

# On the Misspecification of Linear Assumptions in Synthetic Control

Achille Nazaret, Claudia Shi, and David M. Blei

*Columbia University*

## Abstract

The synthetic control (SC) method is a popular approach for estimating treatment effects from observational panel data. It rests on a crucial assumption that we can write the treated unit as a linear combination of the untreated units. This linearity assumption, however, can be unlikely to hold in practice and, when violated, the resulting SC estimates are incorrect. In this paper we examine two questions: (1) How large can the misspecification error be? (2) How can we limit it? First, we provide theoretical bounds to quantify the misspecification error. The bounds are comforting: small misspecifications induce small errors. With these bounds in hand, we then develop new SC estimators that are specially designed to minimize misspecification error. The estimators are based on additional data about each unit, which is used to produce the SC weights. (For example, if the units are countries then the additional data might be demographic information about each.) We study our estimators on synthetic data; we find they produce more accurate causal estimates than standard synthetic controls. We then re-analyze the California tobacco-program data of the original SC paper, now including additional data from the US census about per-state demographics. Our estimators show that the observations in the pre-treatment period lie within the bounds of misspecification error, and that the observations post-treatment lie outside of those bounds. This is evidence that our SC methods have uncovered a true effect.

## 1 Introduction

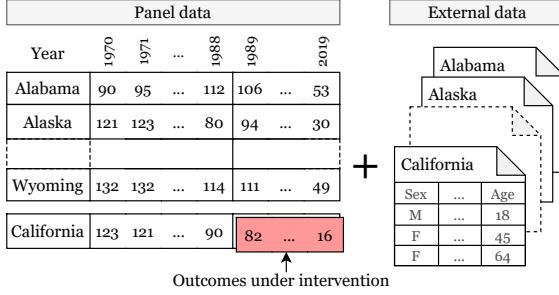
The synthetic control (SC) method is a popular approach for analyzing observational panel data to estimate causal effects [Aba21]. SC has been widely used in science [Pie+16] and social science [HPJ17], as well as for evaluating public policies [DAW19; Pin15; All+17].

The typical SC setup involves measurements of an outcome variable over time. One unit, called the *target*, received an intervention at a certain time. The other units, called *donors*, never received an intervention. The goal of SC is to estimate the target’s counterfactual outcomes. What would have happened had it not received the intervention?

Example: The panel data in Fig. 1 (left) contains cigarette sales across states and time. In 1988 California implemented a program that increased the tobacco tax by 25 cents. After 1988, how much would Californians have smoked had the program not been implemented? Here, California is the target; the other states are the donors.

The idea behind SC is to approximate the target’s control outcomes—the smoking rate in California without its policy—with a weighted combination of the donor’s control outcomes. In the example, SC uses data from the pre-policy periods to fit California’s pre-policy smoking rates as a weighted combination of the other states’ smoking rates. It then uses its fitted weights to estimate the smoking rate in California after 1988, had the policy not been introduced. These estimates, along with California’s post-policy rates, help assess the causal effect of the policy.

What justifies this procedure? In its original formulation, Abadie et al. [ADH10] shows that SC is justified if the control outcomes follow a linear factor model, where a per-period factor linearly combines with a per-unit factor. Following this work, Shi et al. [Shi+22] shows that the linear factor model itself can be justified through assumptions about the individuals within each unit (e.g., people within each state) and invariances around the causal structure of the individual-level outcomes (e.g., whether they smoke). But whether at the aggregate or individual level, these assumptions point to the same requirement: that the target need to be expressed as a linear combination of the donors.



**Figure 1:** Representation of the data used by our estimators. In addition to panel data of annual per-capita cigarette sales (in packs), our estimators leverage external data containing demographic information about each state.

What if this requirement is not satisfied? What if California is not a linear combination of the other states? This paper studies the practical situation where the synthetic control is *misspecified*. We study how to quantify this misspecification error and how to minimize it.

In detail, we derive two bounds on the SC error, the M bound and the James Bound. Both bounds build on the causal framework of Shi et al. [Shi+22]. It assumes a set of *invariant causes*, variables that govern the individual-level outcomes in the same way across units, and where the difference between the units’ outcomes involves different distributions of those causes. For example, whether someone smokes might be caused by their age and education level, and the difference between California’s and Nevada’s smoking rates lies in their different population distributions of those demographic variables. Our theory shows how the similarity between the true target distribution of the causes and its synthetic distribution, induced by the SC weights, helps bound the error of the corresponding SC estimates.

We then consider a situation where we additionally observe external data about the invariant causes, such as demographic information about each state. We show how to use such data to estimate the misspecification interval for a fixed set of SC weights, and we develop two new algorithms for estimating SC weights that explicitly minimize the width of this interval. (One algorithm assumes we observe all invariant causes; the other does not make that assumption, but provides wider misspecification intervals.)

Thus this paper provides a new form of SC analysis, one where we analyze panel data and demographic data together to estimate the target counterfactual and

assess its robustness to misspecification. Figure 2 illustrates this analysis on the California tobacco data, now also using additional data from the U.S. census about per-state demographics. Our estimators show that the observed outcome in the pre-policy period lies within the bounds of misspecification error, and that the observed outcome post-policy lies outside of those bounds. These results suggest that California’s 1988 anti-tobacco program had a true effect, despite possible misspecification of the synthetic control.

**Related Work.** This paper contributes to the literature on synthetic controls [AG03; ADH10]. The M-bound and James-bound estimators of Sections 3 and 4 contribute to research on novel SC estimators [AI11; ADH15; DI16; Xu17; ASS18; Amj+19; Li20; Ath+21; IKM21]. Notably, several estimators penalize the optimization objective [AL21] or adjust the weights [Kel+21] to select donors with outcomes similar to the target. Including covariates in the estimator is recommended in Abadie and Gardeazabal [AG03] but is not mathematically justified. In contrast, the M and James bounds justify why selecting similar donors is important for estimation.

This paper also contributes to the literature on SC methods that are robust to misspecifications. Many existing works make additional assumptions; for example, outcomes follow linear factor models [BMFR21], latent factors are perfectly matched [ASS18; Pow18; FP21], treatments are assigned randomly [ADH15], or many time points are observed [CWZ21]. In contrast, we leverage external data to quantify the errors.

## 2 Synthetic Controls and Misspecification

We review the assumptions behind synthetic controls, the fine-grained model of Shi et al. [Shi+22], and formulate the problem of how to characterize the misspecification error induced by violation of the linearity assumptions.

### 2.1 Panel Data and a Causal Question

Consider a panel dataset containing outcome measurements  $y_{jt}$  for units  $j \in \llbracket 0, J \rrbracket$  over time periods  $t \in \llbracket 0, T \rrbracket$ . Unit  $j = 0$  is the *target*. It has received an intervention at  $T_0$  that may have affected its outcomes  $y_{0t}$  for  $t \geq T_0$ . The remaining units are *donors*. They did not receive an intervention.

For each unit and time, define a pair of potential outcomes  $(Y_{jt}, \tilde{Y}_{jt})$ , where  $\tilde{Y}_{jt}$  is the potential outcome of unit  $j$  at time  $t$  in the world where  $j$  received inter-

vention at  $T_0$ , while  $Y_{jt}$  is the potential outcome in a world with no intervention. For  $j = 0$  and  $t \geq T_0$ ,  $y_{jt} = \tilde{Y}_{jt}$ , otherwise,  $y_{jt} = Y_{jt}$ .

Our causal question is, what would the target counterfactual be, had the intervention not occurred? We would like to estimate  $Y_{0t}$  for  $t \geq T_0$ .

## 2.2 Synthetic Controls and their Assumptions

Synthetic control methods estimate the counterfactual outcomes of the target  $Y_{0t}, t \geq T_0$  with a weighted combination of the outcomes of the donors:  $Y_{0t} = \sum_j w_j y_{jt}$ . The SC weights are fitted from the pre-intervention outcomes,

$$w = \arg \min_{w \in \Delta^J} \sum_{t=0}^{T_0-1} \left( y_{0t} - \sum_j w_j y_{jt} \right)^2. \quad (1)$$

The validity of SC relies on two conditions: (1) During the pre-intervention period, the target's outcomes can be written as a weighted combination of the control units' outcomes. (2) The weighted combination from the pre-intervention periods generalizes to the post-intervention periods.

As a first step to obtain these two conditions, Abadie et al. [ADH10] and most other works assume that the outcomes under no intervention are generated by a linear factor model [Bai09; ADH10; BMFR21; FP21]. We call it assumption A1.

**A1. (Linear Factor Model)** Under no intervention, the outcomes are generated from a linear factor model,

$$Y_{jt} = \mu_j^\top \lambda_t + \epsilon_{jt}, \quad (2)$$

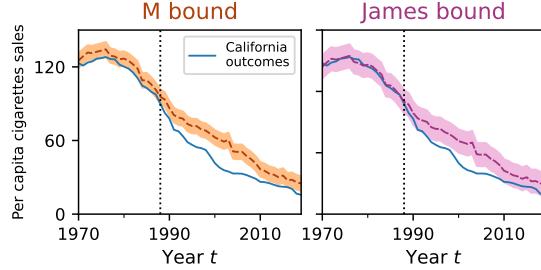
where  $\mu_j$  is a unit-specific latent factor,  $\lambda_t$  is a time-dependent factor, and  $\epsilon_{jt}$  is independent random noise.

