedure generates a synthetically-weighted counterfactual estimate. We then fit a synthetic control model to the data, matching on all pre-period outcome values, and shade the donor units accordingly, with their thickness varying according to the weight they are assigned by the algorithm (including light gray lines for zero weight). The blue line is squarely within the distribution of both potential and realized donor units for the entirety of the pre-period, and gently, but noisily, trends upwards with no apparent or programmed break in the trend or level when the placebo "treatment" occurs in period 6.

[4] These variance, group number and time period parameters are the baseline values used throughout Section 3. From Section 4 on, our baseline model is similar to Eq. 4 but we increase the trend function complexity and set $T_0 = 10$ in order to allow for higher dimensional trend functions to reveal their shape over the pre-period.



(a)

(b)

(c)

Figure 1: Graphical Demonstration of Bias

We then calculate $Y_{S,t}$ and graph it in a dashed black line in panel (b), repeating the treatment group blue line from panel (a). The synthetic control estimate is perfectly matched in the pre-period and then diverges from the synthetic control group in the post period, despite the lack of any true treatment effect in the DGP. The distance between the treated unit and control group increases nearly linearly over the post-period. Taken at face-value, the estimate would suggest a positive treatment effect that increases in magnitude over the post-period.

We then contextualize the magnitude of the implied treatment effect in the bottom panel (c) using the standard inference strategy and a second, simulation based approach. Panel 1c preserves the data from panels (a) and (b) but graphs the difference between the treatment and synthetic control groups instead of the levels of both groups separately (blue line). The gray lines are the "placebo" estimates proposed in Abadie et al (2010), which are the synthetic control estimates for the 30 donor units in the top panel. These trace out an approximation of a "sampling distribution" of synthetic control estimates when the null effect of 0 is assumed to be true, treatment is assumed to be exchangeable and the estimate is unbiased. The first two conditions hold by construction, but, as we show in red in panel (c), the third condition does not.

The thin red lines in panel (c) represent results from 20 additional simulation runs with the same DGP settings and the treated unit again chosen to be the one with the 8th ranked μ_0 and 26th ranked μ_1 in a sample of 30 control units from this DGP. These "alternative treatment" estimates are not centered on zero. The mean of 500 similarly-constructed alternative treatment simulation runs is traced out by the dashed thick red line, which begins the post-period at 0 but then increases linearly over the post-period. The slope of the dashed red line can be interpreted as a (somewhat noisy) estimate of the average degree to which the synthetic control estimator systematically underestimates the underlying trend of the treatment group when the treatment group happens to have the 26th of 31 ranked trend and 8th ranked intercept from this random-effects DGP. A simple interpretation of these results is that about half of the treatment effect seen in Figure 1 is due to bias and half due to sampling and fit variation (and none due to any real treatment effect).

3 Noise Induced Bias

What is going wrong in the example of Figure 1 above? The problem arises because the match in the pre-period of Y_S to $Y_{i=treated,t\leq t_0}$, is based in part on matching noise in the treatment and control units' outcome variables (that is, on $\epsilon_{i,t<t_0}$). This induces a systematic deviation between $\mu_{i=treated}$ and $\mu_S = \sum_{i=1}^{G} w_i^* * \mu_i$. And this in turn leads to the realized bias in the post-treatment periods when mean reversion ensures that ϵ averages 0 in the post period.

As discussed in Section 2, the synthetic control algorithm is designed to match on X between the treatment group's values ($X_{i=treated}$) and the synthetically weighted average X_s (computed as $X_s = \sum_{i=1}^{G} w_i^* * X_i$). When X includes pre-treatment values of the outcome, the weights can in part capture μ_i, the vector of unobservable heterogeneity that drives secular trends in the outcome. A good match on both Z and $Y_{i,t<t_0}$, can then be interpreted as evidence that $\mu_{i=treated} = \mu_S$ (via the asymptotic properties outlined in Abadie et al., 2010). This is how the synthetic control weights allow researchers to control for a rich unobserved factor structure.

The problem arises when the pre-treatment outcomes embody not only the unobserved factor loadings, but also idiosyncratic noise, ϵ_{it}. In fact, in the commonly modeled empirical settings we explore, the algorithm will match in part on this idiosyncratic variation in the process of optimizing model fit in the pre-period. This process of matching to noise induces a systematic relationship between the error terms in the donor unit observations and the synthetic weights assigned to those units. Mean reversion in the post-period assures that the synthetic control group then projects a distorted counterfactual to the treatment group into the post-period.

The magnitude of the bias depends on several factors and we investigate these in turn in Section 4. In this section, we first provide intuition into the nature of the bias using a stylized thought-experiment that demonstrates the fundamental role of noise in the matching process. We then use simulation techniques to demonstrate how the bias depends on the variance of idiosyncratic error terms and the location of the treated unit's unobserved factor loadings among the distribution of the control unit factor loadings. Finally, we connect the bias to the magnitude of the difference between $\mu_{i=treated}$ and μ_S.

3.1 Illustrative Example

We consider a simple, stylized case-study to provide mathematical intuition for understanding the bias. Let the factors λ_t be a single constant $\lambda_t = 1$, for all t. This implies that the effect of unobserved variables μ_i can be interpreted as i-level group effects and Eq. 1 reduces to:

$$Y_{it} = \mu_i + \delta * Treated_{it} + \epsilon_{it} \tag{5}$$

First, we consider a setting where there is exactly one pre-treatment time period, and there are exactly two "donor" units, whose true factor loadings μ_i are on either side of that of the treated unit, and with small enough error realizations that their observed pre-treatment $Y_{i,1}$ are also on the same side of the treated unit as μ_i. In this simplified setting, the weights produced by matching on the pre-treatment $Y_{i=treated,1}$ will be based on how close the treated unit is to each of the two donor units.

Suppose the two donor pre-treatment groups are ordered so that $y_{i=1,t=1} < y_{i=treated,t=1} < y_{i=2,t=1}$. The ideal weights would then be:

$w_1^{target} = (\mu_{i=2} - \mu_{i=treated})/(\mu_{i=2} - \mu_{i=1})$ and $w_2^{target} = (\mu_{i=treated} - \mu_{i=1})/(\mu_{i=2} - \mu_{i=1})$

The actual assigned weights would be:

$w_1^{estimated} = (\mu_{i=2} + \epsilon_{i=2,t=1} - \mu_{i=treated} - \epsilon_{i=treated,t=1})/(\mu_{i=2} + \epsilon_{i=2,t=1} - \mu_{i=1} - \epsilon_{i=1,t=1})$ and $w_2^{estimated} = (\mu_{i=treated} + \epsilon_{i=treated,t=1} - \mu_{i=1} - \epsilon_{i=1,t=1})/(\mu_{i=2} + \epsilon_{i=2,t=1} - \mu_{i=1} - \epsilon_{i=1,t=1})$

If the error for the treated unit $\epsilon_{i=treated,t=1}$ is positive, then w_2 will be larger than ideal, since the observed treated unit $y_{i=treated.t=1}$ is closer to unit 2 than is the underlying $\mu_{i=treated}$. Because unit 2 has a larger factor loading μ_2, this increased weight will have the consequence that the counterfactual prediction in the post treatment period will be larger than the true counterfactual.

This will induce a negative bias in the estimated treatment effect.

A similar situation arises for the error terms in the donor units. If for example $\epsilon_{i=2,t=1}$ is positive, then donor unit 2 will be farther away from the treated unit, and so will get less weight than it should. This will give a downward bias to the synthetic counterfactual, and an upward bias to the estimated treatment effect.

The situation in the preceding two paragraphs is symmetric with respect to positive or negative draws of the error terms. We might then think that in expectation this would balance out between donor units above and below the treated unit, leading to no bias. However, this symmetry is broken when the distribution (density) of the donor units is asymmetric around the treated unit. For example, if among the donor units, unit 1 is closer to the treated unit than is unit 2. This will then lead to a systematic bias. To demonstrate this mathematically, we examine how the weight for each unit varies with error realizations for each of the pre-treatment units.

Consider the effects of an increase on each error on the weight assigned to unit 2. For ϵ_2:
$\frac{\partial w_2^{estimated}}{\partial \epsilon_2} = \frac{\partial w_2^{estimated}}{\partial y_2} \cdot \frac{\partial y_2}{\partial \epsilon_2} = \partial w_2^{estimated} / \partial y_2 = (y_{i=treated,t=1} - y_{i=1,t=1}) / (y_{i=2,t=1} - y_{i=1,t=1})^2$

And for ϵ_1:
$\frac{\partial w_2^{estimated}}{\partial \epsilon_1} \cdot = \frac{\partial w_2^{estimated}}{\partial y_1} \cdot \frac{\partial y_1}{\partial \epsilon_1} = \partial w_2^{estimated} / \partial y_1 = -(y_{i=2,t=1} - y_{i=treated,t=1}) / (y_{i=2,t=1} - y_{i=1,t=1})^2$

If the distance $(y_{i=2,t=1} - y_{i=treated,t=1})$ is larger than $(y_{i=treated,t=1} - y_{i=1,t=1})$, similar sized and signed errors will have asymmetric effects. Additionally, the impact of the noise around each true μ_i will be different for the two donor units. As a consequence, noise around unit 2 will matter less than noise around unit 1. This will lead to on average an under-weighting of unit 2, a downward-biased counterfactual for the treated unit, and thus an upward-biased estimated treatment effect.

