On the Misspecification of Linear Assumptions in Synthetic Controls

Achille Nazaret
Columbia University

Claudia Shi
Columbia University

David Blei
Columbia University

Abstract

The synthetic control (SC) method is popular for estimating causal 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. In practice, this assumption may not hold, and when violated, the resulting SC estimates are incorrect. This paper examines two questions: (1) How large can the misspecification error be? (2) How can we minimize 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 develop new SC estimators specially designed to minimize misspecification error. The estimators are based on additional data about each unit. (E.g., if the units are countries, it might be demographic information about each.) We study our estimators on synthetic data; we find they produce more accurate causal estimates than standard SC. 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 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 method for estimating causal effects from panel data (Abadie, 2021). SC has been widely applied in sciences (Pieters et al., 2016), social sciences (Heersink et al., 2017), and for evaluating public policies (Pinotti, 2015; Allegretto et al., 2017; Donohue et al., 2019). 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 the *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. In 1988, California implemented a policy that increased the tobacco tax by 25 cents. How much would Californians have smoked had the policy not been implemented? The panel data in Fig. 1 (left) contains cigarette sales across states and time. Here, California is the target; the other states are the donors.

The idea behind SC is to approximate the target's control outcomes—the cigarette sales in California without its policy—with a weighted combination of the donor's outcomes. In the example, SC uses data from the pre-policy periods to fit California's pre-policy cigarette sales as a weighted combination of the other states' cigarette sales. It then uses its fitted weights to estimate the counterfactual cigarette sales 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. (2010) shows that SC is justified if the control outcomes follow a linear factor model, in which a per-period factor linearly combines with a per-unit factor, and where the target unit's factor is a linear combination of the donor units' factors. Alternatively, SC can be justified as a matrix completion method where the outcomes form a low-rank matrix where rows (representing the units) are linearly dependent (Athey et al., 2021). But whether as a factor model or a matrix completion method, both justifications point to the same requirement: that the target needs to be expressed as a linear combination of the donors. What if this requirement is not satisfied? What if California is not a *linear*

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

combination of the other states? This paper studies the situation where the linear assumption of synthetic control is *misspecified*. We study how to quantify this misspecification error and how to minimize it.

To understand where the misspecification error may come from and how to bound it, we build on recent work from Shi et al. (2022), which justifies SC and explains why and how a factor model could emerge. When each unit is constituted of multiple individuals (e.g., people within each state), each with individual-level outcomes (e.g., whether they smoke) and with individual traits (e.g., demographics), then the unit-level outcomes (per-capita cigarette sales) follow a factor model. In this model, the per-unit factors have a meaning: they correspond to the probability of individuals' traits in each unit's population. And the linear assumption implies that SC weights can create a *synthetic* mixture of the donors' population distributions to match the target's population distribution.

Using the formulation of Shi et al. (2022), we frame the misspecification as a mismatch between the target and its synthetic population distributions. We then derive two bounds on the SC error, the M bound and the James Bound. Both bounds confines the misspecification error by quantifying the mismatch between the target distribution and its synthetic distribution.

We then consider a situation where we additionally observe external data about the unit's population distribution, such as demographic information about each state. We show how to use such data to estimate the misspecification error for a fixed set of SC weights, and we develop two new algorithms for estimating SC weights that explicitly minimize the amount of error. (One algorithm assumes we have access to the full population distribution; the other does not make that assumption but provides wider misspecification bounds.)

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. Fig. 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 outcomes pre-policy lie within the bounds of misspecification error and that the observed outcomes post-policy lie outside those bounds.

1.1 Related Work

This paper contributes to the growing literature on synthetic controls (Abadie, 2021; Abadie et al., 2010; Abadie and Gardeazabal, 2003). The M-bound and James-bound estimators of § 3 and 4 contribute to research on novel SC estimators (Athey et al., 2021;

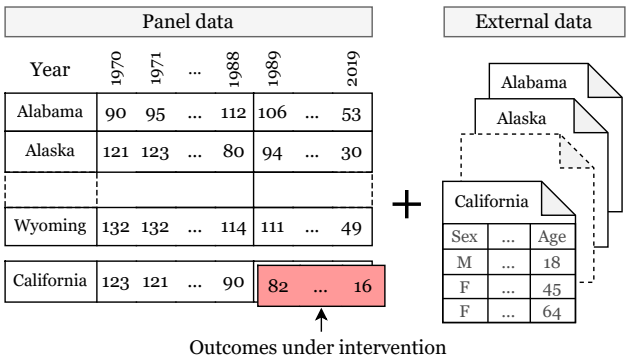


Figure 1: Representation of the data used by our estimators. In addition to panel data of annual per-capita cigarette sales (in packs), our estimators leverage external demographic information about each state.

Abadie and Imbens, 2011; Abadie et al., 2015; Doudchenko and Imbens, 2016; Xu, 2017; Amjad et al., 2018, 2019; Li, 2020; Imbens et al., 2021). Notably, several estimators penalize the optimization objective (Abadie and L'Hour, 2021) or adjust the weights (Kellogg et al., 2021) to select donors with outcomes similar to the target. Including covariates in the estimator is recommended in Abadie and Gardeazabal (2003) but not mathematically justified. Our bounds justify why selecting similar donors is important for estimation.

This paper closely relates to existing works that build robust SC estimators. Some operate under a factor model setup, (Athey et al., 2021; Amjad et al., 2018; Ben-Michael et al., 2021; Powell, 2018; Ferman and Pinto, 2021), while other propose and assume nonlinear models, (Arbour et al., 2021; Chen et al., 2023). However, instead of making additional assumptions, we use external data to quantify the errors empirically.

This paper also contributes to the growing literature that draws connections between other fields and SC. Bottmer et al. (2021) takes an experiment design perspective on SC. Chen (2023) draws a connection between online learning and synthetic control. Shi et al. (2021) formalizes identification and inference for synthetic control through a proximal causal inference perspective. Shi et al. (2022); Zeitler et al. (2023) develops identification results for SC by drawing connections to invariance. We extend these lines of work by drawing connections between available external survey data and bounds on the misspecification error.