Then, Abadie et al. [ADH10] assumes that the target's outcomes can be written as a convex combination of the donors' outcomes. This implies that the target's latent factor is a convex combination of the donors' latent factors, we it call A2.

**A2. (Convex Combination)** The target unit's latent factor is a convex combination of the donors' latent factors,

$$\exists w \in \Delta^J, \quad \mu_0 = \sum_j w_j \mu_j,$$

where  $\Delta^J$  is the simplex over  $J$  coordinates.



**Figure 2:** Visualization of the observed California outcomes (solid lines), the SC estimates (dotted lines), and the misspecification intervals (shaded area), as calculated by the M-bound estimator (left) or the James-bound estimator (right). California's outcomes lie within the error bounds prior to the intervention but escape the error bounds after the intervention. This suggests that the tobacco program had a causal effect despite possible misspecification of SC.

**Remark 1.** To be precise, Abadie et al. [ADH10] assumes that  $A := \sum_{t < T_0} \lambda_t^\top \lambda_t$  is nonsingular. Consequently, once they assume  $Y_{0t} = \sum_j w_j Y_{jt}$  for  $t < T_0$ , then the invertability of  $A$  implies  $\mu_0 = \sum_j w_j \mu_j$  (intuitively, we “invert” the factor model). The target is a convex combination of donors.

With assumptions A1 and A2 in hand, estimator (1) will identify the weights of A2. These weights will then estimate the untreated potential outcomes using the factor model A1. With assumptions A1 and A2, synthetic control is possible.

## 2.3 A Fine-grained Model for SC

When does the linear factor model assumption A1 hold in practice? Shi et al. [Shi+22] explores a justification.

First, the authors notice that SC often considers large units composed of multiple individuals (states, countries) and aggregated outcomes that are averages of individual-level outcomes (per-capita cigarette consumption). Therefore, they propose a “fine-grained” model of synthetic controls, which introduces individual-level variables. The variable  $Y_{ijt}$  denotes individual  $i$ 's outcome in unit  $j$  at time  $t$  (their cigarette consumption at time  $t$ ). The outcome of each unit  $Y_{jt}$  is the average of the individual outcomes in the unit.

Second, Shi et al. [Shi+22] posits the idea of *invariant causes*, which we denote as  $x_{ijt}$ . The invariant causes  $x_{ijt}$  are individual-level variables that follow two invariance assumptions. (1) When conditioned

on the invariant causes, the individual-level outcomes do not depend on which unit the individual is from. This way,  $x \mapsto \mathbb{E}[Y_{ijt} | x_{ijt} = x]$  is the same function for all  $i$  and  $j$ ; we write it  $\mathbb{E}_t[Y|x]$ . (2) The distribution of the invariant causes in each unit can change from unit to unit but remain the same across time. The distribution of invariant causes over the individuals in unit  $j$  is denoted  $x \mapsto p_j(x)$ , with no dependence in  $t$ .

Finally, with these individual-level variables, Shi et al. [Shi+22] shows that the unit-level outcomes are

$$\mathbb{E}[Y_{jt}] = \int_x \mathbb{E}_t[Y | x] p_j(x) dx. \quad (3)$$

Further, if the distributions of invariant causes  $p_j$  are *discrete* and *finite* then Eq. 3 becomes a finite sum,

$$\mathbb{E}[Y_{jt}] = \sum_{x_k} \underbrace{\mathbb{E}_t[Y | x_k]}_{(\lambda_t)_k} \underbrace{p_j(x_k)}_{(\mu_j)_k} = \lambda_t^\top \mu_j. \quad (4)$$

Thus Equation 4 justifies the linear factor model A1 and provides context about what the latent factors might represent.

Regarding A2, Shi et al. [Shi+22] still assumes it; it rewrites,

$$\exists w \in \Delta^J, \quad p_0 = \sum_j w_j p_j,$$

where the equality is in the space of probability distributions.

## 2.4 Misspecification in SC

Shi et al. [Shi+22] explains where the linear factor model A1 can come from. However, even with the fine-grained approach, the factor model arises from the strong assumption of *discrete* and *finite* causes, which might not hold in practice. In other words, the factor model A1 is not guaranteed.

The convex combination A2 is also a key assumption that is unlikely to hold in practice. With a limited number of donors, the  $p_j$  (or  $\mu_j$ ) can be linearly independent. And with continuous invariant causes, the  $p_j$  are densities, making it impossible to match  $p_0$  even with infinitely many donors.

In this work, we relax A1 and A2. To relax A1, we use the fine-grained model of Shi et al. [Shi+22] but do not assume that causes are *discrete* and *finite*. To relax A2, is the focus of the paper. When A2 is violated, the target is not a convex combination of the donors. The

synthetic control is *misspecified*, which leads to errors in the estimation of causal effects. Formally, we define the misspecification error as the absolute difference between the expected counterfactual outcome and the synthetic outcome (after intervention):

$$\left| \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] \right| \text{ for } t \geq T_0.$$

We turn to the problem of how to bound and minimize this error. We will show how to leverage external data about the invariant causes – such as demographic data from the census – to estimate a *misspecification interval* for the synthetic outcomes. We then derive new ways to fit the SC weights that minimize the width of that interval. The result is a new type of estimate of the SC counterfactual, and an assessment of its sensitivity to the linearity misspecification of A2.

## 3 The M Bound and its Estimator

In this section, we derive an exact bound that quantifies errors induced by violation of A2. Using this bound, we develop an estimator minimizing A2 misspecification errors.

### 3.1 The M Bound

Let  $\hat{p}_0$  be the synthetic distribution, defined as  $\hat{p}_0(w) = \sum_j w_j p_j$ . We examine the difference between the distribution of the target unit  $p_0$  and the synthetic distribution  $\hat{p}_0$ . If  $p_0 \neq \hat{p}_0$  but  $\hat{p}_0$  remains “close” to  $p_0$ , we expect the synthetic control estimate to remain approximately correct,

$$\mathbb{E}[Y_{0t}] \approx \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}].$$

We formalize this intuition by bounding the errors resulting from the misspecification of A2.

**Bound 1** (M bound). *For any  $t$ , assume that  $x \mapsto \mathbb{E}_t[Y|x]$  is  $\ell$ -Lipschitz<sup>1</sup>. Then for any weights in the simplex  $w$ , we have the Misspecification error bound (M-bound):*

$$\left| \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] \right| \leq \ell \cdot W_1(p_0, \hat{p}_0), \quad (5)$$

where  $\hat{p}_0 = \sum_j w_j p_j$  and  $W_1$  is a  $\ell_1$ -Wasserstein distance.

<sup>1</sup>See Appendix A for details about Lipschitz functions.

The proof is in Appendix A.

The Wasserstein distance  $W_1$  is a distance between probability distributions [Vil09]. It quantifies the differences between the true population distribution  $p_0$  and the synthetic population distribution  $\hat{p}_0$ .

For any set of weights  $w$ , the M bound (Bound 1) circumscribes the error of the SC estimate by a function of the weights, the population distributions of each unit (the  $p_j$ ), and the sensitivity of the outcome variables to the variation of the causes (the Lipschitz constant  $\ell$ ).

If  $p_0 = \hat{p}_0$ , then the Wasserstein distance  $W_1(p_0, \hat{p}_0)$  between the true and the synthetic distribution is zero. The M bound recovers that the SC estimate is correct. When  $p_0 \neq \hat{p}_0$ , the M bound shows that the estimation error is proportional to the distance  $W_1(p_0, \hat{p}_0)$ .

The intuition behind Eq. 5 is that when a misspecification occurs, a portion of the population  $p_0$  is approximated with an incorrect portion of the synthetic population  $\hat{p}_0$ . It is unpredictable how these populations will behave. In the worst case, their outcomes can differ by at most the distance between them (captured by  $W_1$ ) and the maximum possible variation of the conditional outcome (captured by  $\ell$ ). Hence, the M bound proves (theoretically) that a small misspecification induces a small estimation error.

### 3.2 The M-bound Estimator

We established the M bound, which quantifies the misspecification error for any set of weights  $w$ . To find weights with minimal misspecification error, we develop the M-bound estimator. See Algorithm 1.

The M-bound estimator takes population distribution data  $p_j$  for each unit as input and returns a set of weights that minimizes the M bound. To obtain the weights, it uses projected gradient descent with the following objective,

$$(w_1, \dots, w_J) \mapsto W_1\left(p_0, \sum_j w_j p_j\right).$$

Notice it computes the SC weights using the population distribution of each unit. It does not use the outcomes data.

After obtaining a set of weights from Algorithm 1, we can use Eq. 5 with an estimated constant  $\ell$  to create a

---

#### Algorithm 1: Minimization of the M bound

---

```

Input: Distributions  $p_0, \dots, p_J$ ; learning rate  $\alpha$ ;
number of epochs  $E$ .
Output:  $(w_j)$  minimizing the M bound.
 $(w_1, \dots, w_J) \leftarrow (\frac{1}{J}, \dots, \frac{1}{J})$ 
for  $e = 1$  to  $E$  do
     $\hat{p}_0 \leftarrow \sum w_j p_j$ 
     $\text{grad} \leftarrow \nabla_w W_1(p_0, \hat{p}_0)$ 
     $w \leftarrow w - \alpha \cdot \text{grad}$ 
     $w \leftarrow \text{project\_simplex}(w)$ 
end for
return  $w$ 

```

---

misspecification interval around the synthetic control estimate,

$$\mathbb{E}[Y_{0t}] \in [\hat{y}_{0t} - M, \hat{y}_{0t} + M] \quad \forall t, \quad (6)$$

where  $\hat{y}_{0t} := \sum_{j=1}^J w_j y_{jt}$ ,  $M := \ell \cdot W_1(p_0, \hat{p}_0)$ . Thus, the M bound, with its associated estimator and misspecification interval, can be used to discover causal effects.

