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

Alex Hollingsworth and Coady Wing*

May 3, 2020

**Abstract**

We describe identification assumptions underlying synthetic control studies and offer recommendations for key—and normally ad hoc—implementation decisions, focusing on model selection; model fit; cross-validation; and decision rules for inference. We outline how to implement a Synthetic Control Using Lasso (SCUL). The method—available as an R package—allows for a high-dimensional donor pool; automates model selection; includes donors from a wide range of variable types; and permits both extrapolation and negative weights. In an application, we employ our recommendations and the SCUL strategy to 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; machine learning; marijuana legalization; high-dimensional data

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



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; in which, pseudo-treatment effects are estimated for many untreated placebo units, and the distribution of pseudo-estimates represents the null distribution of no treatment effect.

The method is very popular, but can be somewhat opaque in application. Synthetic control studies are usually framed as quasi-experimental studies and are included in the modern canon of design-based empirical strategies. A hallmark of the design-based movement in empirical social science is careful discussion about the link between identifying assumptions and the social or economic outcome of interest. One concern with the synthetic control method is that the scientific community does not always have a clear understanding of the technical and practical conditions under which the method is apt to work well. Another is that implementing synthetic control studies involves a series of operational decisions that are rarely well documented or explained. Moreover, 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). A goal of this paper is to make the conceptual and practical challenges associated with synthetic controls more vivid to other applied researchers. As such, we highlight identification assumptions and offer recommendations applicable to key decisions that arise in any synthetic control study.

It is well known that difference-in-difference studies depend on the common trends assumption, and applied studies are full of efforts to assess the credibility of this assumption in specific situations (Wing et al., 2018). Likewise, it is standard knowledge that instrumental variables must satisfy detailed versions of assumptions about relevance, monotonicity, independence, and exclusion. While these restrictions may be empirically untestable, well-executed studies explain how the abstract assumptions apply in a given situation and often give examples of behaviors that would violate the core assumptions. Synthetic control studies generally do not follow this practice, often omitting both description and tests of the core identifying assumptions. We outline two simple

identification assumptions required for a synthetic control design to identify causal treatment parameters: (1) conditional independence of treatment exposure and potential outcomes after matching on an underlying factor structure, and (2) no dormant factors during the pre-treatment period. While neither of these assumptions is directly testable, we offer perspectives and strategies that may help in interpreting the validity of such assumptions in applied work.

A useful way to think about synthetic controls is as a procedure that attempts to match donor series to target series based on the unobserved factors that determine the data generating process in the pre-treatment period. If treatment affects the data generating process for the target unit, then deviations from the matched donor series following treatment may represent a treatment effect. When framed in this manner, identification assumptions and strategies for model selection are more salient. One contribution of this paper is the insight that including donor variables of different types than the target series may help better identify the underlying factors driving pre-treatment data generation. For example, if one were trying to predict cigarette sales in a particular state, out-of-state cigar sales, out-of-state e-cigarette use, and out-of-state cigarette prices would all be reasonable donor variable types since each could help identify unobserved factors that affect in-state cigarette demand. As such, 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, as is common practice. However, increasing the size of the donor pool is not without issue; large donor pools raise overfitting concerns and an identification problem arises when the number of donor series exceeds the number of observations. For this reason, we use a method to construct cross-validated donor weights that avoids overfitting and can handle a high-dimensional donor pool.

Most applied synthetic control studies choose donor weights using the approach developed in the seminal work of Abadie et al. (2010). This approach imposes several ancillary restrictions that are not required by the identifying assumptions. For example, the classic method does not allow for a donor pool larger than the number of time periods, and it requires that synthetic groups be constructed using non-negative weights that sum to one. The weight restriction forces the synthetic control group to be a convex combination of donor groups. This prevents extrapolation beyond the common support of the donor pool (i.e., convex hull). While this can be a desirable property, we demonstrate simple and empirically relevant situations in which this restriction prevents the synthetic control method from selecting the ideal donor series, and other situations in which it prevents the synthetic control from incorporating information from donors that are reliably negatively correlated with the target series.

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). In a similar vein, we use a method called Synthetic Control Using Lasso (SCUL) to construct donor weights. This method is a flexible, data-driven way to construct synthetic control groups. It

relies on lasso regressions, which are popular in the machine-learning literature, and favor weights that predict well out of sample. In general, the approach allows for a high-dimensional donor pool that may be larger than the number of time periods, extrapolation from the donor pool, counter-cyclical weights, and the same model selection procedure to be used for target and placebo series. We discuss the advantages and disadvantages of these features throughout the paper. We view the combination of this statistical approach and following our recommendations outlined below as the SCUL procedure

Regardless of how the weights are constructed, every synthetic control study must make a number of decisions related to the donor pool, model selection, and decision rules for statistical inference. These decisions are often ad hoc and undocumented. To serve as a guide for other applied research, we catalogue these decisions and offer accompanying recommendations and the reasons behind each recommendation. For example, we consider the conditions under which a unit should be discarded in advance of data analysis. We recommend discarding any series (treated or placebo) when a pre-specified, unit-free threshold indicating poor model fit is exceeded. The current status quo is to eliminate only candidate placebo series when the synthetic prediction offers a worse pre-treatment fit than that of the synthetic estimate for the treated series. When using rank-based p-values, the status quo approach makes it more likely that the estimated treatment effect will be an outlier in the placebo distribution, overstating statistical significance. Moreover, when root mean squared error is used to measure fit, candidate placebo series that have larger nominal variance, but not necessarily worse fit, are penalized. We propose a standardized Cohen's-D-based criterion for model fit that is unit free. We also show how changing the model-fit inclusion threshold affects the precision of synthetic control estimates by changing the shape of the sampling distribution.

In all, we recommend: 1) using the same model selection procedure for both target and placebo products; 2) using a unit-free measure to evaluate model fit; 3) discarding both potential target series and placebo series that do poorly on a pre-specified threshold of model fit; 4) incorporating a rolling-origin cross-validation procedure to determine optimal weights, which helps guard against over-fitting and likely improves out-of-sample prediction; 5) reporting synthetic control weights as the average contribution to the synthetic prediction rather than as a numeric coefficient to better compare the influence of variables with different magnitudes; 6) using a unit-free measure of the treatment effect estimate to compare estimated treatment effects to the placebo distribution; and 7) reporting the minimum treatment effect size for a given significance level that the placebo distribution used for inference would consider statistically different than zero.

This paper also makes an empirical contribution to the literature on the effects of marijuana legalization. Specifically, 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 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. 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. The strongest evidence suggests that recreational marijuana laws reduce consumption of alcoholic beverages that have both lower total cost to purchase and lower cost per gram of alcohol (i.e., that are cheap and strong).

