Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data

Alex Hollingsworth and Coady Wing*

June 2022

Abstract

Most applied synthetic control studies do not explain the identifying assumptions supporting causal inference. We describe these assumptions, illustrate how they can fail, and examine the consequences of such failures. We offer recommendations for discretionary implementation decisions, connecting each to a core identifying assumption. We also show how to implement a Synthetic Control Using Lasso, which allows a high-dimensional donor pool, automates model selection, allows donors from multiple variable types, and permits extrapolation and negative weights. In an application, we estimate how recreational marijuana legalization affects sales of alcohol and over-the-counter painkillers, finding reductions in alcohol sales.

JEL: C01, C55, C81, I18

Keywords: synthetic controls; marijuana legalization; high-dimensional data; identification

*Hollingsworth: hollinal@indiana.edu. O'Neill School of Public and Environmental Affairs. Indiana University, 1315 E. Tenth Street, Bloomington, IN, 47405 and National Bureau of Economic Research. Wing: cwing@indiana.edu. O'Neill School of Public and Environmental Affairs. Indiana University, 1315 E. Tenth Street, Bloomington, IN, 47405.

We thank Jonathan Kolstad, David Powell, Justin Ross, Jay Ryu, and participants at the ASHEcon Annual Conference, the Ohio State University, the University of Arizona, Miami University, Indiana University, and the IU/VU/Louisville Health Economics and Policy Conference for valuable comments. We thank Shyam Raman, Patrick Carlin, and Ashley Bradford for valuable research assistance.

Disclaimer: Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

An R package and example code using publicly available data are available here: https://hollina.github.io/scul.

1 Introduction

The synthetic control methodology is a strategy for estimating causal treatment effects for idiosyncratic historical events. In the typical application, researchers observe time series outcomes for both a treated unit and a number of untreated units. A weighted average of the untreated series is used as a counterfactual estimate of the treated series, which is referred to as a synthetic comparison group. Weights are chosen to minimize discrepancies between the synthetic comparison group and the treated unit in the pre-treatment time period. Treatment effect estimates are the difference between observed outcomes and the synthetic counterfactual. Statistical inference is normally organized around a placebo analysis.

The method is very popular and has enabled researchers to make progress in a range of research domains. Studies that present evidence from a synthetic control design are usually framed as quasi-experiments, and the method is increasingly considered part of the modern canon of design-based empirical strategies. However, there are some ways in which the synthetic control method does not fit in with the other study designs in the canon.

A hallmark of studies in the design-based tradition is a careful and transparent discussion of the nature and credibility of the study's identifying assumptions. For example, compelling applications of the difference-in-differences design include efforts to assess the credibility of the common trends assumption. Likewise, it is standard knowledge that instrumental variables must satisfy assumptions about relevance, monotonicity, independence, and exclusion. Although these restrictions are often empirically untestable, well-executed studies explain how the abstract assumptions apply in a given situation and often provide examples of behaviors that would violate the core assumptions.

Synthetic control studies generally do not follow this practice. Papers using a synthetic control estimator typically do not include a clear description of core identifying assumptions, even in an abstract form. They rarely discuss the types of real world behaviors or events that might threaten the validity of the treatment effect estimates or approach to statistical inference. Empirical tests (even partial tests) of the robustness of the estimates to violations of key assumptions are uncommon. One possible reason for these omissions is that the scientific community does not yet have a clear understanding of the technical and practical conditions under which the method is apt to work well. Another concern is that implementing a synthetic control study involves a series of operational decisions that are not documented or explained. These ambiguities are expanding as researchers apply the synthetic control technique to a broader range of situations involving multiple treated units (Abadie and L'Hour, 2019; Cavallo et al., 2013; Hainmueller, 2012; Robbins et al., 2017; Xu, 2017), microdata (Abadie and L'Hour, 2019; Robbins et al., 2017), extrapolation (Arkhangelsky et al., 2018; Doudchenko and Imbens, 2017), imperfect controls (Powell, 2019), and poor pre-treatment fit (Ben-Michael et al., 2018).

Our goal in this paper is to make the conceptual and practical challenges associated with synthetic controls more vivid to applied researchers. We make three main contributions. First, we state the core identification assumptions underlying the design, illustrating some of the ways the assumptions may fail in practice, and explaining the consequences of such failures. Second, we give a structured account of the main implementation decisions involved in a synthetic control design, highlighting the ways these decisions are connected to core identifying assumptions, and offering recommendations of tactics that seem useful in practice. Third, we present an original empirical study of the causal effects of Colorado's recreational marijuana law on the sales of other legal psychoactive substances, alcohol and over-the-counter pain killers.

Abadie et al. (2010) show that the synthetic control estimator is biased, but they also show that the expected (absolute) value of the bias is bounded by a function that decreases with the number of pre-treatment time periods included in the analysis. Although it is a slight abuse of terminology, we argue that the core identifying assumptions of the synthetic control estimator should be understood as those required for the estimator to have *bounded bias*. In Section 2 we provide an overview of the general synthetic controls procedure, outline the five "identification asumptions" that are needed for a synthetic control estimator to be bound satisfying, and connect each assumption to the proof of the bias bound.

The five assumptions are somewhat technical. To help clarify the way they matter in theory and in practice, Section 3 provides examples of violations, showing how each can lead to a violation of the bound. In addition, we follow the tradition developed in Campbell and Cook (1979) and describe some threats to validity that may often lead to failures of identifying assumptions in applied work. In particular, we argue that applied synthetic control studies should be concerned about the possibility of five main threats to validity: spillover and anticipation effects, structural breaks in the factor structure model, non-common factors that are systematically correlated with unit specific factor loadings (omitted variable bias), dormant or low-frequency factors, and overfitting.

In addition to the five identifying assumptions, many synthetic control studies also invoke additional ancillary assumptions that are not needed for identification but may have other advantages. These decisions are often ad hoc and undocumented. To serve as a guide for applied research, we catalogue these decisions and offer accompanying recommendations and reasons behind each recommendation, connecting them when possible to the core identification assumptions. Section 4 outlines the procedure used in traditional synthetic controls, shows how to conduct a synthetic control using lasso, introduces our construction of the average treatment effect, and discusses statistical inference. Throughout we attempt to outline general implementation recommendations and other quality control issues.

We make six recommendations that may help researchers implement their synthetic control studies more effectively: 1) use the same model selection procedure for both target and placebo products; 2) use a unit-free measure to evaluate model fit; 3) trim from the study, any target series or placebo series that fares poorly on a pre-specified threshold of model fit; 4) incorporate a rolling-origin cross-validation procedure to determine optimal weights, which helps guard against over-fitting and likely reduces bias; 5) use a unit-free measure of the treatment effect estimate to compare estimated treatment effects to the placebo distribution; and 6) report the minimum detectable effect size given the placebo distribution and a specified significance level.

In the empirical application in this paper, we use a wide range of donor variable types to construct our synthetic control groups – not just variables of the same type as the target variable. This leads to a high dimensional donor pool in which the number of control units exceeds the length of the pre-period. The conventional strategy for choosing synthetic control weights breaks down in this high dimensional case. In addition, a high dimensional donor pool may raise concerns about over-fitting in the pre-period data. To manage these problems, we pursue a Synthetic Control Using Lasso (SCUL) approach and rely heavily on a rolling cross-validation method. Our use of the SCUL strategy is in line with recent methodological work has proposed a number of alternative strategies for estimating synthetic control weights (Arkhangelsky et al., 2018; Doudchenko and Imbens, 2017; Powell, 2019). Our application illustrates the value these strategies can have as solutions to the the problem of high dimensional donor pools, and also explains why high dimensional donor pools may be valuable in practice.

The empirical results of our study are presented in Section 5 and represent an original contribution to the literature on the the regulation of psychoactive substances. We examine how the legalization of recreational marijuana in Colorado has affected the sales of alcohol and over-the-counter pain medication. The data for our analysis come from a large retail scanner database. We estimate treatment effects by comparing observed sales to a counterfactual synthetic time series for narrowly defined groups of products. The setting is complex because product-level sales data are detailed and highly variable. In addition, the scanner data allows for an extremely large donor pool of 2,340 donors containing both the focal products (alcohol and painkillers) and non-focal products (e.g., toilet paper and soda) sold in other states. The set of donor products is so large that traditional synthetic control methods are infeasible because the number of donors is greater than the number of pre-treatment time periods (357). This large set of candidate comparison groups also creates a huge pool of possible placebo products, which seems desirable for statistical inference, but also makes model selection for each placebo analysis more challenging. The incorporation of lasso regressions into the synthetic group construction alleviates both of these issues, allowing for a large donor pool and automating model selection. We find no statistically significant evidence that recreational marijuana laws affect the sale of over-the-counter pain relievers in Colorado. In contrast, we find evidence that recreational marijuana legalization reduces alcohol sales.

2 Synthetic control: Estimands, estimators, and identifying assumptions

In this section, we lay out notation to describe the data structure and overall research design that is typical in a synthetic control study. We define the causal estimand of interest in this setting, and we introduce a generic definition of a synthetic control estimator. A fundamental question at this stage is: what set of assumptions are required for the generic synthetic control estimator to perform well?

Empirical researchers typically work with estimators that are *unbiased* and/or *asymptotically consistent* under a set of assumptions on the data generating process. Abadie et al. (2010) present the seminal theoretical work on the synthetic control estimator. Interestingly, this paper does not prove conditions under which the estimator is unbiased or asymptotically consistent. Instead Abadie et al. (2010) show that the standard estimator is biased and inconsistent, but that expected value of the magnitude of the bias can be bounded. The assumptions required for the synthetic control estimator to respect the bias bound are the probably what most applied researchers should think of as the identifying assumptions for their research, although this is a slight abuse of the concept of identification.

After introducing some framing and notation, we lay out the set of five pseudo-identifying assumptions under which the bias of the synthetic control estimator will satisfy the bias bound. To ensure our discussion has the broadest applicability across different synthetic control methodologies, we focus on a simplified setting with no observable covariates and with no convexity restriction on the synthetic control weights. Appendix Section A provides an annotated derivation of the bias bound for this case, and it points out the role that each assumption plays in the proof. We also discuss the nature of the bias bound, and how it may be relevant to planning and interpreting synthetic control studies.

2.1 Notation and causal estimands

The synthetic control literature is primarily concerned with data generated by a version of the "comparative interrupted time series" design (Campbell and Cook, 1979). In the typical case, a collection of $s = 0...S$ units are observed over $t = 1...T$ time periods. At the outset, all units are untreated and one unit is exposed to treatment after period T_0 while the others remain untreated. Throughout, let $s = 0$ represent the treated unit and units $s = 1...S$ represent the untreated donor pool of candidate comparison units. Let $D_s = 1(s = 0)$ be an indicator of whether the unit is the treated unit, and let $1(t > T_0)$ indicate that period t is part of the post-treatment period. Then $D_{st} = D_s \times 1(t > T_0)$ is a binary treatment variable set to 1 if unit s has been exposed to treatment as of period t. Let $Y_{st}(0)$ and $Y_{st}(1)$ represent potential outcomes that record the outcome of unit s in period t under the control and treatment conditions. The observed outcome is $Y_{st} = Y_{st}(0) + D_{st}(Y_{st}(1) - Y_{st}(0))$. The standard causal parameter of interest in synthetic control studies is the effect of treatment on unit $s = 0$ in a given post treatment time period $t' > T_0$:

$$\beta_{0t'} = Y_{0t'}(1) - Y_{0t'}(0) \tag{1}$$

To estimate β_{0t}, researchers need some method of estimating $Y_{0t'}(0)$, the counterfactual untreated outcomes the treated unit would have experienced in the post-treatment time periods had treatment not occurred.

2.2 The synthetic control estimator

At a generic level, a synthetic control, Y_t^*, is a weighted combination of donor units that is meant to serve as an estimate of the treated unit's untreated outcomes. For example, suppose that $\pi_1, ..., \pi_S$ is a collection of weights for each unit in the donor pool. The synthetic control defined by this particular set of weights is:

$$Y_t^* = \sum_{s=1}^{S} Y_{st}\pi_s \tag{2}$$

The *synthetic control estimator* of $\beta_{0t'}$ for some post-treatment $t' > T_0$ is:

$$\widehat{\beta_{st}} = Y_{0t'} - Y_{t'}^* = Y_{0t'}(1) - \sum_{s=1}^{S} Y_{st'}\pi_s \tag{3}$$

Presenting a generic synthetic control unit as *some* linear combination of controls, makes it clear that are an infinite number of ways to form a synthetic control unit and a corresponding treatment effect estimator. One option, for example, is to set $\pi_s = \frac{1}{S}$. In that case, the synthetic control is simply the average outcome in the donor pool, and the synthetic control estimator is simply the difference between the treated unit's outcome and the average outcome in the donor pool in period $t' > T_0$. Of course, a simple average is unlikely to work well in most situations. In practice, good synthetic control studies choose weights in a more sophisticated fashion. We will use an asterisks—i.e. π_s^*—to differentiate a weight that has been selected by some procedure from some potential weight, π_s.

2.3 Identifying assumptions

The credibility of empirical research based on synthetic control strategies depends on the claim that the bias of the synthetic control estimator will be small or negligible in a particular application. The bound on the absolute value of the bias of the synthetic control estimator relies on the following five assumptions:

Assumption 1 *No Interference Between Units. Each unit's potential outcome in a period depends only on that unit's own treatment exposure in that period.*

Assumption 2 *Factor Structure Model. The time series of untreated potential outcomes for each unit is generated by: $Y_{st}(0) = a_t \alpha_s + \varepsilon_{st}$. In the model, a_t is a $1 \times F$ vector of unmeasured "common factors" that may vary across periods but are constant across units. α_s is an $F \times 1$ vector of unit specific coefficients on the common factors; these time-invariant coefficients are often called factor loadings. ε_{st} is a non-common factor that varies across periods and across units and is usually referred to as a "transitory shock".*

Assumption 3 *Exogenous Shocks. The non-common factors represented by ε_{st} are assumed to satisfy three conditions: (a) $E[|\varepsilon_{st}|^g] < \infty$ for some even integer $g \geq 2$ for all s and t, (b) ε_{st} is independently but not necessarily identically distributed across units and periods, and (c) $E[\varepsilon_{st}|\alpha_s, D_s = 1] = E[\varepsilon_{st}|\alpha_s, D_s = 0] = E[\varepsilon_{st}] = 0$.*

Assumption 4 *No Pre-Period Perfect Multicollinearity of Common Factors. The vector of common factors a_t must be linearly independent during the pre-period. Let A_P be the $T_0 \times F$ matrix of common factors from only the pre-period, and let ξ_{A_P} be the smallest eigenvalue of $\frac{1}{T_0} A_P^T A_P$. Assume that $\xi_{A_P} > 0$. This is a full rank condition that ensures that $A_P^T A_P$ is invertible.*

Assumption 5 *Existence of Weights. There is some collection of weights $\pi_1^*, ..., \pi_S^*$ that creates a synthetic control $Y_t^* = \sum_{s=1}^{S} Y_{st} \pi_s^*$ such that $Y_{0t} = Y_t^*$ for all $t = 1...T_0$.*

Assumption 1 is implicit in the potential outcomes notation. It rules out the possibility of spillover effects from the treated unit to one or more of the comparison units. It also implies that there are no anticipation effects associated with future treatment exposures in any period. Assumption 2 is an interactive fixed effects model for the conditional expectation function linking untreated potential outcomes over time for each unit. Under the interactive fixed effects model, the untreated potential outcomes can be decomposed into two unobserved components: unit specific responses to time varying common factors $(a_t \alpha_s)$ and non-common factors that differ across units and time periods (ε_{st}). The standard two way fixed effects model is a restricted special case of the interactive fixed effects model. To see this, suppose that $a_t = [1 \ \gamma_t]$ and $\alpha_s = [\theta_s \ 1]^T$. Then the model in assumption 2 simplifies to: $Y_{st}(0) = \theta_s + \gamma_t + \varepsilon_{st}$. Under the two way fixed effects model, the common trends assumption is valid and treatment effects can be estimated consistently using difference-in-differences and fixed effects estimators. But these standard estimators fail if the data are generated by the general case interactive fixed effects model–in which a_t is an unrestricted $1 \times F$ vector of common factors. Thus, the more flexible model specification in Assumption 2 is one of the main reasons to pursue synthetic control studies in applied work. Under the model in Assumption 2, units may follow complicated differential trends over time because each unit responds differently to a common set of underlying time varying factors.

There are, however, some limitations on the sources of differential trends across units that are allowed in the synthetic control framework. In particular, Assumption 3 imposes three restrictions on the distribution of non-common factors, ε_{st}. Assumption 3(a) is a regularity condition that would be satisfied, for example, if the variance of $|\varepsilon_{st}|$ is finite. The independent sampling condition in Assumption 3(b) rules out serial correlation or spatial correlation in the non-common factors.[1] Finally, Assumption 3(c) is a conditional mean independence assumption. When it holds, the non-common factors have mean zero for each unit and period, and are not correlated with the treatment status of the unit. This is a kind of matching assumption that implies that–on average–two units with the same α_s will experience the same untreated potential outcomes in any given period.

Assumption 4 requires that the common factors generating the unit specific trends are linearly independent during the pre-period.[2] Finally, Assumption 5 is that the researcher has found some collection of weights that creates a synthetic control that perfectly matches the pre-treatment time series of the treated unit. Assumptions 1-4 do not guarantee that such weights exist. However, Assumption 5 is empirically verifiable in application: researchers can simply compare the treated time series with the candidate synthetic control and measure discrepancies in each period. While Assumption 5 requires a perfect fit, in practice it will only hold approximately and must be balanced against concerns of overfitting (i.e., matching on ε_{st} rather than underlying factor structure).

2.4 Restrictions on weights

In addition to Assumptions 1-5, many synthetic control studies impose additional restrictions on the weights used to form the synthetic control. In particular, many follow the approach in Abadie et al. (2010), constraining the synthetic control weights to be non-negative and sum to 1. This restriction is appealing because it forces the synthetic control to be a convex combination of the donor units, preventing extrapolation outside the convex hull of the donor pool. However in general, such weight restrictions are not needed to create a synthetic control and many recent papers propose synthetic control estimators with relaxed weight restrictions (Doudchenko and Imbens, 2017; Ben-Michael et al., 2018; Abadie and Gardeazabal, 2003), including the synthetic control using lasso (SCUL) approach we pursue in this paper.

It is important to note that the convexity and non-negativity weight restrictions play only a minor role in the derivation of the bias bound presented in Abadie et al. (2010). In Appendix Sections A and B, we extend the bias bound proof to show that the bias bound result continues to hold without these restrictions.[3]

[1] Note that this does not mean that the outcomes, Y_{st}, must be independent across units and time. But it does imply that any such dependence must come through the common factors part of the model – $a_t \alpha_s$ – rather than the non-common factors—ε_{st}—part of the model.

[2] Abadie et al. (2010) state this condition by requiring that smallest eigenvalue of the pre-treatment matrix of common factors is bounded away from zero. This ensures that the matrix of pre-treatment common factors has an inverse.

[3] The logic is straightforward. For any synthetic control $Y_t^* = \sum_{s=1}^{S} Y_{st} \pi_s^*$ with unrestricted weights (i.e. some or all $\pi_s^* < 0$, and

Although this point is simple, it implies that the bias bound remains valid for a broader class of synthetic control estimators which allow for negative weights and weights that do not sum to 1.

2.5 The bias of the synthetic control estimator

Under Assumptions 1-5, the synthetic control estimator can be written:

$$\widehat{\beta_{0t}} = Y_{0t'}(1) - \sum_{s=1}^{S} Y_{st'}\pi_s^* = (\beta_{0t'} + a_t\alpha_0 + \varepsilon_{0t}) - \left(\sum_{s=1}^{S}(a_t\alpha_s + \varepsilon_{st})\pi_s^*\right) \tag{4}$$

In Appendix Section A, we show that under Assumptions 1-5 the expected value of the synthetic control estimator is:

$$E[\hat{\beta}_{0t'}] = \beta_{0t'} + E[a_{t'}A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P] \tag{5}$$

In the equation, $a_{t'}$ is the value of the vector of common factors in post-period $t = t'$, and A_P is the $T_0 \times F$ matrix of the common factors that prevailed during the pre-period. ε_s^P is the $T_0 \times 1$ vector of non-common factors realized by unit s during the pre-period.

The expression shows that—under Assumptions 1-5—a synthetic control estimator that perfectly matches the pre-treatment outcome time series in the treated unit is equal to the true effect *plus* a bias term equal to $E[a_{t'}A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P]$. The bias comes from the pre-treatment (finite sample) association between the non-common factors (ε_S^P) and the common factors (A_P). The bias survives even after taking expectations because the synthetic control weights are chosen using the pre-treatment data itself.

In essence, bias occurs due to overfitting the pre-treatment time series of the treated unit. This creates an important tension in applied work because researchers will tend to seek a very close pre-treatment fit in order to ensure that Assumption 5 holds. However, choosing the closest possible pre-treatment fit may actually increase bias if the improved fit is achieved by matching on a spurious pre-treatment pattern that does not persist in the post-treatment period.

A related point is that the bias of the synthetic control estimator is tightly connected to the relative importance of the common factors and non-common factors in the overall variance of the outcome variable. Under the interactive fixed effects model, the variance of the untreated outcome can be decomposed into two unobserved components as: $V(Y_{st}(0)) = V(a_t\alpha_s) + V(\varepsilon_{st})$. If $V(\varepsilon_{st}) = 0$ then the synthetic control estimator is unbiased. Moreover, the risk of overfitting induced bias will tend to be larger when $V(\varepsilon_{st})$ represents a larger share of the overall variance of $Y_{st}(0)$.

Abadie et al. (2010) shows that under Assumptions 1-5, the absolute bias of the synthetic control estimator is bounded by a function that will be close to zero when the number of pre-treatment time periods is large. The bound is a somewhat complicated object and depends on particular measures of the scale of the common factors in the pre-period, the degree of multicollinearity between the common factors in the pre-period, the

$\sum_{s=1}^{S} \pi_s^* \neq 1$), the donor units can be rescaled so that $Y_t^* = \sum_{s=1}^{S} Y_{st}\pi_s^* = \sum_{s=1}^{S}(\frac{1}{c_s}Y_{st})(c_s\pi_s^*) = \sum_{s=1}^{S}\tilde{Y}_{st}\tilde{\pi}_s^*$, where the $\tilde{\pi}_s^*$ will be non-negative and sum to 1. In Appendix Section B we show that using $c_s = \frac{1}{sign(\pi_s^*)\times\sum_{s=1}^{S}|\pi_s^*|}$ enables the proof to proceed without loss of generality.

number of units in the donor pool, the scale of the non-common factors, and the number of pre-treatment time periods. Appendix Section A derives the bound for the model we use in this paper, and spells out the details of each of the factors that contribute to the bound. To present the bound in a relatively simple form, we take the regularity condition in Assumption 3(a) to hold for $g = 2$. Then the absolute bias of the synthetic control estimator satisfies:

$$|E[\hat{\beta}_{0t'} - \beta_{0t'}]| \leq \sqrt{CS} \times \frac{\bar{a}^2 F}{\xi_{A_P}} \times \sqrt{\frac{\bar{m}}{T_0}} \tag{6}$$

In this expression, C is a constant and S is the number of donor units. $\xi_{A_P} > 0$ is the smallest eigenvalue of $\frac{1}{T_0} A_P^T A_P$, which is a measure of the linear independence of the common factors in the pre-treatment period. $\bar{a}$ and $\bar{m}$ are measures of the scale of the common factors and the non-common factors. Where $\bar{a}$ is the largest absolute value realized by any of the F common factors in any period and $\bar{m} = max_{s=1...S}(\frac{1}{T_0}\sum_{t=1}^{T_0} E[|\varepsilon_{st}|^2])$ is maximum pre-period variance of the non-common factors across all units.

The structure of the bound is more complicated and potentially tighter when $g > 2$, but the main insights of the bound are not changed. The key point is that the final term on the right hand side multiplies the bound by $\sqrt{\frac{\bar{m}}{T_0}}$. This term will be nearly zero when the number of pre-treatment periods is large relative to the scale of the non-common factors. The remaining components of the bound depend on parameters that are usually unknown in applied work, but it is worth noting a few conceptual points. The bound will be larger when there are more units in the donor pool and when there are more common factors. The bound will also be larger when the common factors are more variable and when the common factors are highly collinear with one another in the pre-period.

The bias bound is a repeated sampling concept. Imagine repeatedly sampling S units and T periods from a fixed data generating process that complies with Assumptions 1-5. The expected (absolute) bias of the synthetic control estimator across repeated samples will not exceed the bound. The magnitude of the maximum expected bias depends on features of the data generating process, and the estimator will be nearly unbiased if the number of pre-treatment periods is sufficiently large compared to the scale of the non-common factors (transitory shocks). The observation that the bias bound is shrinking in the number of pre-treatment periods does not imply that the synthetic control estimator is asymptotically consistent. Even in a setting where the bias bound is small (perhaps because the number of pre-treatment periods is large), the treatment effect estimate in a given realized sample could be quite different than the true treatment effect.[4]

2.6 Advantages of synthetic control estimators

The discussion so far has introduced the synthetic control estimator, described its assumptions, and explained the bias bound. But why should researchers want to use the synthetic control estimator in practice? And what other approaches to identification and estimation might be viewed as competitors?

Given a setting with time series outcomes on a group of untreated units and a smaller group of treated

[4] Consistency would imply $Pr(|\hat{\beta_{st'}} - \beta_{st'}| > \delta) \rightarrow 0$ as $T_0 \rightarrow \infty$ for any positive number, δ. This would imply that the probability of even a small estimation error would decline to zero as the number of time periods grew to infinity. This is not what happens with the synthetic control estimator. Even with a large number of pre-periods, the bias bound implies only that the average absolute bias will be small.

units (or perhaps a single treated unit), the two most obvious alternative approaches are: (i) panel data regressions with additive fixed effects (Chamberlain, 1984), and (ii) factor model estimation strategies (Bai, 2009; Gobillon and Magnac, 2016).