In Section 5.2, we revisit the California tobacco example. We use demographic data of each US state to form the invariant causes distributions  $p_j$  and fit the M-bound estimator with these  $p_j$ . Like standard SC, the weights returned by the estimator are used to form the synthetic outcomes. In addition, the M bound provides misspecification intervals accounting for the A2 misspecification error. Fig. 2 illustrates the synthetic control estimate with its misspecification interval generated by the M-bound estimator. We see that California's observed outcomes lie within the interval before intervention and escape it after the intervention. This suggests that a causal effect is present, even in case of misspecification.

### 4 The James Bound and its Estimator

In Section 3, we derived a theoretical bound on misspecification error and showed how to use the M-bound estimator to detect a causal effect. In theory, the true outcome is guaranteed to lie within the M bound. In practice, the misspecification interval produced by the M bound is only valid if we observe the distribution of all invariant causes  $p_j$ . Observing all invariant causes is a strong assumption that may not hold.

Here, we consider the setting where the invariant causes are only partially observed. We first derive

a new error bound, the James bound, that accounts for misspecification on both the observed and unobserved causes. The James bound leverages the pre-intervention outcome data to estimate the influence of the unobserved causes on the outcome variable. To find the weights that minimize the James bound, we develop the James-bound estimator. Finally, we discuss when it is appropriate to use the M bound versus the James bound.

#### 4.1 The James Bound

So far, we have used  $x$  to denote all the invariant causes. With a redefinition of notation, we now refer to the *observed causes* as  $x$ , and the *unobserved causes* as  $z$ . Such that Eq. 3 becomes  $\mathbb{E}[Y_{jt}] = \int_{(x,z)} p_j(x, z) \mathbb{E}_t[Y|x, z] dx dz$ .

In general, we cannot bound the effect of unobserved variables without further assumptions. Here, we assume that the unobserved causes and observed causes are independent and that their respective effect on the outcome can be decomposed into two distinct terms, this is A3. We note that A1, which we relaxed, was more restrictive than A3.

##### A3. Independence of Observed and Unobserved Causes.

For each unit  $j$ , the variable  $x$  and  $z$  are independent,

$$p_j(x, z) = p_j(x)p_j(z),$$

and for each time  $t$ , there exist functions  $g$  and  $h$  such that:

$$\mathbb{E}_t[Y|x, z] = g_t(x) + h_t(z).$$

We note that the distributions of the observed causes  $x \mapsto p_j(x)$  and the unobserved causes  $z \mapsto p_j(z)$  remain arbitrary, and so are  $g_t$  and  $h_t$ .

With A3, we have ‘‘just another misspecification error’’ bound, the James bound.

**Bound 2** (James bound). *For  $t \geq T_0$ , assume that  $x \mapsto \mathbb{E}_t[Y|x]$  is  $\ell$ -Lipschitz. Then for any weights  $w \in \Delta^J$ ,*

$$\left| \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] \right| \leq \ell \cdot W_1(p_0(x), \hat{p}_0(x)) \quad (7)$$

$$+ \max_{u < T_0} \left| \mathbb{E}[Y_{0u}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{ju}] \right| \quad (8)$$

$$+ \inf_{\alpha \in \Delta^{T_0}} \left| \int_z (p_0(z) - \hat{p}_0(z)) \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) dz \right|. \quad (9)$$

The proof is in Appendix A.

The first term (7) mirrors the M bound. It quantifies the similarity between the target and synthetic unit’s distributions of observed causes,  $x \mapsto p_0(x)$  and  $x \mapsto \hat{p}_0(x)$ .

The second term (8) measures the goodness of fit of the pre-intervention outcomes. It indirectly estimates the similarity between the target and the synthetic unit’s distributions of unobserved causes,  $z \mapsto p_j(z)$  and  $z \mapsto \hat{p}_j(z)$ .

The last term (9) contains the remaining error terms. We cannot compute this term directly because it contains unobserved quantities. In Appendix A, we argue that this term is small, and we may ignore it in practice. We also present two explicit models in which the term is null and show that it is null in the standard SC factor model.

#### 4.2 The James-bound Estimator

Building on the James bound (Bound 2), we derive the James-bound estimator. The estimator identifies the weights that minimize the following objective,

$$w \mapsto \max_{t < T_0} \left| y_{0t} - \sum_{j=1}^J w_j y_{jt} \right| + \lambda \cdot W_1(p_0, \sum_j w_j p_j), \quad (10)$$

where  $\lambda$  is a hyperparameter. We update Algorithm 1 into a minimization algorithm for this objective function, it is reported in Appendix A.

If hyperparameter  $\lambda$  is set to  $\ell$ , and if term (9) is effectively negligible, then Eq. 10 is precisely the James bound. Otherwise, it can be viewed as the pre-intervention errors (first term), regularized by the Wasserstein distance over external data (second term). With this perspective, the James-bound estimator finds weights that minimize pre-intervention errors while favoring donors that are similar to the target.

#### 4.3 Choosing between M and James Bound

We introduced two bounds, along with associated estimators and misspecification intervals. M bounds are tighter but require data about all the invariant causes. James bounds can be wider but require fewer data.

As a practical guide, we recommend using the James-bound estimator first. It is indeed more prudent to

assume that some invariant causes might be unobserved. If the post-intervention target outcomes fall outside the misspecification interval, we have discovered a causal effect robust to A2 misspecification (see Fig. 2).

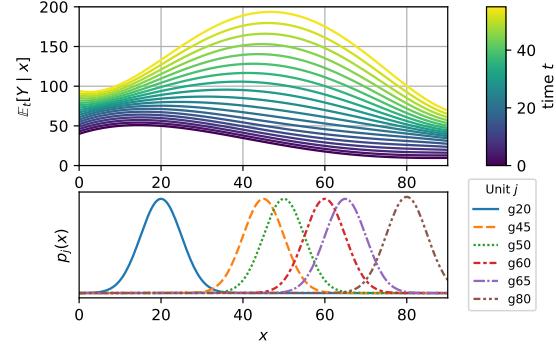
If the post-intervention misspecification interval is too wide to detect a causal effect, then it could be that there is no causal effect. But it could also be that there is too much misspecification to use SC or that the James bound is too loose. We cannot conclude in favor of a causal effect in the first two cases. To check if the James bound is too loose and find a tighter bound, we can use the M bound.

The M bound is guaranteed valid if all invariant causes are observed. Since the M estimator does not use outcome data, the target’s pre-intervention outcomes can be used as a validation set. If the observed pre-intervention outcomes fall outside the predicted misspecification interval, not all invariant causes were observed, and we cannot apply the M bound. Otherwise, we may use the M bound.

## 5 Empirical Studies

We examine the M-bound and James-bound estimators using synthetic and tobacco consumption data. With synthetic data, we demonstrate that the M-bound and James-bound estimators produce better estimates in case of misspecification, and show that their misspecification intervals contain the counterfactual outcomes correctly. Using the tobacco consumption case study, we demonstrate how to collect external data and how to choose between M-bound and James-bound estimators. We find that the post-intervention California outcomes escape the misspecification error bound, suggesting that there is an actual causal effect.

**Implementation Details.** To implement the algorithms we need to manipulate probability distributions and calculate Wasserstein distances with their gradients. Our implementation expects the input  $p_j$  to be non-parametric distributions represented by a collection of atoms and associated probabilities:  $p_j = \sum_{x \in X} \delta_x \cdot p_j(x)$ , where  $X$  is the set of atoms and  $\delta_x$  is a point mass at  $x$ . If  $p_j$  is discrete, such as a histogram, then the atoms are the possible values of the causes, and  $p_j(x)$  their associated probabilities. If  $p_j$  is continuous, then the atoms are samples of  $p_j$ , and  $p_j(x)$  is the normalized density at  $x$ . For all experiments, we compute the gradients of  $(w_1, \dots, w_J) \mapsto W_1(p_0, \sum_j w_j p_j)$ , using the Python Optimal Transport library [Fla+21] coupled with PyTorch [Pas+19]. We use gradient descent with a learn-



**Figure 3:** Visualization of the synthetic data generating process. (top) Each colored line represents the expected conditional outcome  $E_t [Y|x]$  for a different time  $t$  (one line per time), as a function of the cause  $x$  (on the x-axis). As time progresses (from darker to lighter), the expected conditional outcomes increase for all values of  $x$ . For different  $x$ , the rate of increase over time is different. (bottom) Distributions densities of the causes  $x \mapsto p_j(x)$  for each unit  $j$ . Different lines represent different units. The target unit is  $g45$ , which overlaps mostly with unit  $g50$ .

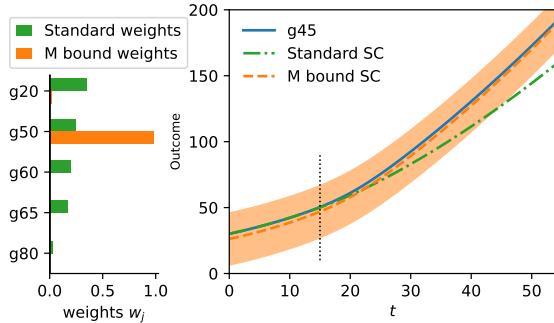
ing rate  $\alpha = 5 \cdot 10^{-6}$  and 200,000 epochs.