Our paper proceeds as follows. Section 2 provides conceptual motivation for synthetic controls and discusses identification assumptions. Section 3 outlines the procedure used in traditional synthetic controls as well as the lasso extension used in our analysis. Section 4 introduces our construction of the average treatment effect, discusses statistical inference, and outlines other recommendations and quality control issues. Section 5 presents our application, including descriptions of marijuana policy and related research, the retail scanner data used in our analysis, and our results. Section 6 concludes.

# 2 Model

The synthetic control method is a way of analyzing data using a Comparative Interrupted Time Series (CITS) research design (Shadish et al., 2002). In a CITS, a time series of outcomes is observed for multiple units. None of the units is exposed to treatment in the early time periods, and one or more units is exposed to treatment after a known "interruption" date. The loose idea is that we may be able to infer something about the treatment effect by comparing the behavior of the treated units to that of the comparison units, before and after the former are exposed to treatment. The most common way to analyze data from a CITS is to perform a difference in difference (DID) comparison. In practice, researchers often go further and estimate the DID comparison using a two-way fixed effects regression model. These DID approaches invoke strong assumptions about common trends across groups. The synthetic control methodology is an alternative way to analyze data from a CITS and is the most valuable in situations where the assumptions required for the

DID and two-way fixed effects approaches are not credible.

## 2.1 Notation

Use $s = 0 \ldots S$ to index the units of analysis. In our application the units are product-state sales, such as ounces of wine sold in Indiana or packages of razor blades sold in Texas. In other settings, the units might be the same outcome or product across geographical territories (e.g., states or countries). In general, each $s \in S$ is either a donor unit or treated unit. For simplicity, suppose there is a single treated unit, denoted by $s = 0$, and a number of untreated units, each denoted by an $s > 0$. Let $t = 1 \ldots T$ index time periods, which are weeks in our application. Next, assume that treatment exposure occurs in period $T_0 + 1$. Finally, set $D_{st} = 1[t > T_0] \times 1[s = 0]$ to be a binary variable equal to 1 if unit $s$ is exposed to treatment in period $t$.

Let $y(0)_{st}$ and $y(1)_{st}$ represent potential outcomes that record the outcome of unit $s$ in period $t$ under the control and treatment conditions. In our application, $y(0)_{st}$ is the quantity sold in product-state $s$ during period $t$ in the absence of a recreational marijuana law, and $y(1)_{st}$ is the quantity sold in the same product-state under the recreational marijuana law. The difference between the two is $\beta_{st} = y(1)_{st} - y(0)_{st}$, which is the causal effect of treatment on unit $s$ at time $t$. The realized outcome is $y_{st} = y(0)_{st} + D_{st}\beta_{st}$. However, this introduces a natural identification problem because untreated outcomes are not observable for the treated unit following exposure to treatment. That is, after period $T_0$, we are only able to observe values of $y(1)_{0t}$ for the treated unit.

The basic goal of the synthetic control strategy is to estimate values of $y(0)_{0t}$ in the post-treatment time periods. With those counterfactual estimates in hand, it is possible to estimate $\beta_{st}$ for the post-treatment periods $t > T_0$. Often the focus will be on multi-period average treatment effects rather period-specific estimates. For example, the average treatment effect on the treated unit (ATT) over the entire post-treatment period is $ATT(T_0 + 1, T) = \frac{1}{(T - T_0 - 1)} \sum_{t=T_0+1}^{T} \beta_{0t}$.

## 2.2 What is a synthetic control group?

A synthetic control is a weighted average of outcomes from a collection of candidate untreated control units. Suppose that $x_t = (y_{1t}, \ldots, y_{St})$ is the $1 \times S$ vector of the outcomes that prevailed in each of the candidate comparison units at time $t$. Let $\omega = (\pi_1, \ldots, \pi_S)^T$ be a $S \times 1$ vector of weights. A synthetic control group for the outcome of the treated unit is:

$$\begin{aligned} y_t^* &= \sum_{s=1}^{S} y_{st}\pi_s \\ y_t^* &= x_t\omega \end{aligned} \tag{1}$$

The definition a synthetic control group is straightforward. A basic problem is that there are an infinite number of ways to form a synthetic comparison unit from a set of candidate comparison units. For example, when setting $\pi_s = \frac{1}{S}$ for each $s$, the synthetic control is simply the average outcome in the donor pool. Weighting each donor unit equally is an arbitrary choice, and in most settings would likely produce a poor estimate of the counterfactual time series of interest. Assigning a smaller weight to some donor units and a larger weight to other units might provide a better estimate of the counterfactual time series. But on what basis should we assign different weights to different donor units? The heart of any synthetic control study is the procedure for determining exactly what weights to use. However, before we discuss different strategies for assigning weights, we first examine the conditions under which a proposed synthetic control will provide a consistent estimate of the counterfactual time series.

## 2.3 What are the identifying assumptions?

Synthetic control groups are essentially an extrapolation based on the assumption that cross-sectional correlations between the treated and donor units during the pre-treatment period would have remained the same in the post-treatment time period had it not been for the treatment. As such, synthetic control studies assume that a synthetic control group closely matching the treated time series during pre-treatment would continue to be a close match for the untreated potential outcome of the treated unit post-treatment. But under what conditions is this claim likely to hold?

To help answer this question, we follow Abadie et al. (2010) and specify an interactive fixed effects model of the potential outcomes. The model is very flexible and allows for situations with complex unit-specific time trends that would create problems for strategies based on difference in difference and two-way fixed effects regressions. We use the model to define and explain the two key assumptions required for the synthetic control methodology to identify causal effects. The first assumption is a conditional independence assumption familiar from the literature on propensity score matching designs. The second assumption is that there are no "dormant factors" in the pre-period that become active during the post-period.

## 2.4 Interactive fixed effects

In an interactive fixed effects model, the untreated potential outcomes are generated by:

$$y(0)_{st} = \delta_t \alpha_s + \varepsilon_{st} \tag{2}$$

In the model, $\delta_t$ represents a $1 \times K$ vector of unmeasured variables, which may change over time but do not vary across groups. $\alpha_s$ is a $K \times 1$ vector of group-specific coefficients on the unmeasured variables in $\delta_t$. The coefficients may vary across groups, but do not change over time. $\varepsilon_{st}$ is a residual error term that represent exogenous sources of variation in the untreated outcome.

The model provides a flexible way of thinking about how units might evolve according to a common trend, and situations under which a common trend assumption might fail. For example, in the special case where $\delta_t = [1\ \gamma_t]$ and $\alpha_s = [\theta_s\ 1]^T$, the model simplifies to $y_{st}(0) = \theta_s + \gamma_t + \varepsilon_{st}$, which is the familiar two-way fixed effects model. More generally, the model allows for time trends to vary across units when units have different coefficients on the vector of common factors.

