INFERENCE FOR SYNTHETIC CONTROLS VIA REFINED PLACEBO TESTS

Timothy Sudijono*, Lihua Lei†

September 25, 2023

Abstract

The synthetic control method is often applied to problems with one or a few treated units and a small number of control units. Inference procedures that are justified asymptotically are often unsatisfactory due to (1) small sample sizes that render large-sample approximation fragile and (2) simplification of the estimation procedure that is actually implemented in practice. An alternative is design-based inference, which is closely related to the placebo test, a widely used diagnostic tool in practice. It provides valid Type-I error control in finite samples without artificial simplifications of the method when the treatment is assigned uniformly among units. Despite the robustness, it suffers from low resolution since the null distribution is constructed from only N reference estimates, where N is the sample size. Inspired by a connection to the conformal inference literature, we propose a novel leave-two-out procedure that bypasses this issue, providing $O(N^2)$ reference estimates while still maintaining the finite-sample Type-I error control under uniform assignments. Unlike the placebo test whose Type-I error always equals the theoretical upper bound, our procedure often achieves a lower Type-I error than theory suggests and a higher power when the effect size is reasonably large. To account for deviation from uniform assignments, we generalize our procedure to allow for non-uniform assignments and show how to conduct sensitivity analysis based on quadratic programming.

*Department of Statistics, Stanford University: 390 Jane Stanford Way, Stanford, CA 94305, USA. Email: tsudijon@stanford.edu

†Graduate School of Business, Stanford University

1 Introduction

Synthetic control methods were first introduced by Abadie and Gardeazabal [AG03] to analyze the effects of terrorism in the Basque Country on the economy of the region. Since then, the method has developed into a powerful tool in comparative case studies with very few treated units, used widely in practice. The synthetic control framework is typically used in panel data settings with one treated unit; suppose we observe data with N units, T time periods, where the Ith unit is treated from time $T_0, \dots, T$.

The synthetic control method models the treated unit by a *synthetic control*, a convex combination of the other control units which resembles the treated unit as closely as possible according to pre-treatment covariates. That is, we seek a vector of weights $W = (w_1, \dots, w_{I-1}, w_{I+1}, \dots, w_N)$ such that $w_j \geq 0$ for all j and $\sum_{j=2}^{N} w_j = 1$, which corresponds to a weighted average of control units. The method chooses the weights W by minimizing the norm

$$\|X_I - X_0 W\|_V = \sqrt{(X_I - X_0 W)^\top V (X_I - X_0 W)}$$

where X_I are covariates for the treatment unit and X_0 is a matrix whose rows are covariates for the control units. V is a matrix which weights the different covariates, which may be either pre-specified or learned from the data [ADH11]. Importantly, in many flagship applications of the method, there is one treated unit and a small number of control units. Moreover, the outcomes are typically measured over a short time horizon. For example in [AG03], $N = 17, T = 40, T_0 = 15$ and in [ADH10], $N = 38, T = 30, T_0 = 18$.

There are two major approaches to do inference. The first imposes factor-model type structure on the outcomes, and does asymptotic inference. Many works in the literature, such as [ADH10, ABD+21, ASS18] impose this assumption to study properties of the synthetic control estimator. The second is a design-based approach, where the data are seen as fixed, and the treatment is assumed to be uniformly assigned from 1 to N. The placebo test of [AG03, ADH10] is an example; it was further extended in [FP18].

There are pros and cons for each approach. Asymptotic inference is difficult to justify for small N, T, T_0, as with many applications of the synthetic control method. Moreover, simplifications of the procedure are usually made in order to do asymptotics. This throws out many important practical considerations when running the method, such as data driven ways of choosing the weight matrix V. On the other hand, design based frameworks make strong assumptions on the treatment, although sensitivity analyses can be done to ameliorate this issue. We

will focus on the design based framework in this paper, but concede that the problem is difficult and that no solution is perfect.

The standard inference procedure in the design based framework is the placebo test. It was introduced in [AG03, ADH10] as a diagnostic tool related to permutation inference and Fisher's randomization test. The procedure was expanded upon in [FP18] to do inference under the assumption that I is uniformly assigned. It creates a p-value via the following recipe. For every unit i, create its synthetic control $\hat{Y}_{i,\bullet}$ using units $\{1, \ldots, N\} \setminus i$. Measure the discrepancy with the actual data via $R_i = |Y_{i,t} - \hat{Y}_{i,t}|$. Compare these "placebos" against I using the usual randomization p-value

$$p_{\text{placebo}} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1} \{R_I \leq R_j\} .$$

However, the p-value obtained from such tests are coarse: they take values in the grid $\left\{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\right\}$. When N is relatively small like in the premier applications of the synthetic control method [AG03, ADH10, ADH15], the smallest value $1/N$ of the p-value may give statistical significance, while the $2/N$ may switch this conclusion. This granularity hinders the application of the placebo test and related inference methods in practice. In some applications, the value of $1/N$ may be too large to even reject at a pre-specified level α.