In applied work, one of the biggest advantages of the synthetic control strategy is that it offers a way to weaken the functional form assumptions required by additive fixed effects models. The standard two way fixed effects model allows for two specific types of unobserved confounders: unit fixed effects that vary across units but do not change over time, and time fixed effects that change over time but do not vary across units. Together these restrictions represent the common trends assumption. In contrast, the functional form assumption in Assumption 2 implies that the synthetic control estimator is applicable in settings the common trends assumption in the additive fixed effects model does not hold. Notice that the synthetic control estimator does not require the researcher to have complete knowledge of the specific list of variables represented by the a_t and α_s components of the model. But it does require that – whatever these unobserved terms are – they must satisfy restrictions described in Assumptions 3 and 4. In this sense, the synthetic control estimator is an alternative to the two way fixed effects workhorse that is robust to a particular class of violations of the common trends assumption. There are, of course, alternative strategies for coping with possible violations of the common trends assumption. Researchers sometimes fit fixed effects models that include unit specific linear time trends, for example, which is another way to weaken the common trends assumption. Future work might consider the relative advantages of synthetic control vs other strategies.

Factor structure models that attempt to directly estimate or eliminate an interactive fixed effect provide another approach to estimation in settings where the additive fixed effects model is not appropriate. This kind of approach also proceeds by imposing restrictions on the nature of the unobserved component of the model (Holtz-Eakin et al., 1988; Bai, 2009), and often requires researchers to assume or estimate the number of common factors (Stock and Watson, 2002). These methods appear to be more widely used in macroeconomics and finance than in empirical microeconomics. Future research examining the differences in assumptions, data requirements, and performance of factor model estimators and synthetic control estimators may be productive.

3 Threats to validity

The five bias bound assumptions are somewhat abstract because they refer to unobserved random variables: common factors, non-common factors, and unit specific loadings. In this section, we describe a collection of "threats to validity" that may occur in applied research and that would violate one or more of the key synthetic control assumptions. The five threats we focus on are: spillover and anticipation effects, structural breaks, omitted variable bias, dormant factors, and overfitting. This list is not meant to be an exhaustive catalogue of the ways that a synthetic control study can be misleading in practice. Our purpose is to describe stylized versions of behavioral patterns and situations that come up in regularly in empirical practice and that would violate one or more of the assumptions required for the synthetic control estimator to respect the bias bound. Thinking carefully about these kinds situations – and others like them – may help researchers assess the credibility of a given synthetic control study, recognize possible sources of bias, and develop robustness checks and sensitivity analysis.

3.1 Spillover and anticipation effects

Spillover effects are an important threat to the validity of the synthetic control estimator. For example, in some situations a policy in one state may lead to changes in behaviors and outcomes in geographically adjacent states. This kind of geographical spillover violates Assumption 1, which requires that there is no interference between units. Assumption 1 also fails in situations where a policy change is anticipated and begins to affect outcomes in advance of the nominal implementation date. In the absence of Assumption 1, the synthetic control estimator may have bias that is larger than suggested by the bias bound.

In practice, researchers do not know for sure whether spillover or anticipation effects have occurred or how far they are apt to extend. Good applied work should state these concerns and present arguments and perhaps supporting analysis that probes the sensitivity of the results to violations of these assumptions. For example, if geographic spillovers are a concern researchers might consider restricting the donor pool to units that are not geographically adjacent to the treated unit. It may be more likely that Assumption 1 holds for the restricted donor pool. Likewise, concerns about anticipation effects can be explored by varying the nominal start date of the policy in the analysis. These kinds of sensitivity checks seem logical but they are not without costs. For example, restricting the donor pool or the length of the pre-treatment period may make it more difficult to find synthetic control weights that match the pre-treatment outcomes of the treated unit.

3.2 Structural breaks

Assumption 2 implies that the data for each unit is generated by an interactive fixed effects model that is stable throughout the study period. A structural break in which the parameters of the interactive fixed effects model change over the course of the study period is an important threat to validity in synthetic control studies. To see the problem, suppose that instead of the factor structure model described in Assumption 2, the data are actually generated by:

$$Y_{st}(0) = a_t\alpha_s + a_t \times 1(t \geq t^*)\theta_s + \varepsilon_{st} \tag{7}$$

In this specification, the unit specific factor loadings change from α_s to $\alpha_s + \theta_s$ after period t^*. Assumption 2 fails if the factor loadings change for the treated unit or any of the donor units at any point during the study window. Without Assumption 2, the bias of the synthetic control estimator could exceed the bias bound.

The intuition for why a structural break may create problems is fairly clear. Suppose a synthetic control is formed by finding weights that closely match the pre-period outcomes of the treated unit. The idea is that by matching the outcome series, the synthetic control is implicitly matching the treated unit's factor loadings. However, if there is a structural break during the pre-period, a perfectly matched comparison unit will be matched on some mixture of the pre-break and post-break factor loadings. In contrast, post-treatment outcomes will be governed by only the post-break factor loadings. Thus unbounded bias is possible. A similar problem arises if a structural break occurs in the post-period. In that case, the pre-treatment factor loadings that lead to the match are no longer a sound basis for extrapolation during the post-period.

Structural breaks could occur for a variety of reasons in practice. Roine and Waldenström (2011), for example, study the time series of top income shares across a set 18 countries over the course of the 20th century. The find evidence of a shared structural break at the end of World War II in 1945, which marked a slowdown in the decline in top income shares. They also find evidence of a shared break in 1980, when top income shares stopped falling and began to rise. In addition to these common breaks, they also find evidence of additional structural breaks that are unique to specific countries or groups of countries. In another recent example, Evans et al. (2019) find evidence that there was a structural break in Oxycodone consumption/prescriptions in August 2010 and in Heroin overdose deaths in September 2010, which coincides with the introduction of abuse-deterrant Oxycodone. They find no evidence of a corresponding structural break for other illegal drugs.

Neither of these papers were using time series data as the basis for synthetic control studies. They simply represent two situations where structural breaks appear to have occurred. Changes like these ones would complicate the use of synthetic control studies. Applied researchers should think carefully about the possibility of structural breaks in their data. The possibility of structural breaks may act as a check on the length of the pre-treatment and post-treatment time periods used in a given study. Other things equal, adding more pre-treatment time periods to a synthetic control study is desirable because it should reduce the size of the bias bound. However, it is plausible that very long pre-treatment periods may be more likely to contain one or more structural breaks. In that case, a longer pre-treatment period could lead to more bias because of violations of Assumption 2. Future research should examine the the use of data driven tools to assess concerns about structural breaks, and perhaps to guide decisions about the length of a study window or the construction of a valid donor pool.

3.3 Omitted variable bias

Omitted variable bias is one of the most common threats to validity in observational research. But how should we think about omitted variable bias in the context of the synthetic control estimator? As a starting point, it is useful to note that Assumptions 1-3 represent a version of the conditional independence assumption, which plays an important role in the literature on matching. Matching estimators (including multivariate regressions) turn on the assumption that treated and untreated observations with the same observable covariates are not systematically different with respect to unobserved covariates that might also affect the outcomes under study. In the synthetic control setting, Assumptions 1-3 imply that units that are matched with respect to the *unit specific factor loadings* (α_s) do not differ systematically with respect to non-common factors (ε_{st}) that may also affect outcomes. Specifically, the expected value of the untreated potential outcome for the treated unit in period t' is $E[Y_{0t'}(0)] = a_{t'}\alpha_0$. The expected value of the untreated outcome among untreated donor units that have the same factor loading parameter as the treated unit is:

$$\begin{aligned} E[Y_{st'}(0)|D_s = 0, \alpha_s = \alpha_0] &= E[a_{t'}\alpha_s + \varepsilon_{st}|D_s = 0, \alpha_s = \alpha_0] \\ &= a_{t'}\alpha_0 + E[\varepsilon_{st'}|D_s = 0, \alpha_s = \alpha_0] \\ &= a_{t'}\alpha_0 \end{aligned} \tag{8}$$

The first line re-writes the potential outcomes in terms of the factor structure model. In the second line, the common factors and the factor loadings are taken as constant because of the conditioning. The third equality follows from Assumption 3, which implies that ε_{st} has mean zero and is mean independent of treatment status and factor loadings. The upshot is that if the α_s were observed characteristics of each unit and Assumptions 1-3 were valid, we could estimate the treatment effect using a standard matching estimator. The key omitted variable bias concern would be that, even after matching on the information in α_s, the treated and comparison groups were still unbalanced with respect to some variable contained in ε_{st}.

In synthetic control studies, of course, the α_s are unobserved so direct matching on the factor loadings is not feasible. The synthetic control estimator *works* because – under Assumptions 1-5 – a comparison group that matches the treated unit's pre-treatment outcome history is implicitly matching on the underlying factor loadings. The close connection between matching estimators and the synthetic control estimator makes it clear that a kind of omitted variable bias is a potential threat to validity in synthetic control studies. The concern in a synthetic control study is that the treated unit's outcomes are partly driven by time varying factors that do not also affect the donor units. In that case, Assumption 3 fails and the bias of the synthetic control estimator may exceed the bias bound.

Omitted variable bias might arise in a synthetic control study if the treated unit adopts more than one policy change during the study period or even at the same time. For example, the empirical application we present later in the paper is a study of the effects of Colorado's recreational marijuana law on the sales of alcohol, tobacco, and over the counter pain medication. Suppose that, in addition to passing a recreational marijuana law during our study period, Colorado had also made changes to the taxes and regulations it applies to alcohol sales. If these other policy changes affect product sales in Colorado then they would be an example of non-common factors that violate Assumption 3. Other policy events are merely one example of how omitted variable bias may occur in synthetic control studies. The key point is that the design is not robust time-varying factors that only matter for the treated unit and have no effect on the donor units.

3.4 Dormant factors

Assumption 4 requires that the the common factors in the factor structure model are not perfectly multicollinear during the pre-period. This is one of the more inscrutable conditions in the set of synthetic control identifying assumptions. The common factors are unobserved variables. And in practice, researchers usually don't have concrete ideas about how many common factors are part of the model or about what kinds of real world variables they represent. To make sense of Assumption 4, it helps to consider scenarios that would violate the condition. One example is a situation in which one of the common factors contained in a_t varies with low frequency relative to the outcome variable. In the extreme, the value of the low frequency factor might be constant across the entire pre-treatment period and then (finally) change to a new value in one or more post-treatment time periods. If one of the factor loadings is taken to be a unit specific intercept, then the intercept will be perfectly collinear with the "dormant factor" during the pre-treatment period. More generally, a dormant factor might be perfectly correlated with another factor during the pre-period and then vary independently in the post-period. Assumption 4 fails in this example and, as a result, the synthetic control estimator can have bias that exceeds the bias bound.

Appendix Section A.3 shows the role that Assumption 4 plays in deriving the bias bound, but that argument is fairly abstract. To see how a dormant factor leads to problems, consider a simplified example in which the potential outcomes are generated by a simple two factor model with no non-common factors, such that $Y_{st}(0) = \alpha_s^x x_t + \gamma_s^z z_t$. This is a nested special case of the model in Assumption 2 in which $a_t = [x_t\ z_t]$, $\alpha_s = [\alpha_s^x\ \gamma_s^z]^T$, and $Var(\varepsilon_{st}) = 0$. Removing the non-common factors means that under Assumptions 1-5 the synthetic control estimator is actually unbiased in the conventional sense. (The bias term is determined by spurious associations between the common factors and non-common factors in the pre-period. When no non-common factors are present, there is no bias.)

Now suppose that the first common factor in each period is a random draw from a normal distribution so that $x_t \sim N(0, \sigma_x)$. In contrast, suppose that the second dormant factor is generated so that $z_t = 1(t < T_0) \times x_t + 1(t >= T_0)N(0, \sigma_z)$. In other words, $x_t = z_t$ during the pre-period, but $x_t \neq z_t$ in the post-period. We refer to z_t as a "dormant factor" because it does not vary independently during the pre-period but then "wakes up" and begins to vary independently in the post-period.

To see why a dormant factor is a threat to validity in a synthetic control study, let $Y_t^d = \sum_{s=1}^S \pi_s^d Y_{st}$ be a candidate synthetic control for the treated unit in the dormant factor case. And suppose that Assumption 5 holds exactly so that $Y_{0t} = Y_t^d$ for all $t = 1...T_0$. When Assumption 4 fails because of the dormant factor, a perfect match on the pre-period outcomes does not guarantee a perfect match on the underlying factor loadings. To see this write the difference between the treated unit and the synthetic control unit in any given pre-period as:

$$\begin{aligned} 0 &= Y_{0t} - \sum_{s=1}^S \pi_s^d Y_{st} \qquad (9) \\ &= (\alpha_0 x_t + \gamma_0 z_t) - \sum_{s=1}^S \pi_s^d (\alpha_s x_t + \gamma_s z_t) \\ &= x_t(\alpha_0 + \gamma_0) - \sum_{s=1}^S \pi_s^d x_t(\alpha_s + \gamma_s) \\ &= x_t \theta_0 - x_t \sum_{s=1}^S \pi_s^d \theta_s \\ &= x_t \theta_0 - x_t \theta_0 \end{aligned}$$

The first equality defines the period specific residual for any of the pre-treatment time periods, where the residual is equal to zero due to the perfect matching assumption. We substitute the factor structure model for the potential outcomes, and then impose the dormant factor restriction from the data generating process, which holds that $x_t = z_t$ during the pre-period. The fourth equality substitutes $\theta_s = \alpha_s + \gamma_s$, which is an equivalent parameterization during the pre-period because of the dormant factor.

Since the pre-period residuals are equal to zero and x_t is a common factor, the weighted sum of θ_s terms must be equal to θ_0. This shows that a synthetic control that perfectly matches the pre-treatment time series of the treated unit can fail to match the treated unit's underlying factor loadings when Assumption 4 fails.

Matching on the "wrong loadings" – as in the example above – leads to bias in the post-period if the dormant factor wakes up. The difference between the realized treated outcome in the treated unit and the synthetic control unit a given post-treatment time period is:

$$
\begin{aligned}
\hat{\beta_{0t}} &= Y_{0t}(1) - \sum_{s=1}^{S} \pi_s^d Y_{st} \\
&= (\beta_{0t} + \alpha_0 x_t + \gamma_0 z_t) - \sum_{s=1}^{S} \pi_s^d (\alpha_s x_t + \gamma_s z_t) \\
&= (\beta_{0t} + \alpha_0 x_t + \gamma_0 z_t) - x_t \theta_0 \\
&= (\beta_{0t} + \alpha_0 x_t + \gamma_0 z_t) - (\alpha_0 + \gamma_0) x_t \\
&= \beta_{0t} + \alpha_0 x_t + \gamma_0 z_t - \alpha_0 x_t - \gamma_0 x_t \\
&= \beta_{0t} + \gamma_0 (z_t - x_t)
\end{aligned}
\tag{10}
$$

Here, the synthetic control estimator is the difference between the treated unit and the synthetic control unit in a given post-treatment period. We replace potential outcomes with the underlying factor structure in the second line. Because the estimator is applied in the post-period, the treated unit's outcome includes the treatment effect $\beta_{0t'}$. The third line shows that the synthetic control is matched on the response to $x_t = z_t$ and θ_0. We substitute $\theta_0 = \alpha_0 + \gamma_0$ and cancel the $\alpha_0 x_t$ terms. The final line shows that the post-treatment gap between the treated outcome and the synthetic control unit consists of the treatment effect plus a bias term equal to $\gamma_0(z_t - x_t)$. Since $z_t \neq x_t$ in the post-period, the estimator is biased unless $\gamma_0 = 0$ so that the dormant factor is irrelevant to the treated unit.

At a basic level, the dormant factors problem is similar to a structural break in that the patterns driving the outcome variable is different in the pre-period and post-period. The conceptual difference is that in the dormant factor case, the same model holds throughout the study period but there is not enough independent variation during the pre-treatment time period used to construct the synthetic control. In contrast, a structural break refers to a change in the parameters (factor loadings or functional form) that determines the outcomes.

Fundamentally, there may be a tension between minimizing the chance of structural break and ensuring that all dormant factors factors are captured. Increasing the number of pre-treatment time periods may mitigate concerns about missing a dormant factor. But a longer pre-period may raise concerns that a structural break may have occurred in the study window.

3.5 Overfitting

In practice, synthetic control weights are chosen so that the resulting synthetic control closely (perfectly) matches the treated unit's pre-treatment time series. In the absence of any association between the common factors (a_t) and the non-common factors (ε_{st}), a tight match on the pre-treatment outcomes implies a close match on the underlying factor loadings. Assumption 3 implies that the non-common factors are not associated with the common factors in the population data generating process. However, the population restriction does not guarantee that there will be no spurious associations between the common factors and the non-common factors during the pre-treatment time periods used in the study.

In Section 2.5, we explained that under Assumptions 1-5, the expected value of the synthetic control treatment effect estimator is equal to the true treatment effect plus a bias term equal to $E[a_{t'} A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P]$.

The bias term reflects a kind of overfitting that occurs because of spurious correlations between the common factors and the series of non-common factors that the treatment group experienced during the pre-treatment periods. In other words, the synthetic control weights inadvertently match on the spurious trends in the treated unit's error process. This creates bias in the estimator because the spurious patterns do not persist in the post-period.

The overfitting bias term that is built into the estimator is bounded under Assumption 1-5 and the bound will be small when the number of pre-periods is large relative to the scale of the non-common factors. Thus, one way to mitigate concerns about overfitting is to include a large number of pre-treatment time periods in the analysis. However, we think it will often be wise for applied researchers to present evidence that overfitting bias appears small in the application, and also to adopt methods like cross-validation that try to directly mitigate overfitting during the selection of synthetic control weights.

There is a large literature on cross-validation procedures that are appropriate in time series setting (Hyndman and Athanasopoulos, 2020). In Section 4.3, we outline how to use rolling-origin k-fold cross-validation in a synthetic control study, which is one method that we think works well in practice.

4 Implementing a synthetic control study

4.1 Classical synthetic control weights

Abadie et al. (2010) (ADH) developed the original and mostly widely used method of constructing a synthetic control group.[5] Their empirical goal was to estimate the effects of a California tobacco control policy implemented in 1988. The outcome of interest was cigarette sales per capita, measured annually at the state level from 1970 to 2000. The donor pool of comparison units consisted of cigarette sales in other states and the District of Columbia. Following our notation above, Y_{0t} would be cigarette sales in California in period t, and $(Y_{1t},\ldots,Y_{50t})$ would be the donor pool of $S = 50$ candidate control outcomes in period t. As in Equation 2, a synthetic control group is formed by applying a set of fixed weights, $\pi = (\pi_1,\ldots,\pi_S)$ to the donor pool.

The ADH method uses two types of pre-treatment data to construct the weights for each donor unit. The first type data is the time series of outcomes for the treated unit, Y_{0t} and the donor pool, Y_{st} for $s \in 1\ldots S$. The second type of data is a set of summary statistics on which the researcher desires balance between the treated and synthetic unit.[6] These summary statistics must be available for every donor unit, but need not be time series data. The ADH method finds a single, fixed weight for each donor unit that is applied to both the donor time series outcomes and the corresponding statistics of interest for each donor. The two types of data are not equally important in determining synthetic control weights. The method finds a second set of weights, called importance weights that trade off the importance of balance between the treated unit and synthetic series, and the importance of balance between each statistic of interest and each synthetic analogue.

Let Z_0 be the $L \times 1$ vector of summary statistics for the treated unit. In the original ADH example, the statistics included measures like average log GDP per capita and mean cigarette prices over a given time

[5] See Alberto Abadie (2020) for an overview of the traditional synthetic control method as well as recent extensions.

[6] In the literature these are often referred to as predictors, rather than summary statistics (Alberto Abadie, 2020). We opt for the phrase summary statistics since it more clearly illustrates the role these values have in the procedure.

period. Analogous statistics for each donor state are contained by Z_C, which is an $L \times S$ matrix in which each column represents a donor unit and each row contains a different statistic of interest. Thus $Z_0 - Z_C\pi$ is a $L \times 1$ vector of differences in pre-treatment statistics between the treatment group and the synthetic control defined by the $S \times 1$ vector of weights, π. ADH summarize the vector of differences with a single summary discrepancy score, $H = \sqrt{(Z_0 - Z_C\pi)^T V (Z_0 - Z_C\pi)}$, where V is a diagonal $L \times L$ matrix of importance weights with values $(v_1, \ldots, v_L)$ along the diagonal.[7]

The importance weights provide a way to penalize baseline treatment vs synthetic control discrepancies differently for each statistic. Given any choice of importance weights, V, the classical synthetic control estimator chooses π (donor weights) to minimize H, subject to the restriction that all of the donor weights are non-negative and the weights sum to one across all donors. In other words, an important additonal restriction of the traditional ADH method is that the weights satisfy a convexity constraint that $0 \leq \pi_s \leq 1, \forall s \in S$ and $\sum_{s=1}^{S} \pi_s = 1$. The convexity constraint, coupled with the need for the weights to balance the importance weighted summary statistics (i.e., minimizing H) encourages sparse synthetic control groups in which many weights are set to zero.

For clarity, let $\pi^*(V)$ be vector of $H - minimizing$ weights for a given importance weight matrix $V = diag(v_1, \ldots v_L)$. Re-writing H in vector form shows how $\pi^*(V)$ is the solution to:

$$\pi^*(V) = arg\ min_\pi \left(\sum_{l=1}^{L} v_l \left(Z_{0l} - \sum_{s=1}^{S} Z_{sl}\pi_s \right)^2 \right)\ s.t.\ 0 \leq \pi_s \leq 1, \forall s \in S\ \&\ \sum_{s=1}^{S} \pi_s = 1 \tag{11}$$

With these $H - minimizing$ weights in hand, the next step is estimate a candidate synthetic control as in Equation 2. ADH propose choosing V to minimize the mean square prediction error of the synthetic control during the pre-treatment period. We denote this minimizing importance weight matrix by V^*.

$$V^* = arg\ min_V \left(\sum_{t=1}^{T_0} \left(Y_{0t} - \sum_{s=1}^{S} Y_{st}\pi_s(V) \right)^2 \right) \tag{12}$$

In this way, the original ADH method contains two nested optimization procedures. In the inner procedure, for a given set of importance weights $(v_1, \ldots, v_L)$, one finds the set of donor weights π that minimizes H. Then, in the outer procedure, one finds the set of importance weights, V^*, that results in donor weights π^*. The two together minimize the pre-treatment mean squared prediction error of the synthetic control relative to the actual data.[8]

We use the notation $\pi^*(V^*)$ to denote the set of weights that minimize the mean squared error between the synthetic control estimate and actual data that arise from running this whole procedure. For simplicity we refer to this as π^* to indicate that these are the weights selected by some method or by π^*_{scm} when contrasting to these weights to weights chosen by a different synthetic control method.

The ADH method is the standard way of choosing synthetic control group weights in applied research.

[7] In Abadie et al. (2010) V is left sufficiently general so it need not be a diagonal matrix, however in applied work V is often assumed to be a diagonal matrix.

[8] Note that in practice the criterion need not be the mean squared error, but this is a common choice in empirical applications.

However, it is complex and imposes restrictions that are not always easy to interpret. For example, what are the advantages and disadvantages of requiring the weights to be non-negative and to sum to one across donor units?

4.2 Synthetic control using lasso (SCUL)

One alternative method for choosing synthetic control weights is a simple regression framework. For example, we could choose synthetic control weights by implementing an ordinary least squares regression on only pre-treatment data, choosing weights that minimize the sum of squared differences between the pre-treatment treated time series and the synthetic control group time series:

$$\pi^*_{OLS} = arg\ min_{\pi} \left(\sum_{t=1}^{T_0} \left(Y_{0t} - \sum_{s=1}^{S} Y_{st}\pi_s \right)^2 \right) \tag{13}$$

Here, the weights are simply the coefficients that arise from a regression of outcomes for the treated unit on the outcomes from each of the comparison units using only the $t = 1 \ldots T_0$ observations from the pre-treatment period. With the coefficients in hand, the synthetic control group is the predicted value from the regression for each period. In post-treatment time periods, the predicted values represent estimates of the counterfactual outcome based on the pre-treatment cross-sectional partial correlations between treated unit outcomes and each donor pool outcome. If the policy does induce a treatment effect on the outcomes, then the connection between treated outcomes and donor unit outcomes should change in the post-treatment period. That pattern will be measurable as an emerging difference between observed outcomes in the treated unit and the synthetic control series.

Notice that the simple regression based synthetic control in Equation 13 is quite similar to the outer function in the traditional synthetic control method from Equation 12. The key difference in the how the weights are selected is that in the simple regression procedure there are no importance weights V, summary statistics to be balanced on Z, or an additional inner optimization procedure (i.e., Equation 11) to consider.

Although it is familiar and intuitive, the OLS method may not be ideal for choosing synthetic control weights. Since there is no additional optimization procedure enforcing balance on other summary statistics, or convex weight restrictions that favor sparsity the OLS method may overfit the pre-treatment outcome data by emphasizing idiosyncratic correlations that are not a part of the true data-generating process. In that case, the synthetic control may have poor out-of-sample predictive performance. Another limitation is that the OLS estimator does not provide a unique set of weights in cases where there are more comparison units than pre-treatment observations (i.e., when $T_0 \leq S$).

An alternative approach is to choose synthetic control weights using a penalized regression method, such as the lasso. A lasso regression chooses synthetic control weights to solve:

$$\pi^*_{lasso} = arg\ min_{\pi} \left(\sum_{t=1}^{T_0} (Y_{0t} - \sum_{s=1}^{S} Y_{st}\pi_s)^2 + \lambda|\pi|_1 \right) \tag{14}$$

The lasso objective function consists of the same squared prediction error as OLS, but with an additional penalty that rises with the complexity of the vector of weights. In the expression, $|\pi|_1$ is the sum of the absolute values of the coefficients associated with each candidate control series. The penalty means that coefficients that are large in an unconstrained OLS regression shrink toward zero. Coefficients that are relatively small may shrink all the way to zero. Since some coefficients are set to zero, the lasso is able to estimate coefficients that minimize the penalized sum of squares even when the number of independent variables exceeds the number of observations.[9]

The regression framework underlying SCUL is different from the classical synthetic control in a few ways. First, the SCUL approach does not require a V-matrix of importance weights, and it does not make use of or attempt to match on any summary statistics other than the pre-treatment outcome series. Second, the regression framework relaxes the restriction that weights must be non-negative and sum to one. It is straightforward, for example, to add an intercept to the model by including a comparison unit that is simply equal to a constant in every period.