The goal of a synthetic control study is closely related to studies based on covariate matching. If the outcome of the treated unit is $y(0)_{0t} = \delta_t \alpha_0 + \varepsilon_{0t}$, then the task is to scour the donor pool for comparison units with values of $\alpha_s$ that are a close match for $\alpha_0$. Direct matching on the $\alpha$-parameters is not feasible because $\alpha_s$ is unobserved in practice. But the model suggests that two time series with a close match on the history of $y(0)_{st}$must be a close match on $\alpha_s$.

### 2.4.1 Conditional independence assumption

The first core assumption in synthetic control studies is that treatment exposure is statistically independent of the untreated potential outcome, conditional on the unit-specific coefficients, $\alpha_s$. This assumption is the same one that motivates studies based on propensity score matching and regression adjustment. Formally, the conditional independence assumption (CIA) implies that:

$$y_{st}(0) \perp\!\!\!\perp D_{st} \mid \alpha_s \tag{3}$$

Conditional mean independence is a slightly stronger form of the CIA that may be easier to understand in many applications. Conditional mean independence occurs when the population average potential outcome does not depend on treatment status among units with the same $\alpha_s$, that is, when $E(y_{st}(0)|\alpha_s, D_{st} = 1) = E(y_{st}(0)|\alpha_s, D_{st} = 0)$. The CIA can also be expressed as a restriction on the residual error term: $E[\varepsilon_{st}|\alpha_s, D_{st} = 1] = E[\varepsilon_{st}|\alpha_s, D_{st} = 0]$.

The CIA implies that if we can observe the vector $\alpha_s$ for each unit in the study population, then we can build a comparison group for the treated unit by selecting donor units with $\alpha_s$ values that are the same as $\alpha_0$. It is a strong assumption. It means that among units with the same value as the $\alpha$ vector, treatment is as good as randomly assigned, and that the two units should follow the same time trends, with differences arising only because of idiosyncratic and independent random shocks represented by $\varepsilon_{st}$. An implication is that the only reason why two units should follow systematically different trends is because they have different values for $\alpha_s$; it is this implication that motivates forming matches on the pre-treatment outcome history.

What are the implications of the CIA for applied researchers? One is that it may be useful to ask whether a structure like the interactive fixed effects model, in which treatment is exogenous conditional on the $\alpha$-coefficients, is a plausible way of thinking about the treated unit and the donor pool. The assumption may be more plausible in settings where the donor pool consists of outcomes that likely respond to a similar collection of time-varying common factors, even if the responses to some of these factors are muted for some units and amplified for others. If treatment is correlated with these factors, or if the procedure matches on some feature other than the underlying factor structure (e.g. statistical noise), then a synthetic control estimate risks being unidentified.

The prospect of overfitting is an important practical concern for identification. The idea is that by forming matches on the observed outcome history, researchers can implicitly construct a comparison group that matches the treated unit on the structural $\alpha$-coefficients. However, the observed values of $y(0)_{st}$ are generated by both the $\delta_t \alpha_s$ and the idiosyncratic errors, $\varepsilon_{st}$. Synthetic control methods are akin to matching on covariates that are measured with error (Ben-Michael et al., 2018). Concerns of overfitting are most severe for small samples, short pre-treatment time periods, and settings where the error variance of the model is large relative to its structural variance. In general, synthetic control studies should be based on long pre-periods, when the relative share of variance explained by $\varepsilon_{st}$ is large.

The CIA is not directly testable. To assess the credibility of the CIA and to guard against concerns of overfitting, we recommend several implementation strategies and supplementary analyses. First, researchers should use a cross-validation procedure when determining synthetic control weights to guard against overfitting. We describe one possible approach in more detail in Section 3.2.1. We also recommend performing an event-study analysis on the pre-treatment difference in fit. The difference should lack pre-trends and be centered around zero. Finally, match quality should be based on a pre-specified and unit-free measure of fit in the pre-treatment period.

### 2.4.2 No dormant factors assumption

The interactive fixed effects model implies that outcomes for each unit respond to changes in $\delta_t$, a $1 \times K$ vector of time-varying covariates. Researchers usually know neither the identity nor the number of variables contained in $\delta_t$. Nevertheless, the logic of the model is that the combination of these common time-varying factors with unit-specific coefficients is responsible for the trend of each unit over time.

Let $\delta_P$ be the $T_0 \times K$ matrix of pre-treatment values of the unmeasured common factors. Here, the $t^{th}$ row of $\delta_P$ is equal to $\delta_t$, and there is a row for each pre-treatment time period from $t = 1...T_0$. In their analysis of the statistical properties of the synthetic control design, Abadie et al. (2010) require the assumption that $\delta_P^T \delta_P$ is an invertable matrix. This restriction may seem somewhat opaque, but it is the second key assumption required for causal identification in synthetic con-

trol studies. Roughly speaking, it means that none of the time-varying common factors from the structural model are perfectly multicollinear during the pre-treatment time period.

One interpretation is that synthetic control designs work as long as all elements of $\delta_t$ that vary independently during the post-treatment time period also vary independently during the pre-treatment period. That is, none of the time-varying common factors are "dormant" during the pre-treatment period and then "wake up" during the post-treatment period. The intuition for the no-dormant-factors assumption is simple: the variation in pre-treatment outcomes only contains information about $\alpha$, the coefficients on time-varying factors that actually change during the pre-period. If a time-varying factor is dormant in the pre-period, then it will not be possible to separately identify two units with different coefficients on that factor. If that same factor begins to vary during the post-period, then the two groups may begin to diverge for reasons that have nothing to do with the effect of the treatment.

The no-dormant-factors assumption is not directly testable. Because of this it is important to understand the kinds of conceptual arguments and supplementary analysis researchers can use to assess the credibility of this assumption. One concern is that some units may adopt new policies or experience novel/unique economic or social events. In principle, these time-varying shocks may always have been part of the data-generating process for each unit. That is, each unit has always had a coefficient for a given shock, such as the adoption of "new policy x." However, when a unit has never previously experienced a particular shock, there is no information in the historical record to uncover how that unit responds to the change. Thus, applied researchers should be on the lookout for events that occur in the post-treatment period that do not have much precedent in the pre-treatment data.

More broadly, it may be the case that the vector of time-varying factors in the structural model may consist of high-frequency factors that change over short periods of time and low-frequency factors that change very rarely or very slowly. Low-frequency factors may be a threat to the no-dormant-factors assumption if they happen to change during the post-period and not in the pre-period. For example, a presidential election season might be a low-frequency factor in a study based on weekly time series data observed for a three-year pre-period and a three-year post-period. If there is an election year in the post-period and not in the pre-period and the outcome responds to the electoral cycle, then a longer pre-period might be required.

In addition to careful reasoning about the likely elements of the time-varying factors in the structural models, we recommend using only donor and placebo units that seem to plausibly depend on the same collection of common factors and where no known violations of the dormant factors have occurred. For example, in our setting we exclude all products from states, such as Oregon, that legalized recreational marijuana in the period after Colorado's legalization.