### 5.1 Experiments with Synthetic Data

**Data Description.** We generate synthetic data by defining the conditional distribution  $E_t [Y|x] = f(x, t)$  and the causes distributions  $p_j(x)$ . We create six different units (called  $g20$ ,  $g45$ ,  $g50$ ,  $g60$ ,  $g65$ ,  $g70$ ), and consider that a single cause  $x \in \mathbb{R}$  impacts the outcome  $Y$ . The units can be thought of as different groups of people (e.g. cities), and the cause  $x$  as the age of each individual in these groups. The six units have different distributions of age (group  $gX$  has an average age of  $X$ ). The target group is  $g45$ , the panel duration is  $T = 50$ , and the intervention time is  $T_0 = 15$ .

The closed form equations of  $(t, x) \mapsto E_t [Y|x]$  and  $(j, x) \mapsto p_j(x)$  are in Appendix B while Figure 3 shows the evolution of  $x \mapsto E_t [Y|x]$  over time  $t$  as well as the distributions  $x \mapsto p_j(x)$  for each unit  $j$ . The expected outcome  $E_t [Y|x]$  varies over time, in different ways for each  $x$ .

We input the distributions  $p_j$  to Algorithm 1 and obtain the weights that minimize the M bound. As a comparison, we calculate the weights obtained from the standard SC in Eq. 1. We report the weights and the induced synthetic outcomes in Fig. 4. Furthermore, we compute  $\ell = 4.0$  from  $x \mapsto E_t [Y | x]$



**Figure 4:** Comparison of the M-bound estimator and the standard SC estimator on synthetic data. (left) Weights returned by each estimator. The M-bound estimator selects donors ( $g_{20}$  and  $g_{50}$ ) that are most similar to the target ( $g_{45}$ ). (right) Synthetic outcomes of each estimator, compared to the true outcome. Under misspecification, the M-bound estimator provides more accurate estimates than the standard SC, despite a poorer pre-intervention fit.

(valid for all  $t$ ). This way, we obtain the exact value of the M-bound and we can form the misspecification interval of Eq. 6, shaded on Fig. 4.

**Analysis.** As shown in Fig. 4, the standard SC places a large weight on donor  $g_{20}$ , which is a unit whose individuals are very different from  $g_{45}$  but with similar pre-treatment outcomes. When time increases, the individuals in  $g_{20}$  and  $g_{45}$  evolve differently and the synthetic outcome of the standard SC weights deviates away from the true outcome. In contrast, the M-bound estimator places most of the weight mass on the donor  $g_{50}$ , which contains individuals with similar  $x$  as the target  $g_{45}$ . By doing so, the synthetic outcomes might not exactly match the  $g_{45}$  outcomes, but they generalize better over time. We also verify that the true outcome is always contained in the misspecification interval (Eq. 6). We repeat the analysis with the James bound and obtain the same conclusions, reported in Appendix B.

With external data, we estimated the misspecification error and limited it using the M-bound and James-bound estimators. Without external data, standard SC was incorrect.

## 5.2 A Case Study on Real Data

We revisit the tobacco study from Abadie et al. [ADH10] to illustrate how to collect external data, apply different estimators, and calculate the misspecification intervals.

**Prop 99.** A tobacco control program was passed in California in 1988, which increased tobacco taxes by 25 cents. The tax revenue was used to fund anti-tobacco campaigns. Our goal is to estimate the causal effect of the tobacco control program on California’s tobacco consumption.

The tobacco panel dataset (Fig. 1) is from the Centers for Disease Control and Prevention [Cen19], which provides the per capita tobacco consumption for 50 states from 1970 to 2019. The intervention of interest is the tobacco program, Prop 99. The observed outcomes for California after 1988 are under intervention. All the other observed outcomes in the dataset are assumed to be under no intervention.

**External Data Collection.** First, we identify the potential causes of smoking. According to Turner et al. [TMF04], smoking is heavily influenced by societal and cultural factors. While these factors are difficult to measure directly, they are often correlated with demographics. Several studies have found that cigarette consumption varies significantly by age, gender, race, and ethnicity [Sak+16; Cor+22]. As a result, we use *age*, *sex*, and *ethnicity/race* as proxies for the causes of smoking.

We use the American Community Survey (ACS) to formulate a distribution of causes for each unit. The ACS is a demographics survey program conducted continuously by the U.S. Census Bureau [Cen20a]. It reports population demographics at different geographical scales, from city boroughs to states. We accessed the ACS data with the Census Reporter API [Cen20b]. For each state, the ACS provides the joint distribution of the variables *age*, *race*, *sex*. Each variable is discretized into multiple bins: *age* into 14 bins (e.g. 15 to 17, 20 to 24 years old), *race* takes 8 values (Asian, Black, Native American, Pacific Islander, White non-Hispanic, White Hispanic, Mix, and Other), and *sex* takes 2 values (Male, Female). The joints  $x \mapsto p_j(x)$  over these variables are defined for each state on these  $14 \times 8 \times 2 = 224$  demographics combinations (atoms).

We estimate  $\ell$  using additional survey data from the Tobacco Use Supplement to the Current Population Survey. This independent study collects individual demographic information along with tobacco consumption. We form the expected tobacco consumption given each invariant cause and compute the induced  $\ell$ . More details about the computation of Lipschitz constant can be found in Appendix B.

**The M-bound Estimator.** The M-bound estima-

tor uses our newly formed distributions  $p_0, \dots, p_J$  to compute a set of SC weights. We report the weights in Appendix B.4 and the SC outcomes with the misspecification interval in Fig. 2. Among the set of 50 potential donors, five obtained non-zero weights: New Mexico, Nevada, D.C, Hawaii and Texas. As expected, the M-bound estimator selected states that are similar to California. New Mexico and Nevada are geographically close and have similar demographics. Both D.C. and California have a relatively young active population. And, California is the number one destination for Hawaiians moving to the US mainland (from US census).

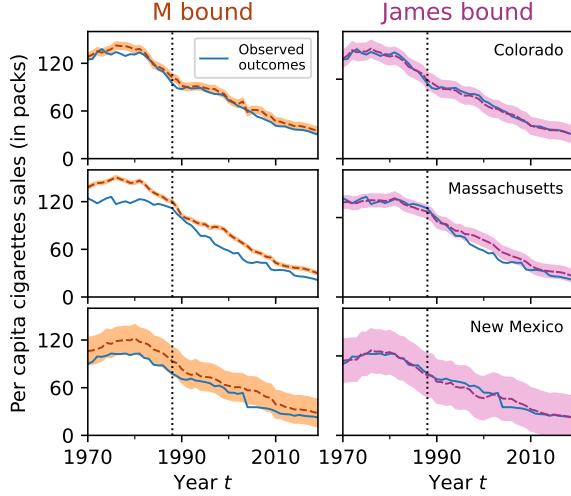
In Fig. 2, the solid and dotted lines denote the observed and synthetic California outcomes. The shaded areas represent the misspecification intervals. California pre-intervention outcomes fall within the estimated M-bound interval, but synthetic California is not a perfect fit, there is misspecification. In spite of the misspecification, Fig. 2 shows that the post-intervention outcomes are outside of the bounds, suggesting a causal effect.

**The James-bound Estimator.** As discussed in Section 4.3, the M-bound misspecification interval is only valid if we observe all the invariant causes. According to Fig. 2, California’s pre-intervention outcomes fall within the M-bound intervals. We find, however, that when some other states are considered as the target unit, their observed outcomes before the intervention are not always within the interval.

We perform *placebo tests* [ADH10] where each donor is considered to be the target and a synthetic control is constructed using the other donors. Because the donors did not receive the intervention, we expect synthetic outcomes to match observed outcomes. In Fig. 5, we illustrate the comparisons for three states, Colorado, Massachusetts, New Mexico. For comparisons on all states, see Appendix B.

Fig. 5 (left) shows the synthetic outcome estimates by the M-bound estimator. Both Colorado and Massachusetts’s pre-intervention outcomes are outside of the misspecification interval. This suggests that not all invariant causes are observed. While New Mexico’s pre-intervention outcomes lie within the misspecification interval of the synthetic New Mexico, the error bound is too large to use synthetic control.

Fig. 5 (right) shows the synthetic outcome estimates using the James-bound estimator. We observe that the pre-intervention outcomes across states now all fall within the James-bound misspecification inter-



**Figure 5:** Placebo study of the M-bound estimator (left) and the James-bound estimator (right), on Colorado, Massachusetts, and New Mexico. The M-bound synthetic outcomes are outside of the misspecification interval before the intervention. This suggests that not all invariant causes are observed and that the James bound should be used. The James-bound estimator accounts for the missing causes, with wider misspecification intervals.

vals, which are also wider than the M-bound intervals. After the intervention, the observed tobacco consumption in Colorado remains in the James-bound misspecification interval, suggesting the intervention had no effect. This is expected because Colorado did not implement an anti-tobacco program like California. For Massachusetts, the James-bound interval is narrow enough to detect a decrease in tobacco consumption that is not due to misspecification. In fact, this is consistent with the policies taken by this state in 1993 to raise taxes and increase its Massachusetts Tobacco Control Program.