The SCUL estimator will be most useful when a researcher is faced with a pool of theoretically valid donor units that is "high dimensional" in the sense that the number of donor units is greater than the number of pre-treatment time periods available for analysis. SCUL will also be useful in the situation where a researcher believes it may be beneficial to allow for negative weights and weights that do not sum to one. For example, if one suspects that the synthetic control could benefit from including donors that are negatively correlated with the treated unit, or from creating a synthetic control that falls outside the convex hull of the donor pool. In addition, we think the SCUL estimator is often a clearer way to describe the synthetic control strategy than the classical estimator. Admittedly, this is partly an issue of tastes and preferences, but we think the regression-based SCUL presentation strips away complicated features such as the importance matrix that may obfuscate understanding of what is actually going on behind the scenes in a synthetic controls analysis.

4.3 Cross-validation

A key choice parameter in the lasso regression method is the penalty parameter, which is represented by λ in Equation 14. As λ increases, each weight in π^*_{lasso} will attenuate and the set of donors with non-zero weight will become more sparse as many weights are driven to zero. At one extreme, the penalty parameter could be so large that every weight is set to zero. At the other extreme, the penalty parameter could be set to zero, which would simply be the OLS estimator. Every choice of λ in between these extremes will result in a different set of unique weights. For each lasso regression, a series of λ choices are considered in a grid that starts at $\lambda = 0$ and increases until it reaches the smallest positive value of λ such that every coefficient is forced to be zero. The search stops at this point because the coefficients will continue to be set to zero for every λ greater than this stopping value.

If the goal of the regression is to maximize in-sample fit, then the λ that minimizes the root-mean square difference between the actual data and the synthetic series will be chosen from the set of candidate penalty parameters. However, maximizing in-sample fit almost certainly over-fits the model to the data and likely results in a prediction that would perform poorly out of sample. Since the goal of synthetic control studies is

[9] Tibshirani (2013) shows that the lasso solution for any given lambda (penalty parameter) is unique so long as there are no discrete predictor variables in the covariate set.

to estimate treatment effects following a treatment, our goal is to create a prediction that performs well out of sample. In the SCUL procedure, we select λ by using rolling-origin cross-validation, a procedure that rewards out-of-sample prediction and minimizes issues related to auto-correlation.[10]

Cross-validation is a simple procedure where the pre-treatment dataset is partitioned into multiple subsets that include training data (in-sample) and test (out-of-sample) data; multiple analyses are performed on the in-sample training data, providing a set of candidate penalty parameter values and associated weights; and from this set of candidate penalties, a single λ is determined using the out-of-sample test data. The surviving λ is called the cross-validated (denoted λ_{CV}) and is used as the penalty parameter in the post-treatment time period

Importantly, all data used in the cross-validation procedure (i.e., both the in-sample training and the out-of-sample test data) must be from the pre-treatment time period. This is because our goal is to create a synthetic time series that represents the counterfactual as if no treatment had occurred. Cross-validation's usefulness is not limited to the specific case of the lasso. For example, it could be used to select optimal donor weights in the traditional ADH synthetic control method.

Selecting the cross-validated penalty: The decision rule for selecting a single cross-validated penalty parameter using the out-of-sample data is heuristic. We start by creating various pairs of training and testing data that are subset from the whole pre-treatment time series; we discuss in the next section how these subsets are determined. We iterate over each training-testing pair.

Each of the K folds of this rolling cross-validation computes the out-of-sample mean square prediction error for each of a grid of candidate values for λ. Let λ_{k^*} be the choice of λ that produces the smallest cross-validation squared error in the k^{th} fold. In our empirical work, we choose the median of the collection of fold-specific cross-validation error minimizing λ's.

There are, of course, other ways to select a "best" penalty parameter from the set of fold-specific parameters. For example, one could select the λ that produces the minimum average mean squared error across all out-of-sample predictions. Another option is to select the largest λ whose associated average mean square error is within one standard error of the minimum average mean square error. This third approach might be motivated by a desire for additional sparsity since increasing the penalty will induce additional shrinkage (Hastie et al., 2009, pg. 244).

Figure A2 in the appendix provides a visual depiction of how the three penalty selection decisions rules apply for two different outcomes of interest from our empirical application. For the purposes of better understanding how the decision rules operate, it is not necessary to understand the empirical setting; it is enough to know that each panel comes from an analysis of a separate target outcome. Across the x-axis is the negative natural log of the penalty parameter, we follow this approach to match convention (Hastie et al., 2015). Note that this means sparser models with larger penalty parameters are on the left side and the penalty increases as we move along the x-axis to the right. Just above the x-axis are light gray tombstones with some transparency, these indicate a penalty parameter that provided the minimum mean squared error

[10] A similar method has been proposed by Kellogg et al. (2020). In general this type of method can be thought of as a type of model averaging designed to reduce overfitting and the influence of noise (Athey et al., 2019). Other smoothing or averaging procedures could perform similar functions (Amjad et al., 2018).

for a particular cross-validation iteration. The median minimum λ is denoted by the solid green vertical line.

The black points depict the average mean square error taken across all cross-validation iterations, for each candidate penalty parameter. The brackets around each point denote the uncertainty about this mean estimate showing the mean ± one standard error. The blue dashed line depicts the penalty parameter associated with the lowest average error across all cross-validation iterations. The red dashed line shows the largest λ whose associated average mean squared error is within one standard error of the minimum average mean squared error.

The mean square error for the outcome in the top-panel is more variant than the analogue in the bottom-panel. The λ selected by each decision rule is influenced by this variance and the difference between the three selected λs is larger in the top panel than in the less variant bottom-panel.

There is not a clear theoretical justification for preferring one of these decision rules to another. As such, it would be problematic if model fit, treatment effect estimates, or statistical significance varied greatly across different, but otherwise justifiable decision rules. Regardless of which decision rule is chosen as the preferred method, we consider it to be a reasonable robustness check to examine sensitivity of outcomes to using the other two choices. Specifically considering how pre-treatment model fit, treatment effect estimates, and p-values each vary with each decision rules.

Determining training data: The most common approaches to cross-validation work by excluding randomly selected observations or blocks of observations. However, we are not interested in finding a model that performs well at back forecasting or interpolating the time series between two points in time. The goal of our synthetic control strategy is to make out-of-sample forecasts for a time series. To pursue this goal, we use a cross-validation procedure in which the hold-out data always come from time periods after the training data in calendar time. This guards against an overfit synthetic control estimator that only performs well when it is able to use future information to forecast past events.

For example, in our application, our goal is to create a synthetic control that extends 165 weeks into the post-treatment time period.[11] Accordingly, we use a cross-validation procedure in which the test data is always at least 165 weeks long. To implement the method, we create a sequence of subsets of the pre-treatment dataset. Each of the individual datasets in the sequence covers a progressively longer time period. To demonstrate, let $k = 1 \dots K$ index the datasets in the sequence. The first dataset ($k = 1$) covers the period from January 2006 to May 2009. Each subsequent dataset adds one additional week of data until the Kth dataset containing all of the pre-treatment data up to October 2009. The final dataset stops at October 2009 because that is the latest data after which there are still 165 weeks of pre-treatment data left for the out-of-sample test. Constructing the sequence in this way means that we are able to use a total of K = 27 datasets for cross-validation purposes.[12] We present a visual example of this procedure in Figure 1.

We perform the same cross-validation procedure for every outcome and placebo series used in our analysis. In other words, the procedure we use to choose the lasso penalty is fixed across our entire analysis,

[11] This is the number of weeks from the date of recreational marijuana legalization by voter referendum in Colorado until the last week of data. That is, it is the maximum post-treatment time in our dataset.

[12] Because the cross-validation procedure is iterative, variation in each donor series is required in each set of training data for it to be separately identified from the intercept.

but the specific penalty is allowed to vary across each outcome variable. Once we have chosen the λ_{CV} penalty parameter for a given time series, we fit a lasso regression using data from the entire pre-period. The coefficients from that regression are then used as weights to construct the synthetic control for that target unit.

4.4 Interpretation of synthetic control weights

Neither the traditional ADH synthetic control weights nor the weights from the SCUL procedure can be directly interpreted as the share of the synthetic prediction composed by each donor series. The ADH weights are constrained to sum to one across the donor units. Weights therefore only represent the fraction of the total weight that is given to a particular donor series; they do not reflect the size and variability of the outcome for each unit across time periods. The SCUL weights, which are lasso regression coefficients, are not constrained to sum to one and are not naturally interpreted as the share given to a particular donor series. In both methods, the fraction of the synthetic prediction a given donor unit is responsible for changes with the value of the donor unit across time.

Applied researchers often like to understand the relative importance of each donor unit to the synthetic control estimation strategy. Weight shares are one way to help gain such understanding. In cases where outcomes vary substantially over time, it may be useful to present information on both the synthetic control weights and on each unit's share of the synthetic prediction in a selection of time periods. We give an example of this approach in the empirical section of the paper.

4.5 Evaluating synthetic control fit

There is no guarantee that the lasso regressions (or any other approach) can find a weighted mixture of donor units that closely mimics the treated unit during the pre-period. Therefore, researchers need a practical method of deciding whether a proposed synthetic control is "good enough" for proceeding with the analysis.

The existing literature on synthetic control estimation proposes a variety of methods for evaluating pre-period fit. But the focus is usually on deciding whether synthetic controls produced for placebo outcomes are of such low quality that they should be excluded from use in statistical inference. Methods for determining whether the synthetic control for the focal treatment unit is of sufficient quality appear to be entirely informal. For example, Abadie et al. (2010) do not explain how they decided that the synthetic control for California was good enough to justify their subsequent analysis. However, when they consider making statistical inferences based on placebo distributions, they report placebo distributions under alternative admissibility rules. In particular, they show the four placebo sampling distributions: one with all available placebos and then three restricted sets of placebos. The restricted distributions consider only placebos with a pre-period mean square prediction error (MSPE) that is less than 20 times, 5 times, and 2 times the pre-period prediction error observed for the treated unit.

There is logic to these methods. However, they rely on the performance of the synthetic control for the treated unit to guide quality control for the placebos. These ad hoc procedures do not provide an objective standard that researchers can use to determine the quality of a synthetic control for the treatment unit itself. In practice, most researchers likely judge quality by visually inspecting the graph of the realized outcome and the synthetic control in the pre-period. We draw on the cross-sectional matching literature to guide our assessment of the quality of a proposed synthetic control for any given outcome. In matching studies, researchers often assess covariate balance before and after matching using the Cohen's D statistic, which is simply the standardized mean difference in a baseline covariate between the treatment and control group.

One rule of thumb is that a covariate is out of balance if the Cohen's D statistic is greater than .25, which means that the imbalance between the groups is more than a quarter of a standard deviation for a particular variable (Ho et al., 2007; King and Zeng, 2006; Cochran, 1968). The specific choice of a Cohen's D threshold is arbitrary in most applications; in general, the smaller the discrepancy the better. However, the Cohen's D statistic is a unit-free, standardized metric that is comparable across different variables. To see why this is important, consider if the mean square prediction error across the pre-treatment time period relative to the treated unit was used as an inclusion threshold. Since the units of a MSPE depend on the specific outcome variable and sample of data under analysis, comparing the MSPE across very different outcome variables could be uninformative. Moreover, it would be meaningless to choose a single numeric MSPE standard across many different dependent variables.

We apply a modified version of the Cohen's D to evaluate pre-period fit. Specifically, we let $\sigma_s = \sqrt{\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_{st} - \overline{Y_s})^2}$ be the standard deviation of outcome s during the pre-treatment period. The pre-treatment average Cohen's D statistic for a proposed synthetic control is $D_s = \frac{1}{T_0}\sum_{t=1}^{T_0} |\frac{Y_{st} - Y_{st}^*}{\sigma_s}|$. We compute D_s for each synthetic control candidate in our study. If $D_s > 0.25$, we do not report a synthetic control estimate of the effect of the marijuana law on that outcome. We apply the same standard to the placebo products we use to conduct statistical inference. We describe the consequences of different Cohen's D inclusion thresholds for statistical precision and inference in Section 4.7.

4.6 Synthetic control, extrapolation, and convex hulls

Abadie et al. (2015) describe a different regression-based strategy for estimating the weights required to form a synthetic control group. Doudchenko and Imbens (2017) study several different ways of constructing synthetic control weights, including a strategy based on elastic net regression that is similar to the lasso approach we pursue in this paper. Similarly, Arkhangelsky et al. (2018) combine a regression based difference-in-differences strategy with synthetic controls to form weights. Abadie et al. (2010, 2015) argue that a key limitation of regression strategies is that they allow for negative weights that can facilitate extrapolation outside the support of the range of data in the donor pool. By requiring weights that are non-negative and sum to one, the Abadie et al. (2010) method limits the search for a good synthetic control to the subset of possible synthetic controls that can be formed as a convex combination of the donor units. This means that the value of the synthetic control at any pre-period time point must lie within the range of outcomes experienced by the donor units in the pre-period.

Protection from extrapolation is a desirable property, but the convex hull restriction is not a requirement for identification; simply because a counterfactual estimate is not extrapolated does not mean it is well identified. Convex weight restrictions allow for any amount of interpolation, no matter how extreme, and for no extrapolation, no matter how minor. Extreme interpolation can be just as undesireable as extreme extrapolation (King and Zeng, 2006; Kellogg et al., 2020).

Consider the two examples depicted in Figure 2. In Panel A, there are two potential synthetic control estimates: one (the blue square) lies close to other points, but is just outside of the convex hull. The other (the orange circle) lies within the convex hull, but in a region where there is no other data. Despite the blue square being more representative of the data, it would not be valid a synthetic control in the classic ADH method because it is extrapolated; the orange circle, meanwhile, would be considered valid despite the extreme interpolation needed to construct it. Similarly, Panel B displays time series from two groups of donor pools, one with a mean of five and another with a mean of negative five. Despite there being no data mean a mean of zero, the target time series in orange would be considered a valid synthetic control because it lies between these two groups. Moreover, if either donor group was to be removed, the target series would no longer be within the convex hull.

In some circumstances the convex hull restriction can even prevent the traditional synthetic control procedure from selecting a perfect donor series (Powell, 2019). With non-negative weights that sum to one, there is no way for the synthetic control outcome to be larger than the largest donor outcome or smaller than the smaller donor outcome. In addition there is no way for a synthetic control weight that is positive to gain information from two series that are counter-cyclical.

For example, imagine a classical difference-in-differences study in which the states have common time trends but different intercepts. Further suppose that there is only one treated state, which is also the state with the highest intercept. An extreme version of this case is depicted in Figure 3, Panels A and C. In this setting, it would be impossible to find a convex combination of donor states that provides a close match to the treated unit. Allowing for unrestricted weights – as our lasso regression does – can solve the problem by "extrapolating" outside the convex hull established by the pre-period outcomes in the donor states. This example is not contrived. Many studies (including this one) examine outcomes that often scale with the state population. If the treated state has a very large population (e.g., California) or a very small population (e.g., Wyoming) then there is a good chance that the outcome will lie beyond or near the boundary of the convex hull. In some applications, researchers sidestep this problem by rescaling the outcome variable to reflect outcomes per capita. This amounts to using non-convex weights on the original outcome variable, which undermines the original goal of avoiding extrapolation outside the convex hull.

Another scenario that creates problems for the restricted weight method arises when the treated unit is negatively correlated with some or all donor units. For simplicity, suppose that the treated unit follows a linear time trend with a slope of α. And suppose further that one donor unit follows a slope of $-\alpha$, while the rest following idiosyncratic random paths. A visual representation of this case is depicted in Figure 3, Panels B and D. With non-negative weights that sum to one, it will be impossible to match on the perfect donor unit with the appropriate weight. This is an extreme example, but the idea that two time series might be negatively correlated is not unrealistic. The pattern of seasonality could be different in some geographical areas than it is in others. Two financial assets could be negatively correlated. And indeed, the sales of substitutes might be negatively correlated with one another. Is extracting information from negatively correlated data obviously worse than extracting information from positively correlated data? It is not clear why this would be the case.

The lasso method we use in this paper does not restrict the weights to be non-negative or sum to one. Our model also allows for an intercept and coefficients on each of the donor units that serve as independent variables. Each of these parameters could be positive or negative, and the coefficients are not required to sum to one.[13] In the regression framework we pursue in this paper, large intercepts and coefficients can allow the synthetic control to take on a value that lies outside the range of outcomes observed in the donor pool. The estimated coefficient (weight) on a negatively correlated donor unit will simply be a negative number. This seems entirely natural when viewed through the lens of a regression model, even though it may seem odd in the context of a weighted average.

The concept of the convex hull is most theoretically appealing when the untreated donor series are the same focal variable and are in the same scale and units as the treated target series. This is the case in most applications of the traditional ADH method. However, synthetic control methods do not require that the donor pool be composed of exactly the same variable or that the variable be in the same units as the target time series. For instance, in the original ADH method cigarette sales in states other than California are used to predict counterfactual cigarette sales in California. However, one could imagine including the sales of cigars or (if studying a similar question in a modern setting) e-cigarette vaping devices in other states as a predictor of California cigarette sales. It is possible that donor variables that differ from the focal time series may better capture the underlying factors contributing to the time series variation and produce a better out-of-sample fit. As the set of potential donor pool variables grows larger and more distinct from the focal product, the concern that the predicted synthetic value be within the common numeric support of the donor pool holds less appeal.

4.7 Treatment effect estimation and inference

Once the counterfactual synthetic control time series is estimated, computing period-specific treatment effects is straightforward. To compactly summarize the results, we consider multi-period average treatment effects rather than the treatment effect for each post-treatment time period. For example, the average treatment effect for the treated group for the entire post-treatment period is $ATT(T_0+1,T)_{p0} = \frac{1}{(T-T_0)}\sum_{t=T_0+1}^{T}(Y_{st} - Y_{st}^{*})$. In principle, one could compute an average treatment effect for any post-treatment period of interest. For example, in our empirical application, we consider both the average treatment effect for the entire post-treatment period and the average treatment effect in each year following treatment.

To perform statistical inference on our ATT estimates, we construct a rank-based, two-sided p-value using randomization inference (Cavallo et al., 2013; Dube and Zipperer, 2015). We compare the absolute value of the standardized ATT estimate to the absolute value of the standardized ATT estimate from a number of placebo series. The estimates from the placebo distribution serve as the null distribution that assumes no treatment effect. In our setting, we use the same units compose both our donor pool and candidate placebo time series. In practice, however, the donor and placebo pools need not overlap. We limit the target time series considered, and placebo time series used for inference, to those that have synthetic control estimates that fit the data reasonably well during the pre-treatment period. We standardize using the pre-treatment period standard deviation for each respective time series, so that the respective ATT estimates are unit-free and comparable. We construct a two-sided p-value by comparing the rank of the absolute value of the standardized treatment effect for the target series against the absolute value of the estimated standardized pseudo-treatment effect for each untreated unit. The p-value is simply the percentile of the rank. For smaller

[13] The same is true of the elastic net regression approach described in Doudchenko and Imbens (2017).

placebo pools, it may make sense to report a bounded p-value. For example, if there are one treated unit and 49 placebos, then a rank of 2 out of 50 represents a p-value of between .02 and .04

To make the analysis manageable and coherent, we have attempted to impose a study design that is consistent across outcomes, and that provides a credible platform for statistical inference. That is, we have applied the same model selection procedure to each placebo. For example, we do not include any series as potential donors if they are from the same state as the focal product; we do this for identification purposes. Thus when constructing synthetic predictions for each series in the placebo pool, we ensure that an adaptable version of this restriction was put in place when computing each synthetic control. Ensuring that the process is similar minimizes the chance of accidental bias creeping into our analysis. Again in our application, the inclusion of in-state series may result in a better fit between the synthetic control group and the time series data. If the differential inclusion of in-state donor units resulted in better fit—and thus smaller pseudo-ATT estimates—for the placebo analyses, then this would bias our p-values toward zero. This is because our p-values are constructed by comparing relative magnitudes of target ATT estimates to placebo pseudo-ATT estimates.

Along a similar vein, we use—and recommend that others consider using—a pre-specified, unit-free threshold for model fit that equally applies to target variables of interest and each candidate placebo series. Our choice for that threshold is a pre-treatment Cohen's D of 0.25, but this need not be the only metric or threshold used. What is important is that this metric not be directly tied to the Cohen's D of the target series. Other work in the literature enforces similar model fit restrictions on the placebo pool. However, the threshold is often tied to the mean-squared prediction error of the target variable. Under the reasonable assumption that synthetic predictions with relatively worse pre-treatment fit are also likely to have relatively worse post-treatment fit, the difference between the actual and synthetic series for the target variable will be be biased toward being larger than the differences in the surviving placebo pool. This biases p-values toward zero. Moreover, when root mean squared error is used as a measure of fit, this additionally penalizes candidate placebo series that have larger nominal variance.

Restrictions on placebo series that are selected for inference can have a large effect on inference itself. Tighter Cohen's D restrictions will result in fewer placebo series contributing to the estimate of the null distribution, but the surviving series will—by construction—have a smaller difference between actual and synthetic predictions. If those series with better pre-treatment fit also have better fit post-treatment, then the null distribution from a tighter Cohen's D should be more compact than the null distribution from a more relaxed Cohen's D. A more compact null distribution means that the rejection region is larger, allowing smaller treatment effect estimates to be considered statistically different than zero. This does not, however, mean that the ideal Cohen's D is zero. As the Cohen's D restriction becomes more binding, fewer placebo series survive, and fewer target series survive as well. Thus there is a trade-off between the quality of model-fit and the size of the surviving placebo pool and set of target time series fit for study. In Section 5.6, we discuss this trade-off for our own application and present the null distribution under various Cohen's D thresholds in Figure 7.

5 Application: The effect of recreational marijuana legalization on alcohol and painkiller sales

Marijuana possession and consumption is illegal under federal law. Nevertheless, a number of states have recently adopted medical and recreational marijuana laws that expand legal access to marijuana. Medical marijuana laws allow people with qualifying health conditions to consume marijuana (ProCon, 2018a). Recreational marijuana laws allow people to use marijuana without qualifying conditions (ProCon, 2018b). Over thirty states have adopted medical marijuana laws and ten have approved marijuana for recreational use. To date, no state has legalized recreational marijuana without first approving medical marijuana.

In a 2012 statewide election, Colorado voters approved a ballot initiative to amend the state constitution legalizing recreational marijuana use for adults. The initiative passed with 55% of the vote, and it made Colorado the first recreational marijuana state. Over the next year, the state developed regulations governing the consumption, production, and distribution of marijuana. Starting in December of 2012, it became legal to possess home-grown marijuana in Colorado. In January of 2014, licensed facilities began selling recreational marijuana.

Our empirical application focuses on the effects of recreational marijuana adoption in Colorado. We limit the study to Coloardo in part because focusing on a single treatment and a single treated unit keeps the key econometric and methodological problems in clear view. Colorado also has the longest post-treatment time series of any recreational marijuana state. The long post-treatment time series allows us to study substitution patterns more credibly. In addition, Colorado's medical marijuana status does not change over our study period (January 2006 to December 2015), alleviating concerns related to multiple treatment effects. Four other states voted to adopt recreational marijuana policies during this time period: Oregon (2014), Alaska (2014), Washington (2012), and DC (2014). We exclude these states from our entire analysis.

5.1 Marijuana legalization and marijuana use

The empirical goal of our analysis is to measure the causal effects of Colorado's recreational marijuana law on the sale of alcohol and over-the-counter pain medications. Recreational marijuana laws might affect sales of other psychoactive substances if they are complements or substitutes for marijuana. This suggests that a first order question is whether recreational marijuana laws have any effect on marijuana consumption. If marijuana use does not change following legalization, it would be unreasonable to assume that our analysis could uncover resulting changes in sales of other psychoactive substances. To this end, Hollingsworth et al. (2020) show that recreational marijuana adoption increased the prevalence of past year use by 15 percent for younger adults and 25 percent for adults over age 25. In particular, they find that recreational adoption increases marijuana use as soon as possession and home cultivation are legal, and that access to dispensaries further increases marijuana use. These findings provide justification for the claim that if marijuana use has an effect on the consumption of other psychoactive substances, then recreational marijuana adoption would induce a large enough change in marijuana use to plausibly uncover such relationships.

5.2 Marijuana legalization and the use of other substances

The connection between marijuana legalization and the use of other psychoactive substance has important implications for policies that are designed to mitigate externalities and social harms associated with drug use. If marijuana consumption produces lower external costs and less harm than some other drug and the two drugs are substitutes, then legalizing marijuana may produce net social benefits. Similarly, if marijuana consumption is complementary to other psychoactive substances, it could be a net harm. Substitution patterns also have fiscal consequences. For example, if marijuana use crowds out or increases alcohol use, but has a differential tax rate, state tax revenue could change substantially.

5.2.1 Alcohol

In the lead up to the ballot initiative proposing recreational marijuana legalization, supporters of the law suggested that legalizing marijuana would be a welfare-improving, harm reduction policy. The premise of the argument was that people would substitute marijuana for alcohol consumption and that alcohol use has greater external costs than marijuana use (Johnson, 2012). After the measure passed, a formal marijuana market developed in Colorado's economy. Alcohol sales increased over the same period, and some observers suggested that marijuana tourism increased alcohol sales in Colorado (Moore, 2014). These anecdotes suggest that legal recreational marijuana may serve as either a substitute for or complement to gross alcohol sales. Of course, marijuana may be a substitute for alcohol in some situations and not others, for some alcohol products and not others, and for some consumers and not others.