# 3 Choosing synthetic control groups

## 3.1 Classical synthetic control weights

Abadie et al. (2010) (ADH) developed the original and mostly widely used method of constructing a synthetic control group.[1] 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 $x_t = (y_{1t}, \ldots, y_{50t})$ would be the donor pool of candidate control outcomes in period $t$. As in Equation 1, a synthetic control group is formed by applying a set of fixed weights to the donor pool.

This method uses two types of pre-treatment data to construct the weights for each donor unit. The first type is the set of time series for the treated unit and the donor pool, $y_s$ for $s \in 0 \ldots S$. The second type is a set of statistics of interest on which the researcher desires balance between the treated and synthetic unit. These 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. In practice, the fit of the synthetic unit compared to the treated unit usually receives the majority of the weight.

These statistics of interest for the treated unit are denoted by $Z_0$, which is an $L \times 1$ vector when there are $L$ statistics of interest. In the original ADH example, seven such statistics were included, such as average log GDP per capita and cigarette prices. 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 $Z_0 - Z_C\omega$ 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, $\omega$. ADH summarize the vector of differences with a single summary discrepancy score, $H = \sqrt{(Z_0 - Z_C\omega)^T V (Z_0 - Z_C\omega)}$, where $V$ is the $M \times M$ matrix of importance weights. 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 $\omega$ (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. With donor weights

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

in hand, the synthetic control is formed as in Equation 1.

ADH propose choosing $V$ to minimize the mean square prediction error of the synthetic control during the pre-treatment period. Suppose that $\omega(V)$ is the $H-minimizing$ vector of donor weights when the importance weight matrix is set to $V$. Then the classical synthetic control donor weights are:

$$\widehat{\omega}_{Synth} = \omega(V^*) = arg\ min_{\omega} \left( \sum_{t=1}^{T_0} (y_{0t} - x_t \omega)^2 \right). \tag{4}$$

This method is the standard way of choosing synthetic control group weights in applied research. 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?

## 3.2 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:

$$\widehat{\omega}_{OLS} = arg\ min_{\omega} \left( \sum_{t=1}^{T_0} (y_{0t} - x_t \omega)^2 \right) \tag{5}$$

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 $t = 1...T_0$ 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.

Although it is familiar and intuitive, the OLS method may not be ideal for choosing synthetic control weights. It 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:

$$\widehat{\omega}_{lasso} = arg\ min_{\omega} \left( \sum_{t=1}^{T_0} (y_{0t} - x_t \omega)^2 + \lambda |\omega|_1 \right) \tag{6}$$

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, $|\omega|_1$ 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. In addition, 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.

### 3.2.1 Cross-validation

A key choice parameter in the lasso regression method is the penalty parameter, which is represented by $\lambda$ in Equation 6. As $\lambda$ increases, each weight in $\widehat{\omega}_{lasso}$ 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 $\lambda$ in between these extremes will result in a different set of unique weights. For each lasso regression, a number of $\lambda$ choices are considered in a grid from zero to the smallest value of $\lambda$, which forces every weight to be zero.

If the goal of the regression is to maximize in-sample fit, then the $\lambda$ 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 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 our

optimal lambda by using rolling-origin cross-validation, a procedure that rewards out-of-sample prediction and minimizes issues related to auto-correlation.[2]

Cross-validation is a simple procedure where a dataset is partitioned into multiple subsets that include training data and test data; multiple analyses are performed on the training data; and the optimal analysis is determined using the test data. In our setting, lasso regressions across a grid of penalty parameters are performed for each subset of training data. The series of optimal weights is stored for each candidate penalty parameter. The test data are then used to evaluate which set of weights (i.e. which penalty parameter) produces the best out-of-sample prediction. Importantly, all data used in the cross-validation procedure (i.e., both the training and test data) must be from the pre-treatment period. This is because our goal is to create a synthetic time series that represents the counterfactual as if no treatment had occurred. In principle, a similar cross-validation procedure could be used to select optimal donor weights in the traditional ADH synthetic control method.

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.[3] 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 $k = 1 \ldots K$ 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.[4] We present a visual example of this procedure in Figure 1.

We select the median penalty parameter from the set of optimal $\lambda$ penalty parameters that

[2]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).

[3]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.

[4]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.

minimize the mean squared error for the test data in each cross-validation run. 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, but the specific penalty is allowed to vary across each outcome variable. Once we have chosen the $\lambda$ 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.

## 3.3 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. Suppose, for example, that there are two donor series, A and B, observed in two periods, 1 and 2, and each unit receives a weight equal to $\frac{1}{2}$. The donor values for the first time period are $y_{A,1} = 10$ and $y_{B,1} = 1$. The donor values for the second time period are $y_{A,2} = 1$ and $y_{B,2} = 10$. The resulting synthetic prediction is $y_1^* = 5.5$ in period 1 and $y_2^* = 5.5$ in period 2. In period 1, unit A represents $100 \times \frac{\frac{1}{2} \times 10}{5.5} = 91\%$ of the synthetic prediction. In period 2, unit B represents 91% of the synthetic prediction. Despite each donor weight being 50%, neither unit contributes 50% to the synthetic unit in either period.

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.

## 3.4 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. Cavallo et al. (2013) go further by limiting inference to placebos with an MSPE that is as good (or better) than the MSPE 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. The current standard in the literature is to measure correspondence using the mean square prediction error (MSPE) across the pre-treatment periods. Unfortunately, the units of an MSPE depend on the specific outcome variable and sample of data under analysis. Comparing an MSPE across very different outcome variables is uninformative; 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.

## 3.5 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 difference 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 $\alpha$. 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.[5] 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

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

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 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 $ATT(T_0+1,T)_{p0} = \frac{1}{(T-T_0-1)} \sum_{t=T_0+1}^{T} (y_{st} - y_{st}^*)$. 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 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 5.5, 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 allows 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 difference 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. 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, 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.[6] In contrast, wine and beer may help satisfy the demand for social drinking or may have other desirable product attributes beyond low-cost 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

[6]The top panel of Figure A1 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.

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.[7] 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

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

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 bulbs, liquid soap, cigarettes, bar soap, bread, and butter.[8] 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 placebo products 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. Panel A of Figure 5 displays each target unit time series and the synthetic control group time series on a separate graph with its own scale. In each of the graphs in panel A, actual sales for the target unit are shown as a thin black line. The line clearly demonstrates the substantial seasonality and other within-product sales variation.

## 5.5 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.

### 5.5.1 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

[8] 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.

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.

### 5.5.2 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 units. 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, even 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. Recall that 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 market factors 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.

### 5.5.3 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 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. A key difference in this analysis 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. We 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 develop 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 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.

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

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.

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.

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*.

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.

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.