The placebo test provides further evidence that the tobacco control program in California indeed had a causal effect on tobacco consumption. In states without tobacco control programs, their outcomes fall within the misspecification interval, whereas California’s outcome does not.

## 6 Discussion

We address the problem of the misspecification of linear assumptions in synthetic controls. We relax assumptions commonly assumed in the literature (A1, A2), derive two misspecification bounds, and propose corresponding estimators. The key idea is to leverage

external data to bound and minimize misspecification. Each bound comes with requirements: data must be available for the M bound (observe all causes), and we identify a modeling assumption for the James bound (A3). As an area of future research, we can explore other SC estimation procedures that might be enabled by these two requirements.

## References

- [Aba21] Alberto Abadie. “Using synthetic controls: feasibility, data requirements, and methodological aspects”. In: *Journal of Economic Literature* (2021).
- [ADH10] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. “Synthetic control methods for comparative case studies: estimating the effect of California’s tobacco control program”. In: *Journal of the American Statistical Association* (2010).
- [ADH15] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. “Comparative politics and the synthetic control method”. In: *American Journal of Political Science* (2015).
- [AG03] Alberto Abadie and Javier Gardeazabal. “The economic costs of conflict: a case study of the Basque Country”. In: *American Economic Review* (2003).
- [AI11] Alberto Abadie and Guido W Imbens. “Bias-corrected matching estimators for average treatment effects”. In: *Journal of Business & Economic Statistics* (2011).
- [AL21] Alberto Abadie and Jérémie L’Hour. “A penalized synthetic control estimator for disaggregated data”. In: *Journal of the American Statistical Association* (2021).
- [All+17] Sylvia Allegretto, Arindrajit Dube, Michael Reich, and Ben Zipperer. “Credible research designs for minimum wage studies: a response to Neumark, Salas, and Wascher”. In: *ILR Review* (2017).
- [Amj+19] Muhammad Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. “Mrsc: multi-dimensional robust synthetic control”. In: *ACM on Measurement and Analysis of Computing Systems* (2019).
- [ASS18] Muhammad Amjad, Devavrat Shah, and Dennis Shen. “Robust synthetic control”. In: *Journal of Machine Learning Research* (2018).
- [Ath+21] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. “Matrix completion methods for causal panel data models”. In: *Journal of the American Statistical Association* (2021).
- [Bai09] Jushan Bai. “Panel data models with interactive fixed effects”. In: *Econometrica* (2009).
- [BMFR21] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. “The augmented synthetic control method”. In: *Journal of the American Statistical Association* (2021).
- [Cen20a] Census Bureau. *2016-2020 5-year American Community Survey*. 2020.
- [Cen20b] Census Reporter. *Census Reporter: Making Census Data Easy to Use*. 2020.
- [Cen19] Centers for Disease Control and Prevention. *The Tax Burden on Tobacco, 1970-2019*. 2019.
- [CWZ21] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. “An exact and robust conformal inference method for counterfactual and synthetic controls”. In: *Journal of the American Statistical Association* (2021).
- [Cor+22] Monica E Cornelius, Caitlin G Loretan, Teresa W Wang, Ahmed Jamal, and David M Homa. “Tobacco product use among adults—United States, 2020”. In: *Morbidity and Mortality Weekly Report* (2022).
- [DAW19] John J Donohue, Abhay Aneja, and Kyle D Weber. “Right-to-carry laws and violent crime: a comprehensive assessment using panel data and a state-level synthetic control analysis”. In: *Journal of Empirical Legal Studies* (2019).
- [DI16] Nikolay Doudchenko and Guido W Imbens. *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*. Tech. rep. National Bureau of Economic Research, 2016.
- [FP21] Bruno Ferman and Cristine Pinto. “Synthetic controls with imperfect pretreatment fit”. In: *Quantitative Economics* (2021).
- [Fla+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. “Pot: python optimal transport”. In: *Journal of Machine Learning Research* (2021).
- [HPJ17] Boris Heersink, Brenton D Peterson, and Jeffery A Jenkins. “Disasters and elections: estimating the net effect of damage and relief in historical perspective”. In: *Political Analysis* (2017).
- [IKM21] Guido Imbens, Nathan Kallus, and Xiaojie Mao. “Controlling for unmeasured confounding in panel data using minimal bridge functions: from two-way fixed effects to factor models”. In: *arXiv:2108.03849* (2021).

- [Kel+21] Maxwell Kellogg, Magne Mogstad, Guillaume A Pouliot, and Alexander Torgovitsky. “Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation”. In: *Journal of the American Statistical Association* (2021).
- [Li20] Kathleen T Li. “Statistical inference for average treatment effects estimated by synthetic control methods”. In: *Journal of the American Statistical Association* (2020).
- [Pas+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems* 32. 2019.
- [Pie+16] Hannah Pieters, Daniele Curzi, Alessandro Olper, and Johan Swinnen. “Effect of democratic reforms on child mortality: a synthetic control analysis”. In: *The Lancet Global Health* (2016).
- [Pin15] Paolo Pinotti. “The economic costs of organised crime: evidence from southern italy”. In: *The Economic Journal* (2015).
- [Pow18] David Powell. “Imperfect synthetic controls: did the massachusetts health care reform save lives?” In: *Available at SSRN 3192710* (2018).
- [Sak+16] Kari-Lyn K Sakuma, Jamie Quibol Felicitas-Perkins, Lyzette Blanco, Pebbles Fagan, Eliseo J Pérez-Stable, Kim Pulvers, Devan Romero, and Dennis R Trinidad. “Tobacco use disparities by racial/ethnic groups: california compared to the united states”. In: *Journal of Preventive Medicine* (2016).
- [Shi+22] Claudia Shi, Dhanya Sridhar, Vishal Misra, and David Blei. “On the assumptions of synthetic control methods”. In: *International Conference on Artificial Intelligence and Statistics*. 2022.
- [TMF04] Lindsey Turner, Robin Mermelstein, and Brian Flay. “Individual and contextual influences on adolescent smoking”. In: *Annals of the New York Academy of Sciences* (2004).
- [Vil09] Cédric Villani. *Optimal Transport: Old and New*. 2009.
- [Xu17] Yiqing Xu. “Generalized synthetic control method: causal inference with interactive fixed effects models”. In: *Political Analysis* (2017).

## A Technical Details

### A.1 Lipschitz function

A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be  $\ell$ -Lipschitz if

$$\forall (x, y) \in (\mathbb{R}^n)^2, \quad |f(x) - f(y)| \leq \ell \cdot \|x - y\|_1 \quad (11)$$

More precisely, the function  $f$  can be from and to any metric spaces, with their associated distance. Here we use the  $L_1$  norm to measure the distance between  $x$  and  $y$  in  $\mathbb{R}^n$ , and we use the absolute value to measure the distance between  $f(x)$  and  $f(y)$  in  $\mathbb{R}$ .

A Lipschitz function is limited in how fast it can change. In the context of synthetic control, if  $\ell$  is small, it implies that a change in the causes induces a small change in the outcomes. Having a small  $\ell$  suggests that misspecification of the causes will have a limited impact on the outcome.

### A.2 Proof of the M Bound

**Bound 1** (M bound). *For any  $t$ , let assume that  $x \mapsto \mathbb{E}_t[Y|x]$  is  $\ell$ -Lipschitz, then for any weights in the simplex  $w$ , we have the Misspecification error bound (M-bound):*

$$\left| \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] \right| \leq \ell \cdot W_1(p_0, \hat{p}_0), \quad (12)$$

where  $\hat{p}_0 = \sum_j w_j p_j$  and  $W_1$  is a  $\ell_1$ -Wasserstein distance.

*Proof.* Notice that for any unit  $j \in \llbracket 0, J \rrbracket$ , the expected outcome writes  $\mathbb{E}[Y_{jt}] = \int_x \mathbb{E}_t[Y|X=x] p_j(x) dx$ .

Fix some weights  $w_{1:J} \in \Delta^J$  and then by the linearity of the integral,

$$\begin{aligned} \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] &= \int_x \mathbb{E}_t[Y|x] \left( p_0(x) - \sum_j w_j p_j(x) \right) dx \\ &= \int_x \mathbb{E}_t[Y|x] (p_0(x) - \hat{p}_0(x)) dx \\ &\leq W_1(p_0, \hat{p}_0) \cdot \ell \end{aligned}$$

where the inequality comes from Kantorovich duality (Theorem 5.10, Villani [Vil09]) about Wasserstein distances.  $\square$

### A.3 Proof of the James Bound

**Bound 2** (James bound). *For  $t \geq T_0$ , let assume that  $x \mapsto \mathbb{E}_t[Y|x]$  is  $\ell$ -Lipschitz, then for any weights in the simplex  $w \in \Delta^J$ , we have Just Another Misspecification Errors bound (James bound) :*

$$\begin{aligned} \left| \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] \right| &\leq \ell \cdot W_1(p_0(x), \hat{p}_0(x)) + \max_{u < T_0} \left| \mathbb{E}[Y_{0u}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{ju}] \right| \\ &\quad + \inf_{\alpha \in \Delta^{T_0}} \left| \int_z (p_0(z) - \hat{p}_0(z)) \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) dz \right|. \end{aligned}$$

Before proving the James bound, we need one lemma which follows from assumption A3.