Previous research on the connection between marijuana and alcohol has mostly relied on survey data to measure outcomes. One line of work examines the way measures of marijuana use respond to changes in alcohol prices, with some finding evidence of substitution (Chaloupka and Laixuthai, 1997; Cameron and Williams, 2001) and others finding evidence of complementarities (Cameron and Williams, 2001; Pacula, 1998). Another line of work studies how marijuana use responds to changes in the availability of alcohol, from minimum age restrictions to outright prohibition. The majority of this research finds that the two are substitutes (Brecher, 1972; Crost and Guerrero, 2012; DiNardo and Lemieux, 2001; Williams et al., 2004). But some research finds no relationship (Crost and Rees, 2013) and even evidence of complementarity (Yörük and Yörük, 2011).

Other work has studied the effects of medical marijuana laws on other substances using a difference-in-differences framework. Wen et al. (2015) find that among those over age 21, medical marijuana laws increase the average number of binge drinking days in the past month, increase the fraction of people who engaged in both marijuana use and binge drinking in the past month, and increase the fraction of people who used marijuana and alcohol on the same occasion in the past month. They do not find any effect of medical marijuana on underage drinking or on the consumption of other psychoactive substances. Anderson et al. (2013) find that alcohol-related car accidents and non-hard liquor sales fell after the implementation of medical marijuana, suggesting that people substituted marijuana for alcohol. Dills et al. (2017) studied decriminalization, medical marijuana, and recreational expansion of marijuana from 1977 to 2015; they find no evidence that these policy changes affected measures of alcohol or tobacco use. Pacula et al. (2013, 2015) attempt to rectify many of the inconsistencies in this literature by exploring policy heterogeneity. They find that using only a simple binary indicator for any marijuana law masks important underlying heterogeneity. When they account for policy heterogeneity, they find that both allowing for home cultivation and allowing for legal dispensaries are positively associated with binge drinking and alcohol-related traffic fatalities.

5.2.2 Painkillers

A more recent literature examines the relationship between medical marijuana and prescription opioid use. To our knowledge, no prior study has evaluated the effects of marijuana liberalization on sales of over-the-counter painkillers. Bradford and Bradford (2016, 2017) find that medical laws reduce prescription among Medicare and Medicaid patients. Bradford and Bradford (2018) conclude that the decline in prescriptions in the Medicare population is due to a decline in opioids prescriptions. Wen and Hockenberry (2018) find decreases in opioid prescribing in the Medicaid population following the passage of both medical and recreational marijuana legislation. Shi (2017) show that these laws are associated with a decrease in opioid-related hospitalizations, and Bachhuber et al. (2014) find that medical laws have reduced the opioid mortality rate and Chan et al. (2019) find similar effects for recreational laws. Powell et al. (2018) show that access to marijuana dispensaries reduces opioid prescriptions and associated overdose deaths. We hypothesize that prescription and over-the-counter analgesics will exhibit similar substitution patterns with marijuana.

5.3 Limitations of survey data

One limitation of most of the existing literature linking marijuana with use of other psychoactive substances is the reliance on survey questionnaires to measure consumption. Imperfect recall and concerns about the social desirability of specific answers to sensitive survey questions may be important sources of bias in survey research on drug and alcohol consumption. In addition, typical survey questions focus on the quantity and frequency of consumption and do not distinguish between different types of alcohol products with differential alcohol by volume.

Retail scanner data make it possible to study the exact quantity of alcohol sold in stores, and it eliminates concerns about whether survey respondents have accurate recall and provide truthful responses. In addition, scanner data make it possible to study substitution patterns in a more detailed way than earlier work based on surveys: we examine the sales of multiple types of alcohol (beer, wine, hard liquor, and malt liquor) as well as over-the-counter painkillers. Distinguishing between different alcohol types may provide insight into the underlying preferences that determine substitution patterns. For example, the market for beer likely satisfies more than one underlying consumer preference. Low-cost, small-volume, and high-alcohol-content beers (like single-serving malt liquor) are meant for immediate consumption and may be associated with negative externalities generated by binge drinking and drinking and driving.[14] In contrast, wine and beer may help satisfy the demand for social drinking or may have other desirable product attributes beyond low-cost

[14] The top panel of Figure A4 in the Appendix provides a visual depiction of this theory. The bottom panel shows that malt liquor is the most likely of the alcohol categories we examine to be a substitute for the intoxicating effects of recreational marijuana, as it has the lowest cost to purchase and provides the most alcohol per dollar spent.

intoxication. Marijuana could be a substitute for one alcohol product and a complement to another. Survey measures that lump heterogeneous goods together risk finding a combined relationship that is misleading.

Retail scanner data also plays a role in two other recent papers studying marijuana and tobacco and alcohol consumption patterns. Baggio et al. (2019) study the impact of medical marijuana legalization on aggregate beer and wine sales using retail scanner data to construct measures of aggregate expenditures on alcohol at the county-month level. They look at total county expenditures in three broad categories of alcohol products: beer, wine, and beer and wine combined. They use a differences-in-differences design and find that the adoption of state-level medical marijuana laws reduce aggregate beer and wine sales by 13 percent.

Miller and Seo (2018) also use retail scanner data and administrative data from Washington State to estimate a structural model of the demand for psychoactive substances. The model is derived using a multistage budgeting approach, which assumes that each consumer first decides how much to spend on psychoactive substances, then decides how to allocate consumption across broad classes of substances (e.g., alcohol, tobacco, or marijuana), and then finally decides how to allocate expenditure within each sub-class. Their model allows for three sub-classes of alcohol (wine, beer, and liquor) and only includes data from Washington state in years following its adoption of a recreational marijuana law. The estimates from their model imply that a 1% decrease in the price of marijuana leads to a .16% decrease in alcohol consumption.

5.4 Data

The primary dataset used in our analysis is the Neilsen Retail Scanner Database, which contains weekly sales information for individual products from a set of food, drug, mass-merchandise, convenience, and liquor stores. The data are derived from scanners used at the point of sale. From 2006 to 2015, there were 41,290 unique stores in the Nielsen database. These stores are not a random sample of all retails stores in the country. However, Nielsen estimates that the sales recorded in the database represent more than 50% of total sales of all U.S. grocery and drug stores; there is little reason to believe that the time series of sales outcomes in the Nielsen data systematically differs from the overall population of stores. To mitigate concerns about changes in the composition of the Nielsen database, we limit our analysis to data from a balanced panel of 31,678 stores that are included every year.

In the raw data, product sales information is available at the store-week-UPC-code level, and there are over 2.5 million unique UPC codes observed across all stores in the database. We extract information on UPC codes from a broad group of alcohol, painkiller, and other products. Neilsen groups UPC codes into intermediate product categories. We use these designations to select all beer, wine, hard liquor, malt liquor, and painkiller sales in the database.[15] After grouping individual UPC codes into these broader product categories, we compute the total ounces (or pills) sold in the panel of Nielsen stores in each state and week. To help make the results interpretable, we focus on total ounces sold in each alcohol product category.

For our donor and placebo units, we create a separate alcohol category for each type based on size: single-serving, small, medium, and large. In addition to the alcohol and painkiller products that are the focus of our analysis, we also extract data on a number of other donor/placebo goods: eggs, soda, diet soda, tea, coffee, pasta, lunch meat, shampoo, feminine hygiene products, razor blades, toilet paper, kitty litter, light

[15] Hard liquor is composed of bourbon, whiskey, scotch, gin, vodka, rum, tequila, brandy, and cognac.

bulbs, liquid soap, cigarettes, bar soap, bread, and butter.[16] For each product type, we group individual UPC codes into categories and then compute the total ounces or counts of each product class sold in the panel of Nielsen stores in each state and week. Throughout the analysis, we work with the natural log of the quantity sold in each week for each $product \times state$ unit.

The raw time series data for our products of interest and for a sample of our untreated state-product pairs are displayed in Figure 4. The first vertical dashed lines in the figure denotes December 2012, when the vote to legalize recreational marijuana in Colorado was completed. The second vertical dash line is January 2014, which is when Colorado's first recreational dispensaries opened. The graph gives some idea about the relative range of the donor/placebo pool as well as general trends in alcohol and painkiller sales in Colorado. The sales of all target products trended upward throughout the pre-treatment period. There is substantial within-product variation across time that is attenuated due to the common y-axis.

5.5 Model

The generic notation we developed in the methodological section of the paper considered a simple case where a collection of geographical units is followed over multiple time periods. The empirical setting we consider here is more complicated because we consider the weekly time series of sales of a large number of product categories in multiple states.

To make this point clearly, we use $p = 1...P$ index the set of product categories and $s = 1...S$ to index states. In our setting we have $P = 52$ total products and $S = 45$ donor states for a total of 2,340 total donor series. Each $product \times state$ cell is observed in weekly periods indexed by $t = 1...T_0...T$. In our setting there are $T = 521$ total weeks and treatment begins after $T_0 = 357$. D_{pst} is a dummy variable set to 1 if the state is Colorado, the product is one of the treated products (beer, wine, hard liquor, malt liquor, and painkillers), and the time period occurs after the onset of the recreational marijuana law in Colorado in December 2012. The donor pool consists of the weekly time series of all treated products of various sizes (beer, wine, hard liquor, malt liquor, and painkillers) and placebo products (eggs, soda, diet soda, tea, coffee, pasta, lunch meat, shampoo, feminine hygiene products, razor blades, toilet paper, kitty litter, light bulbs, liquid soap, cigarettes, bar soap, bread, and butter) in the untreated states.

Imposing Assumption 1 (no interference between units), we let $Y_{pst}(0)$ be a potential outcome that represents the log quantity sold for product p in state s in week t in the absence of recreational marijuana law. $Y_{pst}(1)$ represents the log quantity sold in the product × state market in the presence of a recreational marijuana law. Using. $s = 0$ to represent Colorado, $\beta_{p0t} = Y_{p0t}(1) - Y_{p0t}(0)$ is the causal effect of the Colorado recreational marijuana policy on the sales of product p in period t.

We assume that the untreated potential outcomes for all of the products and states in our study sample are generated by the following interactive factor structure model:

$$Y_{pst}(0) = a_t \alpha_{ps} + \varepsilon_{pst} \tag{15}$$

This model is very similar to the one described in Assumption 2. As before a_t represents a $1 \times F$ vector of

[16] Eggs, tea, feminine hygiene products, razor blades, toilet paper, light bulbs, cigarettes, and bar soap are measured as counts of individual units. The other products are measured in ounces.

time varying common factors. But this time, the factor loadings are allowed to vary by both product and state α_{ps}. In essence, the model allows for a very flexible set of $product \times state$ specific time trends under the assumption that all of the trends are ultimately rooted in a potentially large set of common time varying factors that simply matter more or less across different products and states. We assume that the non-common factor represented by ε_{pst} satisfies the regularity and exogeneity conditions described in Assumption 3, and that the common factors vary independently during the pre-period as required by Assumption 4.

We estimate a synthetic control group for each treated product in Colorado using the SCUL procedures described earlier in the paper. We use cross-validation to determine the Lasso penalty. And assess pre-period fit using a Cohen's D statistic. When the synthetic control for a product fails the Cohen's D threshold, we interpret this as evidence that Assumption 5 (existence of weights) fails for that product.

5.6 Results

This section presents treatment effect estimates derived using the SCUL procedure. Our target units are weekly sales of beer, wine, hard liquor, malt liquor, and over-the-counter painkillers in Colorado. The first order goal was to use the SCUL method to estimate counterfactual sales for each treated time series using out-of-state sales data. We selected optimal weights using a rolling-origin cross validation procedure, allowing donor weights to differ for each target product. We estimate the synthetic counterfactual by multiplying the cross-validated weights by the post-treatment values from the donor pool. In our main analysis, the post-treatment period begins in December 2012. However, we also consider the alternative treatment date of January 2014, when dispensaries first opened.

5.6.1 Treatment effect estimates

In Panel A of Figure 5, both the observed time series and SCUL counterfactual are displayed for each target series. The SCUL method appears to perform quite well in the pre-treatment period, providing a close match to the variation in each target series. However, given the volatile nature of each time series, fit is difficult to visually ascertain. To make pre-treatment fit easier to observe, we plot the difference between the observed data and the SCUL counterfactual prediction in Panel B. In addition, we report a measure of pre-treatment fit, the Cohen's D, in Table 1. The Cohen's D statistic in the table is the average weekly difference between the observed values for each unit and the synthetic prediction, expressed in standard deviation units. Each target unit has a measure of fit below our pre-specified threshold of 0.25, with the Cohen's D for the malt liquor series being the closest to this threshold at 0.22.

For each time series, a clear deviation between the observed outcomes and the synthetic counterfactual begins at the start of 2014. The post-treatment gap between the realized sales and the counterfactual is positive for painkillers and negative for each alcohol series. The average deviation—reported in percent—across the entire post-treatment period and in each year is reported in Table 1 Panel A. Following recreational legalization in 2012, we find a 3% increase in the sales of over-the-counter painkillers and, depending on the product, between a 7 and 40% reduction in alcohol sales. The largest changes in sales begin in 2014, when recreational dispensaries first opened.

5.6.2 Statistical inference

To understand if these treatment effect estimates are statistically significant, we compare them to the pseudo-treatment effects we estimated for many untreated placebo product-state pairs. These placebo units are weekly product sales of alcohol, painkillers, and other goods from untreated states. The distribution of pseudo-treatment effect estimates represents the null distribution of no treatment effect. Importantly, the placebo analysis captures how the fit of the SCUL counterfactual deteriorates over time when there is no treatment effect. To be considered sufficiently rare to be statistically significant, any actual treatment effect must be large enough in magnitude to overcome this deteriorating fit.

For the sake of clarity, we outline results from the SCUL procedure using a single treated unit, sales of hard liquor in Colorado. Figure 6 displays the difference between actual ounces of hard liquor sold each week in Colorado and the SCUL prediction in green. The pre-treatment difference between the two series is small and centered around zero. In the figure, the gray lines show the pseuedo differences between each placebo and its synthetic control. The graph only includes placebo lines that survived the Cohen's D screen by having a pre-treatment Cohen's D less than 0.25. As discussed, this same criterion is applied to both the placebos and the target units. The placebo lines in the graph are drawn with some transparency so that the darker areas have a greater density of placebo units than lighter spaces. This shading highlights the deterioration of the counterfactual fit across time and gives the appearance of smoke. As such, we refer to this style of plot as a "smoke plot."

Under the smoke plot, we report the relative contribution [0-1] and the lasso coefficient for each donor unit to the synthetic prediction. The relative contribution is a function of both the lasso coefficients and the donor pool values in a given time period. Since this can change across time, we report the relative contribution for both the first and last time period. In this application, relative contributions appear to be stable across time. The single most important donor unit for hard liquor is single-serving beer sales from Tennessee, followed closely by the intercept, which is a measure of average pre-treatment hard liquor sales in Colorado. The majority of donor units that receive non-zero weight are alcohol or liquor products. Given that the donor pool contains mostly non-alcoholic products, this was by no means guaranteed and indicates that the synthetic control procedure may be selecting on underlying factors of the data generating process rather than idiosyncratic statistical noise.

The smoke plot sheds light on the intuition underlying both our decision to examine only those goods with an adequate pre-treatment fit and our randomization-inference-based approach for statistical inference. Consider the pre-treatment period, from 2006 until possession became legal in November of 2012. Here we can see that the difference between the target and synthetic units (in green) fits about as well as the average placebo product. While there are occasional large deviations, the average pre-treatment difference is centered around zero, with a small standard deviation.

As the training period of our data ends before legalization, the SCUL estimates are not updated to include information after November 2012. Thus, as time since November 2012 increases, model fit for each time series worsens. Since the placebo goods should not be impacted by treatment, they help us determine how we can expect model to worsen over time in the absence of treatment. In the smoke plot, this can be seen as the "dissipating smoke" following initial treatment.

Since statistical inference essentially compares the magnitude of the treatment effect estimate to the cloud of placebo estimates, placebo units with poor model fit will increase the spread of the null distribution. To mitigate the spread of the null distribution, we remove any placebo units with poor pre-treatment fit. However, post-treatment fit worsens with time even for those placebo units with satisfactory pre-treatment fit. This deterioration implies that statistical power will worsen as time from initial treatment increases: as the placebo distribution grows wider, the minimum effect size needed to be considered significant at a given level also grows.

Using placebo data from our application, Figure 7 demonstrates this concept more clearly. Each row in the figure displays the distribution of pseudo-treatment effects defined over different blocks of post-treatment time: the first three rows show placebo distributions from the average effect over the first year, second year, and third year after legalization (2013, 2014, and 2015). The fourth row displays null distributions for average treatment effects taken over the entire post-treatment period. Each column shows the placebo distribution derived from a different pre-treatment Cohen's D exclusion criteria for placebo units: the first column has no exclusion threshold, the second has an exclusion threshold of 0.25, and the third has an exclusion criteria of 0.10. As time since treatment increases, the null distribution becomes noticeably wider. This makes it harder to reject the null hypothesis for effects of forecasts further into the future. Similarly, more restrictive Cohen's D thresholds yield more compact null distributions. In general, a more compact null distribution is desirable because wider null distributions are less able to differentiate small treatment effects from statistical noise. Thus, synthetic control methods have the greatest power to detect small effect sizes in the time periods closest to treatment, and when the synthetic control method also provides a satisfactory fit for the placebo pool used to compose the null distribution.

Maximizing statistical precision in these ways is not without trade-offs. Dynamic treatment effects that grow over time may not be large enough to be detectable in the periods immediately following treatment. Using a smaller Cohen's D threshold can improve precision by eliminating noisy placebo units, but this also may eliminate target units that do not meet the pre-treatment fit quality standard. Consider both time since treatment and the Cohen's D threshold for our example. The largest treatment effects do not begin until 2014 when the power to detect effects is the weakest. And the most compact null distribution is generated by choosing a Cohen's D threshold of 0.10, which would eliminate every target product of interest from consideration.

To help make sense of these issues, we recommend determining the minimum treatment effect size for each time block and Cohen's D threshold that would be statistically different than zero for a given significance level. This may help researchers decide if a particular study has enough statistical power to be useful. In our application, with a Cohen's D threshold of 0.25 , the standardized average treatment effect during the first post-treatment year would need to be at least 0.45 in absolute value in order to reject the null at the 10% level. In contrast, the third year effect size would need to be at least 1.05 in order to reject the null at the at the 10% level. The minimum treatment effect size more than doubles from year one to year three. If a treatment effect is not realized immediately or is dynamic, then the deteriorating model fit may present an insurmountable hurdle for statistical inference. It is possible that the fit of the synthetic prediction will deteriorate at a faster rate than the growth of the treatment effect, resulting in a minimum treatment effect size far larger than any reasonably expected treatment effect could be.

In Table 1 Panel A, we present both estimated average treatment effects (in percent) for different time periods and rank-based, randomization-inference p-values in parentheses. Only the treatment effect estimates for quantity of malt liquor sold are statistically different from zero at the 10% level, although all treatment effect estimates for alcohol are negative, and their respective p-values are mostly below 0.4. Sales of over-the-counter painkillers appear largely unaffected by recreational marijuana adoption, with positive treatment effect estimates that are not statistically distinguishable from zero. Since we are examining multiple products, we also consider two joint tests of whether recreational legalization has any effect across different product groupings.

The first joint test measures if there is any significant effect across all of the products. We perform this test by summing the absolute value of five randomly chosen standardized treatment effect estimates from the placebo pool and comparing this sum to the analogous measure for our target variables in each post-treatment time period. The p-value is the fraction of cases in which that the sum from our five target products is larger than the sum drawn from the placebo pool. The tests cannot reject the possibility that there is no effect of recreational marijuana legalization on any product. This method, however, does not reward similar products for having the same sign, and it is reasonable to expect that sales of alcohol will all either be substitutes or complements. We construct a second joint test for alcohol products that is based upon the absolute value of the sum of the treatment effect estimates rather than the sum of the absolute values. Moving the absolute value penalizes coefficients of different signs, since opposite signed effects of the same magnitude will cancel out. When we use this second joint test, we find that there is a statistically significant joint effect of marijuana legalization on alcohol sales, which is driven by the years 2014 and 2015.

5.6.3 Robustness checks, model comparisons, and alternative treatment dates

Deviations from uniform weights in randomization inference procedure: The randomization inference procedure we use relies on the assumption that there is no treatment effect among the placebos and that Assumptions 1-5 hold for all units and periods. Our approach could be invalid if some of the placebos we use have non-common factors that are driven up or down in ways that make it harder or easier to reject the null hypothesis. A sensitivity analysis proposed by Firpo and Possebom (2018) explores these concerns using methods that are similar to Rosenbaum (2002)'s approach to sensitivity to hidden bias in observational studies.

The randomization inference approach we use implicitly gives equal weight to every placebo when computing the rank based p-value. Firpo and Possebom (2018) point out that a key assumption of such an approach is that such omitted factors are not affecting our procedure. Firpo and Possebom (2018) also demonstrate a clever alternative procedure that serves to bound how this underlying uniform approach, when in the presence of such bias, could affect the p-value. At an agnostic level, this omitted feature could be making it either more or less likely to find a statistically significant effect. Since it is impossible to know precisely the bias of each placebo unit, Firpo and Possebom (2018) suggest that researchers consider the worst possible case (where bias is working in favor of finding a statistically significant effect) and the best possible case (where bias is working against finding a statistically significant effect). In Appendix Section C.1, we conduct construct and conduct a version of their recommended sensitivity check, finding no evidence that our p-values are sensitive to deviations from the uniform weighting assumption.

Alternative penalty parameter decision rules: In Appendix Table A1, we explore the sensitivity of the results in the last column of Table 1 to different selection rules for the cross-validated penalty parameter. Specifically considering the three options outlined in Section 4.3. For most outcomes, the results are extremely stable finding similar treatment effect estimates, measures of pre-treatment fit, p-values. The one exception is for malt-liquor. Here we find large variation in the treatment effect estimates, indicating that these results are sensitive to the selection rule for the penalty parameter. As such we are less confident of our estimates for this outcome than for the other target products.

Comparison to traditional synthetic control method: In Appendix Figure A3 and Table A2, we conduct a traditional synthetic controls analysis in the spirit of Abadie et al. (2010). Since this traditional method cannot use the high dimensional donor pool used in the SCUL analysis, we include only the sales of the identical product in other states (e.g., the donor pool for pain pills in Colorado is the sales of pain pills in every other state that had no recreational marijuana policy during our sample). For the L summary statistics to be balanced on, we selected average sales of the target product in the pre-treatment period, population (in millions), median income ($10,000), murder rater per 100,000, the percent of residents over the age of 18 with at least a high school degree, and average life expectancy.

We compare results from the traditional synthetic control method to those from the SCUL method, examining differences in prediction fit across time, measures of pre-treatmeent model fit, treatment effect estimates, and p-values. For this particular empirical application, the traditional approach performs poorly. For every outcome the Cohen's D measure of pre-treatment fit is above our pre-specified threshold of model fit of 0.25. This is true for the majority of placebo iterations as well. The exact measures of fit are reported in Table A2. Moreover, Figure A3 demonstrates that the traditional approach does not produce a pre-treatent fit that is centered around zero for beer, malt liquor, or pain pills.

We do not attempt to precisely decompose the reason behind this discrepancy, but it stands to reason that the combination of the larger donor pool available to SCUL and the ability to have negatively correlated, non-convex weights may have contributed to the models better pre-period match. Thus—only for this particular empirical application—we conclude that the synthetic control using lasso appears to preform better than the traditional approach. Of course, this does not imply that one approach is superior to the other. On the website for our R-package we similarly compare the traditional synthetic control approach to SCUL for the classic case examining the effect of California's proposition 99 on cigarette sales (Abadie et al., 2010). In this second comparison of the two methods, a nearly identical treatment effect is found, clearly demonstrating that one method is not strictly superior to the other.[17]

Using dispensary openings as an alternative beginning of treatment: In Colorado, recreational dispensaries did not open until 2014. Prior research has found that recreational dispensary access increases marijuana

[17] Comparison available online at `https://hollina.github.io/scul/articles/scul-tutorial.html`.

use (Hollingsworth et al., 2020), and that medical dispensary access affects downstream substitution of prescription painkillers (Powell et al., 2018). Consistent with this logic, neither the visual data presented in Figure 5 nor the analytic results in Table 1 systematically show large deviations until 2014. It may also be the case that people who are likely to substitute alcohol consumption for marijuana use would not do so until a convenient and legal mechanism such as a dispensary is available. Moreover, if dispensary openings cause most of the average treatment effect, forcing the synthetic control's out-of-sample period to begin in 2013 will widen the placebo distribution relative to an estimator that assumes treatment begins in 2014.

Thus, we consider an alternative analysis where our post-treatment period begins in January of 2014. The results of this procedure are reported in Table 1 Panel B. With the exception of malt liquor, all treatment effect estimates are similar to those estimated in our previous analysis. The Cohen's D on malt liquor increased substantially from 0.22 to 0.32 and is above our pre-determined threshold for model fit. Therefore we do not consider this outcome as a viable candidate for our procedure. This sensitivity coupled with the sensitive of the malt liquor treatment effect to the decision rule for the penalty parameter raise concerns that the SCUL method may not be well suited to determine if recreational marijuana legalization has an effect on this particular outcome.

A key difference in this analysis where treatment begins in 2014 is that the null distribution is more compact, meaning that despite having similar treatment effect estimates, p-values in this analysis tend to be lower. Results indicate that alcohol sales as a whole decreased following legalization and that this change is statistically significant at the 5% level.

6 Conclusion

Synthetic control methods represent an increasingly popular strategy for estimating counterfactual treatment effects. Unfortunately, the core assumptions of the design are somewhat opaque and it is often hard to assess their credibility in social science settings. In addition, synthetic control studies require researchers to make a variety of implementation decisions, and the technical literature offers little practical guidance on how to make these choices.