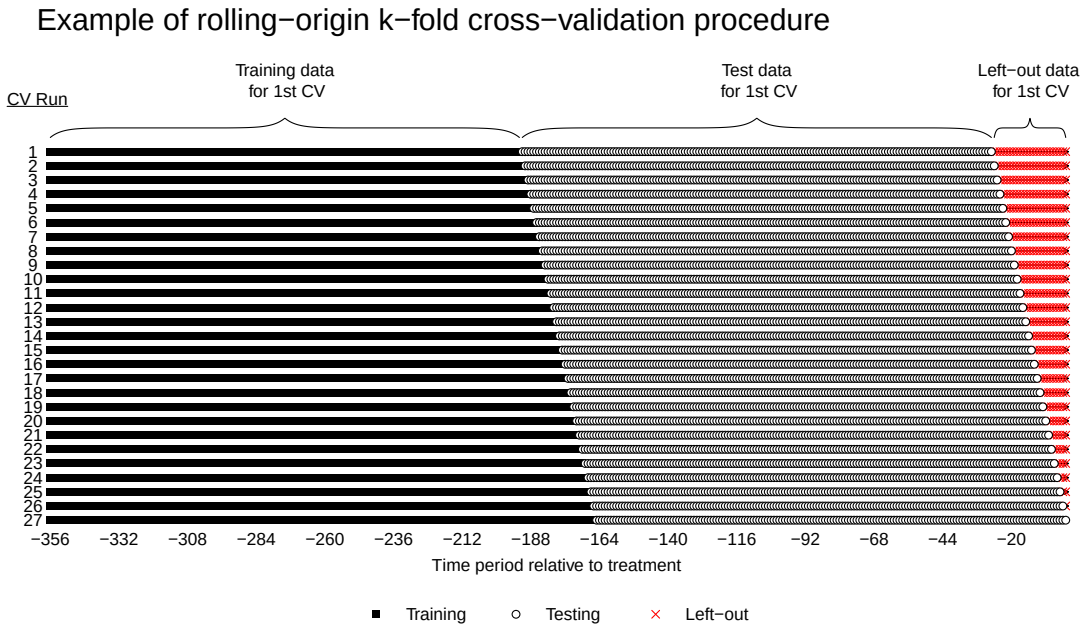
Hollingsworth, Alex, Coady Wing, and Ashley Bradford (2020) "Comparative Effects of Recreational and Medical Marijuana Laws On Drug Use Among Adults and Adolescents," *Working Paper*.

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

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.
Shadish, William R, Thomas D Cook, and Donald T Campbell (2002) "Experimental and quasi-experimental designs for generalized causal inference."
Shi, Yuyan (2017) "Medical marijuana policies and hospitalizations related to marijuana and opioid pain reliever," *Drug and Alcohol Dependence*, Vol. 173, pp. 144–150.
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.
Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez (2018) "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research," *Annual Review of Public Health*, Vol. 39, No. 1, pp. 453–469.
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 3.2.1. 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.