**Lemma 1.** Suppose the distributions of causes  $(x, z) \mapsto p_j(x, z)$  and the expected outcomes  $(x, z) \mapsto \mathbb{E}_t[Y|x, z]$  satisfy assumption A3, then  $\mathbb{E}_t[Y|x] = g_t(x) + \text{constant}$  and  $\mathbb{E}_t[Y|z] = h_t(z) + \text{constant}$  where the constant terms are independent of  $x$  and  $z$  (they can depend on  $t$ ).

*Proof.* Fix  $t$  and  $j$ , and define the function  $f$  by  $f_t(x, z) = \mathbb{E}_t[Y|x, z]$ . We suppose A3, that is  $p_j(x, z) = p_j(x)p_j(z)$  and that there exist two functions  $g_t$  and  $h_t$  such that

$$\forall(x, z), f_t(x, z) = g_t(x) + h_t(z).$$

Then

$$\begin{aligned} \mathbb{E}_t[Y|x] &= \mathbb{E}_{p_j(z|x)}[\mathbb{E}_t[Y|x, z] | x] \\ &= \int_z (g_t(x) + h_t(z)) p_j(z) dz \\ &= g_t(x) + \int_z h_t(z) p_j(z) dz. \end{aligned}$$

We notice that  $\int_z h_t(z) p_j(z) dz$  is a constant independent of  $x$  or  $z$ . (It is the expectation of  $z \mapsto h_t(z)$  with respect to  $z \mapsto p_j(z)$ .)

We obtain a similar result for  $\mathbb{E}_t[Y|z]$ ,  $h_t(z)$  and the constant  $\int_z g_t(x) p_j(x) dx$ .  $\square$

We can now prove the James bound.

*Proof.* Fix some weights  $w \in \Delta^J$ , weights  $\alpha \in \Delta^{T_0}$  then,

$$\begin{aligned} \mathbb{E}[Y_{0t}] - \sum_{j=1}^J w_j \mathbb{E}[Y_{jt}] &= \int_{(x,z)} \mathbb{E}_t[Y|x, z] \left( p_0(x, z) - \sum_j w_j p_j(x, z) \right) dx dz \\ &= \int_{(x,z)} (g_t(x) + h_t(z)) \left( p_0(x) p_0(z) - \sum_j w_j p_j(x) p_j(z) \right) dx dz \\ &= \underbrace{\int_x g_t(x) \left( p_0(x) - \sum_j w_j p_j(x) \right) dx}_A + \underbrace{\int_z h_t(z) \left( p_0(z) - \sum_j w_j p_j(z) \right) dz}_B. \end{aligned}$$

We have  $g_t(x) = \mathbb{E}_t[Y|x] + \text{constant}$ , the constant cancels out in  $A$  and we obtain:

$$|A| = \left| \int_x \mathbb{E}_t[Y|x] \left( p_0(x) - \sum_j w_j p_j(x) \right) dx \right| \leq \ell \cdot W_1(p_0(x), \hat{p}_0(x)).$$

We have  $h_t(z) = \mathbb{E}_t[Y|z] + \text{constant}$ , the constant cancels in  $B$  and we obtain:

$$B = \int_z \mathbb{E}_t[Y|z] (p_0(z) - \hat{p}_0(z)) dz$$

For any  $\alpha \in \Delta^{T_0}$  we have:

$$\begin{aligned} B &= \int_z \left( \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz + \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz \\ &= \sum_{u < T_0} \alpha_u \int_z (\mathbb{E}_u[Y|z] p_0(z) - \sum_j w_j \mathbb{E}_u[Y|z] p_j(z)) dz + \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz \\ &= \sum_{u < T_0} \alpha_u \left( \mathbb{E}[Y_{0u}] - \sum_j w_j \mathbb{E}[Y_{ju}] \right) dz + \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz. \end{aligned}$$

So,

$$\begin{aligned}
|B| &\leq \sum_{u < T_0} \alpha_u \left| \mathbb{E}[Y_{0u}] - \sum_j w_j \mathbb{E}[Y_{ju}] \right| dz + \left| \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz \right| \\
&\leq \left( \sum_{u < T_0} \alpha_u \right) \max_{u < T_0} \left| \mathbb{E}[Y_{0u}] - \sum_j w_j \mathbb{E}[Y_{ju}] \right| dz + \left| \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz \right| \\
&= \max_{u < T_0} \left| \mathbb{E}[Y_{0u}] - \sum_j w_j \mathbb{E}[Y_{ju}] \right| dz + \left| \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz \right|.
\end{aligned}$$

Because the previous inequality hold for any  $\alpha \in \Delta^{T_0}$ , we can “take” the inf on the right term.

We obtain,

$$|B| \leq \max_{u < T_0} \left| \mathbb{E}[Y_{0u}] - \sum_j w_j \mathbb{E}[Y_{ju}] \right| dz + \inf_{\alpha \in \Delta^{T_0}} \left| \int_z \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) (p_0(z) - \hat{p}_0(z)) dz \right|.$$

This proves the James bound.  $\square$

#### A.4 Interpretation of Eq. 9

In the James bound, we can compute the following terms:

- $\ell \cdot W_1(p_0(x), \hat{p}_0(x))$  is estimated with the external data on the subset of observed causes.
- $\max_{u < T_0} \left| \mathbb{E}[Y_{0u}] - \sum_j w_j \mathbb{E}[Y_{ju}] \right| dz$  is estimated from the outcome data.

The last term:  $\inf_{\alpha \in \Delta^{T_0}} \left| \int_z \left( p_0(z) - \sum_j w_j p_j(z) \right) \left( \mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] \right) dz \right|$  cannot be estimated from observed data. Hence, we defined the James-bound estimator without this term. and showed that it was minimizing the James bound only if this last term is negligible.

We give justification as to why this last term might be negligible, at least in comparison to the two other terms. If any of the two following conditions holds, the last term is 0:

1. If  $p_0 = \sum_j w_j p_j$ .
2. If there exists  $\alpha \in \Delta^{T_0}$  such that  $\mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z] = \text{constant}$ .

Naturally, we are not expecting the first condition to hold, but at least we can hope that  $p_0(z) - \sum_j w_j p_j(z)$  is of the same order of magnitude as  $p_0(x) - \sum_j w_j p_j(x)$  (and so of  $W_1(p_0(x), \hat{p}_0(x))$ , the first term). If in addition  $\mathbb{E}_t[Y|z] - \sum_{u < T_0} \alpha_u \mathbb{E}_u[Y|z]$  is small, then the last term (which is a product of two small terms) is negligible compared to the other terms of the James bound.

With this intuition, it seems possible to finding a  $\alpha$  that makes the full integral close to zero should be possible.

More concretely, we give two examples of models for which the term (9) is null.

**Standard SC setting** In the standard SC setting, assumption A2 is made. The practitioner assumes that there exists a set of weights  $(w_j)$  such that  $p_0 = \sum_j w_j p_j$  (with the notations of factor models, each  $p_j$  is a point mass located at the latent factor  $\mu_j$ , such that  $\mu_0 = \sum_j w_j \mu_j$ ). In particular, it implies that  $(p_0(z) - \sum_j w_j p_j(z)) = 0$  and so (9) = 0.

**Arbitrary model with linear conditional expectation.** We now assume arbitrary distributions  $p_j$ . They can be point mass on linear factors as in standard SC or arbitrary continuous distributions. A2 does not need to hold, it may be impossible to write the target as a linear combination of the donors. However, we assume that the response functions ( $z \mapsto \mathbb{E}_t [Y | z]$ )<sub>t</sub> are of the form  $\mathbb{E}_t [Y | z] = \beta_t^\top z$ , with  $\beta_t$  being able to change arbitrarily over time. Actually, we even require the  $\beta_t$  to change enough such that there exists  $\alpha \in \Delta^{T_0}$  such that they are linearly independent and there exists  $\beta_t = \sum_{u < T_0} \alpha_u \beta_u$ . In that case again, (9) = 0.

### A.5 The James-bound Estimator

We adapt the M-bound estimator algorithm from Algorithm 1 into Algorithm 2.

---

#### Algorithm 2: Minimization of the James-bound

---

**Input:** Distributions  $p_0, \dots, p_J$ ; Pre-intervention measurements  $\{(y_{0t}, \dots, y_{jt})_{t=0, \dots, T_0-1}\}$ ; learning rate  $\alpha$ ; number of epochs  $E$ , James parameter  $\lambda$ .

**Output:**  $(w_j)$  minimizing the James-bound.

```

 $(w_1, \dots, w_J) \leftarrow (\frac{1}{J}, \dots, \frac{1}{J})$ 
for  $e = 1$  to  $E$  do
     $\hat{p}_0 \leftarrow \sum w_j p_j$ 
     $\text{grad} \leftarrow \nabla_w \left( \max_{t < T_0} |y_{0t} - \sum_{j=1}^J w_j m_{jt}| + \lambda \cdot W_1(p_0, \hat{p}_0) \right)$ 
     $w \leftarrow w - \alpha \cdot \text{grad}$ 
     $w \leftarrow \text{project\_simplex}(w)$ 
end for
return  $w$ 
```

---

## B Experiment Details

### B.1 Simulation Details

For the synthetic experiment, we generate the outcomes under no intervention by defining the conditional expected outcomes  $f : (x, t) \mapsto \mathbb{E}_t [Y|x]$  and the unit specific causes distributions  $x \mapsto p_j(x)$ . In this experiment,  $x$  is a single scalar variable.