2 Model Misspecification in SC

In this section, we first review the classic synthetic control setup and the fine-grained model in Shi et al. (2022). We then state the assumptions of our work

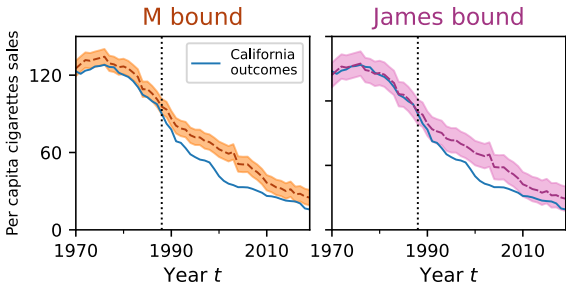


Figure 2: Comparison of California's observed outcomes (solid), our SC estimates (dotted), and the misspecification intervals (shaded), as calculated by the M-bound and James-bound estimators. The outcomes lie within the intervals prior to the intervention, but escape after. This suggests that the tobacco program had a causal effect despite possible misspecification of SC.

and formally define the misspecification error.

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

2.1 Background

For each unit and time, define a pair of potential outcomes $(Y_{jt}, \widetilde{Y}_{jt})$, where $\widetilde{Y}_{jt}$ is the potential outcome in the world where j received intervention at T_0, while Y_{jt} is the potential outcome in a world with no intervention. For $j = 0$ and $t \geq T_0$, we observe the treated potential outcomes ($y_{jt} = \widetilde{Y}_{jt}$); otherwise, we observe the control potential outcomes ($y_{jt} = Y_{jt}$).

A typical assumption in SC is that the control random variable Y_{jt} follows from a linear factor model,

$$Y_{jt} = \mu_j^\top \lambda_t + \epsilon_{jt}, \tag{1}$$

where μ_j is a unit-specific latent factor, λ_t is a time-dependent factor, and ϵ_{jt} is independent noise.

A further assumption is that the target's latent factor is a convex combination of the donors' latent factors[1]. Write Δ^J as the simplex over J coordinates, we have,

$$\exists w \in \Delta^J, \qquad \mu_0 = \textstyle\sum_j w_j \mu_j. \tag{2}$$

SC methods estimate the counterfactual outcomes of the target with a weighted combination of the outcomes of the donors. They use the pre-intervention data to fit the weights,

$$w = \operatorname*{arg\,min}_{w \in \Delta^J} \sum_{t=0}^{T_0 - 1} \Big(y_{0t} - \sum_j w_j y_{jt} \Big)^2. \tag{3}$$

The original SC estimator (Abadie et al., 2010) assumes the weights w to be on the simplex. Other innovations have been proposed to fit w, such as adding regularization (Doudchenko and Imbens, 2016).

2.2 Problem Setup

We consider the setting where the target factor cannot be written as a weighted combination of the donor factors, i.e., Eq. (2) does not hold. This misspecification of Eq. (2) will introduce errors in the SC estimate. To meaningfully bound it, we need to understand where the error could come from. To achieve that, we adopt the fine-grained model in Shi et al. (2022) as a generalization of the linear model Eq. (1).

2.2.1 A Fine-grained Model for SC

Shi et al. (2022) explores the question of when the linear factor model in Eq. (1) hold. 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 $(Y_{ijt}, \tilde{Y}_{ijt})$ denotes the potential outcomes of individual i's outcome in unit j at time t. Shi et al. (2022) assumes each unit's observed outcomes, y_{jt}, is the mean of the individual outcomes in the unit at time t. We write this assumption as A1.

A1. Fine-grained Population Means. The observed outcomes are the population-level expectations,

$$y_{jt} = \begin{cases} \mathbb{E}\left[\widetilde{Y}_{ijt}\right], & \text{if } j = 0 \text{ and } t \geq T_0 \\ \mathbb{E}\left[Y_{ijt}\right], & \text{otherwise.} \end{cases}$$

Given A1, Shi et al. (2022) articulates a set of sufficient assumptions to identify causal effects with SCs. It posits the idea of *invariant causes*, denoted x_{ijt}. The invariant causes x_{ijt} are individual-level variables with 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, $x \mapsto \mathbb{E}\left[Y_{ijt} | x_{ijt} = x\right]$ is the same function for all i and j; we write it $\mathbb{E}_t\left[Y | x\right]$. (2) The distribution of the invariant causes in each unit can change from unit to unit but remain the same across time. It is denoted $p_j(x)$, without dependence on t. The invariance assumptions lead to a fine-grained factor model.

[1] To be precise, Abadie et al. (2010) assumes that $A := \sum_{t<T_0} \lambda_t^\top \lambda_t$ is nonsingular. If $Y_{0t} = \sum_j w_j Y_{jt}$ for $t < T_0$, then the invertability of A implies $\mu_0 = \sum_j w_j \mu_j$.

A2. Fine-grained Factor Model. There exist a set of invariant causes X, such that,

$$\mathbb{E}\left[Y_{jt}\right] = \int_x \mathbb{E}_t\left[Y \mid x\right] p_j(x)\, \mathrm{d}x. \tag{4}$$

Shi et al. (2022) points out that if the causes x are discrete, then Eq. (4) recovers the linear model,

$$\mathbb{E}\left[Y_{jt}\right] = \sum_{x_k} \underbrace{\mathbb{E}_t\left[Y \mid x_k\right]}_{(\lambda_t)_k} \underbrace{p_j(x_k)}_{(\mu_j)_k} = \lambda_t^\top \mu_j. \tag{5}$$

The formulation of Eq. (5) reveals that the time-dependent factors λ_t encode the evolution of the expected outcomes conditional on the causes x, while the unit-specific latent factors μ_j represent information about the population distributions p_j of the causes in each unit. Similarly, the convex assumption on the latent factors μ_j in Eq. (2) translates into a distribution matching assumption on the distributions p_j. We write it Assumption A3.

A3. Convex Combination. The target population is a weighted combination of the donor populations,

$$\exists w \in \Delta^J, \quad p_0 = \sum_j w_j p_j. \tag{6}$$

Hence, the formulation of Shi et al. (2022) with A1 and A2 recovers the standard factor model Eq. (1) and contextualizes its factors. Using continuous distributions in A2 instead of discrete distributions, the fine-grained model is even more general than the factor model. In addition, the key linear combination assumption Eq. (2) of the factor model formulation is now contextualized as A3, a convex combination assumption over the population distributions of each unit. With this SC formulation, we are ready to study misspecification in SC.

2.2.2 Misspecification in SC.

In this paper, we assume the outcomes follow the fine-grained model from Shi et al. (2022), formalized in assumptions A1 and A2. We do not assume A3. We instead study when A3 is violated, in which case the target is not a convex combination of the donors; the synthetic control is *misspecified*. This leads to errors in the estimation of causal effects. The misspecification error is: $\left|\mathbb{E}\left[Y_{0t}\right] - \sum_{j=1}^J w_j \mathbb{E}\left[Y_{jt}\right]\right|$.