What would cause an asymmetric distribution of donor units around the treated unit? This will result from an asymmetric distribution of the density of μ_i around the treated unit. So, for a typical unimodal distribution (such as Gaussian) this will happen whenever the treated unit is not right at the middle. More generally, if the distribution of μ_i is such that the density decreases as you move toward the boundary, and if the treated unit is not in the center of the distribution, then there will be more density of donor units on the side toward the center, and less density on the other side. This will mean that on average the donor units are closer on the side toward the center, which will result in a systematic bias in estimated treatment effects. We illustrate this in detail in our simulations in section 3.2.2 below.

The thought experiment extends to additional pre-periods and choice of matching variables. Consider the case where there are $T_0 > 1$ pre-treatment periods of data, and we are trying to match on pre-treatment average Y. In this case the same problem will apply, only based on $\bar{\epsilon}_{t\leq T_0}$. Because of the averaging over the error term over the pre-periods, the bias should be decreasing as T_0 gets bigger and thus the variance of $\bar{\epsilon}_{t\leq T_0}$. gets smaller. In the next section we demonstrate this empirically, showing that both the variance of ϵ and the location of $\mu_{i=treated}$ in the distribution of μ_i affect the bias as generalized from the thought experiment presented here.

3.2 Simulation Evidence of Noise-Induced Bias

In this section we use Monte Carlo simulations to empirically explore the bias described above in a controllable panel data setting, similar to what is typically used in the literature. We show results from 3 sets of Monte Carlo simulations and demonstrate the bias visually using simple figures graphing mean bias across various specifications of the model parameters and hyper-parameters. We first show that the bias in a basic random-intercepts DGP (Eq. 5) exists only when there is a positive variance on the period-specific error term (noise). We then demonstrate that the biased matches generated by the synthetic control algorithm persist into DGPs with trends, where the method again partly matches on noise and fails to accurately match on the unobserved components of the DGP. We show that in both cases the bias can be explained by the mismatch between $\mu_{i=treated}$. and μ_S.

The underlying DGP and parameter values vary over the three simulations, but all share the same basic structure. In each of a series of realizations of simulated data from Eq. 4, we use synthetic controls to estimate a treatment effect on the simulated data, matching on all pre-period values of Y. The true treatment effect in the DGP is set to 0, and the average estimated treatment effect across realizations is our estimate of the bias. We design the simulation (and all simulations performed in this paper) so that the treatment group is always within the convex hull of the donor group distribution by rejecting simulation runs where the treated unit's outcome variable is the largest or smallest value among all control units within any particular pre-intervention period. We do this because we are not interested in cases where the treated unit is an outlier with no possible pre-treatment match, which then trivially produces poor estimates.

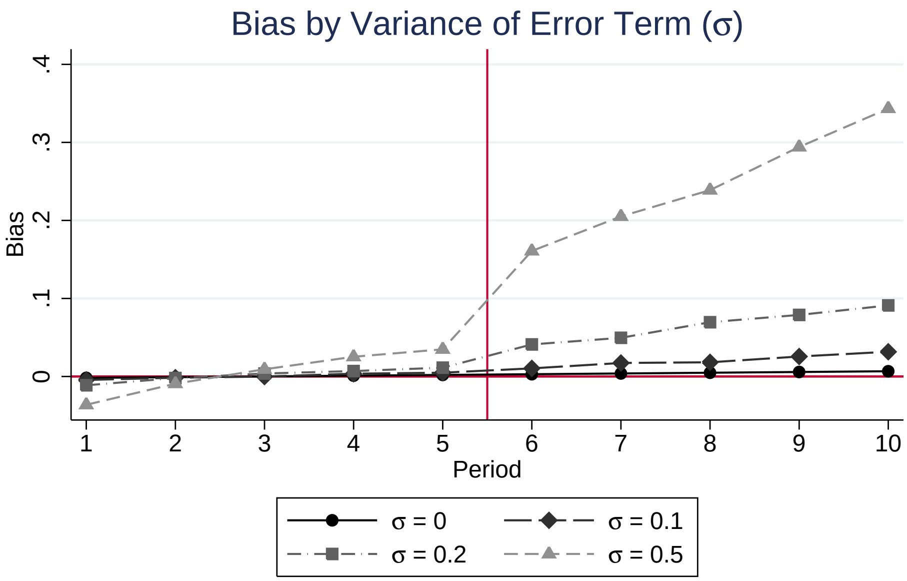
3.2.1 Noise and Bias

We first empirically demonstrate the relationship between the bias and error term variance using the DGP used to generate Figure 1 and described by Eq. 4. We then run the entire Monte Carlo simulation for 4 different values of σ_ϵ and compute the period-specific average estimated treatment effect across realizations.

To demonstrate the bias in action, we choose as the treatment group the unit with the 70th percentile value of the sum of μ_0 and μ_1 from the realized distribution in the simulated data. This generates an asymmetry in the distribution of donor units around the treatment group, and thus should generate the bias described above.[5] By focusing on a particular region of the joint-distribution of μ, we hold fixed the bias and isolate its sign and magnitude for treatment and control units with a similar set of structural trend parameters. We also do so without loss of generalization: any particular region of μ would produce a different bias, but the bias would stem from the same root of matching on noise even as the sign and magnitude of the bias change.

Figure 2 graphs the resulting bias by period. The solid line averages from simulations where σ_ϵ is set to 0 (removing idiosyncratic errors from the DGP). The dashed lines represent variances of 0.2, 0.5 and 2, respectively.

[5] Averaging our simulation results over all ranks of μ would average out the bias. In that sense, unconditional on any particular group being designated the treatment group, the estimator is in fact unbiased. However, conditional on a group actually being the treatment group, the estimate is biased unless the unit is in the center of the realized donor unit distributions of μ and Y.


Bias by Variance of Error Term (σ)

Estimates from matcheing on all Pre−Period Values of Y
Results are average bias estimate from 5th post−period
DGP includes linear time trends and 30 donor groups

Figure 2

There is no bias when σ_ϵ is set to zero. The synthetic control group matches only on the structural components of the underlying DGP, and so produces an unbiased counterfactual. Increasing the variance of ϵ_{it} a relatively small amount (0.2) generates bias: a mismatch between the synthetic control and treatment group trends leads to a bias that increases over time in the post-period. This comes despite a very small loss of precision in the pre-period fit. Increasing the variance further increases the bias a large amount. The higher variances also produce a slight average trend in in the pre-period, despite the fact that we are matching on all pre-treatment periods of the outcome variable. This occurs because in some cases no perfect match is available. This is distinct from the phenomenon we focus on in this paper, which is that a bias occurs even when there is a perfect pre-treatment match. This issue is discussed and illustrated further in Appendix Figure A.1, where we show that the (mean-0) trend over pre-periods disappears when you focus on the realizations that produced low real mean square prediction error (RMSPE) in the pre-period fit more generally. We interpret this as a warning to practitioners that choosing a matching specification based purely on pre-period fit does not guarantee a good counterfactual projected into the post-period.

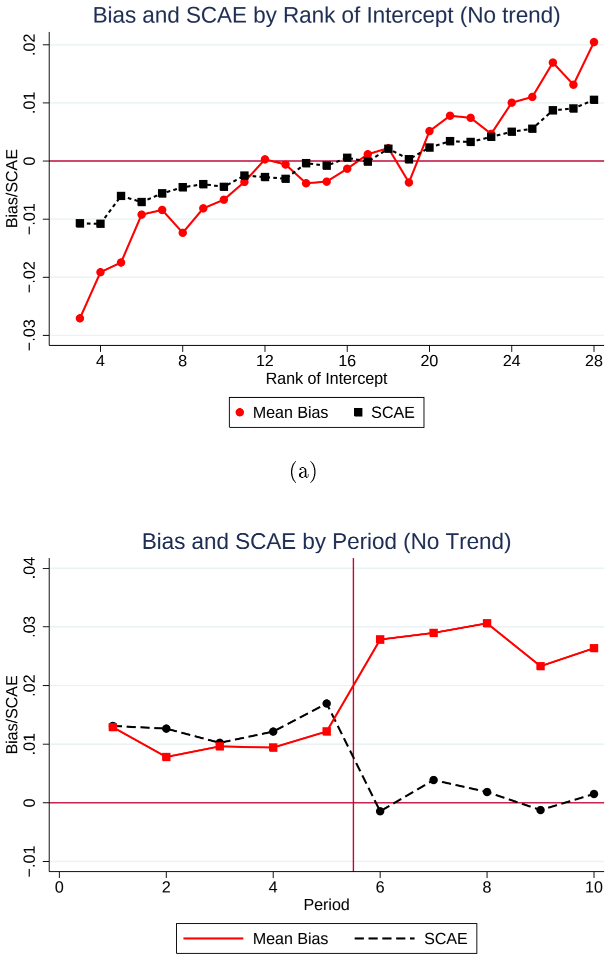
3.2.2 Intercept Rank and SCAE

To illustrate the mechanism of the bias, and how it relates to the treated unit's location in the distribution of control units, we next show that synthetic control weights selectively load onto units with systematically non-zero average error terms over the pre-periods. Here we employ the simplified DGP from Eq. 5 where the unit-specific intercepts are the only structural component and are drawn from a N(0,1) distribution. We then compute the mean bias, as a function of the rank of μ_0. The changing density of donor groups around the treatment unit (with more on one side than the other) at the edges of the outcome distribution (high or low ranks of μ_0) implies that the influence of error terms on the matching should be largest towards the tails of the distribution of μ and smaller towards the center, with no influence at the center where the density is balanced on both sides.