In this paper, we try to articulate the practical meaning of the core synthetic control assumptions. We outline the problems, discretionary choices, and conceptual challenges associated with synthetic controls in a way that we hope will be useful for other applied researchers. Where it seems prudent, we offer advice about how researchers should handle key issues that are apt to apply to many different synthetic control studies. Taken together, we make six recommendations that may help researchers implement their synthetic control studies more effectively: 1) use the same model selection procedure for both target and placebo products; 2) use a unit-free measure to evaluate model fit; 3) trim from the study, any target series or placebo series that fares poorly on a pre-specified threshold of model fit; 4) incorporate a rolling-origin cross-validation procedure to determine optimal weights, which helps guard against over-fitting and likely reduces bias; 5) use a unit-free measure of the treatment effect estimate to compare estimated treatment effects to the placebo distribution; and 6) report the minimum detectable effect size given the placebo distribution and a specified significance level.

We also argue that using donor units from a wide range of variable types can contribute to improved identification of underlying factors driving the pre-treatment data generating process for the treated unit. We also use an extension of the synthetic controls estimator that exploits machine learning to automate model selection, relaxes convexity restrictions, and allows for a high-dimensional donor pool. This approach may be useful in many settings and we provide code and a online statistical package to help others use the method or parts of the method in their own work.

Finally, we apply our recommendations and technique to a policy-relevant question: what is the relationship between recreational marijuana legalization and consumption of alcohol and over-the-counter painkillers? Taken as a whole, our results indicate that recreational marijuana legalization decreases alcohol sales and does not affect the sales of over-the-counter painkillers. This suggests that marijuana and alcohol are likely to be substitutes and—surprisingly—that marijuana and over-the-counter painkillers are not.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, Vol. 105, No. 490, pp. 493–505.

——— (2015) "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, Vol. 59, No. 2, pp. 495–510.

Abadie, Alberto and Javier Gardeazabal (2003) "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, Vol. 93, No. 1, pp. 113–132.

Abadie, Alberto and Jérémy L'Hour (2019) "A Penalized Synthetic Control Estimator for Disaggregated Data," *Working Paper,*, pp. 1–35.

Alberto Abadie (2020) "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, Vol. Forthcoming.

Amjad, Muhammad, Devavrat Shah, and Dennis Shen (2018) "Robust synthetic control," *Journal of Machine Learning Research*, Vol. 19, pp. 1–51.

Anderson, D. Mark, Benjamin Hansen, and Daniel I. Rees (2013) "Medical Marijuana Laws, Traffic Fatalities, and Alcohol Consumption," *The Journal of Law and Economics*, Vol. 56, No. 2, pp. 333–369.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager (2018) "Synthetic Difference in Differences."

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu (2019) "Ensemble Methods for Causal Effects in Panel Data Settings," *AEA Papers and Proceedings*, Vol. 109, pp. 65–70.

Bachhuber, Marcus A., Brendan Saloner, Chinazo O. Cunningham, and Colleen L. Barry (2014) "Medical cannabis laws and opioid analgesic overdose mortality in the United States, 1999-2010," *JAMA Internal Medicine*, Vol. 174, No. 10, pp. 1668–1673.

Baggio, Michele, Alberto Chong, and Sungoh Kwon (2019) "Marijuana and alcohol evidence using border analysis and retail sales data," *Canadian Journal of Economics*, Vol. Accepted, pp. 1–39.

Bai, Jushan (2009) "Panel data models with interactive fixed effects," *Econometrica*, Vol. 77, No. 4, pp. 1229–1279.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2018) "The Augmented Synthetic Control Method," No. November.

Botosaru, Irene and Bruno Ferman (2019) "On the role of covariates in the synthetic control method," *The Econometrics Journal*, Vol. 22, No. 2, pp. 117–130.

Bradford, Ashley C. and W. David Bradford (2016) "Medical Marijuana Laws Reduce Prescription Medication Use In Medicare Part D," *Health Affairs*, Vol. 35, No. 7, pp. 1230–1236.

——— (2017) "Medical Marijuana Laws May Be Associated With A Decline In The Number Of Prescriptions For Medicaid Enrollees," *Health Affairs*, Vol. 36, No. 5, pp. 945–951.

Bradford, Ashley C and W David Bradford (2018) "The Impact of Medical Cannabis Legalization on Prescription Medication Use and Costs under Medicare Part D," *The Journal of Law and Economics*, Vol. 61, No. 3, pp. 461–487.

Brecher, Edward M (1972) *Licit and illicit drugs*, Boston: Little, Brown.

Cameron, Lisa and Jenny Williams (2001) "Cannabis, Alcohol and Cigarettes: Substitutes or Complements?," *Economic Record*, Vol. 77, No. 236, pp. 19–34.

Campbell, Donald T and Thomas D Cook (1979) "Quasi-experimentation," *Chicago, IL: Rand Mc-Nally*.

Casella, George and Roger Berger (2001) *Statistical Inference*: Duxbury Resource Center.

Cavallo, Eduardo, Sebastian Galiani, Ilan Noy, and Juan Pantano (2013) "Catastrophic Natural Disasters and Economic Growth," *Review of Economics and Statistics*, Vol. 95, No. 5, pp. 1549–1561.

Chaloupka, Frank J and Adit Laixuthai (1997) "Do Youths Substitute Alcohol and Marijuana? Some Econometric Evidence," *Eastern Economic Journal*, Vol. 23, No. 3, pp. 253–276.

Chamberlain, Gary (1984) "Panel data," *Handbook of econometrics*, Vol. 2, pp. 1247–1318.

Chan, Nathan W, Jesse Burkhardt, and Matthew Flyr (2019) "The Effects of Recreational Marijuana Legalization and Dispensing on Opioid Mortality," *Economic Inquiry*.

Cochran, W. G. (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, Vol. 24, No. 2, p. 295.

Crost, Benjamin and Santiago Guerrero (2012) "The effect of alcohol availability on marijuana use: Evidence from the minimum legal drinking age," *Journal of Health Economics*, Vol. 31, No. 1, pp. 112–121.

Crost, Benjamin and Daniel I. Rees (2013) "The minimum legal drinking age and marijuana use: New estimates from the NLSY97," *Journal of Health Economics*, Vol. 32, No. 2, pp. 474–476.

Dills, Angela K, Sietse Goffard, and Jeffrey Miron (2017) "The effects of marijuana liberalizations: Evidence from monitoring the future,"Technical report.

DiNardo, John and Thomas Lemieux (2001) "Alcohol, marijuana, and American youth: the unintended consequences of government regulation," *Journal of Health Economics*, Vol. 20, No. 6, pp. 991–1010.

Doudchenko, Nikolay and Guido W. Imbens (2017) "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," oct.

Dube, Arindrajit and Ben Zipperer (2015) "Pooling Multiple Case Studies Using Synthetic Controls: An Application to Minimum Wage Policies," *IZA Discussion Paper No. 8944*.

Evans, William N, Ethan MJ Lieber, and Patrick Power (2019) "How the reformulation of OxyContin ignited the heroin epidemic," *Review of Economics and Statistics*, Vol. 101, No. 1, pp. 1–15.

Firpo, Sergio and Vitor Possebom (2018) “Synthetic control method: Inference, sensitivity analysis and confidence sets,” *Journal of Causal Inference*, Vol. 6, No. 2.

Gobillon, Laurent and Thierry Magnac (2016) “Regional policy evaluation: Interactive fixed effects and synthetic controls,” *Review of Economics and Statistics*, Vol. 98, No. 3, pp. 535–551.

Hainmueller, Jens (2012) “Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies,” *Political Analysis*, Vol. 20, No. 1, pp. 25–46.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The elements of statistical learning: data mining, inference and prediction*: Springer, 2nd edition.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations* in , Monographs on Statistics and Applied Probability, No. 143, Boca Raton: CRC Press, Taylor & Francis Group.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart (2007) “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference,” *Political Analysis*, Vol. 15, No. 3, pp. 199–236.

Hollingsworth, Alex, Coady Wing, and Ashley Bradford (2020) “Comparative Effects of Recreational and Medical Marijuana Laws On Drug Use Among Adults and Adolescents,” preprint, SocArXiv.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S Rosen (1988) “Estimating vector autoregressions with panel data,” *Econometrica: Journal of the econometric society*, pp. 1371–1395.

Hyndman, Rob J and George Athanasopoulos (2020) *Forecasting: principles and practice*, Melbourne, Australia: OTexts, 2nd edition.

Ibragimov, R. and Sh. Sharakhmetov (2002) “The Exact Constant in the Rosenthal Inequality for Random Variables with Mean Zero,” *Theory of Probability & Its Applications*, Vol. 46, No. 1, pp. 127–132.

Johnson, Kirk (2012) “Marijuana Push in Colorado Likens It to Alcohol,” jan.

Kellogg, Maxwell, Magne Mogstad, Guillaume Pouliot, and Alexander Torgovitsky (2020) “Combining Matching and Synthetic Controls to Trade off Biases from Extrapolation and Interpolation,” *National Bureau of Economic Research*.

King, Gary and Langche Zeng (2006) “The Dangers of Extreme Counterfactuals,” *Political Analysis*, Vol. 14, No. 2, pp. 131–159.

Miller, Keaton and Boyoung Seo (2018) “Tax Revenues When Substances Substitute: Marijuana, Alcohol, and Tobacco.”

Moore, Thad (2014) “Legal pot trade not siphoning sales from Colorado brewers, distillers.”

Pacula, Rosalie L., David Powell, Paul Heaton, and Eric L. Sevigny (2015) “Assessing the Effects of Medical Marijuana Laws on Marijuana Use: The Devil is in the Details,” *Journal of Policy Analysis and Management*, Vol. 34, No. 1, pp. 7–31.

Pacula, Rosalie Liccardo (1998) “Does increasing the beer tax reduce marijuana consumption?,” *Journal of Health Economics*, Vol. 17, No. 5, pp. 557–585.

Pacula, Rosalie Liccardo, David Powell, Paul Heaton, and Eric Sevigny (2013) “Assessing the Effects of Medical Marijuana Laws on Marijuana and Alcohol Use: The Devil is in the Details,” *NBER Working Paper No 19302*.

Powell, David (2019) "Imperfect Synthetic Controls," *Unpublished Working Paper*.

Powell, David, Rosalie Liccardo Pacula, and Mireille Jacobson (2018) "Do medical marijuana laws reduce addictions and deaths related to pain killers?," *Journal of Health Economics*, Vol. 58, No. November, pp. 29–42.

ProCon (2018a) "17 States with Law Specifically about Legal Cannabidiol (CBD)."

——— (2018b) "33 Legal Medical Marijuana States and DC."

Robbins, Michael W, Jessica Saunders, and Beau Kilmer (2017) "A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention," *Journal of the American Statistical Association*, Vol. 112, No. 517, pp. 109–126.

Roine, Jesper and Daniel Waldenström (2011) "Common trends and shocks to top incomes: a structural breaks approach," *Review of Economics and Statistics*, Vol. 93, No. 3, pp. 832–846.

Rosenbaum, Paul R. (2002) *Observational Studies*, Springer Series in Statistics, New York: Springer Science and Business Media, 2nd edition.

Shi, Yuyan (2017) "Medical marijuana policies and hospitalizations related to marijuana and opioid pain reliever," *Drug and Alcohol Dependence*, Vol. 173, pp. 144–150.

Stock, James H and Mark W Watson (2002) "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, Vol. 97, No. 460, pp. 1167–1179.

Tibshirani, Ryan J (2013) "The lasso problem and uniqueness," *Electronic Journal of statistics*, Vol. 7, pp. 1456–1490.

Wen, Hefei and Jason M Hockenberry (2018) "Association of medical and adult-use marijuana laws with opioid prescribing for Medicaid enrollees," *JAMA internal medicine*, Vol. 178, No. 5, pp. 673–679.

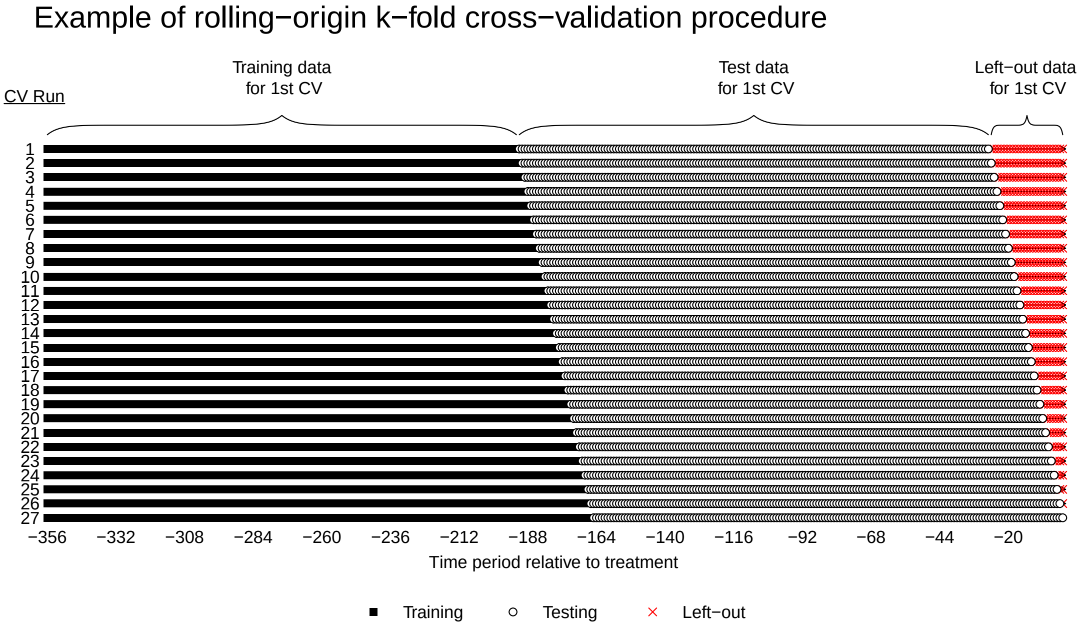
Wen, Hefei, Jason M. Hockenberry, and Janet R. Cummings (2015) "The Effect of Medical Marijuana Laws on Adolescent and Adult use of Marijuana, Alcohol, and Other Substances," *Journal of Health Economics*, Vol. 42, pp. 64–80.

Williams, Jenny, Rosalie Liccardo Pacula, Frank J. Chaloupka, and Henry Wechsler (2004) "Alcohol and marijuana use among college students: economic complements or substitutes?," *Health Economics*, Vol. 13, No. 9, pp. 825–843.

Xu, Yiqing (2017) "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, Vol. 25, No. 1, pp. 57–76.

Yörük, Barış K. and Ceren Ertan Yörük (2011) "The impact of minimum legal drinking age laws on alcohol consumption, smoking, and marijuana use: Evidence from a regression discontinuity design using exact date of birth," *Journal of Health Economics*, Vol. 30, No. 4, pp. 740–752.

Figure 1: SCUL procedure uses rolling k-fold cross-validation to select optimal donor weights, avoiding over-fitting and auto-correllation.



Note: This figure presents a visual depiction of the rolling-origin cross-validation procedure we use to determine the penalty parameter (and therefore synthetic control weights) in our procedure. We only use data from the pre-treatment time period, which in our setting runs for 356 weeks from January 2006 until November 2012. The goal of our application is to create a synthetic control that extends from the date of legalization until the last week of our data, which is 165 weeks. Thus, we use a cross-validation procedure in which the test data is always at least 165 weeks long. For each cross-validation run, we conduct a number of lasso regressions with different penalty parameters using the training data. Training data always come before the test data to avoid using future values to predict past levels. Training and test data are also in contiguous blocks, this forces the method to extrapolate and avoids overfitting (e.g. interpolation). In each run, we choose the penalty parameter that has offers the smallest mean square prediction error for the respective test data. Each subsequent cross-validation run adds one additional week of data until no longer possible. In our setting, we are able to preform a total of 27 runs. We then choose the median lambda penalty parameter from these 27 procedures as our cross-validated penalty parameter. For more details see Section 4.3. Code for this figure was adapted from Section 3.4 of Hyndman and Athanasopoulos (2020).

Figure 2: Restrictions on weights in traditional synthetic control methods prevent any extrapolation and allow for any interpolation, no matter how extreme.

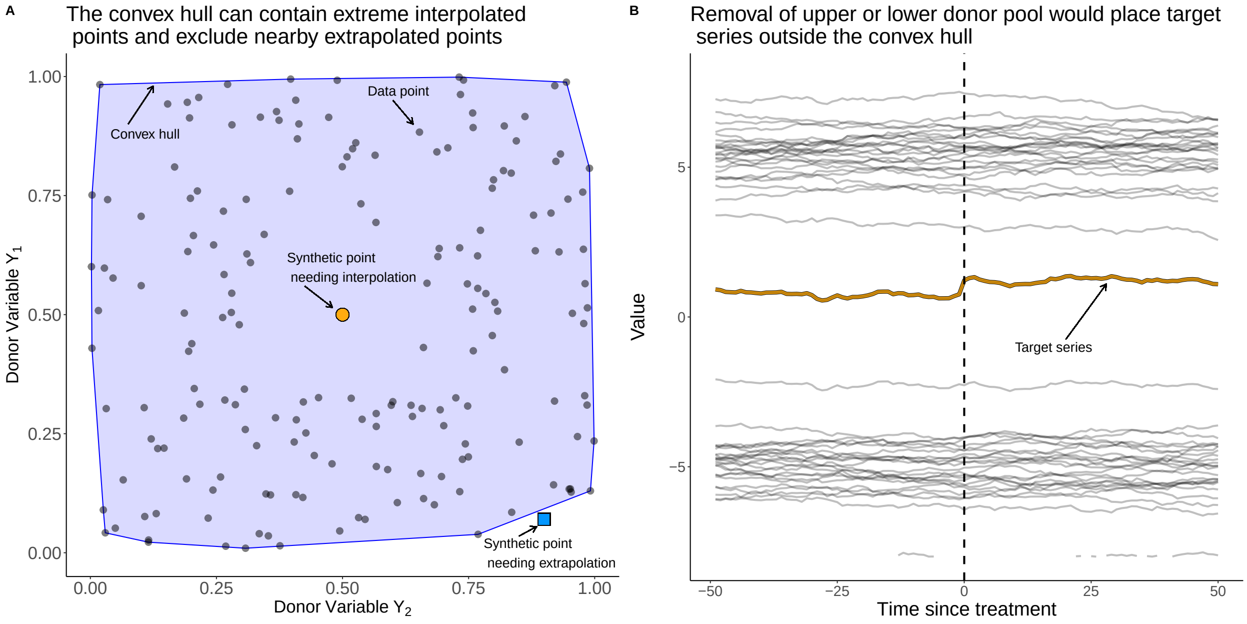
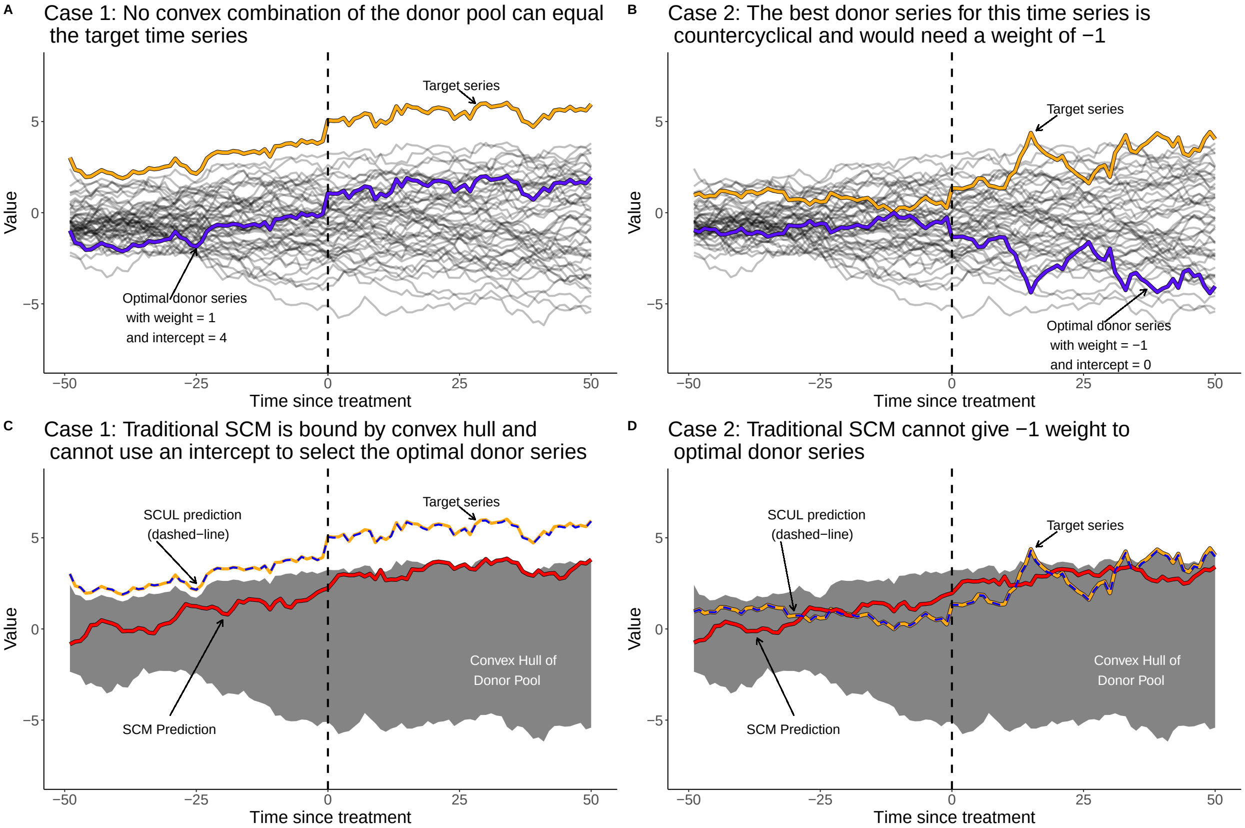
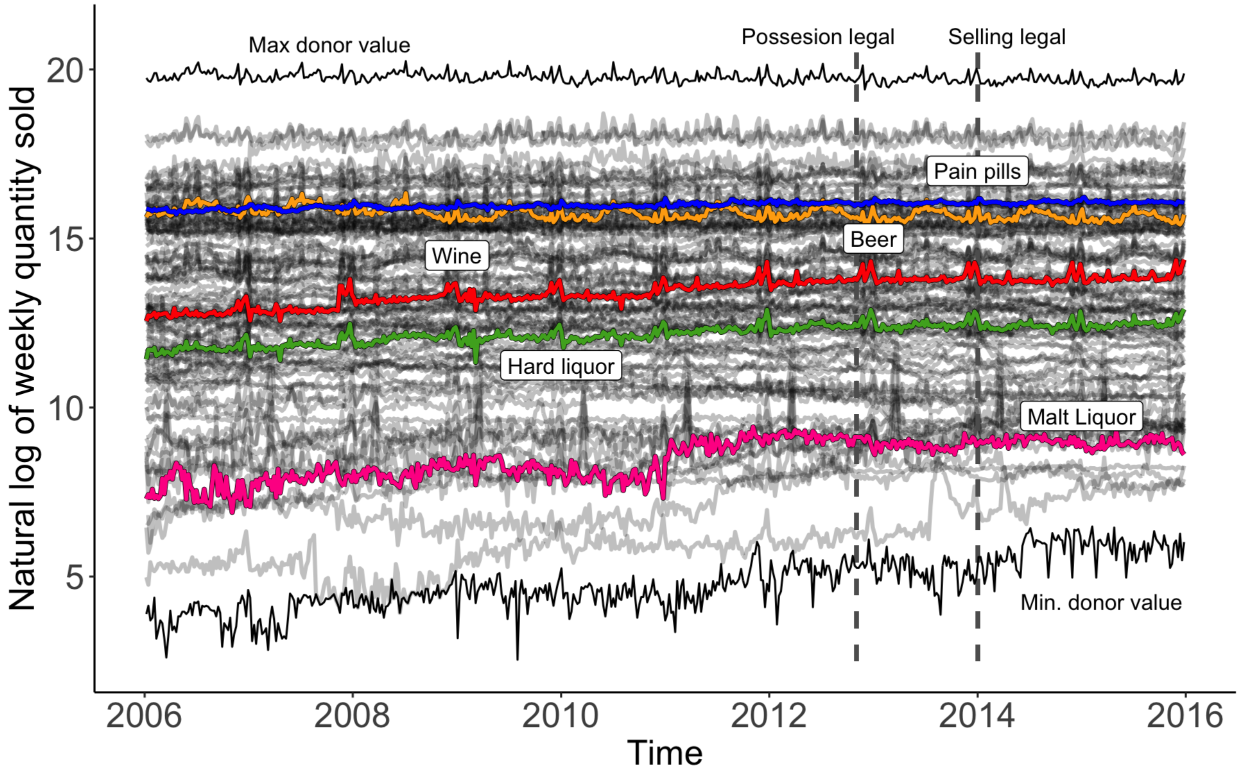


Figure 3: Restrictions on weights prevent the traditional synthetic control methods (SCM) from selecting the optimal donor series in some cases. The synthetic control using lasso (SCUL) procedure preforms well in these settings.