Remark 2.1. By assuming A1, we do not assume the observations are noisy, which can differ from other SC analyses. While noise can also impact the SC estimation, we focus instead on the error due to misspecification. As explained in Shi et al. (2022), the individual-level outcomes are indeed random, but as we increase the number of individuals in each unit, the noise from individual-level outcomes disappears at the unit level.

3 The M Bound and its Estimator

In this section, we derive an exact bound that quantifies errors induced by the violation of A3. Using this bound, we develop an estimator minimizing A3 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, $\mathbb{E}\left[Y_{0t}\right] \approx \sum_{j=1}^J w_j \mathbb{E}\left[Y_{jt}\right]$. We formalize this intuition by bounding the errors resulting from the misspecification of A3.

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

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

where $\hat{p}_0 = \sum_j w_j p_j$.

The proof is in Appendix A.

The Wasserstein distance W_1 is a distance between probability distributions (Villani, 2009). 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) confines the error of the SC estimate with a function of the weights, the population distributions p_j of each unit, and the sensitivity of the outcome variables to the variation of the causes (the Lipschitz constant ℓ).

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. (7) 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 ℓ). Hence, the M bound proves (theoretically) that a small misspecification induces a small estimation error.

Algorithm 1 Minimization of the M bound

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

3.2 The M-bound Estimator

We established the M bound, which quantifies the misspecification error for any set of weights. 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. The weights are learned using projected gradient descent with the following objective, $(w_1, ...w_J) \mapsto W_1(p_0, \sum_j w_j p_j)$. Notice that it computes SC weights using only the distributions p_j. It does not use outcome data.

After obtaining a set of weights from Algorithm 1, we can use Eq. (7) with an estimated constant ℓ to create a misspecification interval around the SC estimate,

$$\mathbb{E}[Y_{0t}] \in [\hat{y}_{0t} - M, \hat{y}_{0t} + M] \qquad \forall t, \tag{8}$$

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

In § 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 A3 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 § 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 it is computed with 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 can be estimated with only partially observed causes, such that it can be used in practice. The 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 or 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. (4) becomes $\mathbb{E}[Y_{jt}] = \int_{(x,z)} p_j(x,z)\mathbb{E}_t[Y|x,z]\,\mathrm{d}x\mathrm{d}z$.

We cannot generally 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 A4. We note that standard the linear model Eq. (1), which we generalize with the fine-grained model Eq. (4), is still more restrictive than A4.

A4. Independence of Observed & Unobserved. For each unit, x and z are independent, and there exist functions g_t and h_t such that:

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

We note that the distributions of the observed causes $p_j(x)$ and the unobserved causes $p_j(z)$ remain arbitrary, and so are g_t and h_t. With A4, we have "just another misspecification error" (James) bound.

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

$$\left|\mathbb{E}[Y_{0t}] - \textstyle\sum_{j=1}^{J} w_j \mathbb{E}[Y_{jt}]\right| \leq \ell \cdot W_1(p_0(x), \hat{p}_0(x)) \tag{9}$$

$$+ \max_{u<T_0} \left|\mathbb{E}[Y_{0u}] - \textstyle\sum_{j=1}^{J} w_j \mathbb{E}[Y_{ju}]\right| \tag{10}$$

$$+ \inf_{\alpha \in \Delta^{T_0}} \left| \int_z \Big(p_0(z) - \hat{p}_0(z)\Big)\Big(\mathbb{E}_t[Y|z] - \sum_{u<T_0} \alpha_u \mathbb{E}_u[Y|z]\Big)\,\mathrm{d}z \right|. \tag{11}$$

The proof is in Appendix A.

The first term (9) mirrors the M bound. It quantifies the similarity between the target and synthetic distributions of observed causes, $p_0(x)$ and $\hat{p}_0(x)$.

The second term (10) measures the fit of the pre-intervention outcomes. It indirectly estimates the similarity between the target and the synthetic distributions of unobserved causes, $p_j(z)$ and $\hat{p}_j(z)$.

The last term (11) 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. For instance, we show that it is zero in the standard SC factor model and in two other models.

4.2 The James-bound Estimator

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

$$w \mapsto \max_{t<T_0} \left| y_{0t} - \sum_{j=1}^{J} w_j y_{jt} \right| + \lambda \cdot W_1\left(p_0, \textstyle\sum_j w_j p_j\right), \quad (12)$$

where λ is a hyperparameter. We update Algorithm 1 to minimize this objective in Appendix A.

If hyperparameter λ is set to ℓ, and if term (11) is effectively negligible, then Eq. (12) 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 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 invariant causes. James bounds are wider but require less 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 A3 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 can be computed only 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. We provide implementation details in Appendix B.1. We provide code at this address: https://github.com/blei-lab/synthetic_controls.

5.1 Experiments with Synthetic Data

Data Description. We generate synthetic data by defining the conditional distribution $\mathbb{E}_t[Y|x] = f(x,t)$ 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 $(t,x) \mapsto \mathbb{E}_t[Y|x]$ and $(j,x) \mapsto p_j(x)$ are in Appendix B.2 while Fig. 3 shows the evolution of $x \mapsto \mathbb{E}_t[Y|x]$ over time t and the distributions $p_j(x)$ for each unit j. The expected outcome $\mathbb{E}_t[Y|x]$ 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. (3). We report the weights and the induced synthetic outcomes in Fig. 4. Furthermore, we compute $\ell = 4.0$ from $x \mapsto \mathbb{E}_t[Y \mid x]$ (valid for all t). This way, we obtain the exact value

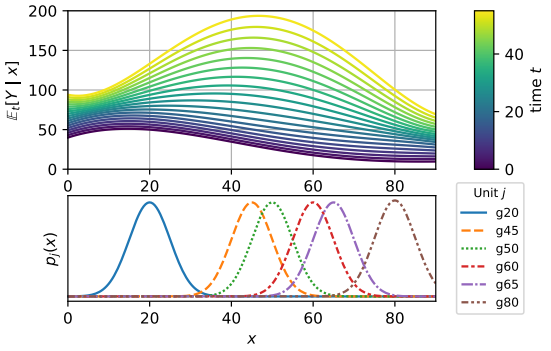


Figure 3: Visualization of the synthetic data generating process. (top) Each line represents the expected conditional outcome $\mathbb{E}_t[Y|x]$ for a different time t, as a function of the cause x. As time progresses (from darker to lighter), the expected conditional outcomes increase for all values of x, but with different rate of increase over time. (bottom) Distributions densities of the causes $x \mapsto p_j(x)$ for each unit j. The target unit is g45, which overlaps mostly with unit g50.