The function  $f$  we choose is represented in Fig. 3 (top). It enjoys a closed-form expression:

$$\begin{aligned}
f(x, t) = & -\frac{13t^2x^4}{2100000000} + \frac{71t^2x^3}{78750000} - \frac{10141t^2x^2}{252000000} + \frac{12521t^2x}{12600000} + t + 4.07142857142857 \\
& \cdot 10^{-6}x^4 \log \left( e^{\frac{t}{3}-\frac{20}{3}} + 1 \right) - \frac{43x^4}{13300000} - 0.000731428571428571x^3 \log \left( e^{\frac{t}{3}-\frac{20}{3}} + 1 \right) \\
& + \frac{1313x^3}{1496250} + 0.0380053571428571x^2 \log \left( e^{\frac{t}{3}-\frac{20}{3}} + 1 \right) - \frac{359953x^2}{4788000} \\
& - 0.500107142857143x \log \left( e^{\frac{t}{3}-\frac{20}{3}} + 1 \right) + \frac{401813x}{239400} + 40.
\end{aligned}$$

It was generated by combining Lagrange polynomials in  $x$  with time varying coefficients.

For each group gXX (g20, g45, g50, g60, g65, g70), their associated distribution of causes is given by a normal distribution centered at XX (e.g. at 20 for g20), and with scale 5 (variance 25). Each distribution is represented in Fig. 3 (bottom).

Because our implementations of the M-bound estimator and James-bound estimator use non-parametric distributions represented by a collection of atoms and associated probabilities, each  $p_j$  is more precisely defined as

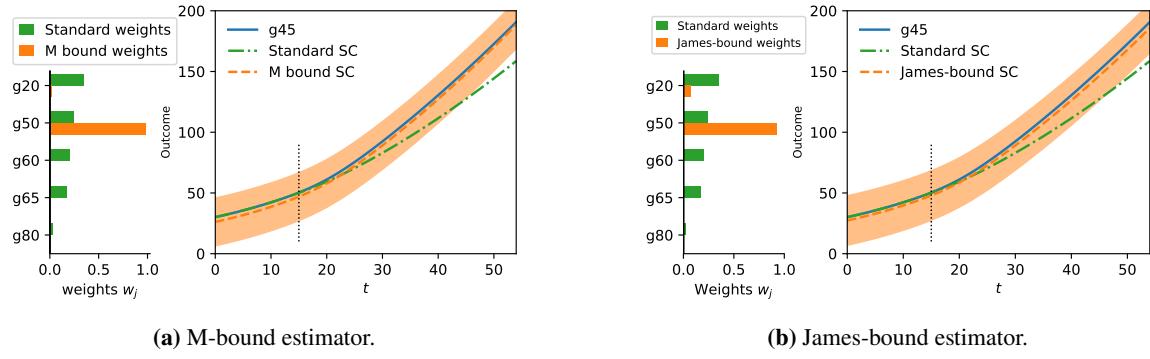
$$p_j \propto \sum_{x \in X} \delta_x \cdot \mathcal{N}(x; \mu_j, 5^2)$$

where  $\mu_j = \mathbb{E}_X X$  for each group  $gXX$ , and the set of atoms  $X$  is  $X = \{90 \cdot k / 199 \mid k \in [0, 199]\}$ .

## B.2 Evaluation of the James-bound Estimator on Synthetic Data

Fig. 6 reports the estimates and weights produced by the M-bound, James-bound, and standard SC estimators. Both M-bound and James-bound estimators select donor units that are more similar to the target. The M-bound estimator favors donor g50, a unit with individuals most similar to the target. Using the standard SC estimator, donor g20 is preferred, as it has similar outcomes, but different individuals, before the intervention. The James-bound estimator chooses mainly donor g50 with a small selection of donor g20 as it trades off between selecting donors with similar outcomes and similar individuals.

Both the M-bound and James-bound estimators produce misspecification intervals that cover the true outcomes. As expected, the James bound estimator produces a wider misspecification interval than the M bound. The James-bound estimator also produces a better fit for the observed data than the M-bound estimator. This is expected since the M-bound estimator does not consider the pre-intervention outcomes, whereas the James-bound estimator does.



**Figure 6:** Comparison of the M-bound estimator, the James-bound estimator and the standard SC estimator on the synthetic data. Both the James-bound and M-bound estimators produce more accurate counterfactual estimates than the standard SC, despite a poorer pre-intervention fit. The M-bound estimator favors donor g50 (which is the unit with individuals most similar to the target). The standard SC estimator favors donor g20, which has similar outcomes before the intervention but have different individuals. The James-bound estimator trades off and selects mostly g50 with a little of g20. Both the M-bound and James-bound misspecification intervals contain the true outcomes.

## B.3 Using Survey Data to Estimate the Lipschitz Constant

To compute  $\ell$  for the tobacco case study, we leverage external survey data. The (smallest) Lipschitz constant of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is by definition,

$$\inf_{x \neq x'} \frac{|f(x) - f(x')|}{\|x - x'\|_1}.$$

(We use the  $L_1$  norm over  $\mathbb{R}^n$ ).

For the M and James bounds, we need to compute the Lipschitz constant of  $x \mapsto \mathbb{E}_t [Y|x]$ , that is, of the expected tobacco consumption given the causes  $x$ . We use additional survey data from the Tobacco

Use Supplement to the Current Population Survey (TUS-CPS). This study collects individual demographic information along with tobacco consumption. We estimate the expected tobacco consumption given the invariant causes  $x$  and compute the induced  $\ell$  with the formula above by computing the pairwise differences, normalized by the differences of causes  $x$ .

**Handling categorical causes.** For both the Wasserstein distance and the Lipschitz constant, we take  $L_1$  norms over the causes  $x$ . Some causes might be categorical. We represent a categorical variable  $C$  which can take  $k$  values  $c_1, \dots, c_k$  as a one-hot encoding with  $k$  different binary variables  $x_1, \dots, x_k$  with values 0 and  $\frac{1}{2}$ , such that  $C = c_r$  is represented by  $(x_1, \dots, x_k) = (\mathbb{1}(i = r))_{1 \leq i \leq k}$ .