To clarify the relationship between bias and matching on noise, we define the *Synthetic Control Average Error* (SCAE) as $\frac{1}{G*T_0}\sum_{t=1}^{T_0}\sum_{i=1}^{G} w_i^* * \epsilon_{it}$, the synthetically-weighted mean error term averaged over the donor units for the entire pre-period. In usual empirical practice, this cannot be calculated since ϵ_{it} are by definition unobserved, but in our simulations we know each ϵ_{it}. Since these errors are on average mean 0 in the post-period by construction, an $SCAE > 0$ in the pre-period will generate bias in the post period. We plot the average (across realizations) SCAE, as well as the mean bias, by treated unit rank of μ in Figure 3a. The solid red line plots the mean bias for each rank of the treated unit intercept $\mu_{0,i=treated}$ and for each rank the bias is estimated across 3000 Monte Carlo runs). The dashed black line overlays a graph of mean SCAE on top of the mean bias graph across the rank of μ.

The two curves follow a similar pattern. Note when the treated unit's rank is in the middle of the distribution of μ, the density of control units is symmetric around the treated unit, and there is little to no bias when the treated unit's rank is near the median. However, the bias increases in magnitude for ranks closer to the extremes as the asymmetry of control units around the treated unit increases.

We speculate that the bias that is not accounted for by SCAE is driven by the fact that model errors in the pre-treatment values for the treated unit leads to distorted weights on the donor units, as discussed above in section 3.1. The figure suggests that the bias in the synthetic control estimates can be explained by the algorithm matching on a string of realized error terms at the expense of the unobserved structural features of the DGP. The resulting effect on the difference between $Y_{i=treated}$ and Y_S is shifted from a match on errors in the pre-period to a bias in the post-period and explains a large fraction of the deviation of the synthetic control treatment effect estimate from 0.



(a)

(b)

Figure 3: Bias and Synthetic Control Average Error (SCAE)

Figure 3b further clarifies the role of matching on error terms in causing bias. Here we plot the mean SCAE and bias by period for the simulations where the treated unit intercept rank is 29th out of 31. The matching algorithm tends to select control units with positive errors during the pre-period, distorting the counterfactual. Synthetic control average errors then revert to roughly mean zero in the post-period, transferring the pre-period SCAE into post-period bias. The small pre-period bias is driven by realizations with poor fit in the pre-period, and much like in Appendix Figure 11, the pre-period bias disappears when we restrict our analysis to matches with relatively low MSE in the pre-period.

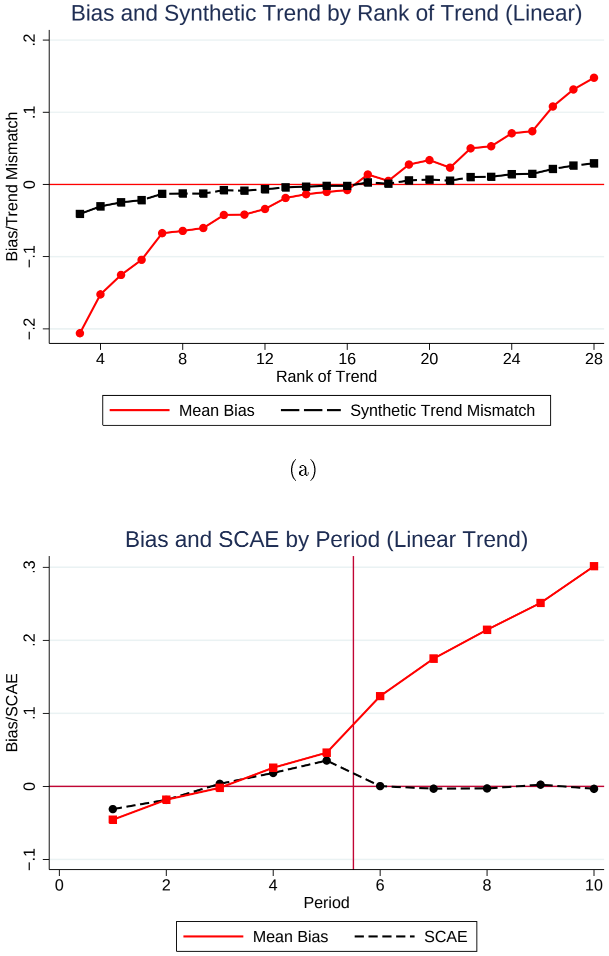
3.2.3 Trend Rank and Synthetic Trends

We next return to the unit-specific linear trends model (as in Figure 1). The presence of unit-level trends induces a second source of bias due to poor matching on unobservables relative to the SCAE graph above - there can be a mismatch on levels, as in Figure 3a, and there can be a mismatch on trends if $\mu_{1,S}$ doesn't match $\mu_{1,i=treated}$. To focus on this second bias, we fix $\mu_{0,i=treated}$ at the median of the realized donor pool so as to mitigate bias of the kind demonstrated in Figure 3 .

Figure 4 is similar to Figure 3, but here the red line plots the average bias across treated unit *trend* ($\mu_{1,i=treated}$) rank. Again, there is little to no bias in the middle of the trend rank distribution, but lower ranked trend ranks are biased downwards and vice-versa. In this case, the bias is generated not as a level-mismatch (as above), but as a trend-mismatch with an increasing bias across the post-period, as we saw in Figure 1.

To demonstrate that this bias is the result of a mismatch between the treated unit's trend and the trend in the synthetic control group, for each Monte Carlo run we also calculate the "synthetic trend" $\mu_{1,S} = \sum_{i=1}^{G} w_i * \mu_{1,i}$. We then subtract $\mu_{1,S}$ from $\mu_{1,i=treated}$ to find the "trend bias", the mismatch between the true trend of the treatment group and the true trend of the synthetic control group. Mirroring Figure 3, the black line in Figure 4 graphs the synthetic trend bias ($\mu_{1,i=treated} - \mu_{1,S}$). When the treated unit trend rank is relatively low, the synthetic trend tends to be too high (the trend bias is negative), and vice-versa. This type of bias becomes more severe (in level terms) longer into the post-treatment period, as the mismatched trends diverge further over time.

The mismatch on trends is itself a function of the algorithm again matching on noise instead of structural parameters. Figure 4b shows the mean SCAE by period for the case where the treated unit trend rank is 29th. The matching algorithm in this case selects control units where the errors happen to trend upwards, and again mean errors revert to roughly zero in the post-treatment period.



(a)

(b)

Figure 4: Bias and Synthetic (Linear) Trend

4 Determinants of Magnitude of Bias

The preceding section generates two key insights regarding the data generating process and bias. First, the bias we describe is driven by unit-period-level error terms. This is because the bias stems from the "over-fitting" that occurs by matching on noise at the expense of structural features of the unit-level time-series. Second, the rank (or deviation) of unobserved parameters for the treatment unit relative to the distribution of control units affects the average bias. These two features - noise and asymmetry - are key determinants of the problem.

In this section we examine the impacts of additional features of the data on the *magnitude* of the bias; including DGP complexity, the number of potential donor units (G), and the number of pre-treatment time periods (T_0). We do this with another set of Monte Carlo simulations.

Up to this point we have restricted the DGP to include only random intercepts and/or linear trends. In this section we extend this to consider higher-order polynomial trend functions. For a polynomial time trend function of degree D, we generate Y according to the generalized DGP:

$$Y_{it} = \sum_{d=0}^{D} \lambda_{d,t} * \mu_{d,i} + \delta * Treated_{it} + \epsilon_{it}, \tag{6}$$

with $\lambda_{d,t} = t^d$. Then $\mu := (\mu_{1,i}, \mu_{2,i}, ..., \mu_{D,i})$ is a vector interpreted as including (intercept, linear trend slope, ... coefficient on D'th degree polynomial).

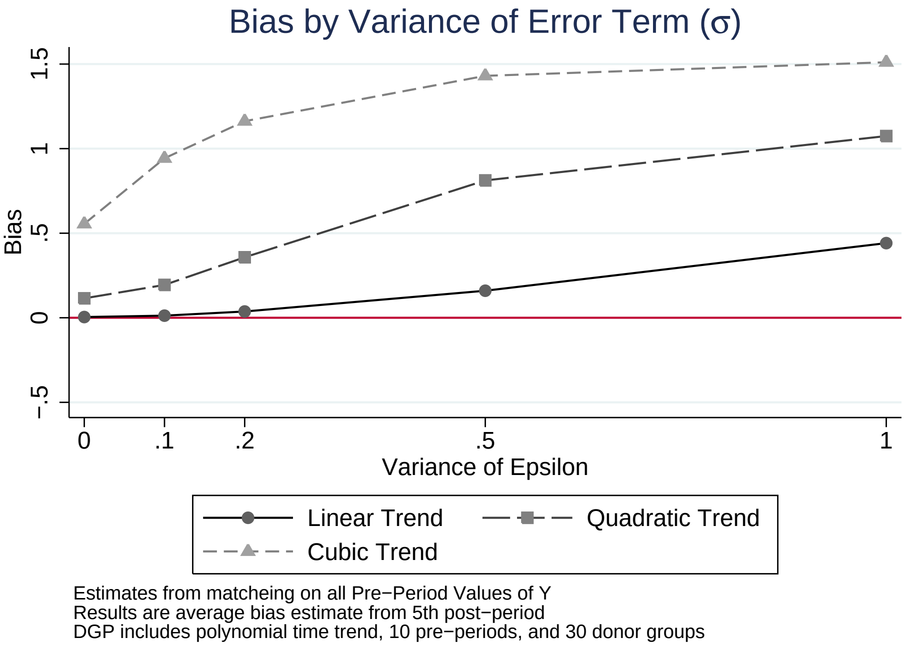
One computational complication with higher-order polynomials is that the higher-order terms may dominate in the later periods and lead to extreme outcome values. To address this, we generate λ_t through a D-degree orthogonal polynomial transformation of the time period variable.[6] Additionally, we now set $T_0 = 10$ so that there is adequate data for higher-order DGPs to reveal their underlying trend function.