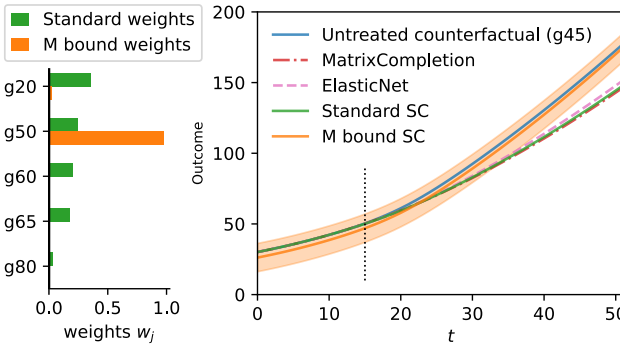


Figure 4: Comparison of the M-bound estimator and the standard SC estimator on synthetic data. (left) Weights returned by each estimator. The M-bound estimator selects donors (g20 and g50) that are most similar to the target (g45). (right) Synthetic outcomes of each estimator, compared to the true outcome. Under misspecification, the M-bound estimator provides more accurate estimates than the standard SC, despite a poorer pre-intervention fit.

of the M-bound and we can form the misspecification interval of Eq. (8), shaded on Fig. 4.

Analysis. As shown in Fig. 4, the standard SC places a large weight on donor g20, which is a unit whose individuals are very different from g45 but with similar pre-treatment outcomes. When time increases, the individuals in g20 and g45 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 g50, which contains individuals with similar x as the target g45. By doing so, the synthetic outcomes might not exactly match the g45 outcomes, but they generalize better over time. We also verify that the true outcome is always contained in the misspecification interval (Eq. (8)). We report the same conclusions in Appendix B with the James bound.

Fig. 4 further includes two more baselines: the penalized ElasticNet SC (Doudchenko and Imbens, 2016) and the matrix completion SC method (Athey et al., 2021). Both methods form incorrect SCs that are similar to the standard SC. It is not surprising since both methods assume the treated unit is a linear combination of the untreated units.

With external data, we estimated the misspecification error and limited it using the M-bound and James-bound estimators. Without external data, standard SC made incorrect predictions, so as did more robust SC methods assuming a linear factor model.

5.2 A Case Study on Real Data

We revisit the tobacco study from Abadie et al. (2010) to illustrate how to collect external data, apply the estimators, and calculate misspecification intervals.

Prop 99. A tobacco control program was passed in California in 1988. It increased tobacco taxes by 25 cents and funded 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 (2019), 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. (2004), 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 (Sakuma et al., 2016; Cornelius et al., 2022). 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 (Census Bureau, 2020). It reports population demographics at different geographical scales, from city boroughs to states. We accessed the ACS data with the Census Reporter API (Census Reporter, 2020). 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 ℓ 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 ℓ. More details about the computation of Lipschitz constant can be found in Appendix B.

The M-bound Estimator. The M-bound estimator uses our newly formed $p_0, ..., p_J$ to compute a set of SC weights. We report the weights in Appendix B.5 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 are 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. Despite the misspecification, Fig. 2 shows the post-intervention outcomes are outside of the bounds, suggesting a causal effect.

The James-bound Estimator. As discussed in § 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* (Abadie et al., 2010) 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.

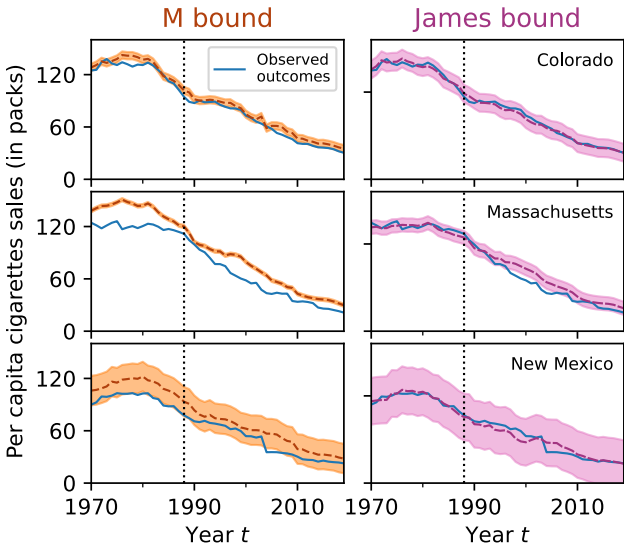


Figure 5: Placebo study of the M-bound estimator (left) and the James-bound estimator (right), on Colorado, Massachusetts, and New Mexico. The M-bound synthetic outcomes are outside of the misspecification interval before the intervention. This suggests that not all invariant causes are observed and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

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 SC.

Fig. 5 (right) shows the synthetic outcome estimates using the James-bound estimator. We observe the pre-intervention outcomes across states now all fall within the James-bound misspecification intervals, 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 as Colorado did not implement anti-tobacco programs 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 program in California had a true causal effect on tobacco consumption. In states without tobacco programs, their outcomes fall within the misspecification interval, whereas California's outcome does not.

Conclusion of the case-study. With the Jamesbond estimator, we confirm the conclusions of Abadie et al. (2010) but now, they come with one extra important guarantee: the discovered causal effect is robust to linear misspecification.

References

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59:391–425, 2021.

Hannah Pieters, Daniele Curzi, Alessandro Olper, and Johan Swinnen. Effect of democratic reforms on child mortality: a synthetic control analysis. *The Lancet Global Health*, 4:e627–e632, 2016.

Boris Heersink, Brenton D Peterson, and Jeffery A Jenkins. Disasters and elections: Estimating the net effect of damage and relief in historical perspective. *Political Analysis*, 25:260–268, 2017.

Paolo Pinotti. The economic costs of organised crime: Evidence from southern italy. *The Economic Journal*, 125:F203–F232, 2015.

Sylvia Allegretto, Arindrajit Dube, Michael Reich, and Ben Zipperer. Credible research designs for minimum wage studies: A response to Neumark, Salas, and Wascher. *ILR Review*, 70:559–592, 2017.

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. *Journal of Empirical Legal Studies*, 16:198–247, 2019.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105:493–505, 2010.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, pages 1–41, 2021.

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.

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93:113–132, 2003.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29:1–11, 2011.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59:495–510, 2015.

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

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

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19:802–852, 2018.

Muhammad Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. mrsc: Multi-dimensional robust synthetic control. *ACM on Measurement and Analysis of Computing Systems*, 3:1–27, 2019.

Kathleen T Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115:2068–2083, 2020.

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. *arXiv:2108.03849*, 2021.

Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116:1817–1834, 2021.

Maxwell Kellogg, Magne Mogstad, Guillaume A Pouliot, and Alexander Torgovitsky. Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American Statistical Association*, 2021.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116:1789–1803, 2021.