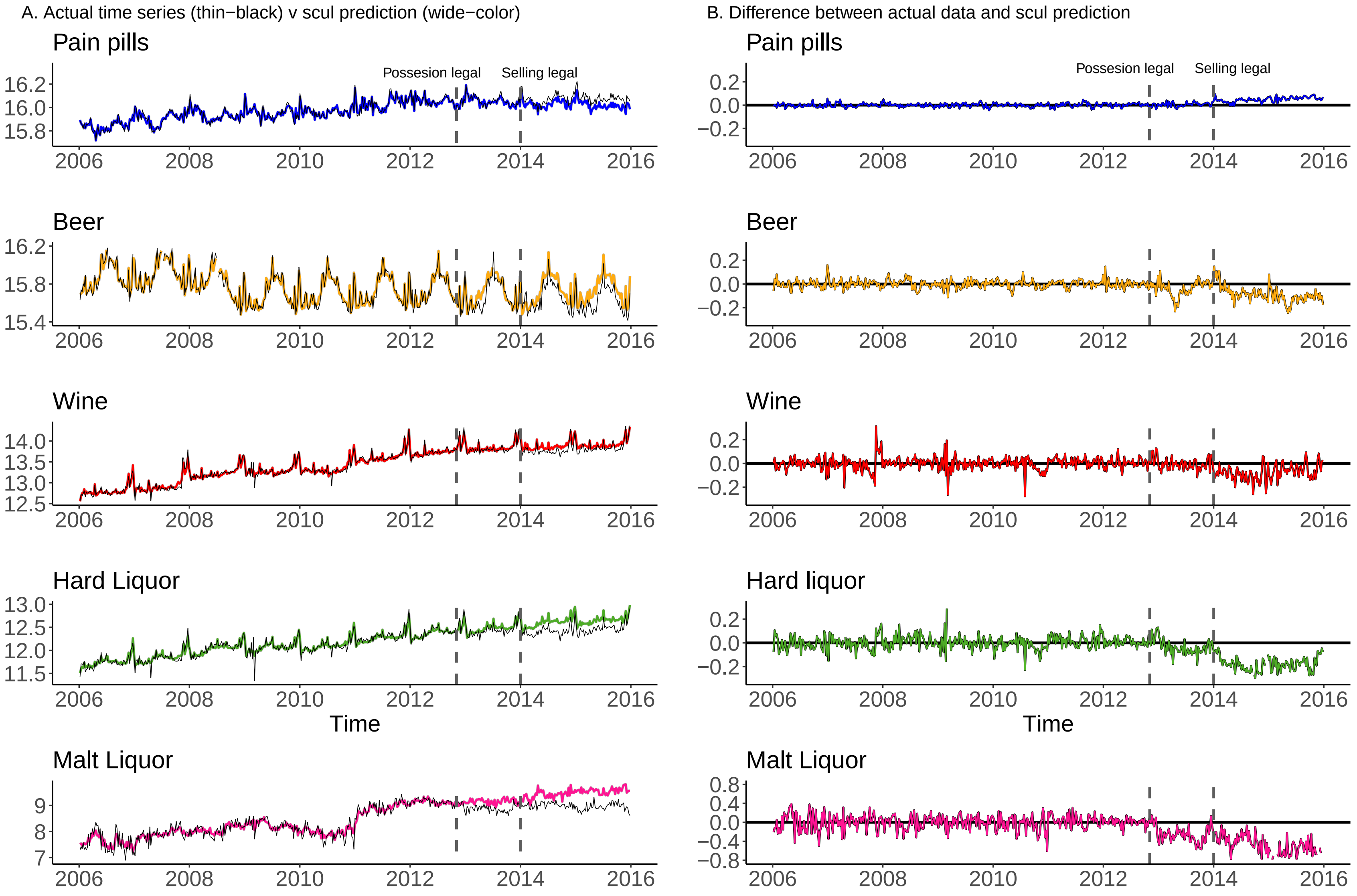
Note: In each case a perfect donor series exists for the target series. All other donor series are unrelated to the target series. In case 1, the target series lies outside of the convex hull of the donor pool with the optimal donor series shifted down by four units. In case 2, the target series has a perfect mirror in the donor pool; that is, it is the negative of the target series. In both cases, the traditional synthetic control method (SCM) cannot select the perfect donor. In case 1, this is because traditional weights cannot extrapolate beyond the support of the donor pool. In case 2, this is because negative weights are not allowed. Our method, the synthetic control using lasso (SCUL), relaxes these two restrictions and selects the perfect donor series in both cases. These cases are not contrived. It is easy to imagine a target series being outside of the convex hull of the donor pool (e.g., U.S. GDP compared to other countries) or two series exhibiting negative correlation (e.g., a price and consumption series or two financial assets).

Figure 4: Sales of five target products in Colorado across time compared to sales from random sample of donor pool.



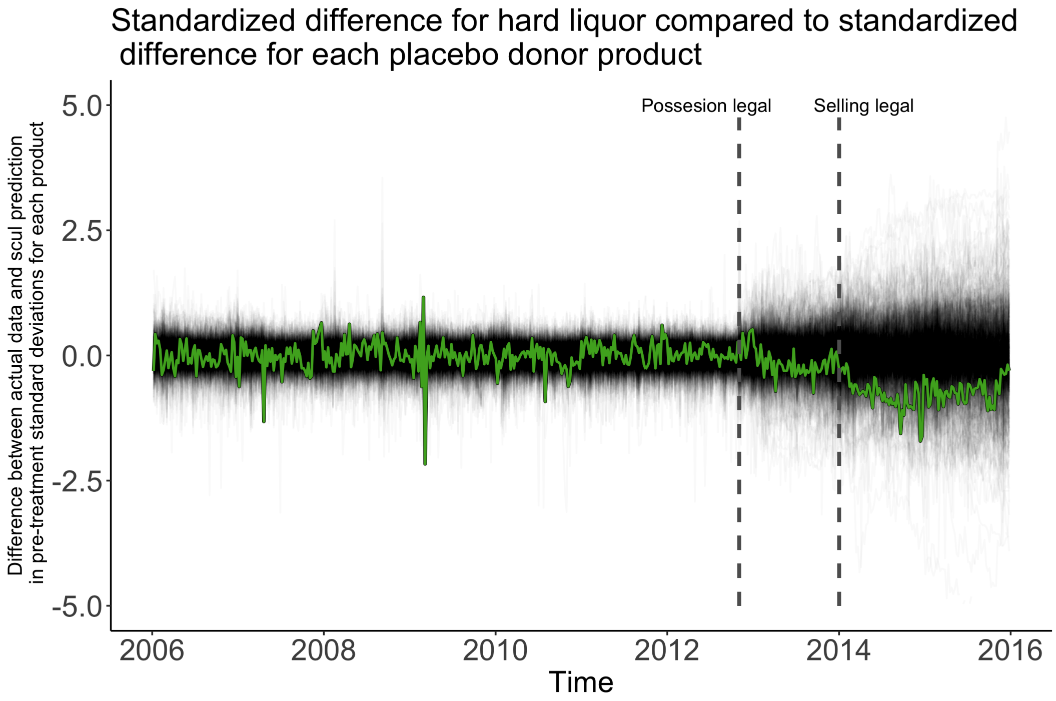
Note: All time series in this figure are the natural log of weekly sales for different state-products. The five target products from Colorado we study are labeled and in color. A random sample of the donor pool—sales of products from states other than Colorado—are displayed with transparency. Darker regions have greater density. We also display the maximum and minimum donor value in each week. The first vertical dashed lines in the figure denotes December 2012, when the vote to legalize recreational marijuana in Colorado was completed. The second vertical dash line is January 2014, which is when Colorado's first recreational dispensaries opened. The graph gives some idea about the relative range of the donor/placebo pool as well as general trends in alcohol and painkiller sales in Colorado. The sales of all target products trended upward throughout the pre-treatment period. There is substantial within-product variation across time that is attenuated due to the common y-axis.

Figure 5: Observed weekly sales data compared to SCUL counterfactual prediction.



Note: Panel A displays the natural log of weekly sales for each target product in Colorado as well as the SCUL counterfactual prediction of that value. In each graph actual sales for the target unit are shown as a thinner black line and the prediction is shown as the wider line in color. Panel B plots the difference between actual sales and the SCUL counterfactual prediction.

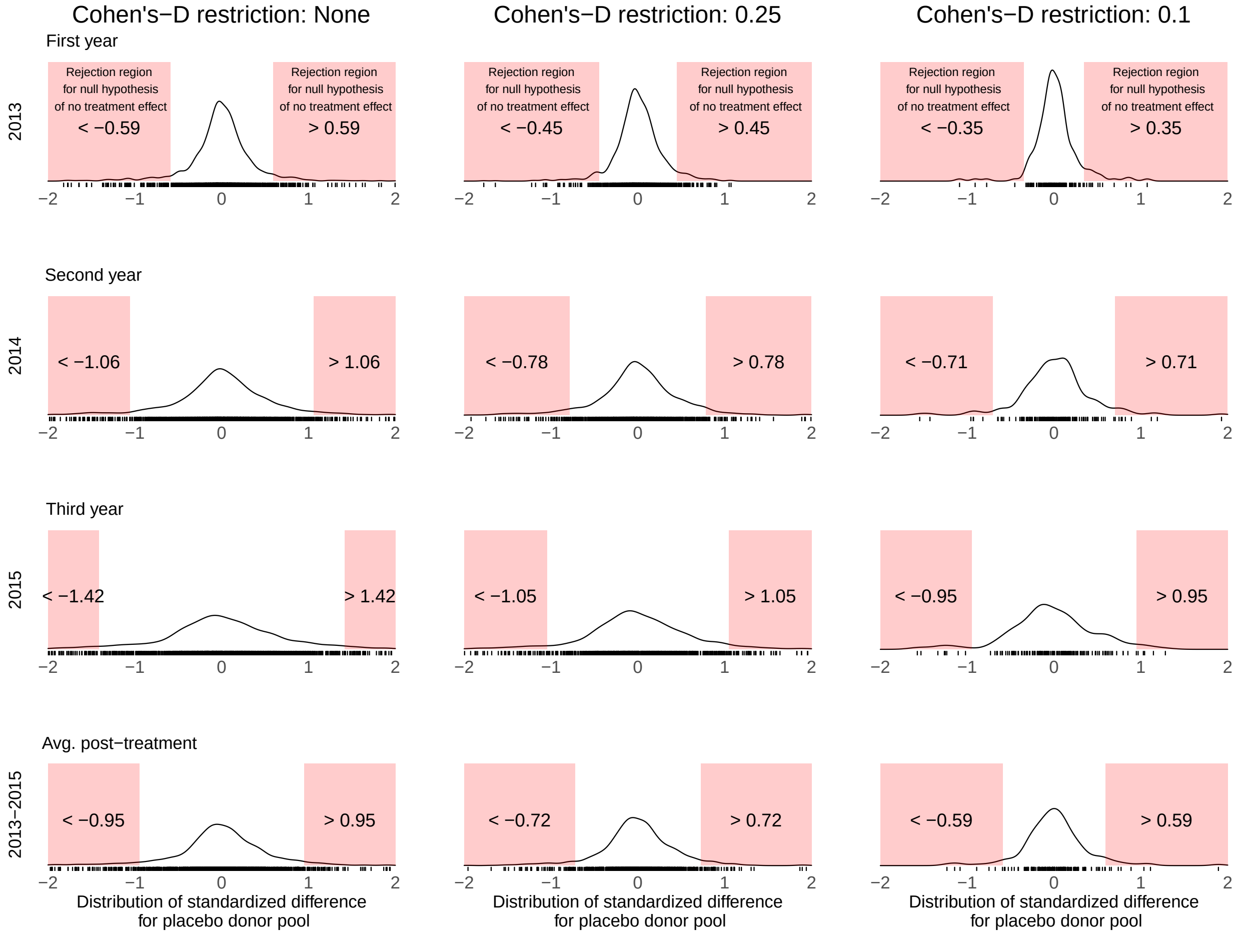
Figure 6: Smoke plot and donor contribution to synthetic control estimates for Colorado hard liquor.



	Share for First Prediction	Share for Most Recent Prediction	Coefficient
TN_oth_beer_oz_0_40_oz	0.12	0.12	0.28
Intercept	0.12	0.12	3.70
MA_cigarettes_total_cnt	0.08	0.08	-0.15
MI_liquor_oz_handle_oz	0.06	0.06	0.13
IN_liquor_oz_fifth_liter_oz	0.05	0.05	0.12
NH_cigarettes_total_cnt	0.05	0.05	-0.09
AZ_bread_total_oz	0.04	0.05	-0.08
KY_liquor_oz_handle_oz	0.04	0.04	0.10
VA_bread_total_oz	0.04	0.04	-0.08
MS_oth_beer_oz_0_40_oz	0.04	0.04	0.12
NY_wine_total_oz	0.04	0.04	0.08
LA_liquor_oz_handle_oz	0.03	0.03	0.06
NY_liquor_oz_handle_oz	0.03	0.03	0.07
VT_fem_hygiene_total_cnt	0.02	0.03	-0.06
MA_kitty_litter_total_oz	0.02	0.02	0.05
VA_soap_lq_total_oz	0.02	0.02	0.05
MI_liquor_oz_fifth_liter_oz	0.02	0.02	0.04
KY_liquor_oz_fifth_liter_oz	0.02	0.02	0.04

Note: In the top-panel, the wide green line displays the difference between actual hard liquor sales each week in Colorado and the SCUL counterfactual prediction as if recreational marijuana had not been legalized. The pre-treatment difference between the two series is small and centered around zero. The gray lines depict the differences between each placebo and its synthetic control assuming those the pre-treatment Cohen's D for that placebo is less than 0.25. The placebo differences are displayed with transparency so that the darker areas have greater density. In the bottom-panel, we report the relative contribution [0-1] and the lasso coefficient for each donor unit to the synthetic prediction. The relative contribution is a function of both the lasso coefficients and the donor pool values in a given time period.

Figure 7: Both time since treatment and Cohen's-D affect the shape of the null distribution, which changes the threshold for statistical significance.



Note: Each null distribution displayed relates to a different post-treatment time period and different Cohen's D inclusion threshold. The blocks of post-treatment time vary by row, with the first three rows showing null distributions over the first (2013), second (2014), and third (2015) year after legalization. The fourth row displays null distributions for average treatment effects taken over the entire post-treatment period. Each column shows null distributions derived from varying pre-treatment Cohen's D exclusion criteria for placebo units: the first column has no exclusion threshold, the second has an exclusion threshold of 0.25, and the third has an exclusion criteria of 0.10. The range of effect sizes (in standard deviation units) that would be considered statistically different from zero at the 10% level are displayed in red for each null distribution. This figure demonstrates that synthetic control methods have the greatest power to detect small effect sizes in the time periods closest to treatment, and when the synthetic control method also provides a satisfactory fit for the placebo pool used to compose the null distribution.

Table 1: The effect of recreational marijuana legalization on sales (0-100%) by of alcohol and over-the-counter painkillers.

Panel A: Treatment begins in 2013 following passage of the recreational marijuana law.

	Pre-treatment fit 2006-2012	First Year 2013	Second Year 2014	Third Year 2015	All Post Treatment 2013-2015
Pain pills	0.15	0.47	3.85	5.96	3.27
		(0.77)	(0.26)	(0.20)	(0.28)
Beer	0.15	–3.44	–6.12	–11.49	–6.82
		(0.33)	(0.35)	(0.21)	(0.28)
Wine	0.12	–0.28	–9.72	–5.48	–4.89
		(0.97)	(0.42)	(0.73)	(0.63)
Hard liquor	0.18	–3.29	–19.44	–17.38	–12.82
		(0.49)	(0.10)	(0.20)	(0.18)
Malt liquor, 0-40oz.	0.22	–24.76	–38.58	–62.75	–41.09
		(0.10)	(0.14)	(0.09)	(0.10)
p-value from joint test of any effect		0.57	0.16	0.19	0.21
p-value from joint test of any alcohol effect		0.15	0.05	0.08	0.06

Panel B: Treatment begins in 2014 following opening of recreational marijuana dispensaries.

	Pre-treatment fit 2006-2013	First Year 2014	Second Year 2015	All Post Treatment 2014-2015
Pain pills, OTC	0.14	2.32	3.44	2.87
		(0.18)	(0.24)	(0.19)
Beer	0.21	–4.31	–7.93	–6.10
		(0.21)	(0.19)	(0.17)
Wine	0.1	–10.90	–10.19	–10.55
		(0.16)	(0.39)	(0.25)
Hard liquor	0.14	–17.50	–16.75	–17.13
		(0.03)	(0.12)	(0.06)
Malt liquor, 0-40oz.	0.32			
p-value from joint test of any effect		0.04	0.17	0.08
p-value from joint test of any alcohol effect		0.03	0.08	0.06

Note: In 2013, possession and home cultivation of recreational marijuana were legal. Dispensaries, physical locations where recreational marijuana can be legally purchased, opened in 2014. Two-sided randomization inference rank based p-values in parentheses. The p-value from the joint test of any effect is an exact test of the sum of the absolute values of effects from five randomly chosen donor products with a Cohen's D of less than 0.25. The p-value from the joint test of any alcohol effect is an exact test of the absolute value of the sum of effects from four randomly chosen donor products with a Cohen's D of less than 0.25. In Panel B, malt liquor is not included in the joint tests and does not have a p-value, since the pre-period Cohen's D for this product is above our threshold for a good fit (0.25). The joint test is adjusted for this exclusion.

Appendix for:
Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data

Alex Hollingsworth and Coady Wing

2022

A The bias bound of synthetic controls and identification

The seminal paper in the synthetic control literature is Abadie, Diamond and Hainmueller (2010) (ADH). Unlike most papers in the modern causal inference literature, ADH does not explicitly lay out a set of identifying assumptions and prove that the synthetic control estimator is an unbiased and/or asymptotically consistent estimator of the treatment effect. The results of the paper are—of course—valid, but they are somewhat obscure and are perhaps hard to interpret for applied researchers.

The key technical material in ADH mainly appears in Appendix B. In this appendix, we work through the analysis in Appendix B, taking a deliberately plodding approach to the derivations, skipping fewer steps, trying to explain how particular expressions arise, and pointing out where specific assumptions are used in the derivations. Whenever possible we link to our main text highlighting identification assumptions. To an interested reader it is also worth exploring the well written appendix section A.1 of Botosaru and Ferman (2019), who present a still technical, but much clearer version of the ADH proof. To enhance clarity for the reader, we reference statistical papers or textbooks that better outline inequalities used. In a similar vein in a few places we refer the reader to portions of the Botosaru and Ferman (2019) version of the proof, where arguments linking one portion of the proof to another are clearly outlined.

In addition to working through this proof, we provide an important extension. Both ADH and Botosaru and Ferman (2019) exploit the restriction of the traditional synthetic control method that the weights are non-zero and sum to one. We demonstrate that this restriction is not necessary for the same proof to be used. We show that for any synthetic control that is a linear weighted combination of a donor pool, that this same proof can be used to show that the bias is bounded and decreases in number of pre-treatment time periods. This simple extension demonstrates that the original ADH proof holds for a much broader collection of synthetic control methods than just the original method.

Throughout, we note steps in the proof that connect to each of the five identifying assumptions outlined in Section 2.3 of the main paper.

A.1 Introduction

The major claim in ADH Appendix B is that—under some assumptions—the absolute value of the bias of the synthetic control estimator will not exceed a particular bound and that the bias declining with the number of pre-treatment periods. We consider meeting these assumptions to be meeting the identification assumptions of ADH. See Section 2.3 of the main paper for more details on each of these identifying assumptions.

A synthetic control estimator will satisfy the ADH bound under five assumptions:

1. No interference between units.

2. The untreated outcomes for each unit are generated by an interactive factor structure model.

3. Any non-common factors are independent over time and units, mean independent of unit specific factor loadings, and possess a finite even g^{th} moment, for some even $g \geq 2$.

4. The time varying common factors are not perfectly multicollinear during the pre-period.

5. There exist weights such that a synthetic control unit consisting of a weighted combination of untreated units can perfectly match the pre-treatment time series of the treated unit.

Note that in the original ADH proof non-negative and convex weights that sum to one are used in a step limit the size of the bias bound. We extend the ADH proof to show that it is not necessary to start with non-negative and convex weights that sum to one in order to estimate a bias bound that decreases in size as the number of pre-treatment periods increases. Thus these weight restrictions, while perhaps desirable properties, are not requirements for identification. Broadly the logic is straightforward; any synthetic control that is a linear combination of a donor pool and is not degenerate can—without loss of generality—have the values of each donor and their respective weights rescaled to enforce the condition that all weights are non-negative and sum to one. For a proof of this claim see Section B of this Appendix.

This extension allows us to proceed with the same proof in ADH without an *a priori* restriction on donor weights save that at least one weight must be non-zero. We also point out that this simple extension implies that the original ADH proof holds for a broader class of synthetic control methods, not just for the classic method.

A.2 Proof of the bias bound

The quantity of interest in synthetic control studies is the period specific treatment effect for the treated unit, which is β_{0t} in the notation developed in Section 2. For each period $t = T_0 + 1, \ldots, T$, the synthetic control estimator of β_{0t} is:

$$\hat{\beta}_{0t}^{\pi^*} = Y_{0t}(1) - \sum_{s=1}^{S} \pi_s^* Y_{st}(0)$$

Here π^* simply indicates the optimal set of weights chosen from a general synthetic control procedure. For our purposes these could be from the traditional ADH synthetic controls, our lasso based approach with cross-validation, or some other procedure. Note that these weights are fixed across time.

Define the bias of the synthetic control estimator as $Bias(\hat{\beta}_{0t}^{\pi^*}) = E[\hat{\beta}_{0t}^{\pi^*} - \beta_{0t}]$, where the expectation is taken over repeated samples from the distribution of ε_{st}.

Under assumptions 1-5, the synthetic control estimator is *not* unbiased because the expected value of the difference between the estimator and the true treatment effect is not equal to zero. Instead, Appendix B in ADH shows that under assumptions 1-5 the absolute value of the bias of the synthetic control estimator is smaller than a specified bound:

$$|Bias(\hat{\beta}_{0t}^{\pi^*})| \ \leq \ C(g)^{1/g} \left(\frac{\overline{a}(T_0)^2 F}{\xi_{A_P}} \right) S^{1/g} max \left\{ \frac{\bar{m_g}^{1/g}}{T_0^{1-1/g}}, \frac{\overline{\sigma}^{1/2}}{T_0^{1/2}} \right\}$$

The bound on the right hand side is fairly complicated and it depends on several unknown parameters. Let's take stock of the pieces of the bound.

- $C(g)$ is the g^{th} moment of a poisson random variable with rate parameter equal to 1. This relates to Assumption 3 stating that $E[|\varepsilon_{st}|^g] < \infty$ for some specified even integer $g \geq 2$.
- ξ_{A_P} is the smallest eigenvalue of $\frac{1}{T_0}\sum_{t=1}^{T_0} a_t^T a_t$. If the common factors are perfectly multicollinear during the pre-period, then
  $xi_{A_P} = 0$. This relates to Assumption 4 stating that there is no pre-period perfect multicollinearity of common factors.
- Let $a_{t,f}$ denote the value of the f^{th} common factor in a_t, where we use $f = 1...F$ to index the F common factors. Then define $\overline{a}(T_0) = \max_{t=1...T_0; f=1...f} |a_{t,f}|$ to be the largest realized value of any of the common factors across the pre-treatment time period. This directly relates to Assumption 2, that the underlying data generating process is a factor structure model.
- S is total number of donor units contributing to the synthetic control. For clarity, for procedures where there is a first stage selection (as is the case with the lasso), this is the number of non-degenerate donors.
- Define $\bar{m_g} = \frac{1}{T_0}\sum_{t=1}^{T_0} E[|\varepsilon_{st}|^g]$, where g is the even integer defined in Assumption 3. This is what ADH refer to as a measure of the *scale* of the non-common factors.
- Finally, $\bar{\sigma} = \frac{1}{T_0}\sum_{t=1}^{T_0} E[|\varepsilon_{st}|^2]$

By supplying a value of g, we can compute $C(g)$. S and T_0 are known features of the study design. In contrast, F, ξ_{A_P}, $\overline{a}(T_0)$ are for all practical purposes, unknown. Likewise $\bar{m_g}$ and $\bar{\sigma}$ are basically unknown since ε_{st} is unknown, although perhaps these terms could be approximated reasonably well.

The bound simplifies when $g = 2$: To see the bound in a relatively simple form, take the regularity condition in Assumption 3(a) to hold for $g = 2$. Then the absolute bias of the synthetic control estimator satisfies:

$$|Bias(\widehat{\beta_{0t'}})| \leq \sqrt{CS} \times \frac{\bar{a}^2 F}{\xi_{A_P}} \times \sqrt{\frac{\bar{m_2}}{T_0}}$$

Here we also rewrite $\bar{m_g}$ to be $\max_{s=1...S} \left[\frac{1}{T_0}\sum_{t=1}^{T_0} E[|\varepsilon_{st}|^g] \right]$ and rely on the fact that when $g = 2$, $\bar{m_2} = \bar{\sigma}$.

In this expression, C is a constant and S is the number of donor units. $\xi_{A_P} > 0$ is the smallest eigenvalue of $\frac{1}{T_0}A_P^T A_P$, which is a measure of the linear independence of the common factors in the pre-treatment period.

$\bar{a}$ and $\bar{m}$ are measures of the scale of the common factors and the non-common factors. Specifically, $\bar{a}$ is the largest absolute value realized by any of the F common factors in any period. And $\bar{m}_2$ is defined above.

The details of the bound are somewhat more complicated and potentially tighter when $g > 2$. But the main insights of the bound for practical purposes are not changed. The key point is that the bias bound will be negligible when the number of pre-treatment periods is large relative to the scale of the non-common factors. It is also be worth noting that the bound will be larger when there are more units in the donor pool, more (and more variable) common factors, and when the common factors are more collinear in the pre-treatment time period (the smaller the ξ_{A_P} the more collinear the common factors are).

Abstracting from $g = 2$: More generally, inspecting the bound does provide a few indications of the things that produce a larger vs smaller bound.

1. As ADH point out, the bound is *decreasing* in the number of pre-treatment time periods. You can see this because T_0 is in the denominator of the final term in the bound.

2. The bound is increasing in the *scale* of ε_{st}. This simply means that for a fixed T_0, the bias bound will be higher if the ε_{st} is more variable than if it is less variable.

3. The bound is increasing in the number of common factors, F. For example, if the outcomes are generated by a two factor model the bias bound of the synthetic control estimator that satisfies assumptions 1-5 will be smaller than the bias bound would be if the data were generated by a five factor model. (In practice, we usually do not know much about F, but the intuition here is easy to understand and also there could be some value in adjusting for observable covariates in some way.)

4. The bias bound is increasing with the number of units in the donor pool. A large donor pool may have many advantages in terms of finding a set of weights with good pre-period fit and could therefore be required in order to satisfy assumption 5. But a downside of a large donor pool is that the bias bound is higher. For the setting of the lasso note, that this is the selected donor pool, not the potential donors that the lasso procedure chooses over. One could imagine deriving a bias bound related to the whole donor choice set, but that is not a feature we have incorporated into the proof.