For all simulations that follow, we draw each component of $\mu_{d,i}$ from a standard normal distribution $N(0,1)$, and ϵ_{it} is drawn from $N(0,.2)$ unless otherwise specified (such as when we vary the

[6] We use the "orthpoly" command in Stata, defining the orthogonal polynomial transformations over the full time period.

error variance in Figure 5). We choose the treated unit by calculating the sum of $\mu_{d,i}$ for each unit individually and then selecting the unit with the 70th percentile of $\sum_{d=0}^{D} \mu_{d,i}$ as treated.[7] We choose the 70th percentile so that the treated unit is some distance away from the center of the control unit distribution, but is still typically able to find reasonable pre-treatment matches.

4.1 Noise (σ_ϵ)



Estimates from matcheing on all Pre−Period Values of Y
Results are average bias estimate from 5th post−period
DGP includes polynomial time trend, 10 pre−periods, and 30 donor groups

Figure 5: Bias and Noise (Variance of Y)

Figure 2 showed simulation results when the DGP had an underlying linear trend ($D = 1$). In this section we expand the scope of the exercise and consider DGPs from Eq. 6 with D=1, 2 and 3 (linear, quadratic, and cubic trends). We graph the average bias in the final post-period (5 periods after treatment) across 4 variance levels of σ_ϵ (0, .5, 2, 5) separately for each degree polynomial. The results are shown in Figure 5. Bias increases with σ_ϵ, though the marginal impact of additional

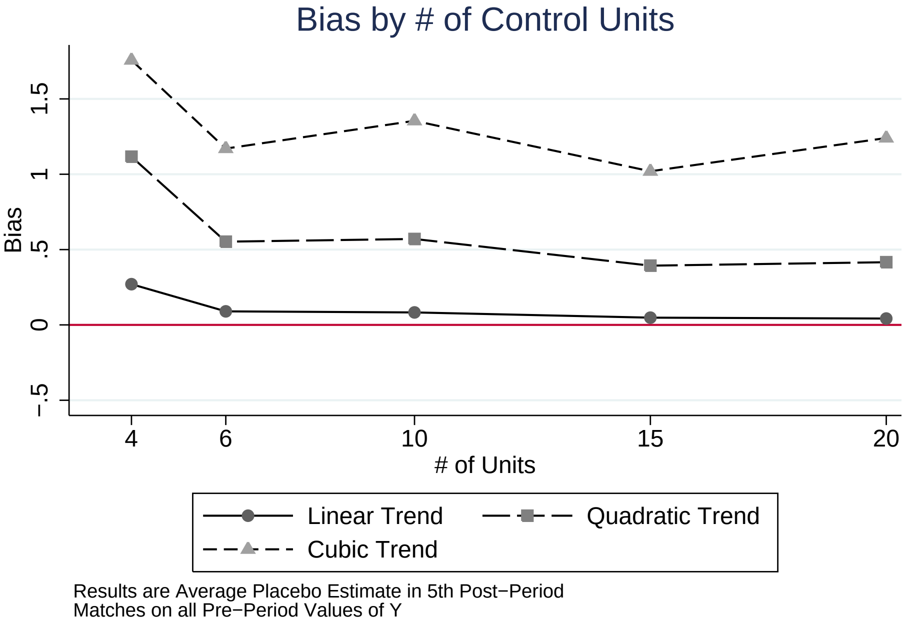
[7] This is a compromise mode of treatment assignment that balances competing interests. Enforcing that the treatment unit contain the 70th percentile on every value of μ_d would ensure perfect comparability of treatment groups across realizations, but would generate a highly selected sample of potential realizations and treatment groups. Instead, under our treatment assignment rule, in any given data realization some group will indeed be assigned "treated" status. That group effectively represents units with relatively positively influencing unobserved characteristics.

variance diminishes at higher values. In addition, the bias increases in D (the degree of the highest polynomial trend term). These are basic themes of the subsequent exercises as well; increases in D and σ_ϵ, tend to increase bias.[8]

4.2 Size of Donor Pool (G)

One candidate approach to mitigate the bias is to increase the number of potential matches, that is, the number of donor units, G. This might increase the chance of matching on a structurally similar set of units instead matching on noise.

This turns out to be less effective at reducing bias than we had expected. Figure 6 graphs mean bias across the number of donor units, separately by D. While increasing the number of groups can reduce bias when there are very few units, in general the gains beyond 10 or so units are modest if they exist at all. Higher dimensionality of D dwarfs the impact of increased donor pool size.



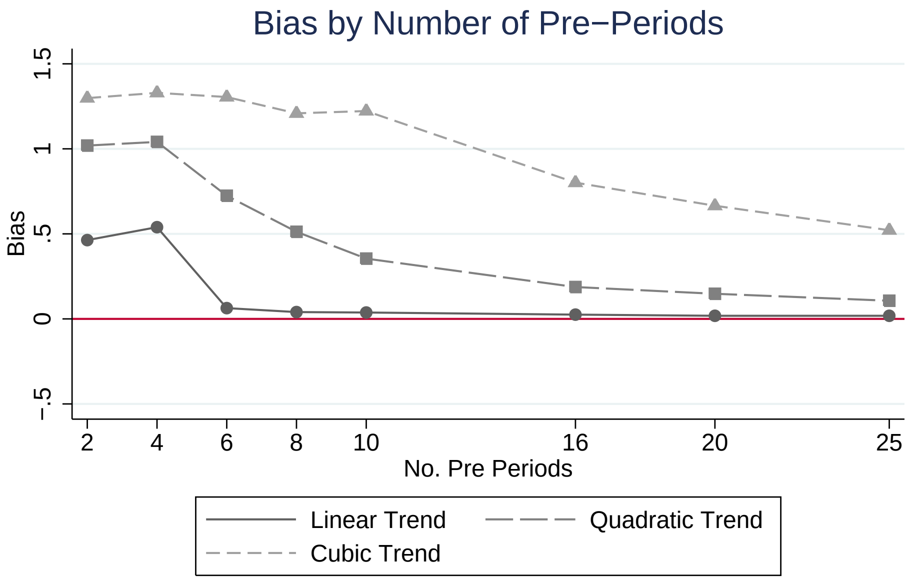
Results are Average Placebo Estimate in 5th Post−Period
Matches on all Pre−Period Values of Y
DGP includes polynomial (degree D) time trend, 5 pre−periods, and 30 donor groups

Figure 6

[8] Note that there is a non-zero bias even with zero error terms for the higher-order DGPs. This occurs because for more complex DGPs, in some cases there is no perfect match available. This is again similar to the issue discussed in Appendix A.1.

4.3 Pre-Period Length (T_0)

We next consider the role of the number of pre-treatment periods, T_0. Figure 7 graphs mean bias in the final-post period against the number of pre-periods. The bias decreases as the length of the pre-period panel increases for all values of D. This result reflects the fact that with longer matching periods the importance of the trend parameters increases relative to idiosyncratic error terms, so that matching on outcomes yields a better match to the underlying trend function. Thus, when (in terms of pre-period length) the gains are realized depends on the the dimensionality of the DGP (D). The relative bias relative reduction going from 10 to 25 periods with a cubic DGP (D=3) is smaller, percentage-wise, than going from 4 to 6 periods with a linear DGP (D=1).



Results are Average Placebo Estimate in 5th Post−Period
Matches on all Pre−Period Values of Y
DGP includes polynomial (degree D) time trend and 30 donor groups

Figure 7

5 Bias Reduction Strategies

The bias that we identify in this paper is a result of a type of "over-fitting" of the weights to match the model errors (noise). This insight provides us with two avenues for potential correction procedures.

The first approach is to select a parsimonious set of variables to match on. We model the functional form of the underlying DGP and estimate $\hat{\mu}_i$ for each unit using a series of unit-specific regressions. We then match exclusively on the estimated trend coefficients. This approach thus provides sufficient information to capture the curvature in the treatment unit time-series, while avoiding including information that allows for matching on noise in the first place.

The second approach is to perform a direct bias correction. Using a parametric bootstrap, we simulate the magnitude of bias that results from matching on noise given the observable features of the data and relative location of the treatment group parameters in the joint distribution of donor unit parameters. We then subtract this estimated bias from the main estimate. This approach involves three steps: (1) estimating the underlying structural model; (2) simulating realizations of the synthetic control estimate when the treatment group is chosen to be similar to the one in the original data relative to the realized sample; (3) averaging the treatment effect estimate across many simulation runs to estimate the bias in the original analysis.

In this section we lay out the procedures for both of the proposed corrections and demonstrate that both reduce bias relative to matching on all pre-period values of Y. The results in this section are based on the unrealistic assumption that the functional form of the underlying DGP (e.g., that the DGP is polynomial time trend of degree D) is known to the researcher. This demonstrates that the methods can work in theory, even if empirically feasible implementation is challenging. In Section 6 we take up empirically feasible estimation when the underlying DGP is unknown but comes from a known or specified family of potential DGPs.

5.1 Matching on Trend Parameters ($\hat{\mu}$)

The discussion in the previous sections frames two considerations for choosing a set of matching variables for the synthetic control function, whether they be pre-period values of Y, observable covariates or estimated structural parameters. First, researchers want to avoid matching on too few moments — the cost of this is not being able to model/span the factor loadings μ_i. In other words, without sufficient moments to match on, the synthetic control matching function will not be rich enough to predict complicated trends in the treated unit, potentially causing a significant mis-specification bias. This could be the case, for example, if the matching variables were chosen such that they could only capture linear trends in a treatment outcome that had quadratic trends (e.g. matching exclusively on 2 pre-period values of Y). On the other hand, if we are matching on a greater-than-necessary number of moments, this may worsen the problem of matching on noise. The goal is to provide a matching model rich enough to capture the dimensionality (D) of the DGP, but not so rich that the model ignores the structural trends in favor of matching on noise.