1.1 Contributions

We present the Leave-Two-Out (LTO) Jackknife+, an approximate p-value which is constructed by leaving out two units at a time (excluding the treated unit). We adopt a design based setup similar to that of the placebo test, assuming that the treated unit I is uniformly assigned. The LTO Jackknife+ has three advantages over the placebo inference method.

Firstly, by leaving multiple data points out at a time in the same procedure, one can create p-values which lie on a much more refined scale than $\left\{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\right\}$.

Secondly, the procedure exhibits desirable empirical properties. In extensive numerical simulations, we observe that p_{LTO} has higher power when the effect size is large. It also has desirable Type I error properties as mentioned below.

Thirdly, our method potentially gives useful inference when $\alpha < 1/N$. In general, we cannot create a nonrandomized p-value which has Type I error less than α. The reason is the following. Supposing that $I \sim \text{Unif}\{1, \ldots, N\}$, if $p(I)$ is a p-value, then $\mathbb{P}_{H_0}(p(I) < 1/N) < 1/N$. But $\mathbb{P}_{H_0}(p(I) < 1/N)$ is equal to k/N for some integer $k \geq 0$; from this we conclude that $\mathbb{P}_{H_0}(p(I) < 1/N) = 0$. We should ex-

pect methods to have zero power in this regime, or have some sort of Type I error inflation. Accordingly, the LTO p-value, p_{LTO}, satisfies $\mathbb{P}(p_{LTO} \leq \alpha) \leq 1/N$. Empirically however, we often observe that Type I error of p_{LTO} is much lower than the purported bound. This makes the method empirically valid while for alternatives such as comparable versions of the placebo test, the Type I error is fixed.

Further, we provide an extension to the LTO Jackknife+ for inference in the $\alpha > 1/N$ setting, where the procedure yields an exact p-value, with no Type I error inflation. In this setting, the resulting p-value empirically has lower Type I error than advertised, and has higher power than the usual placebo test for larger effect sizes. In the above cases, we provide a sensitivity analysis for the procedure. Finally, our work analyzes extensions of the procedure to leaving k units at a time.

1.2 Example: Application to German Reunification Dataset

In [ADH15], the synthetic control method is applied to understand the effects of German reunification on the GDP of West Germany, using 16 other OECD countries as potential controls. The procedure demonstrates a large effect size.

Suppose we were to do inference in this problem with a level constraint $\alpha = 0.05$; in this case, $\alpha < 1/N$ so that the usual placebo inference will be powerless. We may do inference in this example using the LTO Jackknife+ p-value of Eq. (1). The p-value turns out to be 0.0417. On the other hand, the placebo p-value was computed to be 0.0588.

Because the inference is based on the assumption that each unit is equally likely to be treated, it is important to assess the sensivity of the inference to assumption. Section 3 provides a Rubin-Rosenbaum style sensitivity analysis procedure for the LTO statistic which we may conduct on this example. The sensitivity analysis relies on a weighted version of the LTO p-value, which is also an approximate p-value under non-uniform treatment probabilities π_i for unit i. Fixing a Γ, we constrain the non-uniform propensities π_i to satisfy $\pi_i \in [\frac{1}{\Gamma N}, \frac{\Gamma}{N}]$ and calculate the largest possible value of the weighted p-value for propensities in this constraint. Finally, we search for the smallest value of Γ such that the conclusion of significance is overturned. See Section 3 for more details.

As the reported LTO p-value is 0.04167, we want to see how large Γ can be before the conclusion of significance is possibly overturned. The output is shown in Figure 1. From the figure, Γ is around 1.1 before the maximum possible p-value is greater than 0.05. Overall, the curve as a function of Γ does not seem to increase too wildly.

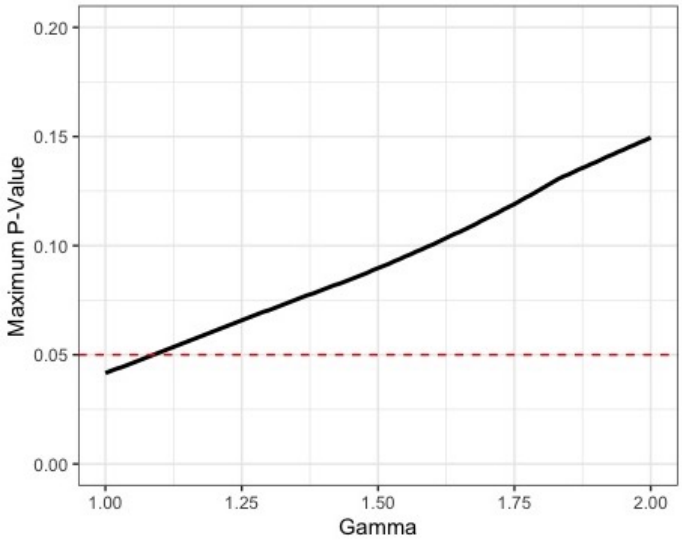

Figure 1: Output of a sensitivity analysis for the German reunification dataset of [ADH15]. The plot displays the maximum weighted LTO p-value as a function of Γ. The red dashed line signifies the level $\alpha = 0.05$.