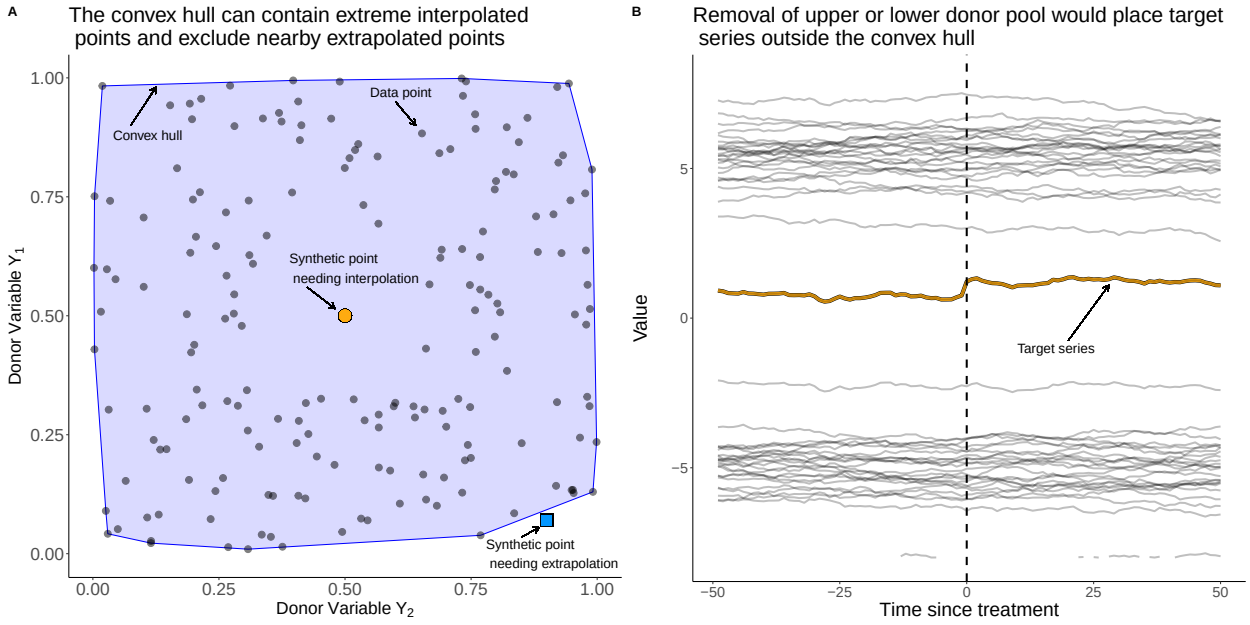
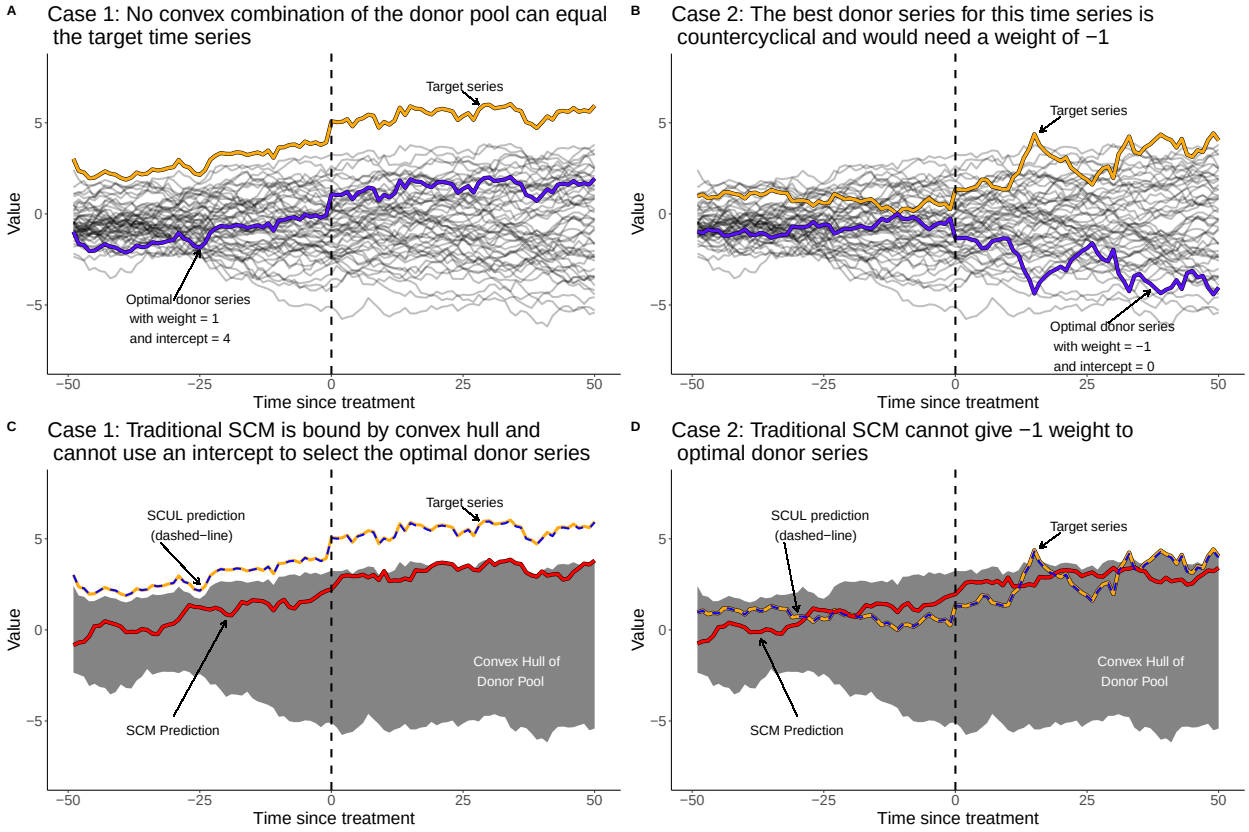
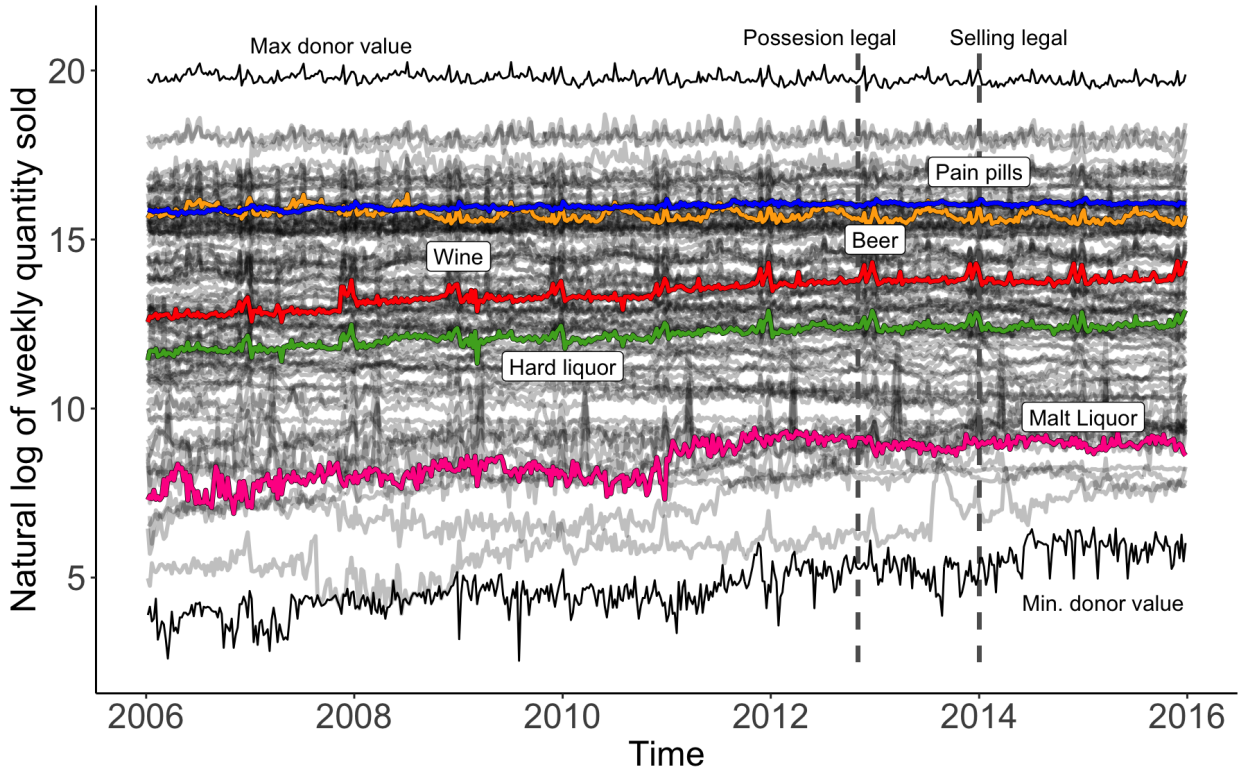


Figure 3: Restrictions on weights prevent the traditional synthetic control methods (SCM) from selecting the optimal donor series in some cases. The synthetic control using lasso (SCUL) procedure preforms well in these settings.