We accomplish this in practice by including in the matching specification estimates of the relevant unobserved unit-level trend parameters. As we do throughout the paper, we limit our focus to the family of polynomial time trend functions indexed by degree D. For each unit separately we obtain $\hat{\mu}_{D,i}$ from a regression of the form:

$$Y_{it} = \sum_{d=0}^{D} \lambda_{d,t} * \mu_{d,i} + \eta_{it} \tag{7}$$

We then include only the estimates of $\hat{\mu}_{D,i}$ in the set variables for the synthetic control algorithm to match on. This essentially allows us to turn μ_i into a component of Z_i (from Eq. 1), making our estimate of the unobserved component an observed predictor for the model. If the underlying DGP is known, and the information included sufficiently summarizes the unit-level outcome time paths, then the algorithm can match exclusively and sufficiently on structural components, and not on noise.

5.2 Computational Bias Correction

Our second bias reduction strategy is to perform a direct bias-correction, using a calibrated Monte Carlo simulation. Our proposal involves estimating the parameters of the assumed DGP using the researcher's actual data, then using this calibrated DGP to run a Monte Carlo simulation that estimates the bias when the treated unit's (estimated) parameters are held fixed.[9] The estimated bias is then subtracted off the estimated treatment effect at each time period to yield an adjusted estimate.

The correction procedure is as follows:

1. Using the original data, estimate the parameters from the DGP.

   (a) Choose $\hat{D}$ based on researcher's belief or knowledge about D (polynomial order)

   (b) Estimate the mean and standard deviation of each $\mu_{d,i}$ (we assume that each is generated from a normal distribution, though other distributions are theoretically viable) using a GMM mixed-effects model.[10] The same model is used to estimate $\hat{\sigma}$, the standard deviation of the error term.

2. Generate artificial data for the control units from the assumed DGP using the estimated hyper-parameter values and setting G, T_0 and T as in the sample.

3. Create a Simulated Treatment Unit.

   Using the actual data, for each parameter μ use Eq. 7 to estimate each unit's parameter value and record the treated unit's rank. Then assign the treated unit the value of μ with the corresponding rank from the simulated parameters generated in step #2. Then generate treated unit outcomes accordingly.

4. Run synthetic control estimation using the same matching specification as used for the real data (in our simulations below we match on all pre-treatment Ys, but again other specifications are possible), and record the estimated treatment effect.

[9] In our simulations, when the true μ is observed, both this Monte Carlo method and the matching specification above produce perfectly unbiased estimates. Later in this section we discuss variations on these methods using alternative forms of the estimated parameters in the correction procedures. For the main specifications presented in this work, though, we use the value of the estimated parameters ($\hat{\mu}$) in the matching specification, and the rank of $\hat{\mu}$ among all units in the Monte Carlo simulation.

[10] We use Stata's xtmixed command with the outcome as the dependent variable and D-degree orthogonal transformations of the period variable as independent variables.

5. Repeat many times and compute the average treatment effect.[11] This is an estimate of the bias.

6. Subtract the estimated bias from the treatment effect estimated from the actual data, resulting in a bias-corrected estimate.

5.3 Performance of Bias Reduction with Known DGP

In this section we run a new series of Monte Carlo simulations to assess the ability of our proposed strategies for mitigating bias. We assume the researcher knows both D (model complexity) and the functional form of the DGP, but does not observe μ. In results not shown because they are both trivial and expected, both procedures can perfectly remove bias when the DGP is known and μ is observed. In this section D is known but μ is not. In Section 6 we relax this requirement using model selection techniques to estimate $\hat{D}$ when D is unknown.

Our simulations continue to use the parameter values used in previous simulations, and we simulate polynomial trend functions of degrees D=1 to 3. For each degree polynomial we estimate a specification that matches on all pre-period values of the outcome variable, one that matches only on the vector $\hat{\mu}_D$ (the moments matching method), and one that applies the simulation-based correction (the Monte Carlo correction). We then graph, separately for each D, the mean difference between the treatment group and the synthetic control group (the bias) across period. For each D, we run 3000 simulations, each with 100 "nested" simulation runs used to estimate bias with the Monte Carlo bias correction procedure.

The results are presented in the left column of Figure 8. A number of lessons discussed previously continue to hold. For the specification that matches on all pre-period outcome values (the red short-

[11] For each "nested" simulation described in steps 2-5, as with all other analysis in this paper, we exclude data realizations where the treated unit has any pre-treatment outcome observations that are the largest or smallest of all units. In rare cases within our simulations, when $\hat{D}$ is mis-specified the treated unit can receive extreme estimated parameter values, so that nearly every simulated data realization is rejected because treated unit outcomes are very large or small. We therefore use the following rule: if out of the first 20 nested simulations, four or fewer are successful (i.e. the treated unit never has the largest or smallest outcome in a pre-treatment period), we terminate the correction procedure and use the original baseline estimate as our "corrected" estimate.

dashed line), increasing D (moving down panels of Figure 8) leads to increased bias. The pre-period fit is on average quite good. The bias increases over time following a functional form determined by the curvature in the underlying trend because it is a result of a mismatch between $\mu_{i=treated}$ and μ_S.

The two bias correction procedures, matching on moments graphed in grey with long dashes and the simulation-based correction graphed in solid black, both lower bias compared to matching on all pre-period outcome values in all cases. The moments matching method generates less biased estimates than matching on all pre-period outcome values in each example. In the case of linear trends (top panel), matching on moments reduces bias by around half relative to matching on all pre-period outcomes. The effectiveness of the matching on moments correction, though, fades strongly as model complexity increases. In the case of the quadratic trend it reduces bias by only about one-third, and for the cubic trend it is not particularly effective, reducing the bias by less than half.

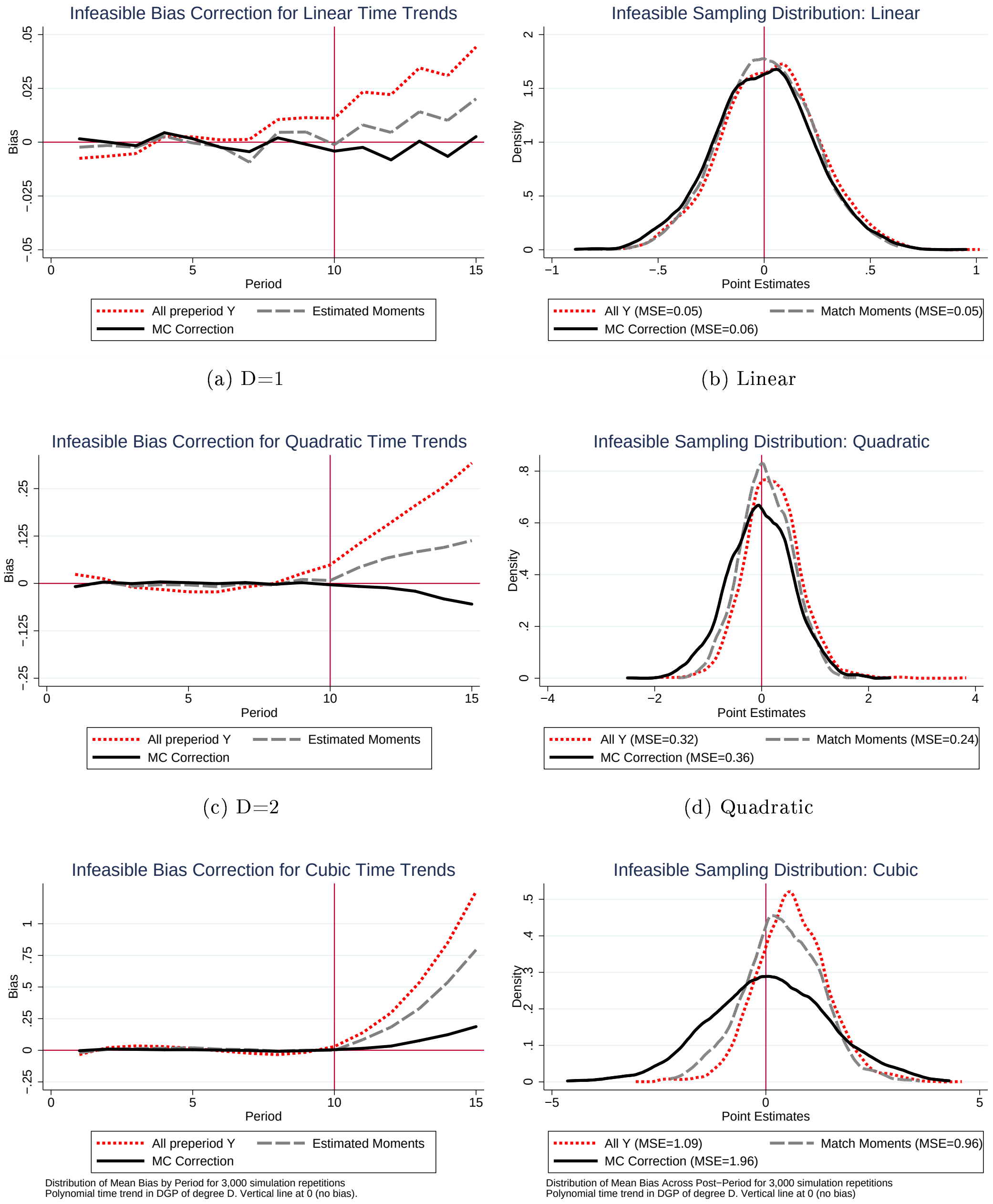
The simulation-based correction has better bias reduction properties. In the case of linear trends the bias is nearly eliminated. For quadratic and cubic trends, the simulation procedure performs better than the other methods but still displays some bias, and slightly over-corrects in the quadratic case.