David Powell. Imperfect synthetic controls: Did the massachusetts health care reform save lives? *Available at SSRN 3192710*, 2018.

Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12:1197–1221, 2021.

David Arbour, Eli Ben-Michael, Avi Feller, Alex Franks, and Steven Raphael. Using multitask gaussian processes to estimate the effect of a targeted effort to remove firearms. *arXiv preprint arXiv:2110.07006*, 2021.

Yehu Chen, Annamaria Prati, Jacob Montgomery, and Roman Garnett. A multi-task gaussian process model for inferring time-varying treatment effects in panel data. In *Artificial Intelligence and Statistics*, pages 4068–4088. PMLR, 2023.

Lea Bottmer, Guido Imbens, Jann Spiess, and Merrill Warnick. A design-based perspective on synthetic control methods. *arXiv:2101.09398*, 2021.

Jiafeng Chen. Synthetic control as online linear regression. *Econometrica*, 91(2):465–491, 2023.

Xu Shi, Wang Miao, Mengtong Hu, and Eric Tchetgen Tchetgen. On proximal causal inference with synthetic controls. *arXiv:2108.13935*, 2021.

Jakob Zeitler, Athanasios Vlontzos, and Ciarán Mark Gilligan-Lee. Non-parametric identifiability and sensitivity analysis of synthetic control models. In *Causal Learning and Reasoning*, pages 850–865. PMLR, 2023.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Centers for Disease Control and Prevention. The tax burden on tobacco, 1970-2019, 2019. URL https://chronicdata.cdc.gov/Policy/The-Tax-Burden-on-Tobacco-1970-2018/7nwe-3aj9.

Lindsey Turner, Robin Mermelstein, and Brian Flay. Individual and contextual influences on adolescent smoking. *Annals of the New York Academy of Sciences*, 2004.

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. *Journal of Preventive Medicine*, 2016.

Monica E Cornelius, Caitlin G Loretan, Teresa W Wang, Ahmed Jamal, and David M Homa. Tobacco product use among adults—united states, 2020. *Morbidity and Mortality Weekly Report*, 71: 397, 2022.

Census Bureau. 2016-2020 5-year American Community Survey, 2020.

Census Reporter. Census reporter: Making census data easy to use, 2020. URL https://censusreporter.org/.

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. *Journal of Machine Learning Research*, 2021.

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*, pages 8024–8035. Curran Associates, Inc., 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] The algorithms are not computationally intensive. We still details efficient implementations in Appendix
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Technical Details

A.1 Lipschitz function

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

$$\forall (x, y) \in (\mathbb{R}^n)^2, \quad |f(x) - f(y)| \leq \ell \cdot \|x - y\|_1 \tag{13}$$

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 ℓ is small, it implies that a change in the causes induces a small change in the outcomes. Having a small ℓ 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 ℓ-Lipschitz, then for any weights in the simplex w, we have the Misspecification error bound (M-bound):

$$\left| \mathbb{E}[Y_{0t}] - \sum_{j=1}^{J} w_j \mathbb{E}[Y_{jt}] \right| \leq \ell \cdot W_1(p_0, \hat{p}_0), \tag{14}$$

where $\hat{p}_0 = \sum_j w_j p_j$ and W_1 is a ℓ_1-Wasserstein distance.

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

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

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

where the inequality comes from Kantorovich duality (Theorem 5.10, Villani (2009)) about Wasserstein distances. □

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 ℓ-Lipschitz, then for any weights in the simplex $w \in \Delta^J$, we have Just Another Misspecification Errors bound (James bound) :

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

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

Lemma A.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 $f_t(x,z) = \mathbb{E}_t[Y \mid x,z]$. 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

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

We notice that $\int_z h_t(z)p_j(z)\mathrm{d}z$ 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 $\int_z g_t(x)p_j(x)\mathrm{d}x$. □

We can now prove the James bound.

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

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

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

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

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

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

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

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

So,

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

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

We obtain,

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

This proves the James bound. □

A.4 Interpretation of Eq. (11)

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.
- $\max_{u<T_0} \Big| \mathbb{E}\left[Y_{0u}\right] - \textstyle\sum_j w_j \mathbb{E}\left[Y_{ju}\right] \Big| \mathrm{d}z$ is estimated from the outcome data.

The last term: $\inf_{\alpha \in \Delta^{T_0}} \left| \int_z \Big(p_0(z) - \sum_j w_j p_j(z) \Big) \Big(\mathbb{E}_t\left[Y|z\right] - \sum_{u<T_0} \alpha_u \mathbb{E}_u\left[Y|z\right] \Big) \mathrm{d}z \right|$ 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 $\mathbb{E}_t\left[Y|z\right] - \sum_{u<T_0} \alpha_u \mathbb{E}_u\left[Y|z\right] = \textit{constant}$.

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 $\mathbb{E}_t\left[Y|z\right] - \sum_{u<T_0} \alpha_u \mathbb{E}_u\left[Y|z\right]$ 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 α that makes the full integral close to zero should be possible.

More concretely, we give two examples of models for which the term (11) 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 μ_j, such that $\mu_0 = \sum_j w_j \mu_j$). In particular, it implies that $\Big(p_0(z) - \sum_j w_j p_j(z) \Big) = 0$ and so (11) $= 0$.

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 $(z \mapsto \mathbb{E}_t[Y \mid z])_t$ are of the form $\mathbb{E}_t[Y \mid z] = \beta_t^\top z$, with β_t being able to change arbitrarily over time. Actually, we even requires the β_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, (11) $= 0$.

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 $p_0, ..., p_J$; Pre-intervention measurements $\{(y_{0t}, ..., y_{jt})_{t=0,...,T_0-1}\}$; learning rate α; number of epochs E, James parameter λ.
Output: (w_j) minimizing the James-bound.
$(w_1, ..., w_J) \leftarrow (\frac{1}{J}, ..., \frac{1}{J})$
for $e = 1$ **to** E **do**
  $\hat{p}_0 \leftarrow \sum w_j p_j$
  grad $\leftarrow \nabla_w \left(\max_{t<T_0} |y_{0t} - \sum_{j=1}^{J} w_j m_{jt}| + \lambda \cdot W_1(p_0, \hat{p}_0) \right)$
  $w \leftarrow w - \alpha \cdot$ grad
  $w \leftarrow$ project_simplex(w)
end for
return w

B Experiment Details

B.1 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 δ_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 $(w_1, ..., w_J) \mapsto W_1(p_0, \sum_j w_j p_j)$, using the Python Optimal Transport library (Flamary et al., 2021) coupled with PyTorch (Paszke et al., 2019). We use gradient descent with a learning rate $\alpha = 5 \cdot 10^{-6}$ and $200,000$ epochs.