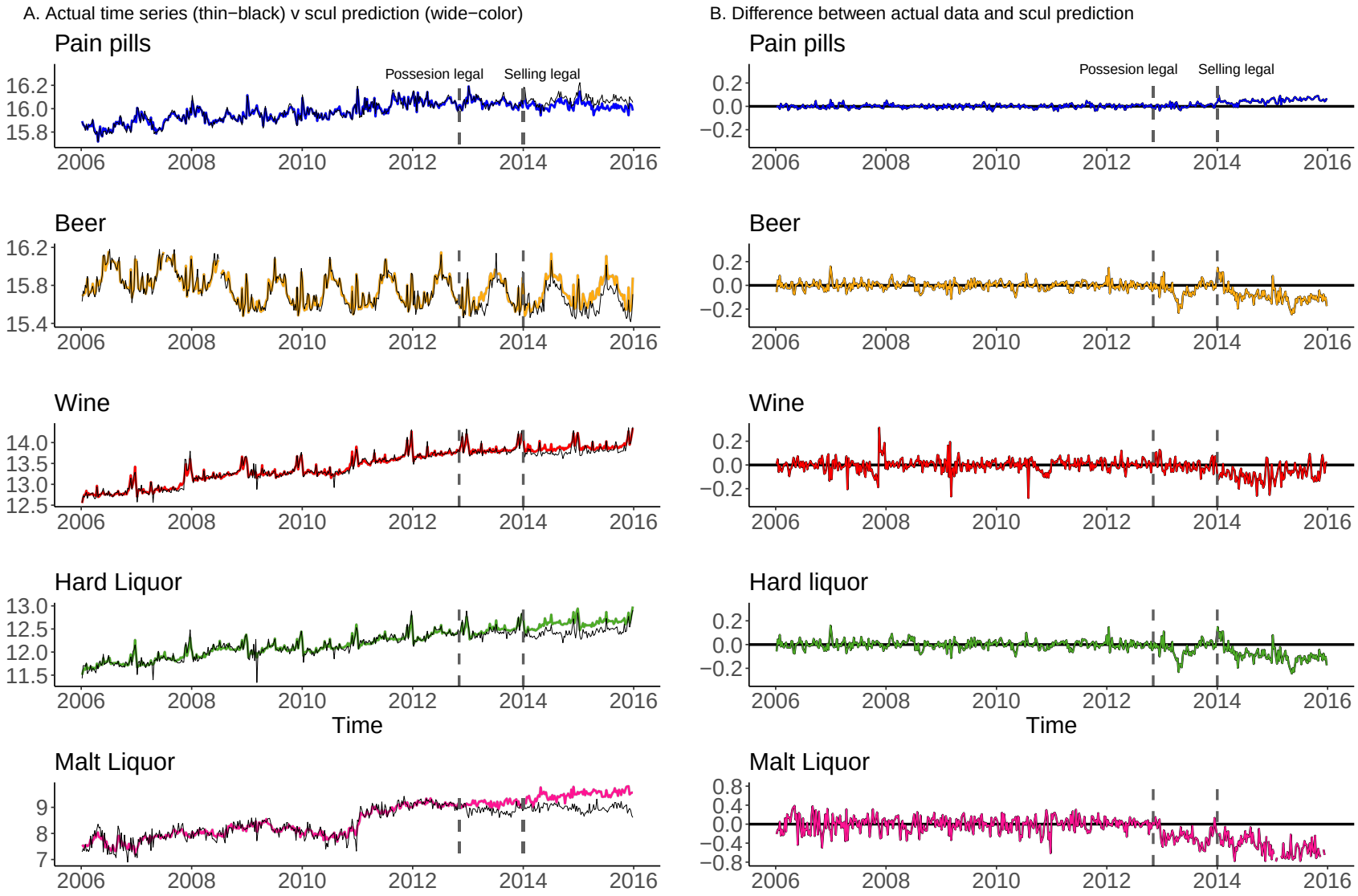

*Note:* In each case a perfect donor series exists for the target series. All other donor series are unrelated to the target series. In case 1, the target series lies outside of the convex hull of the donor pool with the optimal donor series shifted down by four units. In case 2, the target series has 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.



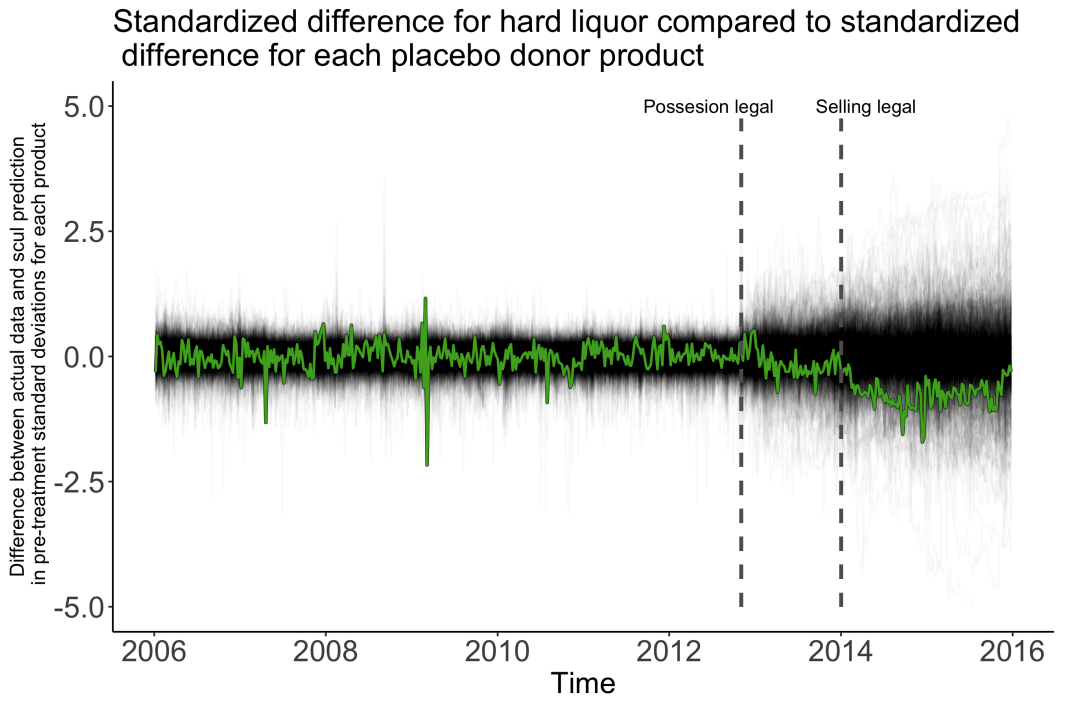

*Note:* All time series in this figure are the natural log of weekly sales for different state-products. The five target products from Colorado we study are labeled and in color. A random sample of the donor pool—sales of products from states other than Colorado—are displayed with transparency. Darker 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.