5. The bound is increasing in the level of multicollinearity between the common factors. ξ_{A_P} is a measure of collinearity between the common factors, where the smaller the ξ_{A_P} the more collinear the common factors are. Since this is in the denominator less collinearity reduces the size of the bound.

A key point in ADH is that if the number of pre-treatment time periods is large *relative* to the scale of ε_{st} then the bias bound is apt to be quite small. In applied work, this is often taken to mean that the bias of the synthetic control estimator is negligible if the number of pre-treatment time periods is large, which is—of course—not quite right. The pre-period must be long relative to the scale of the non-common factors, which in the majority of applied settings is unknowable.

However, an even more important point is that the bound itself is only valid if assumptions 1-5 hold. This is a crucial detail. Even though the bound is shrinking with the length of the pre-period, the result does not mean that as the pre-period goes to infinity a synthetic control estimator is guaranteed to produce the correct counterfactual. The inequality holds if Assumptions 1-5 are met *and* the pre-period is long compared to the scale of the non-common factors; only when all these conditions are met will the bias be negligible. Notice, however, that when assumption 5 holds that the pre-period fit of the synthetic control estimator is perfect; this is an assumption and not a result of the bound. Stated differently, a very long pre-period is not enough to ensure that pre-period fit will be adequate or that any of the other assumptions hold.

A.3 How does the proof work?

The Treated Unit

To make sense of the technical argument in ADH, and understand how the identifying assumptions come into play, start by writing the micro level outcome for the treated unit in a generic period:

$$Y_{0t} = \beta_{0t} D_{0t} + a_t \alpha_0 + \varepsilon_{0t}$$

Now assemble the pre-period data from the treated unit into a collection of vectors and matrices, noting that $D_{0t} = 0$ for $t \leq T_0$ so that the treatment effect drops out of the pre-period model. Let Y_0^P be the $T_0 \times 1$ vector of pre-treatment outcomes for the treated unit. A_P is the $T_0 \times F$ matrix of common factors for the pre-period with t^{th} row a_t. And ε_0^P is the $T_0 \times 1$ vector of non-common factors for the treated unit. In vector-matrix form, the pre-treatment outcomes for the treated unit can be written as:

$$Y_0^P = A_P \alpha_0 + \varepsilon_0^P$$

Now invoke Assumption 4, which implies that the common factors are not perfectly multicollinear in the pre-period and therefore that $(A_P^T A_P)^{-1}$ exists. Multiply both sides of the pre-treatment outcome model by $(A_P^T A_P)^{-1} A_P^T$ to obtain:

$$\begin{aligned}
(A_P^T A_P)^{-1}(A_P^T Y_0^P) &= (A_P^T A_P)^{-1}(A_P^T A_P)\alpha_0 + (A_P^T A_P)^{-1}(A_P^T \varepsilon_0^P) \\
(A_P^T A_P)^{-1}(A_P^T Y_0^P) &= \alpha_0 + (A_P^T A_P)^{-1}(A_P^T \varepsilon_0^P) \\
\alpha_0 &= (A_P^T A_P)^{-1}(A_P^T Y_0^P) - (A_P^T A_P)^{-1}(A_P^T \varepsilon_0^P)
\end{aligned}$$

The Untreated Comparison Units

Now turn to the untreated units. The micro level outcomes for a generic member of the donor pool in a generic time period are given by:

$$Y_{st} = a_t \alpha_s + \varepsilon_{st}$$

And – in matrix form – the pre-treatment data for a generic candidate comparison unit is:

$$Y_s^P = A_P \alpha_s + \varepsilon_s^P$$

Once again invoking Assumption 4, the pre-treatment data can be rewritten as an expression for α_s:

$$\begin{aligned}
(A_P^T A_P)^{-1}(A_P^T Y_s^P) &= (A_P^T A_P)^{-1}(A_P^T A_P)\alpha_s + (A_P^T A_P)^{-1}(A_P^T \varepsilon_s^P) \\
(A_P^T A_P)^{-1}(A_P^T Y_s^P) &= \alpha_0 + (A_P^T A_P)^{-1}(A_P^T \varepsilon_s^P) \\
\alpha_s &= (A_P^T A_P)^{-1}(A_P^T Y_s^P) - (A_P^T A_P)^{-1}(A_P^T \varepsilon_s^P)
\end{aligned}$$

The Estimator

Next, consider the synthetic control estimator, which can be written for a given set of synthetic control weights that come from some procedure, π_s^*. Note that these weights do not need be be non-negative and do not need to sum to one to proceed with the proof. We address this later on and in Section B of this appendix.

$$\begin{aligned}
\hat{\beta}_{0t}^{\pi^*} &= Y_{0t} - \sum_{s=1}^{S} \pi_s^* Y_{st} \\
\hat{\beta}_{0t}^{\pi^*} &= (\beta_{0t} D_{0t} + a_t \alpha_0 + \varepsilon_{0t}) - \left(\sum_{s=1}^{S} \pi_s (a_t \alpha_s + \varepsilon_{st}) \right)
\end{aligned}$$

At this point, substitute the matrix expressions for the factor loadings α_0 and α_s based on the pre-treatment data.

$$\begin{aligned}
\hat{\beta}_{st}^{\pi^*} &= (\beta_{0t} D_{0t} + a_t [(A_P^T A_P)^{-1}(A_P^T Y_0^P) - (A_P^T A_P)^{-1}(A_P^T \varepsilon_0^P)] + \varepsilon_{0t}) \\
&- \left(\sum_{s=1}^{S} \pi_s^* (a_t [(A_P^T A_P)^{-1}(A_P^T Y_s^P) - (A_P^T A_P)^{-1}(A_P^T \varepsilon_s^P)] + \varepsilon_{st}) \right) \\
&= \beta_{0t} D_{0t} + a_t A_P^T A_P^{-1} A_P^T Y_0^P - a_t A_P^T A_P^{-1} A_P^T \varepsilon_0^P + \varepsilon_{0t} \\
&- a_t A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* Y_s^P + a_t A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P - \sum_{s=1}^{S} \pi_s^* \varepsilon_{st} \\
&= \beta_{0t} D_{0t} + a_t A_P^T A_P^{-1} A_P^T (Y_0^P - \sum_{s=1}^{S} \pi_s^* Y_s^P) - a_t A_P^T A_P^{-1} A_P^T \varepsilon_0^P \\
&+ a_t A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P + \varepsilon_{0t} - \sum_{s=1}^{S} \pi_s^* \varepsilon_{st}
\end{aligned}$$

Invoke Assumption 5, which implies that the weights are chosen so that $Y_{0t} = \sum_{s=1}^{S} \pi_s^* Y_{st}$ for all $t = 1 \ldots T_0$. This means that the second term on the right hand side is equal to zero. This gives us:

$$\hat{\beta}_{st}^{\pi^*} = \beta_{0t}D_{0t} - a_t A_P^T A_P^{-1} A_P^T \varepsilon_0^P + a_t A_P^T A_P{}^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P \tag{16}$$
$$+ \varepsilon_{0t} - \sum_{s=1}^{S} \pi_s^* \varepsilon_{st}$$

Now consider the treated period ($t > T_0$) and take expectations over the distribution of ε_{st}:

$$E[\hat{\beta}_{st}^{\pi^*}] = \beta_{0t} - E[a_t A_P^T A_P^{-1} A_P^T \varepsilon_0^P] + E[a_t A_P^T A_P{}^{-1} A_P^T \sum_{s=1}^{S} \pi_s \varepsilon_s^P]$$
$$+ E[\varepsilon_{0t}] - E[\sum_{s=1}^{S} \pi_s \varepsilon_{st}]$$

Under Assumption 3, we know that ε_{st} is distributed independently over time and is mean independent across units. Thus, $E[\varepsilon_{0t}] = 0$ and $E[a_t A_P^T A_P^{-1} A_P^T \varepsilon_0^P] = 0$.

The synthetic control weights are optimized using pre-treatment data. This means that the weights are independent of ε_{st} for $t > T_0$ so $E[\sum_{s=1}^{S} \pi_s^* \varepsilon_{st}] = 0$ because we are considering post treatment time periods.

However, the weights are not independent of pre-treatment values of ε_{st}, and so $E[a_t A_P^T A_P{}^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P] \neq 0$ under assumption 3 or any of the other assumptions. As a result we are left with:

$$E[\hat{\beta}_{st}^{\pi^*}] = \beta_{0t} + E[a_t A_P^T A_P{}^{-1} A_P^T \sum_{s=1}^{S} \pi_s^* \varepsilon_s^P]$$

This shows that – under assumptions 1-5 – a synthetic control estimator that perfectly matches the pre-treatment outcome time series in the treated unit is equal to the true effect *plus* a bias term. The bias comes from the pre-treatment (finite sample) association between the exogenous non-common factors (ε_S^P) and the common factors (A_P). The bias survives *even after* taking expectations because the synthetic control weights are chosen using the pre-treatment data itself.

The intuition for how this happens is not dissimilar from the way that so-called weak instruments bias can create problems for the two-stage least squares estimator. In the weak instruments case, we can imagine a situation where first stage coefficient on the instrumental variable is equal to zero in the population model. Even if the instrument is exogenous (uncorrelated with the population error terms in the first stage and reduced form models), the estimated first stage based on a given (finite) sample of data is unlikely to be exactly equal to zero. The reason is that in the sample data there is likely to be – just by chance – some level of association between the instrument and the first stage error terms. Even though this finite sample association is spurious and will disappear asympototically, it can still lead to important biases in applied research based on the two-stage least squares estimator.

In the synthetic control case, we assume that error (i.e., the non-common factors) in the population data generating process, ε_{st} is independent of the common factors in a_t. This means that there is no systematic association between the values of a_t and ε_{st}. It means that periods in which there happens to be high values of a_t are not more likely to have high (or low) values of ε_{st}. It's important to note that these assumptions are for *the population*—in the sample of available data, there will almost certainly be some association between the common factors and the non-common factors. Look carefully at the bias term to make sense of what is happening:

$$Bias(\hat{\beta}_{st}^{\pi^*}) = E[a_t A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s \varepsilon_s^P]$$

$$Bias(\hat{\beta}_{st}^{\pi^*}) = E[a_t \sum_{s=1}^{S} \pi_s^* A_P^T A_P^{-1} A_P^T \varepsilon_s^P]$$

$A_P^T A_P^{-1} A_P^T \varepsilon_s^P$ is the coefficient vector from a linear projection of the pre-treatment non-common factors onto the pre-treatment common factors for unit s. If the common factors and non-common factors were observable, these are the coefficients you'd get if you regressed ε_{st} on a_t using only data from unit s and the pre-treatment time periods. These coefficients will not be exactly zero in a finite sample even though they are zero in the population data generating process according to Assumption 3. Bringing in the summation sign shows simply shows that the synthetic control weights are used to form a weighted combination of these non-common factor on common factor linear projection coefficients, producing an $F \times 1$ summary set of synthetic first stage spurious associations.

When we use the synthetic control estimator to forecast the counterfactual for a *new* post treatment time period, we multiply these linear projection coefficients by the new value of the common factor vector, a_t. That causes bias, of course, because these pre-period linear projection coefficients are actually spurious associations that won't have any out of sample predictive validity.

One intuitive observation is that with a sufficiently large number of pre-treatment time periods, it's unlikely that the bias will be substantial because more time periods will lead to fewer chance associations between the common factors and the non-common factors. But what else determines the size of the bias?

The bias bound

ADH show that the bias term is bounded, and that the bound does indeed decline with the length of the pre-treatment time period and will be small when the number of pre-treatment time periods is large relative to the scale of the non-common factors. However, the bound is somewhat complicated and depends on multiple unknown quantities and so it's numeric value is generally not estimable. Moreover, the idea that you can establish such a bound is somewhat counter-intuitive to most people. So it may be helpful to work through the logic of the exercise.

Start by recalling that the matrix representation $A_P^T A_P$ can be re-written as $A_P^T A_P = \sum_{m=1}^{T_0} a_m^T a_m$. Similarly, we can write $A_P^T \varepsilon_s^P = \sum_{j=1}^{T_0} a_j^T \varepsilon_{sj}$. Substitute these expressions into the bias term to obtain:

$$a_t \sum_{s=1}^{S} \pi_s^* A_P^T A_P^{-1} A_P^T \varepsilon_s^P = \sum_{s=1}^{S} \pi_s^* a_t \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1} \left(\sum_{j=1}^{T_0} a_j^T \varepsilon_{sj} \right)$$

The final summation sign on the right hand side is cycling over the $j = 1...T_0$ pre-treatment time periods. At each iteration, we can see that it multiplies a_j^T by a constant equal to $a_t \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1}$. Let's reformulate that action by writing $V_{tj} = a_t \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1} a_j^T$, where t is the post-treatment time period index that we are trying to forecast and j is a particular pre-treatment time period. Substitute V_{tj} into the expression to obtain:

$$a_t \sum_{s=1}^{S} \pi_s A_P^T A_P^{-1} A_P^T \varepsilon_s^P = \sum_{s=1}^{S} \pi_s^* \sum_{j=1}^{T_0} V_{tj} \varepsilon_{sj}$$

This means we can re-express the expected bias of the synthetic control estimator for the treatment effect in period t as:

$$\begin{aligned} Bias(\hat{\beta}_{st}^{\pi^*}) &= E[a_t A_P^T A_P^{-1} A_P^T \sum_{s=1}^{S} \pi_s \varepsilon_s^P] && (17) \\ &= E\left[\sum_{s=1}^{S} \pi_s \sum_{j=1}^{T_0} V_{tj} \varepsilon_{sj} \right] && (18) \end{aligned}$$

The first step is to apply the Cauchy-Schwarz inequality to V_{tj}. The need for this will not be immediately clear, but will help create the bounds. Recall that the Cauchy-Schwarz inequality implies that for a pair of random variables X and Y, $E[XY]^2 \leq E[X^2]E[Y^2]$. To see the connection to V_{tj}, think about a_t and a_j as the two random vectors. (Don't focus on $\sum_{m=1}^{T_0} a_m^T a_m$, which is the filling of the sandwich.) We can think about a_t times a_j through the lens of Cauchy-Shwarz. ADH point out that:

$$\begin{aligned} V_{tj}^2 &\leq V_{tt} V_{jj} \\ \left[a_t \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1} a_j^T \right]^2 &\leq \left(a_t \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1} a_t^T \right) \times \left(a_j \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1} a_j^T \right) \\ &\leq \left(\frac{F \times \bar{a}(T_0)^2}{T_0 \times \xi_{A_P}} \right)^2 && (19) \end{aligned}$$

By definition, F is total the number of common factors, T_0 is the number of pre-treatment periods, $\bar{a}(T_0)$ is maximum absolute factor value in the pre-period, and ξ_{A_P} is the minimum eigenvalue of the matrix of common factors in the pre-period.

The first inequality follows above from a direct application of the Cauchy-Shwarz restriction. Botosaru and Ferman (2019) provide a nice demonstration of the second inequality. Here we adapt their proof to our notation, add a few steps, and move in reverse for clarity.

Starting by considering the numerator inside the square in equation 19. It is straightforward to show that it is weakly greater than the following term

$$F \times \overline{a}(T_0)^2 \geq \sum_{m=1}^{F} a_{tm}^2$$

By definition $\bar{a}(T_0) = max(|a_t|)$ for $t \leq T_0$, therefore:

$$\begin{aligned}
\sum_{m=1}^{F} a_{tm}^2 &\leq \sum_{m=1}^{F} max(|a_t|)^2 \\
&= max(|a_t|)^2 \sum_{m=1}^{F} 1 \\
&= max(|a_t|)^2 F \\
&= F \times \bar{a}(T_0)^2
\end{aligned}$$

Given that each side of the inequality is weakly positive. We can divide each side by the same positive expression $T_0 \times \xi_{A_P}$ without affecting the inequality.

$$\begin{aligned}
F \times \overline{a}(T_0)^2 &\geq \sum_{m=1}^{F} a_{tm}^2 \\
\frac{F \times \overline{a}(T_0)^2}{T_0 \times \xi_{A_P}} &\geq \frac{\sum_{m=1}^{F} a_{tm}^2}{T_0 \times \xi_{A_P}}
\end{aligned}$$

Now given that,

$$\frac{\sum_{m=1}^{F} a_{tm}^2}{T_0 \times \xi_{A_P}} = \frac{a_t a_t^T}{T_0 \times \xi_{A_P}}$$

and

$$\begin{aligned}
\frac{a_t a_t^T}{T_0 \times \xi_{A_P}} &\geq \left(a_t \left(\sum_{m=1}^{T_0} a_m^T a_m \right)^{-1} a_t^T \right) \\
&= V_{tt}
\end{aligned}$$

We have that

$$V_{tt} \leq \frac{F \times \overline{a}(T_0)^2}{T_0 \times \xi_{A_P}} \tag{20}$$

Since this is not indexed by t, we also have that

$$V_{jj} \leq \frac{F \times \overline{a}(T_0)^2}{T_0 \times \xi_{A_P}}$$

Therefore,

$$V_{tt} V_{jj} \leq \left(\frac{F \times \overline{a}(T_0)^2}{T_0 \times \xi_{A_P}} \right)^2$$

We use this inequality in a later step. Before employing it, consider the bias equation 18. Let us rewrite this equation collapsing the second term into a transformed error. Specifically, we will rewrite $\sum_{j=1}^{T_0} V_{tj} \varepsilon_{sj}$ to be $\bar{\varepsilon}_{ts}$.

Now we can consider the portion of equation 18 that is inside the expectation, $\sum_{s=1}^{S} \pi_s^* \sum_{j=1}^{T_0} V_{tj} \varepsilon_{sj}$, and rewrite it as $\sum_{s=1}^{S} \pi_s^* \bar{\varepsilon}_{ts}$.

Recall that Hölder's inequality[18] shows that for two random variables X and Y, with $\frac{1}{g} + \frac{1}{q} = 1$ and $g, q > 0$,

$$|E(XY)| \leq E(|XY|) \leq (E|X|^g)^{(1/g)} (E|Y|^q)^{(1/q)}$$

Also not that in the original proof, for the next two steps ADH invoke the convexity assumptions enforced on their synthetic control weights. So long as at least one weight is non-zero, the fact that the synthetic control is a linear combination of donors allows us to transform the weights (by multiplying each by $c_s = \frac{1}{sign(\pi_s^*) \times \sum_i^S |\pi_i^*|}$) and donors (by multiplying by $1/c_s$) so that the transformed weights are non-negative and sum to one. In appendix section B, we show that this initial assumption on the weights is not necessary to achieve the same property.

Continuing with our proof and following ADH, we use Hölder's inequality to show:

$$\begin{aligned} |\sum_{s=1}^{S} \pi_s^* \bar{\varepsilon}_{ts}| &\leq \sum_{s=1}^{S} |\pi_s^*||\bar{\varepsilon}_{ts}| \\ &= \sum_{s=1}^{S} \pi_s^* |\bar{\varepsilon}_{ts}| \end{aligned}$$

The last step follows since $\pi_s = |\pi_s|$ once the weights are transformed as in Section B.

We can further apply Hölder's inequality to show that:

$$\sum_{s=1}^{S} \pi_s^* |\bar{\varepsilon}_{ts}| \leq \left(\sum_{s=1}^{S} \pi_s^{*q} \right)^{1/q} \left(\sum_{s=1}^{S} |\bar{\varepsilon}_{ts}|^g \right)^{1/g}$$

Noting that both g and q must be > 1 and $\frac{1}{g} + \frac{1}{q} = 1$.

[18] Casella and Berger (2001, eq, 4.7.2).

Invoking section B of our appendix allows us to state that the transformed weights are non-negative and sum to one and that $(\sum_{s=1}^{S} \pi_s^{*q})^{1/q}$ will always be less than one. This can be proven in two simple steps. First by realizing that for any $x \in [0,1]$ and for any $q > 1$, $x^q < x$. Thus $\sum_{s=1}^{S} \pi_s^{*q} < \sum_{s=1}^{S} \pi_s^{*} = 1$. Since $\sum_{s=1}^{S} \pi_s^{*q} < 1$ and $q > 1$, $\left(\sum_{s=1}^{S} \pi_s^{*q}\right)^{1/q} < 1$. Thus we know that:

$$\left(\sum_{s=1}^{S} \pi_s^{q*}\right)^{1/q} \left(\sum_{s=1}^{S} |\bar{\varepsilon}_{ts}|^g\right)^{1/g} \leq \left(\sum_{s=1}^{S} |\bar{\varepsilon}_{ts}|^g\right)^{1/g}$$

This gives us that:

$$\sum_{s=1}^{S} \pi_s^{*} |\bar{\varepsilon}_{ts}| \leq \left(\sum_{s=1}^{S} |\bar{\varepsilon}_{ts}|^g\right)^{1/g}$$

Now consider a special case of Holder's inequality with the second random variable set to 1, then for $1 < g < \infty$, $E|X| \leq \{E|X|^g\}^{1/g}$ (Casella and Berger, 2001, eq, 4.7.5).

Therefore

$$\begin{aligned} E\left(\sum_{s=1}^{S} \pi_s^{*} |\bar{\varepsilon}_{ts}|\right) &\leq \left(E\left(\sum_{s=1}^{S} |\bar{\varepsilon}_{ts}|^g\right)\right)^{1/g} \\ &= \left(\sum_{s=1}^{S} E|\bar{\varepsilon}_{ts}|^g\right)^{1/g} \\ &= \left(\sum_{s=1}^{S} E|\sum_{j=1}^{T_0} V_{tj}\varepsilon_{sj}|^g\right)^{1/g} \end{aligned} \tag{21}$$

The second steps follows due to the linearity of expectations. The last step is substituting back out for $\bar{\varepsilon_{ts}}$.

Now consider just the middle component from equation 21, $E|\sum_{j=1}^{T_0} V_{tj}\varepsilon_{sj}|^g$. Both ADH and Botosaru and Ferman (2019) show that you can use Rosenthal's inequality and the inequality from equation 20 to further bound this element.

Before proceeding we will write Rosenthal's inequality from Ibragimov and Sharakhmetov (2002). Here there are a sequence of random variables x_i. We are bounding the expectation of the absolute value of the sum of the random variables raised to t.

$$E|\sum_{i=1}^{N} x_i|^t \leq C(t) \max\left(\sum_{i=1}^{n} E|x_i|^t, \left(\sum_{i=1}^{n} Ex_i^2\right)^{t/2}\right) \tag{22}$$

Ibragimov and Sharakhmetov (2002) demonstrate that under some conditions, $C(t)$ can be directly calculated as the t^{th} moment of a Poisson random variable with rate parameter equal to 1.

Now continuing with our proof and using Rosenthal's inequality, noting again that we are focusing on $E|\sum_{j=1}^{T_0} V_{tj}\varepsilon_{sj}|^g$, which is analogous equation 22 except for the additional V_{tj} component that we bounded earlier.

$$E|\sum_{j=1}^{T_0} V_{tj}\varepsilon_{sj}|^g \leq C(g)\left(\frac{\overline{a}(T_0)^2 F}{\xi_{A_P}}\right)^g max\left\{\frac{\bar{m_g}^{1/g}}{T_0^{1-1/g}}, \frac{\overline{\sigma}}{T_0^{1/2}}\right\}$$

Where,

- Define $\bar{m_g} = \left[\frac{1}{T_0}\sum_{t=1}^{T_0} E[|\varepsilon_{st}|^g]\right]$, where g is the even integer defined in Assumption 3. This is what ADH refer to as a measure of the *scale* of the non-common factors.
- $\bar{\sigma} = \left[\frac{1}{T_0}\sum_{t=1}^{T_0} E[|\varepsilon_{st}|^2]\right]$
- C(g) is the g^{th} moment of a Poisson random variable with rate parameter equal to 1 (Ibragimov and Sharakhmetov, 2002)

Since this is the bound for just the middle term in equation 21, we need to modify the inequality slightly, to accommodate the additional summation (over S total elements) and exponent $(1/g)$.

Once we do this, we have the final bound on the bias.

$$|Bias(\hat{\beta}_{st}^{\pi^*})| \leq C(g)^{1/g}\left(\frac{\overline{a}(T_0)^2 F}{\xi_{A_P}}\right) S^{1/g} max\left\{\frac{\bar{m_g}^{1/g}}{T_0^{1-1/g}}, \frac{\overline{\sigma}}{T_0^{1/2}}\right\}$$

B Rescaling to ensure convexity

Let $Y_{0t}^* = \sum_{s=1}^{S} \pi_s^* Y_{st}$ be a synthetic control estimate for an outcome Y at time t. The estimate is simply a weighted combination of donor variables (denoted by $s \geq 1$). There are no restrictions on π_s^* save that $\forall\ \pi_s \in \mathbb{R}$ and at least one $\pi_s \neq 0$, which directly comes from assumption 5. Because this is a simple linear combination, without loss of generality, we can transform the weights and each respective donor value such that the transformed weights will be convex.

First, we remove the time subscript for clarity. We consider just a single time period for now, since it is straightforward to show that the result generalizes to multiple time periods. Second, let π Be the $(1 \times S)$ vector of weights and $Y = Y_{s\neq 0}$ be a vector $(S \times 1)$ donors values.

We show that there exists an element-wise rescaling of π and Y, denoted by $\tilde{\pi}$ and $\tilde{Y}_s$, where for element s,

- $\tilde{\pi}_s = c_s \pi_s$,
- and $\tilde{Y}_s = \frac{1}{c_s} Y_s$,

such that the following four properties hold:

1. $\sum_s^S \tilde{\pi}_s = 1$

2. $\forall\ \tilde{\pi}_s,\ 0 \leq \tilde{\pi}_s \leq 1$, and

3. $\pi_s Y_i = \tilde{\pi}_s \tilde{Y}_i$

4. $Y^* = \sum_s^S \pi_s Y_i = \sum_s^S \tilde{\pi}_s \tilde{Y}_i$

Below we show that the transform $c_s = \frac{1}{sign(\pi_s) \times \sum_s^S |\pi_s|}$ has these properties. Note that $\frac{1}{c_s}$ will exist $\forall i$ so long as our assumption that at least one $\pi_s \neq 0$. This is to say, our re-weighting is possible so long as the synthetic control is not trivially always equal to zero.