B.2 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:

$$
\begin{aligned}
f(x,t) = &-\frac{13t^2x^4}{2100000000} + \frac{71t^2x^3}{78750000} - \frac{10141t^2x^2}{252000000} + \frac{12521t^2x}{12600000} + t + 4.07142857142857 \\
&\cdot 10^{-6}x^4 \log\left(e^{\frac{t}{3}-\frac{20}{3}} + 1\right) - \frac{43x^4}{13300000} - 0.00073142857142857 1x^3 \log\left(e^{\frac{t}{3}-\frac{20}{3}} + 1\right) \\
&+ \frac{1313x^3}{1496250} + 0.0380053571428571x^2 \log\left(e^{\frac{t}{3}-\frac{20}{3}} + 1\right) - \frac{359953x^2}{4788000} \\
&- 0.500107142857143x \log\left(e^{\frac{t}{3}-\frac{20}{3}} + 1\right) + \frac{401813x}{239400} + 40\,.
\end{aligned}
$$

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

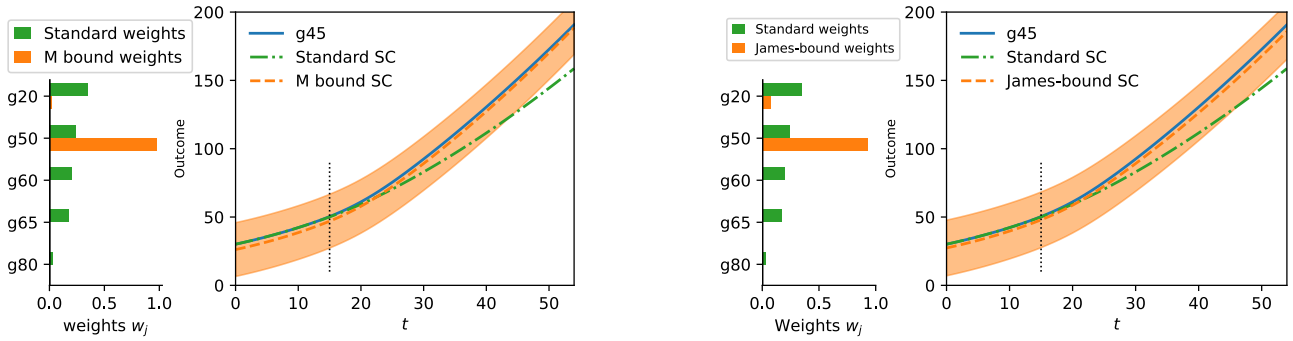
$$p_j \propto \sum_{x \in X} \delta_x \cdot \mathcal{N}(x; \mu_j, 5^2)$$

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

B.3 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.



(a) M-bound estimator.

(b) James-bound estimator.

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.4 Using Survey Data to Estimate the Lipschitz Constant

To compute ℓ for the tobacco case study, we leverage external survey data. The (smallest) Lipschitz constant of a function $f : \mathbb{R}^n \to \mathbb{R}$ 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 ℓ 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 $c_1, ..., c_k$ as a one-half-hot encoding with k different binary variables $x_1, ..., x_k$ with values 0 and $\frac{1}{2}$, such that $C = c_r$ is represented by $(x_1, ...x_k) = (0.5 \cdot \mathbb{1}(i = r))_{1 \leq i \leq k}$. The 0.5 is so that the L1 distance between two different encoding is either 1 or 0.

B.5 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.6 Placebo tests

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

Estimation error for ℓ. We discuss practical consequences of estimation error for ℓ. If ℓ has estimation errors, then the misspecification intervals, which are functions of ℓ, might have errors too. Nevertheless, we can understand and limit this error.

First, the M-and-James bounds/intervals are always defined and they hold for any set of weights, no matter how the weights were obtained (see Bound 2). That is, each set of weights has a valid, theoretical, misspecification interval; defined with the true ℓ. For e.g., if the set of weights was obtained by optimizing the James bound with an incorrect ℓ, then those weights will have a true misspecification interval, simply not the tightest one.

However, computing that interval will incur an error due to ℓ. Fortunately, the estimators are "robust" in the sense that this error is controlled: 5% error on ℓ will give at most 5% error on the interval (see Bound 2). An overestimated ℓ will provide a looser bound, and the interval might be too wide to conclude a causal effect. But the conclusion would remain correct: there is too much misspecification/estimation error, which prevents us from finding a possible causal effect.

To prevent for an under-estimated ℓ, we recommend favoring overestimation of ℓ over understimation.