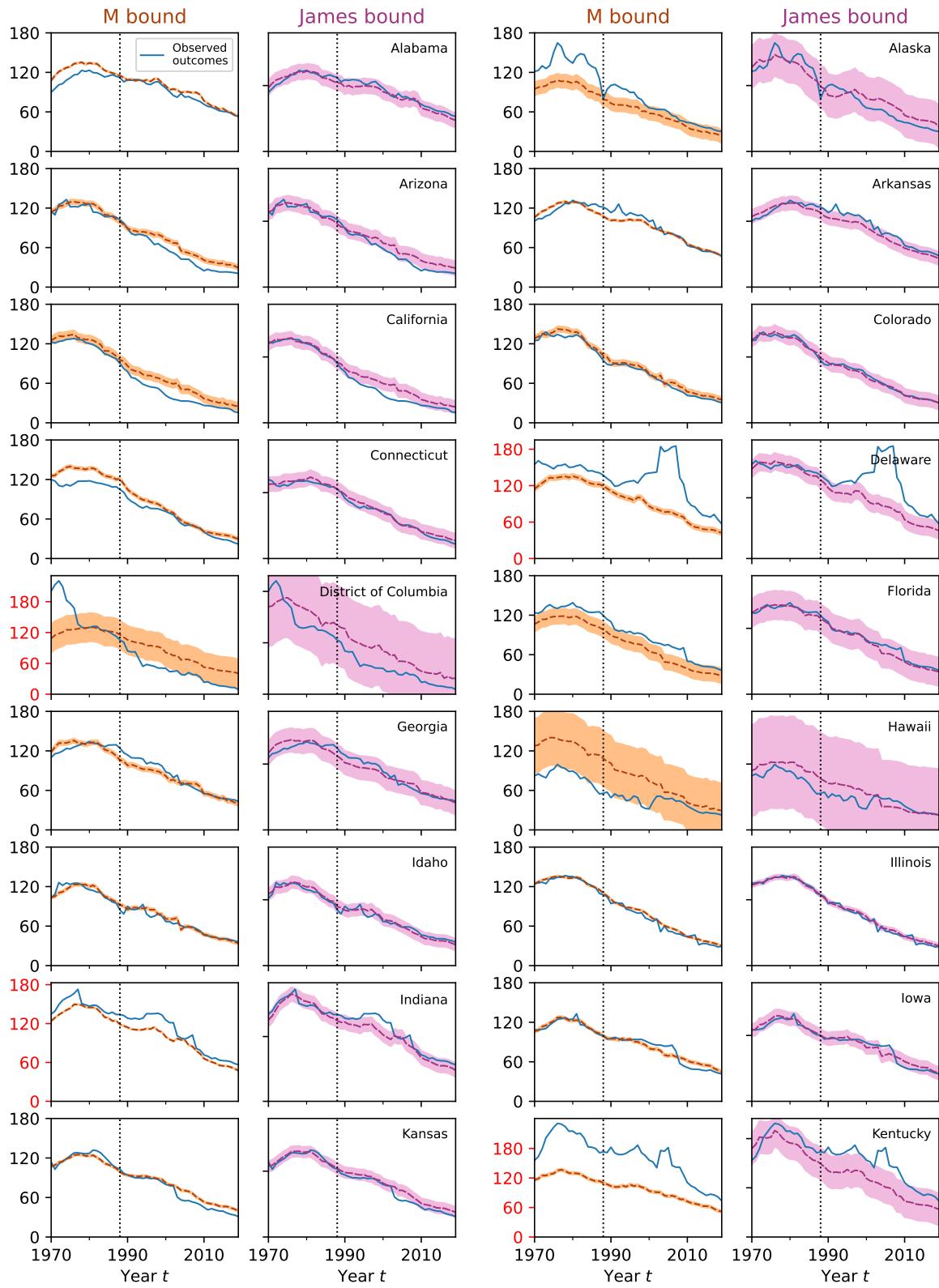
#### B.4 California M-estimator Weights

D.C	Hawaii	Nevada	New Mexico	Texas
0.106	0.166	0.195	0.209	0.324

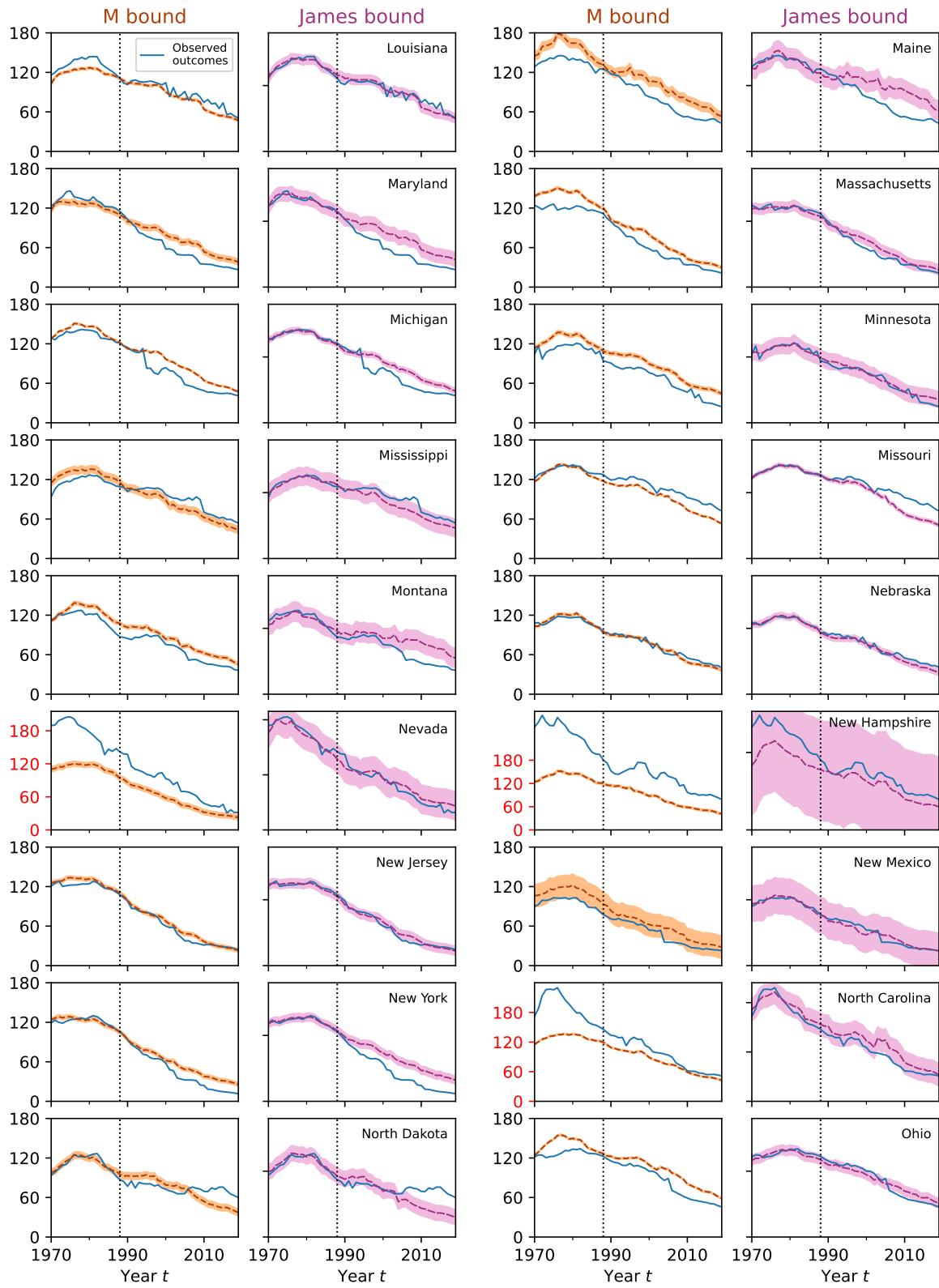
**Table 1:** Non-zero weights returned by the M-bound estimator for the synthetic California of Fig. 2.

#### B.5 Placebo tests

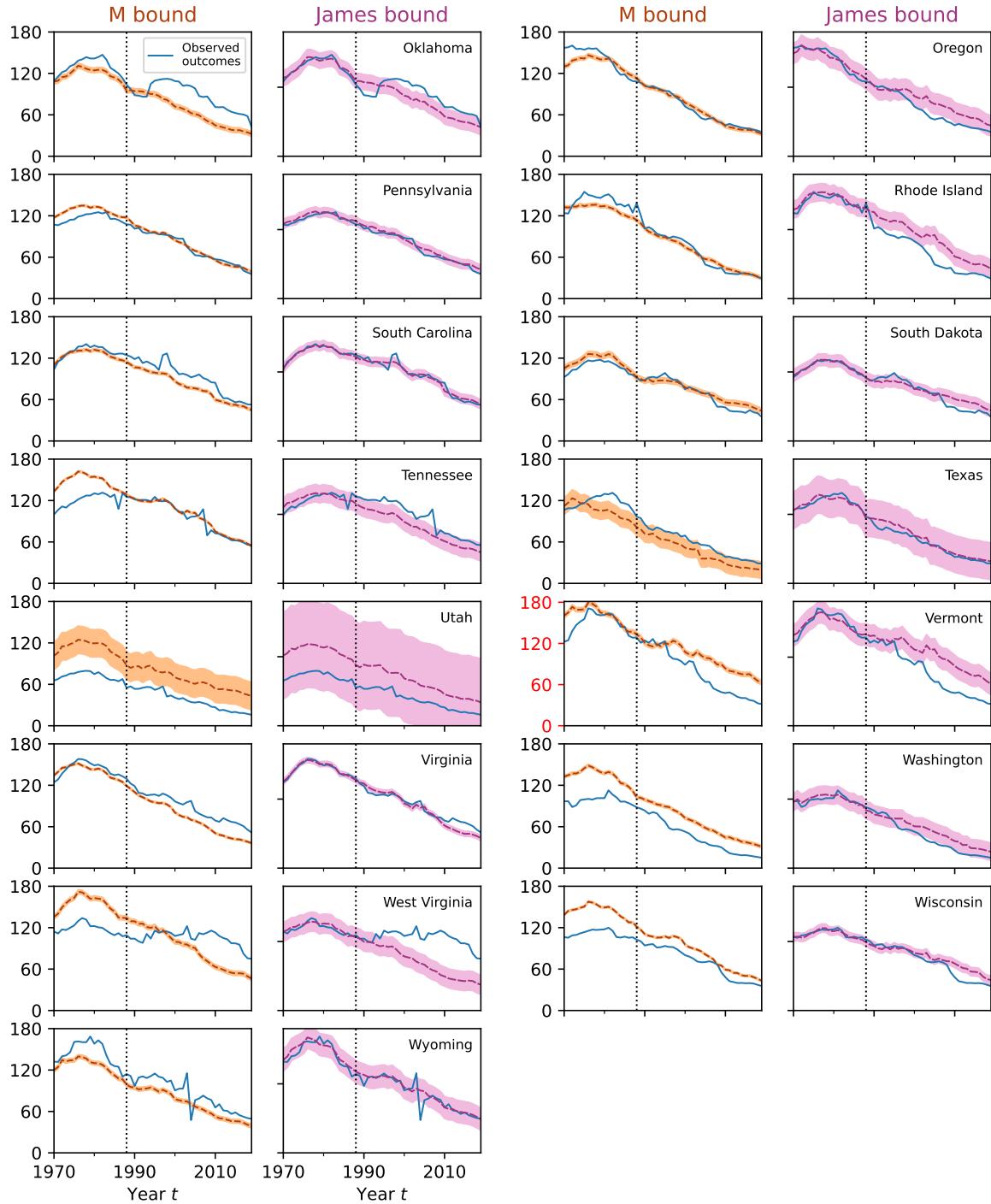
In Figs. 7a to 7c, we report the full placebo study with all the states.



**Figure 7a:** Placebo study, part 1, of the M-bound estimator (left) and the James-bound estimator (right). The y-axis represents the per capita cigarette sales (in packs). The y-axis usually spans from 0 to 180 and is colored in red otherwise.



**Figure 7b:** Placebo study, part 2, of the M-bound estimator (left) and the James-bound estimator (right). The y-axis represents the per capita cigarette sales (in packs). The y-axis usually spans from 0 to 180 and is colored in red otherwise.



**Figure 7c:** Placebo study, part 3, of the M-bound estimator (left) and the James-bound estimator (right). The y-axis represents the per capita cigarette sales (in packs). The y-axis usually spans from 0 to 180 and is colored in red otherwise.