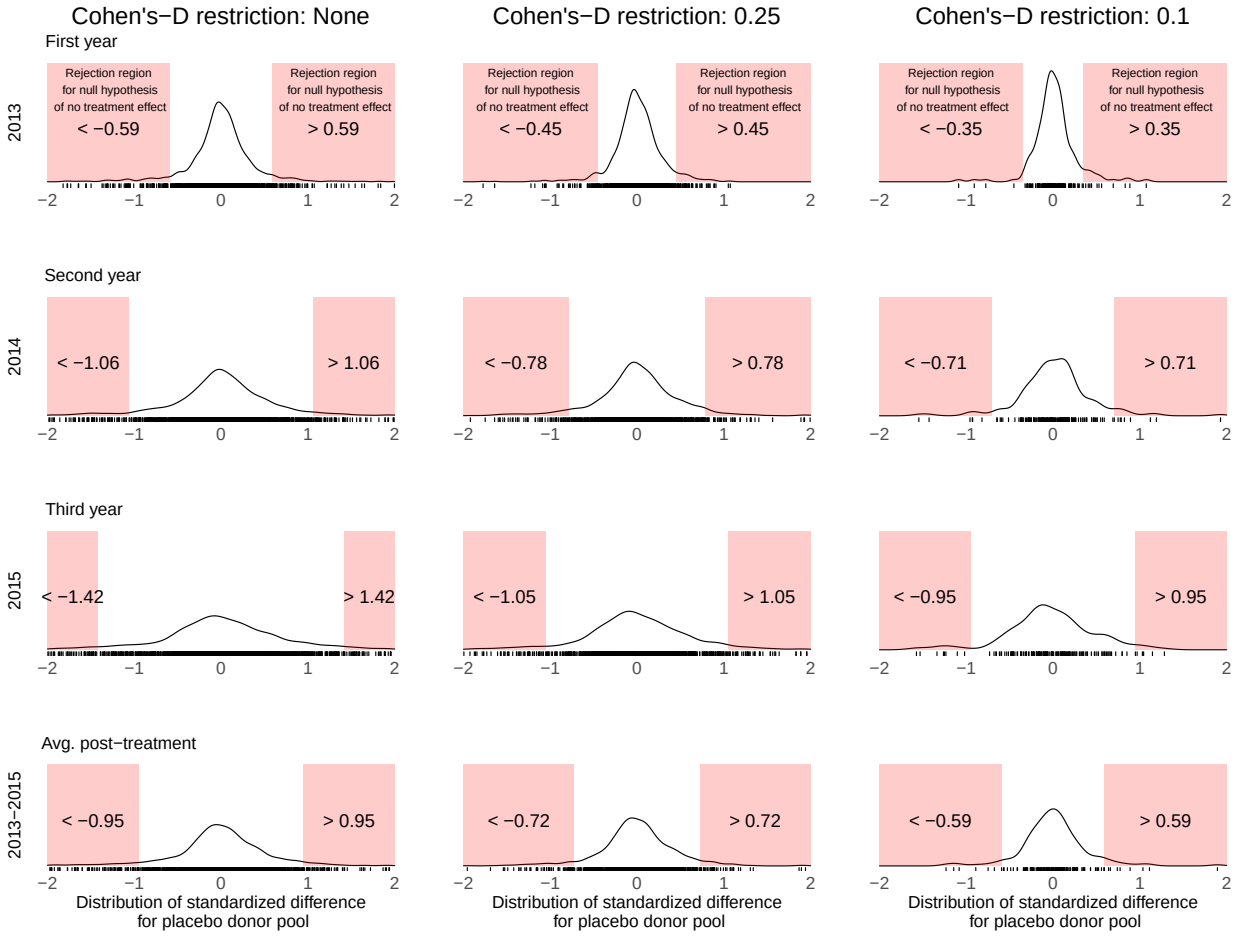
Figure 6: Smoke plot and donor contribution to synthetic control estimates for Colorado hard liquor.




| | 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.

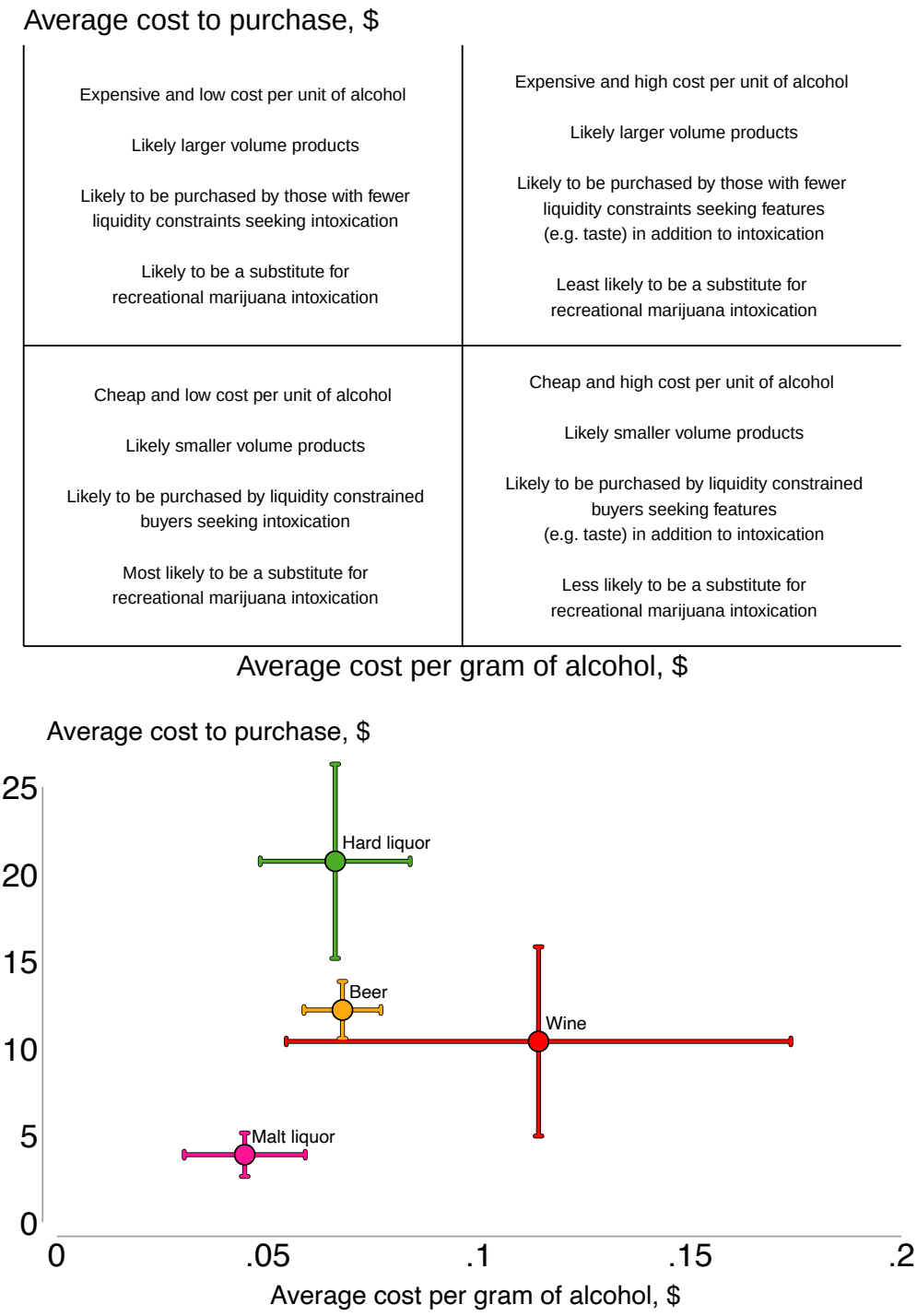
| | Pre-treatment fit 2006-2012 | First Year 2013 | Second Year 2014 | Third Year 2015 | All Post Treatment 2013-2015 |
|---|---|---|---|---|---|
| Pain pills | 0.15 | 0.47 | 3.85 | 5.96 | 3.27 |
| | | (0.77) | (0.26) | (0.20) | (0.28) |
| Beer | 0.15 | –3.44 | –6.12 | –11.49 | –6.82 |
| | | (0.33) | (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.

Figure A1: Malt liquor is the most likely alcohol to be purchased for intoxication, making it the most likely substitute for recreational marijuana intoxication.




*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.