The practitioner should be careful in interpreting the magnitude of the bias reduction in either outcome units or percentage reduction. First, as we will discuss in the next sub-section, the bias reduction properties of the correction procedures are dependent in complex ways on the hyper-parameters of the DGP (D, σ_ϵ). Second, while the bias reduction properties of the corrections are fairly effective, the right side of Figure 8 shows the tradeoff of the proposed corrections in terms of increased sampling variability. Here we graph, for each D, a kernel density estimate of the sampling distribution of $\hat{\delta}$ from each specification. While the correction procedures do indeed reduce bias, they mostly do so at the expense of increased variance.

Maintaining the color and line themes from the time series bias graphs, the dashed red line shows the distribution of estimates when matching on all pre-period values of Y. Relative to matching on all pre-period Y values, the matching on moments specification (gray dashed lines) produces estimates with (weakly) lower MSE (computed as the average squared point estimate across the pre-period averaged across realizations). This is the result of two competing factors, where the variance of the sampling distribution of both estimates are similar, but the matching on moments estimate is less biased leading to reduced MSE.

The same is not true of the Monte Carlo (MC) based correction. The simulation-based correction yields the least biased sampling distribution (it is centered close to 0 for all D) but the largest sampling variance, particularly in the cubic DGP case. In terms of MSE, the MC-based correction sacrifices variance for small reductions in bias.

In the next section we evaluate the correction procedures in settings where the underlying DGP is not known and can only be estimated with significant uncertainty. First, though, we discuss two alternative implementations of the correction procedures involving variations on the way that estimates of $\hat{\mu}$ are employed in the procedures, and alternative importance weights for each parameter.



(a) D=1

(b) Linear

(c) D=2

(d) Quadratic

(e) D=3

(f) Cubic

Figure 8: Performance of Bias Correction Procedures when DGP is Known (Infeasible)

5.4 Robustness and Alternative Implementations of Corrections

As stated previously, the performance of the bias correction procedures depends on the underlying DGP. In general, the corrections work better (in terms of percent of bias removed) at lower variance levels and lower dimensions of D. The moments matching method tends to quickly lose much of its bias reduction power as σ_ϵ increases. And while the Monte Carlo correction also begins to fail at high variance levels using the implementation described above, an alternative implementation we describe here performs well (in terms of bias reduction) even at relatively high values of σ_ϵ. In this section we discuss this alternative implementation of the correction procedures, and demonstrate robustness (or lack of) across different values of D and σ_ϵ.

The correction procedures in our main results use estimated information on μ in two different ways. In the case of the moments matching correction, we simply include the value of $\hat{\mu}$ as a matching variable directly. We refer to this as the "value" method. Alternatively, we could match on the estimated rank of each trend parameter, which we refer to as the "rank" method. In the case of the Monte Carlo correction, for each nested simulation we assign treated unit parameters based on the estimated rank from the "real" data (see Section 5.2), but alternatively we could assign the exact estimated treated unit value of each parameter. The decision on using ranks versus values was based on both theoretical considerations[12] and the performance of the two options in our particular DGP setting. For completeness we present results using both rank and value methods for both corrections across values of D. We also evaluate the bias of each procedure across different values of σ_ϵ. These results are presented in Appendix Figure A.2.

The left column of Appendix Figure A.2 shows results from implementing both correction procedures using the "rank" method, and the right column using the "value" method. In the top row, we compare the moments matching estimate with the baseline specification of matching on all pre-period outcome values. In the second row we repeat this exercise comparing the baseline estimates

[12] Since the treated unit value of $\hat{\mu}$ may have a different rank relative to the donor units when a new realization of data occurs in the simulation, we generally consider the rank method more intellectually consistent with our understanding of the bias.

with the Monte Carlo corrections. The third row plots the fraction of bias removed by each of the two correction procedures.[13]

For the moments matching specification, the value method dominates the rank method at low variance levels, where the rank method can actually induce a large (negative) bias where only a small (positive) bias occurs in the baseline estimates. At higher levels of σ_ϵ the two methods are very similar.

The Monte Carlo correction is more effective using the rank method at low variance levels and the value method at high variance levels. Perhaps the most notable result is that the Monte Carlo correction using the value method removes nearly all bias even at very high variance levels.

We note one further robustness exercise we conducted but do not report, related to the "importance weights" that can be assigned to synthetic controls. These importance weights adjust the distance minimizing algorithm to account for the pre-determined importance assigned to each matching variable. We hypothesized that late-period behavior of polynomial functions is largely driven by the highest order polynomial. Thus, mismatch on the D'th-degree coefficient is likely to be more biasing than mismatch on the (D-1)'th-degree coefficient. We tested a series of importance weights that placed increasing weight on matching to the highest order coefficient and less to the lower order coefficients. However, these weighting choices made almost no difference to our estimates, and we omit these results.

6 Feasible Bias Correction

In this section we test the correction procedures described above when relaxing the assumption that D is known. We consider empirical estimation of D when the true DGP is known to come from the family of polynomial trend functions from degree 0 to 4. We then run a series of Monte Carlo

[13] Note that in Panel (e) for the D=1 matching on moments case, results for lower error variances are not shown. This is because in these cases the matching on moments with rank method produces an over-correction on a near-zero bias (this can be seen in panel (a)), so that the "corrected" bias is many times larger than the original bias and the fraction is impractical to graph with the other cases.

simulations where the two correction procedures are based on using $\hat{D}$ and moments of the empirical distribution of outcomes instead of (the empirically unobservable) D itself. We also provide results on estimate precision, and show that improvements in bias come at the expense of increased variance.

6.1 Estimating $\hat{D}$

To choose $\hat{D}$, in each simulation run we estimate a series of random-intercepts and random-slopes models up to polynomial degree 4. We use the Bayesian information criteria (BIC) as a model selection procedure, choosing the model with lowest BIC.[14]

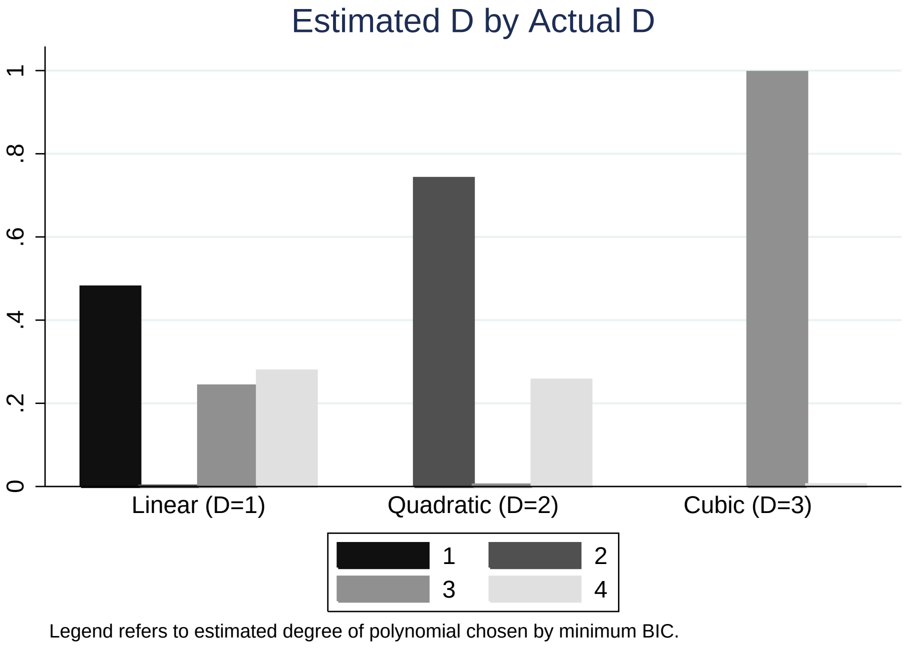


Figure 9: Estimates of D Given Candidate Family of Polynomial Functions up to D=4

Figure 9 shows how accurately the BIC estimate identifies the polynomial order of the DGP ($\hat{D}$). The histogram shows the distribution of $\hat{D}$ under three true values of D (linear, quadratic, and cubic trends) and the parameter settings used in Figure 8. In each case, the BIC chooses the correct degree of heterogeneity more than half the time, increasing from about 50% in the linear case, to

[14] A researcher might well choose other model selection techniques.