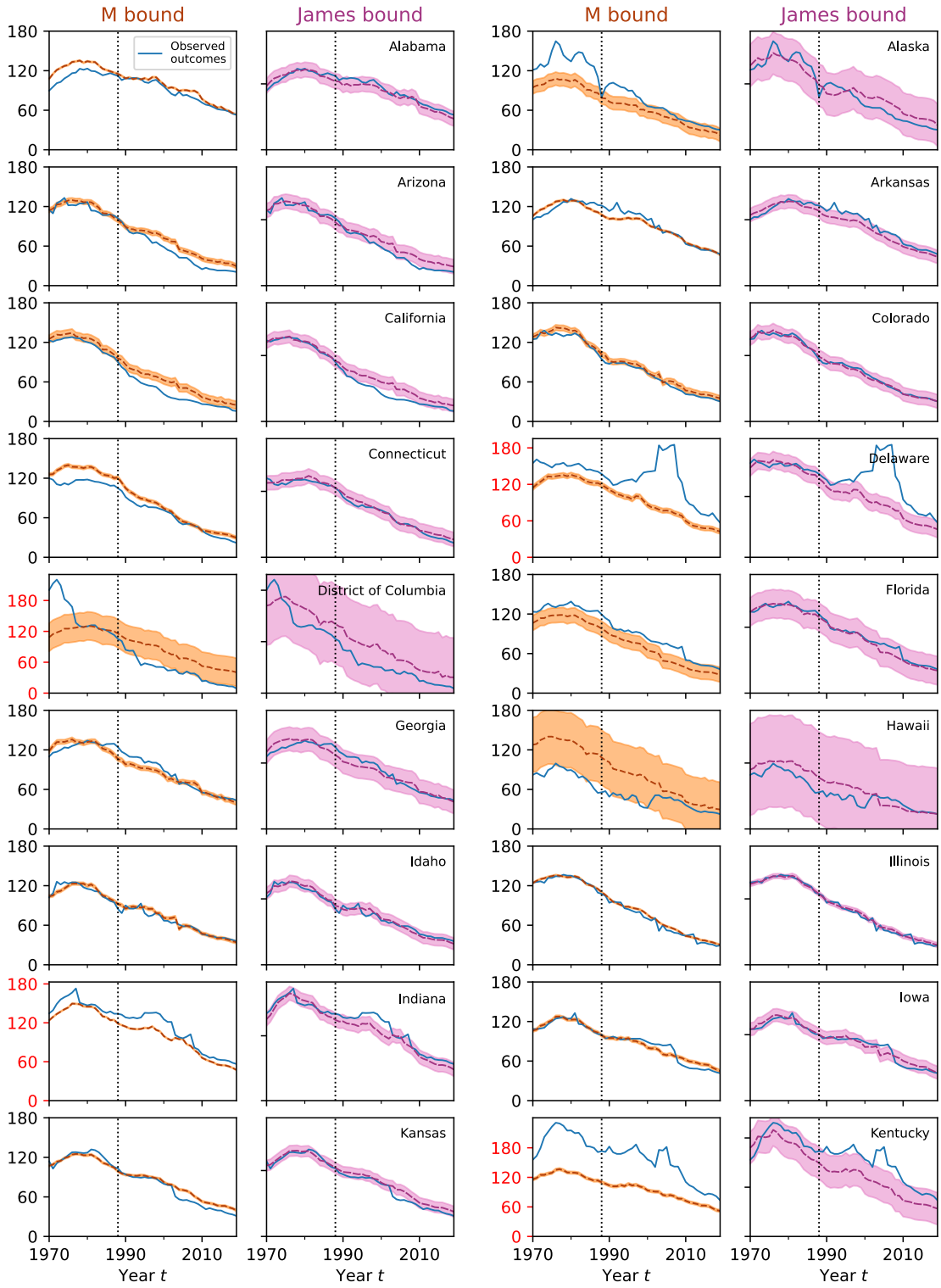


Figure 7a: Placebo study, part 1, of the M-bound estimator (left) and the James-bound estimator (right). The y-axis represents the per capita cigarette sales (in packs). The y-axis usually spans from 0 to 180 and is colored in red otherwise.

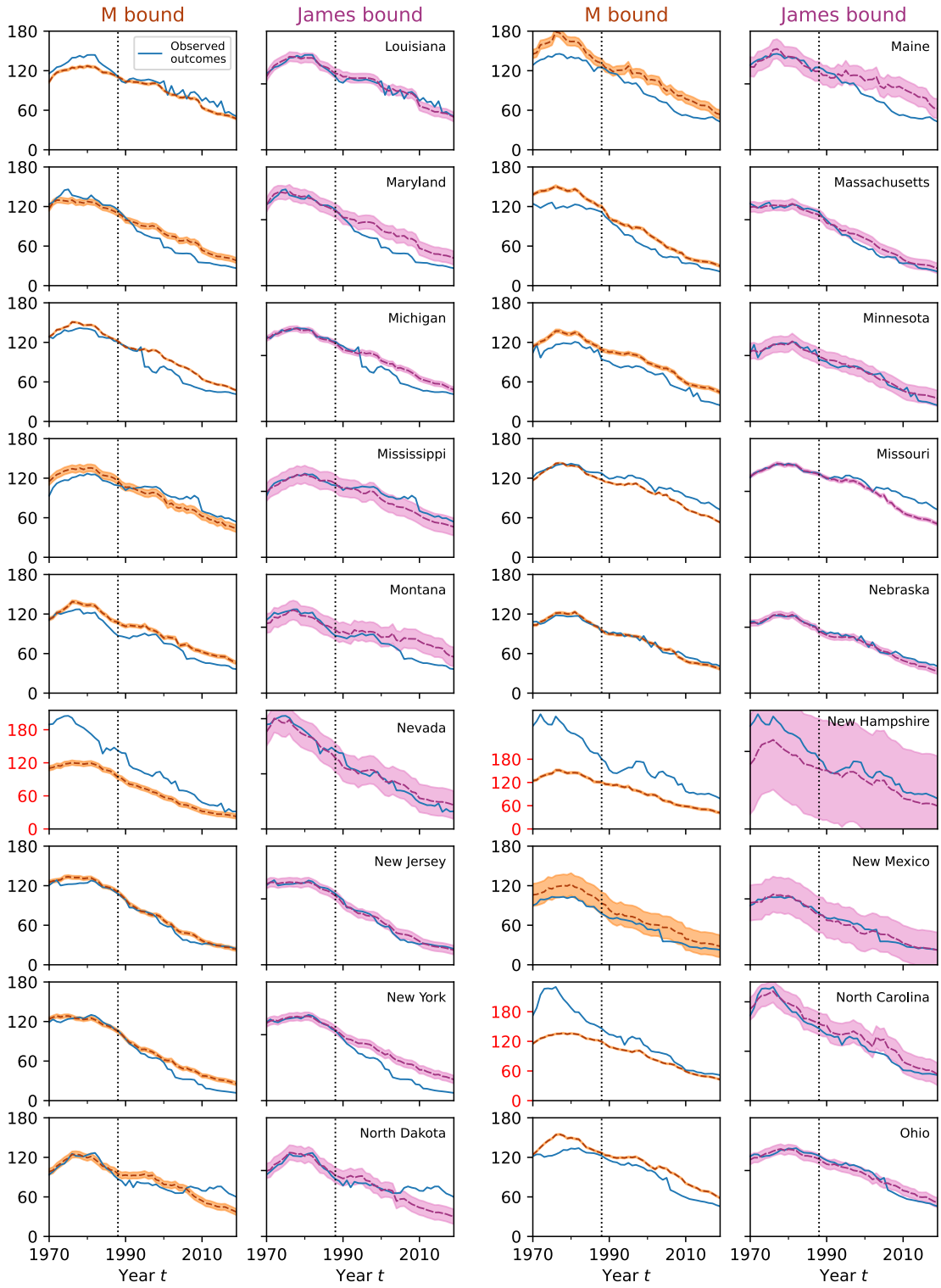


Figure 7b: Placebo study, part 2, of the M-bound estimator (left) and the James-bound estimator (right). The y-axis represents the per capita cigarette sales (in packs). The y-axis usually spans from 0 to 180 and is colored in red otherwise.