B.1 Property 1: Rescaled weights sum to one

$$\begin{aligned}
\sum_s^S \tilde{\pi}_s =& \sum_s^S c_s \pi_s \\
=& \sum_s^S \frac{\pi_s}{sign(\pi_s) \times \sum_s^S |\pi_s|} \\
=& \sum_s^S \frac{sign(\pi_s) \times \pi_s}{\sum_s^S |\pi_s|} \\
=& \sum_s^S \frac{|\pi_s|}{\sum_s^S |\pi_s|} \\
=& \frac{\sum_s^S |\pi_s|}{\sum_s^S |\pi_s|} \\
=& 1
\end{aligned}$$

Note that $\sum_s^S |\pi_s| \neq 0$ since at least one $\pi_s \neq 0$.

B.2 Property 2: Rescaled weights are between zero and one

Let's begin by rewriting the transform for a given element i using absolute values

$$\begin{aligned}
\tilde{\pi}_s =& c_s \pi_s \\
=& \frac{\pi_s}{sign(\pi_s) \times \sum_s^S |\pi_s|} \\
=& \frac{sign(\pi_s) \times \pi_s}{\sum_s^S |\pi_s|} \\
=& \frac{|\pi_s|}{\sum_s^S |\pi_s|}
\end{aligned} \tag{23}$$

Rewriting the transformation using absolute values makes both the lower and upper bound clear.

Lower bound: rescaled weight is greater than zero

By definition, $|\pi_s| \geq 0 \, \forall i$ and the sum of a weakly positive series is also weakly positive. Therefore $\sum_s^S |\pi_s| \geq 0$. This establishes that both the numerator and denominator in equation 23 are weakly positive, which means the expression is also weakly positive. Note that the expression is defined since $\sum_s^S |\pi_s| \neq 0$ because at least one $\pi_s \neq 0$.

$$\begin{aligned} \frac{|\pi_s|}{\sum_s^S |\pi_s|} &\geq 0 \\ \tilde{\pi}_s &\geq 0 \end{aligned}$$

Upper bound: rescaled weight is less than one

By definition,

$$|\pi_s| \leq \sum_s^S |\pi_s|$$

With a simple rearrangement,

$$\begin{aligned} \frac{|\pi_s|}{\sum_s^S |\pi_s|} &\leq 1 \\ \tilde{\pi}_s &\leq 1 \end{aligned}$$

showing that the upper bound for any given $\tilde{w}_s$ is one.

Combining B.2 and B.2, gives us that $\forall i \; \tilde{\pi}_s, \; 0 \leq \tilde{\pi}_s \leq 1$.

B.3 Property 3: Rescaling does not affect the contribution of any one donor to the synthetic prediction.

Consider the original synthetic control, $Y_t^* = \sum_s^S \pi_s Y_{it}$. Here the contribution of any specific donor i to the synthethic prediction is $\pi_s Y_s$, which is simply the weight for i, π_s, multiplied by the value of the donor, Y_s. Multiplying the weight by c_s and the donor value by the inverse $\frac{1}{c_s}$ leaves this value unchanged. That is,

$$\begin{aligned} \tilde{\pi}_s \tilde{Y}_i &= (c_s \pi_s)(\frac{1}{c_s} Y_i) \\ &= \pi_s Y_i \end{aligned}$$

B.4 Property 4: The synthetic prediction is the same after rescaling

Technically this is a restating of the third property in B.3, but for clarity we show here that the value of the synthetic control (i.e., the predicted value for Y) is the same when rescaled as it was before the rescaling.

$$\begin{aligned} Y^* &= \sum_s^S \tilde{\pi}_s \tilde{Y}_i \\ &= \sum_s^S (c_s \pi_s)(\frac{1}{c_s} Y_i) \\ &= \sum_s^S \pi_s Y_i \end{aligned}$$

Therefore our claim holds.

C Sensitivity checks and model comparisons

C.1 Non-uniform weighting in randomization inference procedure

The randomization inference procedure we use relies on the assumption that there is no treatment effect among the placebos and that Assumptions 1-5 hold for all units and periods. Our approach could be invalid if some of the placebos we use have non-common factors that are driven up or down in ways that make it harder or easier to reject the null hypothesis. The sensitivity analysis proposed by Firpo and Possebom (2018) explores these concerns using methods that are similar to Rosenbaum (2002)'s approach to sensitivity to hidden bias in observational studies.

The randomization inference approach we use here implicitly gives equal weight to every placebo when computing the rank based p-value. Firpo and Possebom (2018) point out that a key assumption of such an approach is that such omitted factors are not affecting our procedure. Firpo and Possebom (2018) also demonstrate a clever alternative procedure that serves to bound how this underlying uniform approach, when in the presence of such bias, could affect the p-value. At an agnostic level, this omitted feature could be making it either more or less likely to find a statistically significant effect. Since it is impossible to know precisely the bias of each placebo unit, Firpo and Possebom (2018) suggest that researchers consider the worst possible case (where bias is working in favor of finding a statistically significant effect) and the best possible case (where bias is working against finding a statistically significant effect).

We adapt the Firpo and Possebom (2018) method to our setting and provide sensitivity estimates for the p-value of our treatment effect estimates. The procedure is straightforward for any given target variable. In short, we will consider many different weighting schemes and record how a given treatment effect's p-value changes based on changing these weights. The base weighting scheme we will perturb is where each placebo j receives weight $\omega_j(\phi) = \frac{exp(\phi v_j)}{\sum_{j' \in \Omega} exp(\phi v_{j'})}$. Under uniform weighting $\phi = 0$. We will change v_j to be either 0 or 1 based on the particular scenario we are considering in a deterministic manner. ϕ simply controls how much more weight the placebos whose v_j has been set to one receive relative to those whose v_j has been set to zero.

For any given scenario, where $j = 0$ indicates the target product and J indicates the number of placebo units, the Firpo-Possebom p-value is:

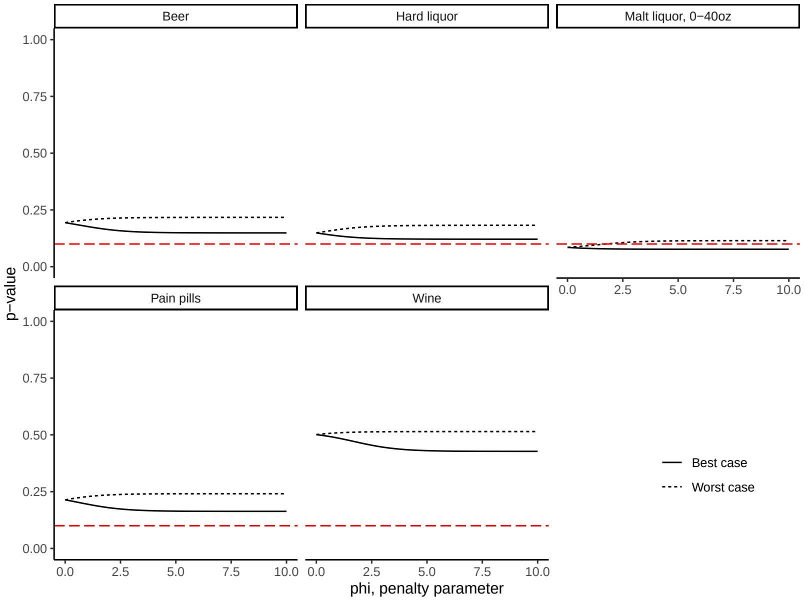
$$p - value_0^{FP} = \frac{1}{J+1} \sum_{j=1}^{J} \omega_j(\phi) 1(|\text{standardized treatment effect}_j| \geq |\text{standardized treatment effect}_0|) \quad (24)$$

In the best case scenario, the omitted variable bias is working against us finding a significant treatment effect. In the worst case scenario, the omitted bias has biased us towards finding a statistically significant treatment effect. For both the best case and worst case, we consider a range of values of ϕ (0 to 10) and report how the estimated p-value changes with ϕ.

We operationalize this robustness check using the following procedure. First we estimate a unit-free measure of pre-treatment fit, Cohen's D for our target unit and for each placebo unit. Second, we compare the each placebo Cohen's D to the target variable Cohen's D, altering that unit's v_j value if the placebo Cohen's D is less than the target Cohen's D; we change v_j to be 1 in the best case scenario and 0 in the worst case scenario. Third, we re-estimate the p-value for the target variable across a range of ϕ values using equation 24.

Figure A1 displays the results from this procedure. Here each panel is for a different target product. Within each panel the black solid line represents the best case scenario, the dashed black line represents the worst case scenario. Recall that when $\phi = 0$, the assumption is uniform (i.e., equal) weights, so the p-value is the same as the last column in Table 1. It is evident from this exercise that our randomization inference based p-values are not sensitive to deviations from the uniform weighting assumption. Even under the worst case scenario with an extremely large penalty parameter, the p-value remains nearly unchanged.

Figure A1: Sensitivity of our randomization inference p-value under non-uniform weighting under the best and worst case scenarios



Note: Each panel is for a different target product and shows how the randomization inference based p-value changes as the relative importance of placebos with a better pre-treatment fit than the target unit are given more (or less) weight than the other placebos. ϕ changes the relative weight. The p-value is calculated as in Equation eq:firpo-possebom and follows Firpo and Possebom (2018). The black solid line represents the best case scenario, the dashed black line represents the worst case scenario. The long-dashed red line indicates $p = 0.10$.

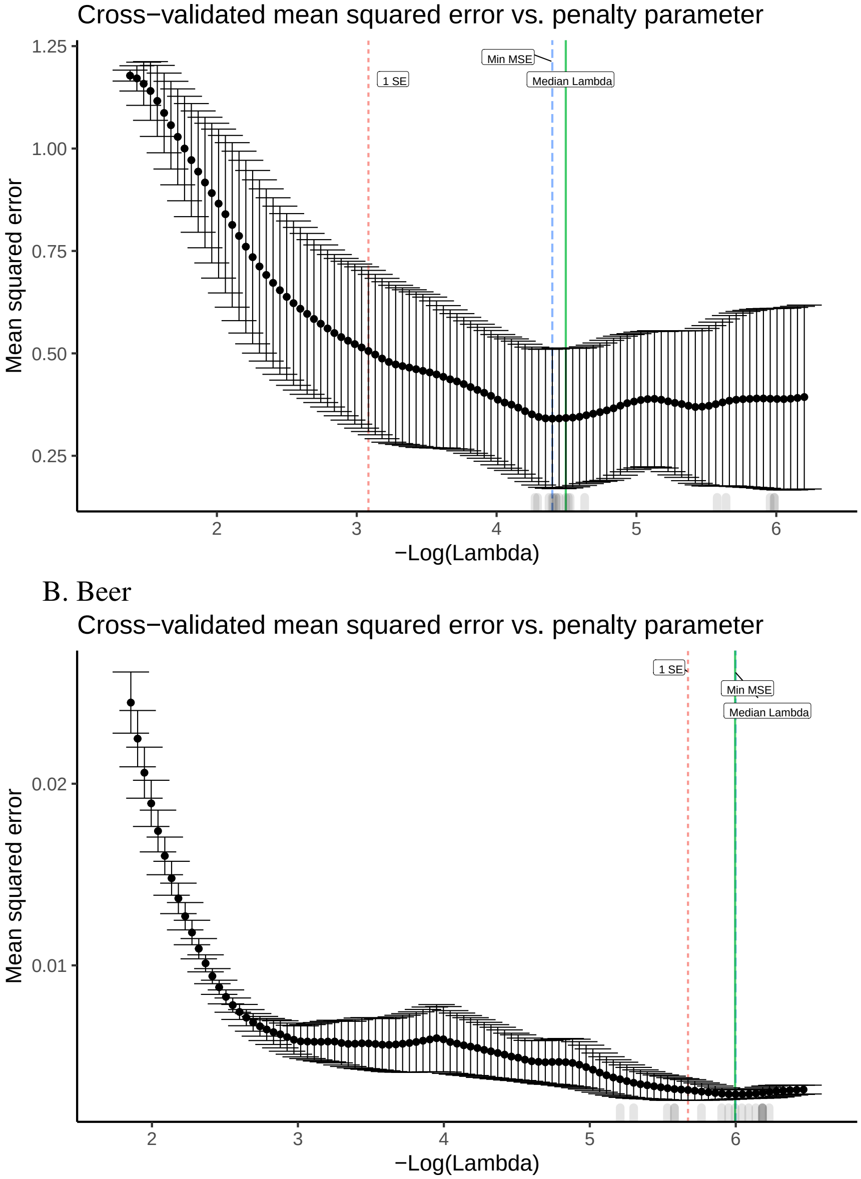
C.2 Using different penalty parameter decision rules:

Table A1: Comparing results across various penalty parameter decision rules

	Median rule	Min rule	+1 S.E. rule
Pain pills			
Cohen's D	0.15	0.14	0.16
Treatment est.	3.27%	2.73%	3.55%
p-value	(0.28)	(0.35)	(0.24)
Beer			
Cohen's D	0.15	0.16	0.18
Treatment est.	-6.82%	-6.26%	-5.39%
p-value	(0.28)	(0.32)	(0.33)
Wine			
Cohen's D	0.12	0.1	0.13
Treatment est.	-4.89%	-3.22%	-8.97%
p-value	(0.62)	(0.74)	(0.39)
Hard liquor			
Cohen's D	0.18	0.13	0.16
Treatment est.	-12.82%	-17.93%	-15.75%
p-value	(0.18)	(0.1)	(0.12)
Malt liquor, 0-40oz.			
Cohen's D	0.22	0.26	0.32
Treatment est.	-41.09%	-25.47%	-23.76%
p-value	(0.1)	(0.23)	(0.24)

Note: See Section 4.3 for more details on each penalty parameter decision rule. See Figure A2 for a visual depiction of this rule for the Beer and Malt Liquor outcomes. For the randomization inference based p-values, we used the same rule as stated in the column. That is, we used the same penalty selection rule as is stated in the column for both the treated and placebo units.

Figure A2: Visual illustration of penalty parameter selection rule for two products.



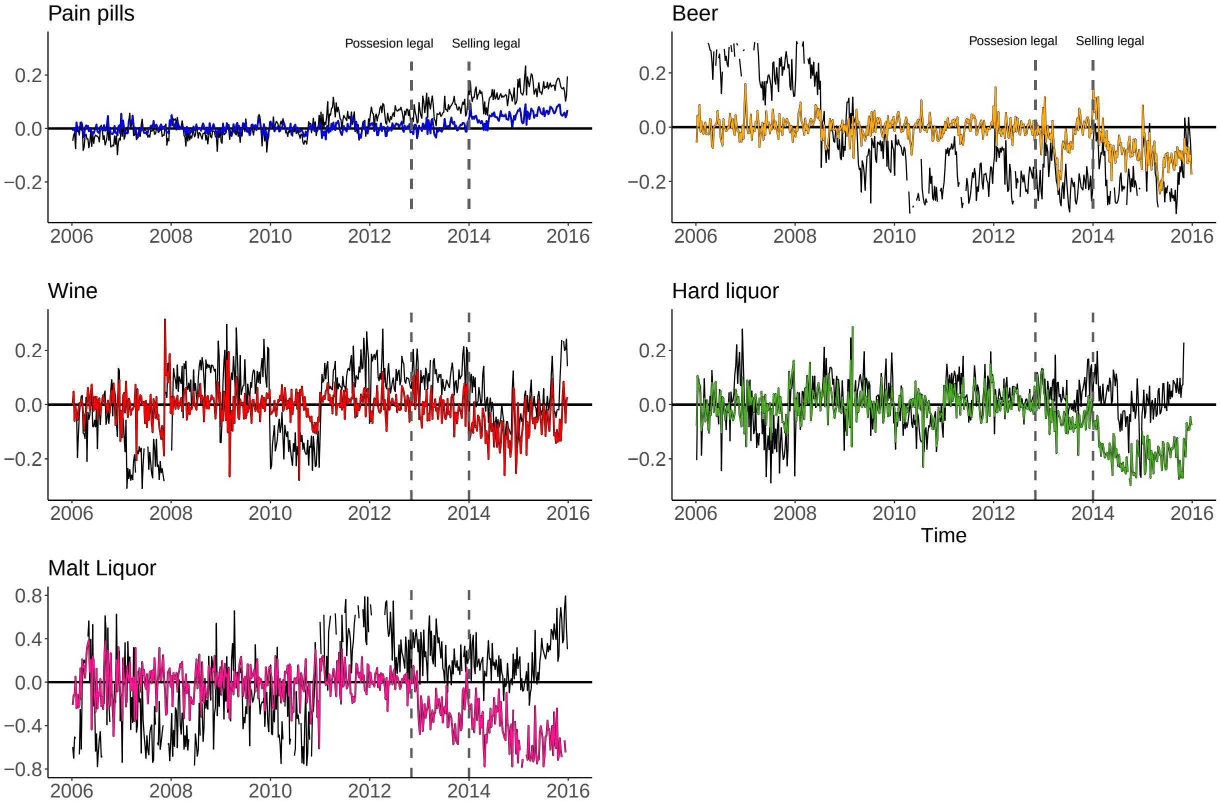
A. Malt liquor

B. Beer

Note: Across the x-axis is the negative natural log of the penalty parameter. Thus larger penalty parameters (and sparser models) are on the left side and the penalty increases as we move along the x-axis to the right. Just above the x-axis are light gray tombstones with some transparency, these indicate a penalty parameter that provided the minimum mean squared error for a particular cross-validation iteration. The median minimum λ is denoted by the solid green vertical line. The black point depict the average mean square error taken across all cross-validation iterations, for each candidate penalty parameter. The brackets around each point denote uncertainty about each mean estimate showing the mean $\pm$ its standard error. The blue dashed line depicts the penalty parameter associated with the lowest average error across all cross-validation iterations. The red dashed line shows the largest λ (most parsimonious model) whose associated average mean squared error is within one standard error of the minimum average mean squared error.

C.3 Comparison to traditional synthetic control method

Figure A3: Difference between actual time series and traditional synthetic control (thin-black) vs difference between actual and SCUL prediction (wide-color)



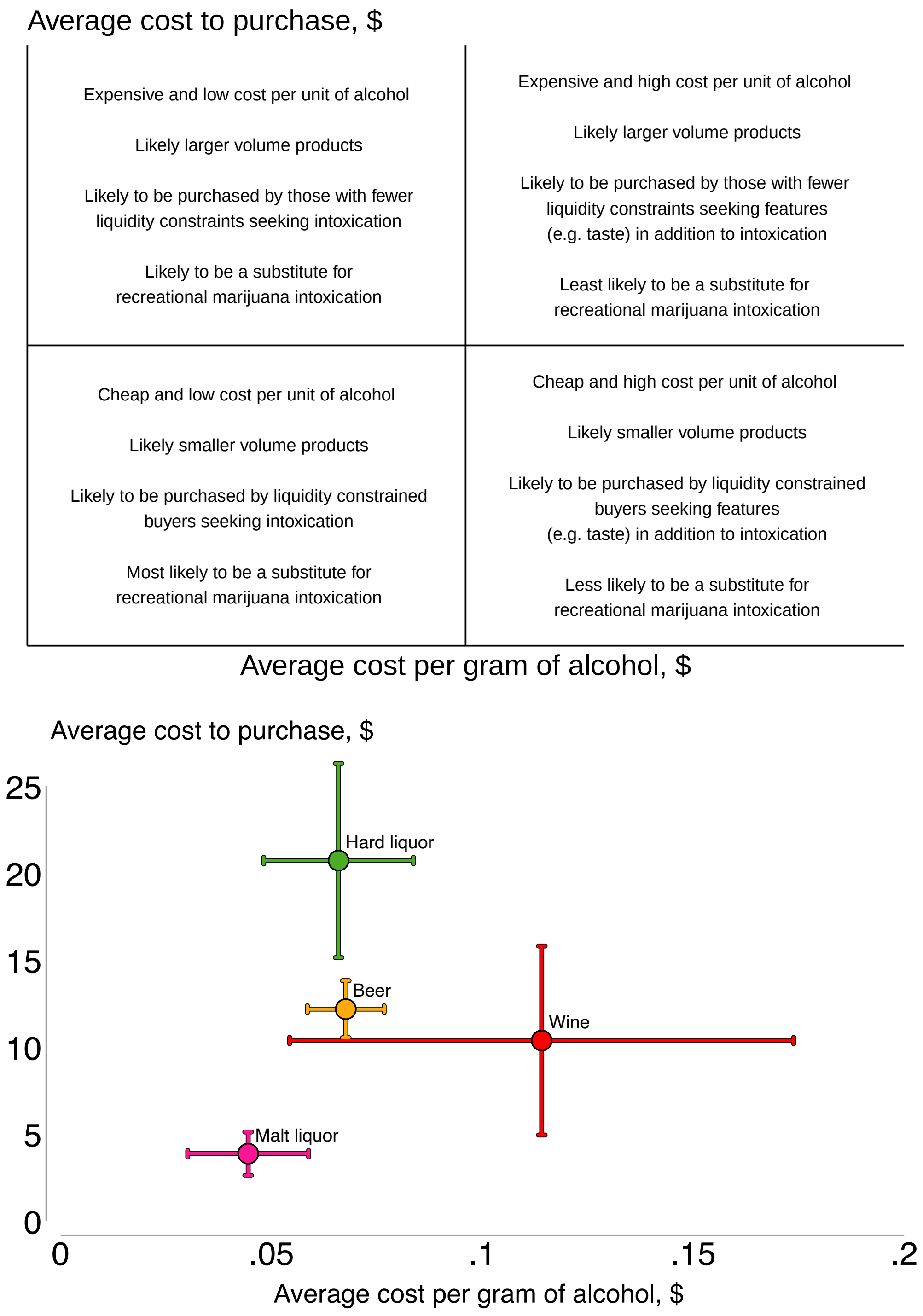
Note: Each panel depicts the results from two synthetic control analyses, one using the traditional approach as in Abadie et al. (2010) (depicted by the thin black line) and another using our SCUL procedure (depicted by the wide colored line). Each line represents the time varying treatment effect estimate, where the post-treatment time period begins just before the beginning of 2013. The SCUL procedure uses the entire high dimensional donor set and as such these treatment effect estimates are the same as those presented in the right column of Figure 5. The traditional synthetic control estimates are constructed using the Synth package in R. In these analyses we include as donors only the sales of the identical product in other states (e.g., the donor pool for pain pills in Colorado is the sales of pain pills in every other state that had no recreational marijuana policy during our sample). We also include six summary statistics to be balanced on: average sales of the target product in the pre-treatment period, state population (in millions), state median income ($10,000), state murder rate per 100,000, the percent of state residents over the age of 18 with at least a high school degree, and state average life expectancy.

Table A2: Comparing results from a traditional synthetic control approach to those from SCUL

	Traditional SCM		SCUL	
	Cohen's D	Treatment effect estimate	Cohen's D	Treatment effect estimate
Pain pills, OTC	0.37	11.39	0.15	3.27
		(0.13)		(0.28)
Beer	1.12	-19.38	0.15	-6.82
		(0.33)		(0.28)
Wine	0.38	4.26	0.18	-4.89
		(0.82)		(0.62)
Hard liquor	0.32	5.08	0.12	-12.82
		(0.87)		(0.18)
Malt liquor, 0-40oz	0.80	0.22	0.22	-41.09
		(0.78)		(0.1)

Note: The traditional synthetic control estimates are constructed using the Synth package in R. In these analyses we include as donors only the sales of the identical product in other states (e.g., the donor pool for pain pills in Colorado is the sales of pain pills in every other state that had no recreational marijuana policy during our sample). We also include six summary statistics to be balanced on: average sales of the target product in the pre-treatment period, state population (in millions), state median income ($10,000), state murder rate per 100,000, the percent of state residents over the age of 18 with at least a high school degree, and state average life expectancy. The placebo analyses are conducted in an analogous manner, where every donor state serves as a placebo. The SCUL procedure uses the entire high dimensional donor set and as such these treatment effect estimates, measures of pre-period fit, and p-values are the same as those presented in the last column of Table 1.

Figure A4: Malt liquor is the most likely alcohol to be purchased for intoxication, making it the most likely substitute for recreational marijuana intoxication.



Expensive and low cost per unit of alcohol Likely larger volume products Likely to be purchased by those with fewer liquidity constraints seeking intoxication Likely to be a substitute for recreational marijuana intoxication	Expensive and high cost per unit of alcohol Likely larger volume products Likely to be purchased by those with fewer liquidity constraints seeking features (e.g. taste) in addition to intoxication Least likely to be a substitute for recreational marijuana intoxication
Cheap and low cost per unit of alcohol Likely smaller volume products Likely to be purchased by liquidity constrained buyers seeking intoxication Most likely to be a substitute for recreational marijuana intoxication	Cheap and high cost per unit of alcohol Likely smaller volume products Likely to be purchased by liquidity constrained buyers seeking features (e.g. taste) in addition to intoxication Less likely to be a substitute for recreational marijuana intoxication

Average cost to purchase, $
Average cost per gram of alcohol, $
Average cost to purchase, $
25
20
15
10
5
0
Hard liquor
Beer
Wine
Malt liquor
0
.05
.1
.15
.2
Average cost per gram of alcohol, $

Note: Top panel is a visual representation of the discussion outlined in Section 5.3. Bottom panel presents data allowing evaluation of which alcohol category is most likely to be purchased for intoxication. Each point displays the average cost to purchase a product against the average cost per gram of alcohol contained in the product. Products composing these averages are taken from a random sample (weighted by annual expenditures) of alcohol products observed in the Nielsen retail scanner data. For each sampled product, authors collected data on alcohol by volume. This was combined with price and volume data from Nielsen to create a measure for average cost per gram of alcohol $= \frac{\text{Average cost to purchase}}{(ABV \times volume)}$. 95% confidence intervals for the mean of each attribute are reported by brackets.