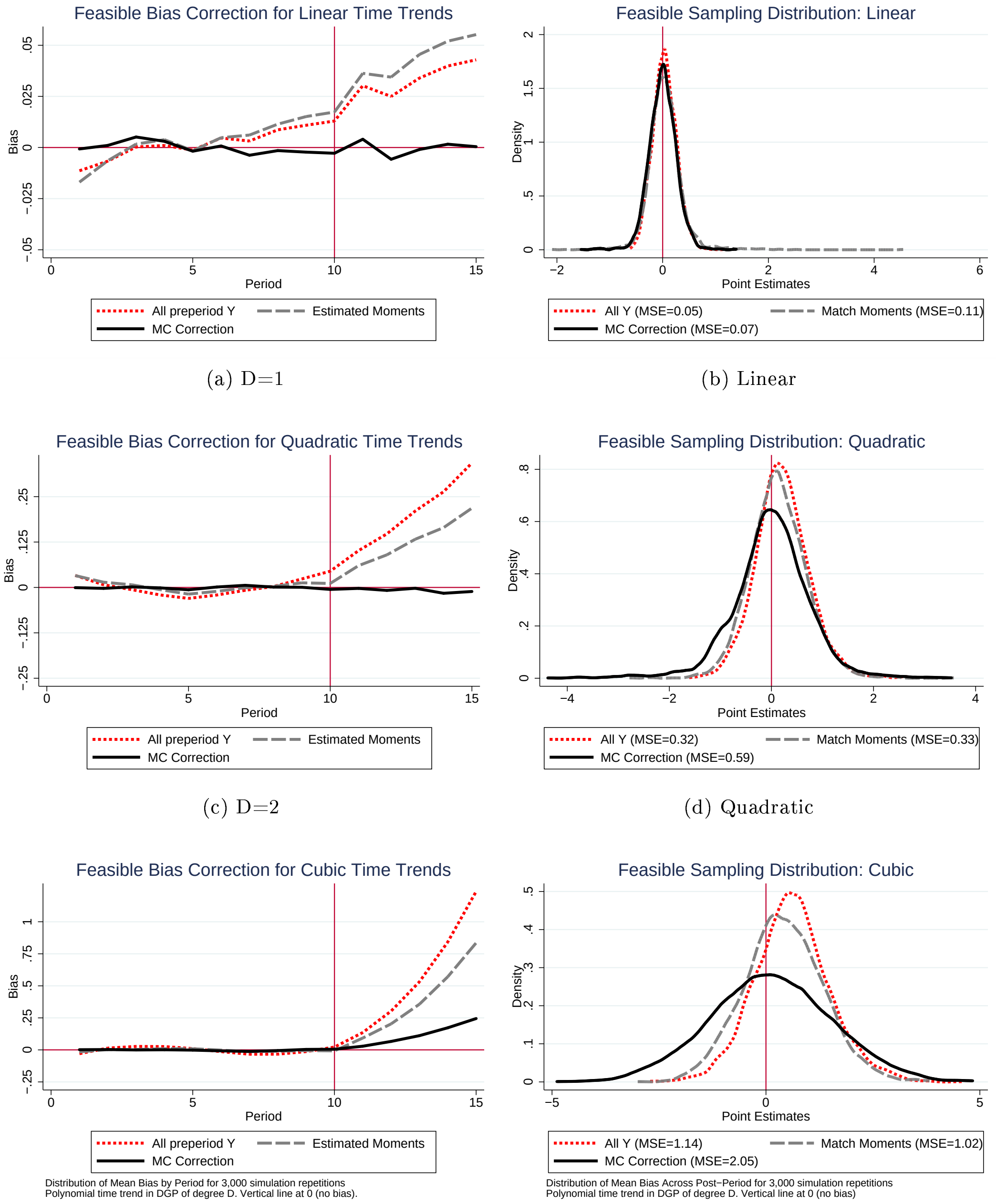
almost 100% in the case of a cubic polynomial. In every case, BIC chooses $\hat{D} \geq D$.

6.2 Performance of Feasible Bias Corrections

We now present results from a simulation similar to that in Section 5.3, but using $\hat{D}$ in place of the true D. After choosing $\hat{D}$ via minimum BIC, we estimate the matching on moments model using a polynomial of degree $\hat{D}$, and separately run the direct bias-correction under the assumption of a $\hat{D}$-dimensional polynomial DGP.

Figure 10 shows the results of this exercise. The bias-reduction is less effective when we choose a DGP using $\hat{D}$ instead of D. Both methods generally continue to reduce bias relative to matching on all pre-period outcome variables, with the exception that the moments matching correction is now more biased in the linear case than the baseline specification. This failure of the feasible implementation is easy to interpret given Figure 9 – the estimates of $\hat{D}$ for D=1 are by far the least accurate and by including extra, meaningless parameters up to degree 4 the moments matching method invites additional matching on noise. Surprisingly, though, this is a much smaller problem for the Monte Carlo correction, which continues to be nearly unbiased in the feasible implementation through degree 2.[15] Results for the cubic case are virtually identical to those in Figure 8, since in this case the BIC criteria correctly chooses $\hat{D}$=3 nearly 100% of the time.

[15] The near-zero bias in the quadratic case is somewhat misleading. Figure 8 shows that the simulation-based correction slightly over-corrects and produces a negative bias when D=2 is known. When $\hat{D}$ is mis-specified it creates a positive bias, and these opposing biases roughly even out in this particular simulation. Note that the MSE is still lower in the infeasible case.



(a) D=1

(b) Linear

(c) D=2

(d) Quadratic

Distribution of Mean Bias by Period for 3,000 simulation repetitions
Polynomial time trend in DGP of degree D. Vertical line at 0 (no bias).

Distribution of Mean Bias Across Post−Period for 3,000 simulation repetitions
Polynomial time trend in DGP of degree D. Vertical line at 0 (no bias)

(e) D=3

(f) Cubic

Figure 10: Performance of Bias Correction Procedures when DGP is Unknown (Feasible)

As in Figure 8 we include density plots of the sampling distributions of the estimates and their MSE. The lessons are broadly the same as with the infeasible results, but exacerbated due to the simulation runs where $\hat{D} \neq D$. None of the corrected estimates have an MSE below that achieved from matching on all pre-period Y values at any dimensionality D. While the corrections do continue to reduce bias, the cost in terms of variance is sufficiently high that their MSE is higher than the more biased estimates from matching on all pre-period outcomes.

7 Conclusion

We document a bias in synthetic controls estimates stemming from matching on idiosyncratic noise in the outcome variable. The magnitude of the bias is a function of the variance of model errors of the outcome variable, the complexity of the underlying DGP, and features of the data sample including number of potential donor units and length of pre-period.

We then present two potential correction procedures; one based on matching on $\hat{\mu}$, and a simulation-based correction that directly estimates the bias. Both corrections can lead to bias improvements under a range of settings for our polynomial-trend family of DGPs. However, both corrections generate higher MSE in the particular simulations we run. While we thus do not necessarily recommend the presented corrections be applied by practitioners except as a robustness exercise, we do hope our discussion serves three valuable purposes for econometricians and empirical researchers. First, our paper sheds light on a heretofore unknown problem with an increasingly popular program evaluation method. Second, our proposed correction procedures do point towards paths for future research and refinements. Finally, we hope researchers employing synthetic control methods to evaluate real world programs will employ similarly-motivated models and specifications in their work to convince themselves and readers that their matches are on meaningful features of the world, and not on noise.

References

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

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

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

Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton (2016). The value of connections in turbulent times: Evidence from the united states. *Journal of Financial Economics 121*(2), 368–391.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Bohn, S., M. Lofstrom, and S. Raphael (2014). Did the 2007 legal arizona workers act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics 96*(2), 258–269.

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

Courtemanche, C. J. and D. Zapata (2012). Does universal coverage improve health. *The Massachusetts Experience. Andrew Young School of Policy Studies.*

Cunningham, S. and M. Shah (2017, 12). Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health. *The Review of Economic Studies 85*(3), 1683–1715.

Donohue, J. J., A. Aneja, and K. D. Weber (2019). Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. *Journal of Empirical Legal Studies 16*(2), 198–247.

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

Eren, O., I. S. Ozbeklik, et al. (2011). Right-to-work laws and state-level economic outcomes: Evidence from the case studies of idaho and oklahoma using synthetic control method. *Department of Economics, University of Nevada, Las Vegas.*

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

Ferman, B. and C. Pinto (2017). Placebo tests for synthetic controls.

Ferman, B., C. Pinto, and V. Possebom (2017). Cherry picking with synthetic controls.

Fitzpatrick, M. D. (2008). Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The BE Journal of Economic Analysis & Policy 8*(1).

Gautier, P. A., A. Siegmann, and A. Van Vuuren (2009). Terrorism and attitudes towards minorities: The effect of the theo van gogh murder on house prices in amsterdam. *Journal of Urban Economics 65*(2), 113–126.

Hinrichs, P. (2014). Affirmative action bans and college graduation rates. *Economics of Education Review 42*, 43–52.

Kaul, A., S. Klößner, G. Pfeifer, and M. Schieler (2015). Synthetic control methods: Never use all pre-intervention outcomes together with covariates.

Kiesel, K. and S. B. Villas-Boas (2013). Can information costs affect consumer choice? nutritional labels in a supermarket experiment. *International Journal of Industrial Organization 31*(2), 153–163.

Klasik, D. (2013). The act of enrollment: The college enrollment effects of state-required college entrance exam testing. *Educational researcher 42*(3), 151–160.

Kreif, N., R. Grieve, D. Hangartner, A. J. Turner, S. Nikolova, and M. Sutton (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics 25*(12), 1514–1528.

Lee, W.-S. (2010). Comparative case studies of the effects of inflation targeting in emerging economies. *Oxford Economic Papers 63*(2), 375–397.

Nannicini, T. and R. Ricciuti (2010). Autocratic transitions and growth.

Peri, G. and V. Yasenov (2015). The labor market effects of a refugee wave: Applying the synthetic control method to the mariel boatlift. Technical report, National Bureau of Economic Research.