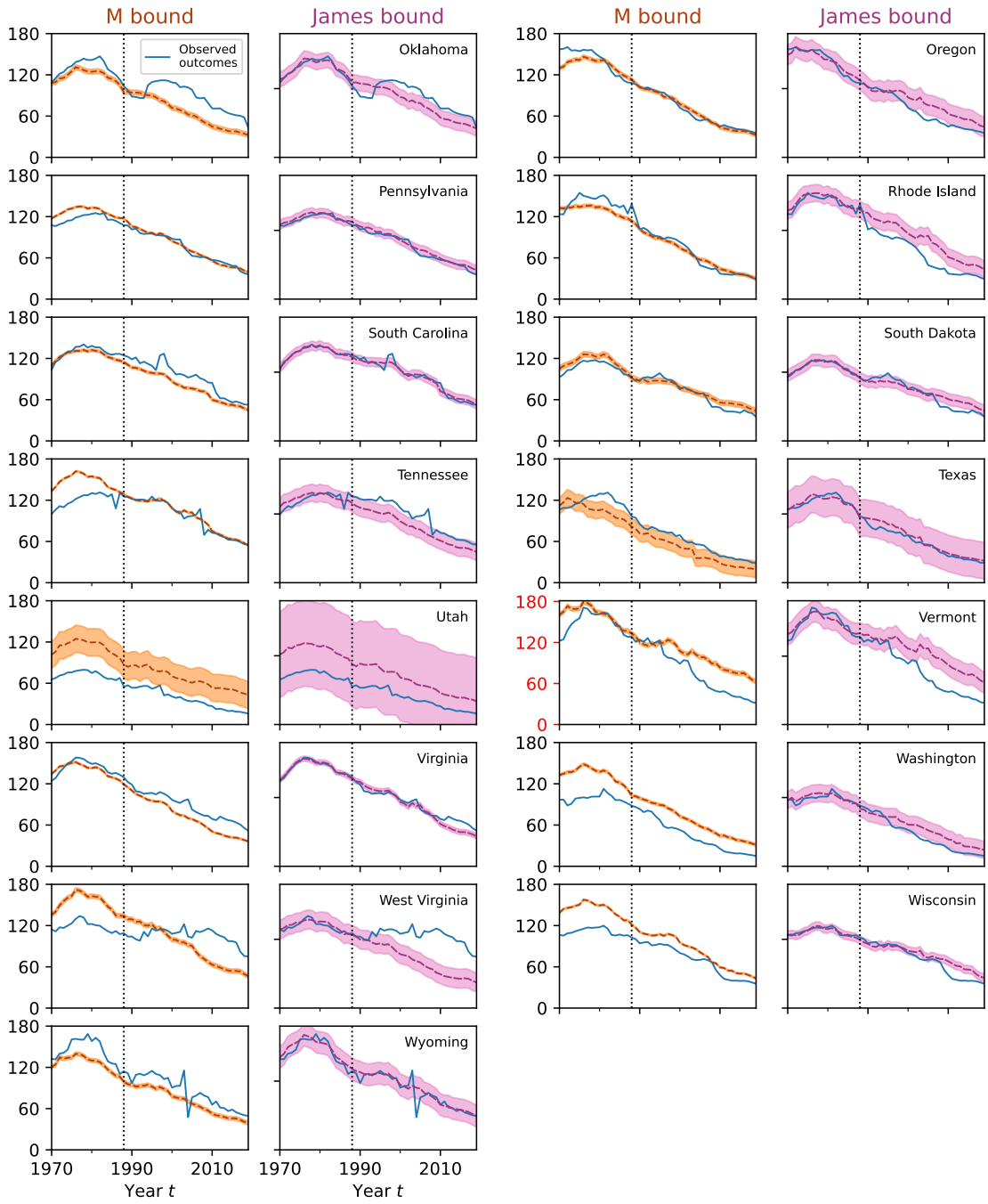


Figure 7c: Placebo study, part 3, of the M-bound estimator (left) and the James-bound estimator (right). The y-axis represents the per capita cigarette sales (in packs). The y-axis usually spans from 0 to 180 and is colored in red otherwise.

The James bound's objective function $J(w;\lambda)$ requires a hyperparameter λ. To choose the hyperparameter λ, we recommend using survey data to compute the Lipschitz constant ℓ (as described in Appendix B.3) and then set $\lambda \leftarrow \ell$. This is necessary to compute the James bound, which involves $J(w;\lambda)$ with $\lambda = \ell$. If external survey data is unavailable for evaluating ℓ, we suggest the following cross-validation method based on the placebo test.

First, notice that for any $\lambda \geq \ell$, we have $J(w;\lambda) \geq J(w;\ell)$, so the James bound (Bound 2) is also valid for any $\lambda \geq \ell$:

$$|\mathbb{E}[Y_{0t}] - \hat{y}_{0t}| \leq J(w;\lambda),$$

where $\hat{y}_{0t} = \sum_{j=1}^{J} w_j \mathbb{E}[Y_{jt}]$ are the synthetic outcomes.

Hence, once one uses $\lambda \geq \ell$, the untreated potential outcomes $\mathbb{E}[Y_{0t}]$ for $t > T_0$ must fall inside the misspecification interval $[\hat{y}_{0t} - J(w;\lambda), \hat{y}_{0t} + J(w;\lambda)]$. But note that if λ is too large, the misspecification interval might become too wide to detect any causal effect for the target. On the other hand, if $\lambda < \ell$, the James bound is not guaranteed to hold and might just be incorrectly too narrow. This can be checked by *placebo* analysis (i.e. *cross-validation*). Since we observe the untreated outcomes of the donors, we can find a λ that is just large enough so that all the donors have their observed outcomes inside the bounds. We then use such a λ for our target unit and check for a causal effect.

In practice, we stop when at least a proportion α (e.g., 0.95) of all the donor outcomes is contained in the bounds. If the target is then outside its bound, we can conclude in favor of a causal effect.

Algorithm 3 Cross validation for selecting λ.

Input: Set Λ of possible λ in increasing order, threshold α (e.g. $\alpha = 0.95$).
Output: Whether a causal effect is detected.
for each donor $j > 0$ **do**
  for $\lambda \in \Lambda$ **do**
    Compute $w^* = \arg\min_w J(w;\lambda)$ where the target is j.
    Compute synthetic outcomes: $\hat{y}_{j,t} = \sum_k w_k^* y_{kt}$.
    Compute $b_{\lambda,j,t} \leftarrow \mathbb{1}(y_{kt} \in [\hat{y}_{kt} - J(w^*,\lambda), \hat{y}_{kt} + J(w^*,\lambda)])$
  end for
  if $\text{mean}(\{b_{\lambda,j,t} \mid t > T_0, j \in \{1..J\}\}) \geq \alpha$ **then**
    return λ
  end if
end for
return You should consider larger λ.