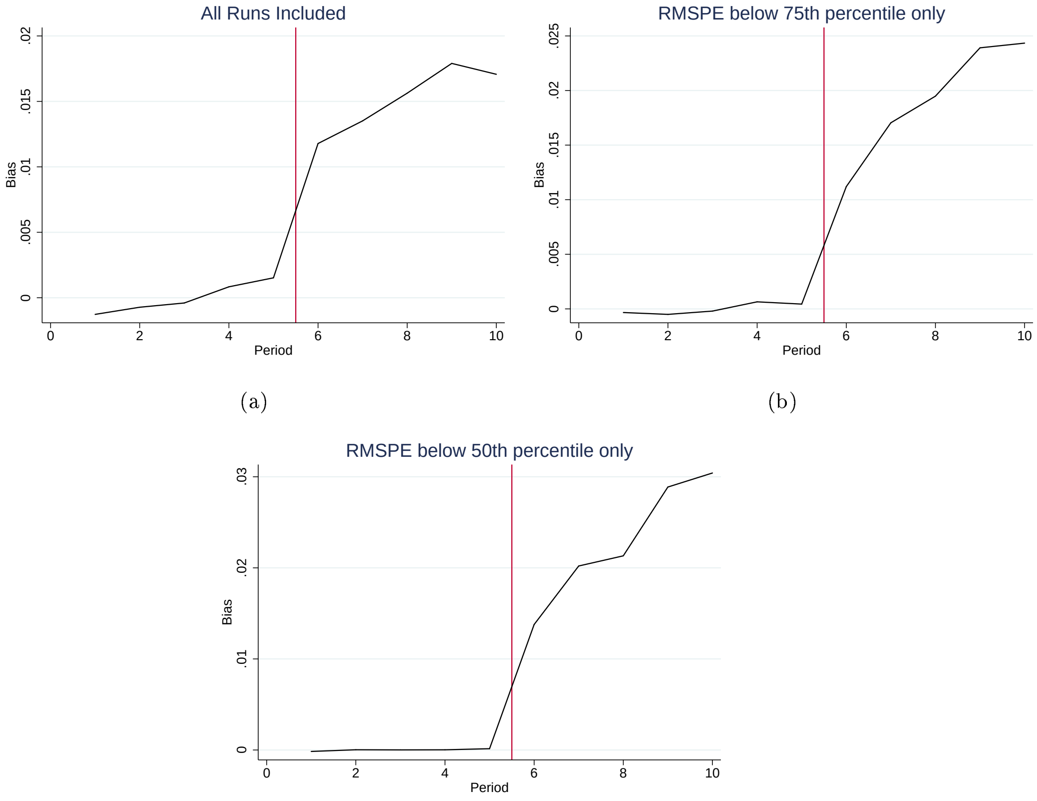
Smith, B. (2015). The resource curse exorcised: Evidence from a panel of countries. *Journal of Development Economics 116*, 57–73.

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

A Pre-treatment Trend in Bias

Figure 2 shows a slight pre-treatment trend in the bias for higher values of error term variance. As mentioned in the main text, this occurs because in some cases no perfect pre-treatment match is available. It is also a less interesting one since it will be apparent to the researcher, in the form of a bad pre-treatment match. Our paper focuses on the case where there is a good pre-treatment match. Figure A.1 demonstrates that the bias is not primarily driven by lack of available pre-treatment matches. Figure A.1a recreates the case in Figure 2, where error variance is equal to 0.2 and the treated unit trend rank is 22nd out of 31. There is a slight pre-trend in the bias. But if we drop Monte Carlo runs where the root mean squared pre-treatment error (RMSPE, a measure of the pre-treatment fit) is below the 75th percentile across all Monte Carlo runs and recreate the same graph, the trend is diminished. If we further limit the RMSPE to cases below the 50th percentile, where all cases are effectively a perfect match, the bias is nearly exactly zero throughout the pre-treatment period. But in all cases there is still a bias in the post period due to the spurious matching on error terms discussed in this paper.

We considered requiring a maximum RMSPE for a given Monte Carlo run for all analysis performed in this paper (recall that for all analysis we do require that the treated unit outcome is never the highest or lowest of all units in any pre-treatment period, making a match impossible). However, given the multitude of empirical settings we analyze, particularly when varying the dimensionality of the DGP, a consistent RMSPE standard that both enforces an excellent match and does not exclude the vast majority of runs proved impractical. Therefore we highlight the issue here and acknowledge that some of the bias we show in our analysis may be driven by mismatches in the pre-period (though also note that in most analyses we use 30 control units and conservatively set the treated unit rank at the 70th percentile, so good matches are theoretically feasible), but a bias exists even in the case of perfect matching).



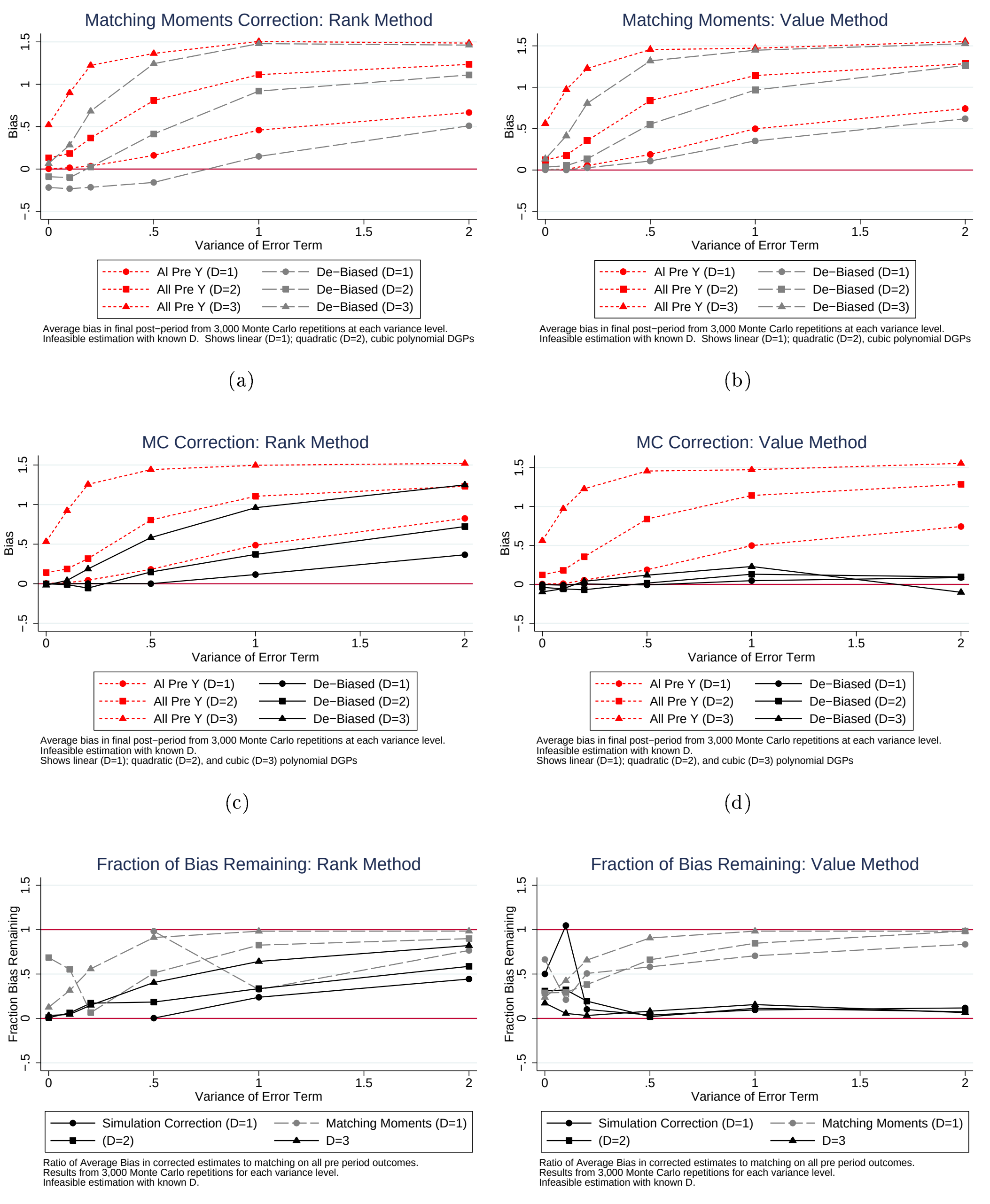
(a)

(b)

(c)

Figure A.1: Bias by Period for Different Match Quality Thresholds

B Alternative Correction Methods Performance by Noise Level



Matching Moments Correction: Rank Method

Average bias in final post−period from 3,000 Monte Carlo repetitions at each variance level.
Infeasible estimation with known D. Shows linear (D=1); quadratic (D=2), cubic polynomial DGPs

(a)

Matching Moments: Value Method

Average bias in final post−period from 3,000 Monte Carlo repetitions at each variance level.
Infeasible estimation with known D. Shows linear (D=1); quadratic (D=2), cubic polynomial DGPs

(b)

MC Correction: Rank Method

Average bias in final post−period from 3,000 Monte Carlo repetitions at each variance level.
Infeasible estimation with known D.
Shows linear (D=1); quadratic (D=2), and cubic (D=3) polynomial DGPs

(c)

MC Correction: Value Method

Average bias in final post−period from 3,000 Monte Carlo repetitions at each variance level.
Infeasible estimation with known D.
Shows linear (D=1); quadratic (D=2), and cubic (D=3) polynomial DGPs

(d)

Fraction of Bias Remaining: Rank Method

Ratio of Average Bias in corrected estimates to matching on all pre period outcomes.
Results from 3,000 Monte Carlo repetitions for each variance level.
Infeasible estimation with known D.

(e)

Fraction of Bias Remaining: Value Method

Ratio of Average Bias in corrected estimates to matching on all pre period outcomes.
Results from 3,000 Monte Carlo repetitions for each variance level.
Infeasible estimation with known D.

(f)

Figure A.2: Performance of Bias Correction Procedures by Rank and Value Methods