2 Leave-Two-Out Jackknife+

2.1 Main Result

The Leave-Two-Out (LTO) Jackknife+ procedure works as follows:

1. For every pair of distinct units (i, j) from $1, \ldots, N$ and unequal to I, create the synthetic control with control units $1, \ldots, N$ excluding points i, j, I to obtain $\hat{\mathbf{Y}}_I^{-(i,j,I)}$, the vector of synthetic control outcomes. For some statistic $S(\cdot, \cdot)$, define the residual

$$R_{i,j,I;k} = |S(\mathbf{Y}_k, \hat{\mathbf{Y}}_k^{-(i,j,I)})|,$$

where $k \in \{i, j, I\}$ and $\mathbf{Y}_k$ is the observed data for unit k.

For example, the statistic function $S(\cdot, \cdot)$ may look at the difference of the data at a fixed time $t, |Y_{k,t} - \hat{Y}_{k,t}^{-(i,j,I)}|$. Another common choice is the pre-period to post-period ratio of Root Mean Square Prediction Error (RMSPE)

statistic proposed in [ADH15]:

$$S_{RMSPE}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{t=T_0+1}^{T}(X_t - Y_t)^2/(T - T_0)}{\sum_{t=1}^{T_0}(X_t - Y_t)^2/T_0}.$$

2. For every pair of indices in the first step, compute the Leave-Two-Out Residual

$$R_{i,j,I}^{LTO} := \max(R_{i,j,I;i}, R_{i,j,I;j}).$$

3. Compute the quantity

$$p_{LTO} := \frac{1}{(N-1)(N-2)} \sum_{\substack{i,j \in [N] \setminus I \\ i \neq j}} \mathbf{1}\left\{R_{i,j,I;I} \leq R_{i,j}^{LTO}\right\}, \tag{1}$$

which is an approximate p-value under the null hypothesis.

Remark 2.1. *Thinking of the LTO Jackknife+ p-value in terms of tournaments is useful. Consider a tournament between all of the N units, where matches are held for every triple of units. A unit i wins a match amongst units i, j, k if its residual $R_{i,jk;i}$ is the largest. Then p_{LTO} after renormalization counts the number of matches that the unit I has won.*

By approximate p-value, we mean to say that the LTO p-value satisfies the following Type I error guarantee:

$$\mathbb{P}(p_{LTO} \leq \alpha) \leq \frac{3 - \frac{3}{N} - \sqrt{9(1 - \frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \alpha(1 - \frac{1}{N})(1 - \frac{2}{N})\right)}}{2}. \tag{2}$$

To gain intuition on this, it is prudent to inspect the Type I error when N is large. In this setting,

$$\mathbb{P}(p_{LTO} \leq \alpha) \lesssim \frac{3 - \sqrt{9 - 12\alpha}}{2}. \tag{3}$$

The right hand side is very close to α when α is small. For $\alpha = 0.05$, the right hand side is around 0.0508. For $\alpha = 0.1$, the bound comes out to 0.104. It is interesting that there is no guarantee for $\alpha \geq 3/4$, but this is usually not needed in practice.

Letting $f(N, \alpha)$ denote the right hand side of (2), the Type I error bound may be improved even further by noticing that $\mathbb{P}(p_{LTO} \leq \alpha)$ is of the form k/N. As a result, we have the following theorem.

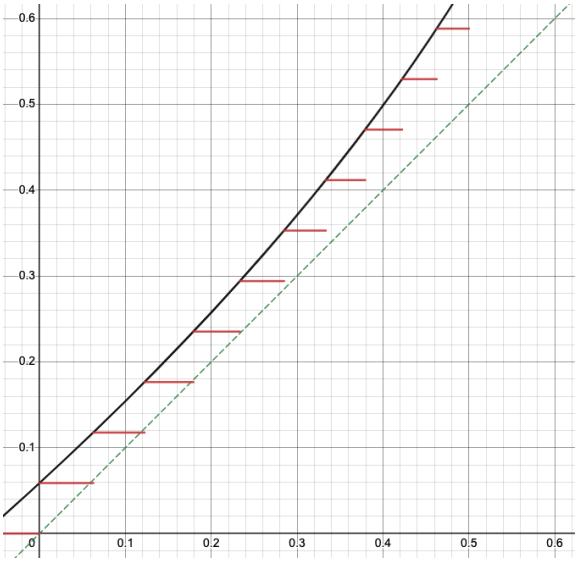

Figure 2: Solid black line shows $f(N, \alpha)$, with $N = 17$, as a function of α. The red line shows the Type I error bounds in Theorem 2.1. The dashed green line is the $y = x$ line.

Theorem 2.1. *Let $f(N, \alpha)$ denote the quantity in Equation (2). Then under the assumption of uniform treatment,*

$$\mathbb{P}(p_{LTO} \leq \alpha) \leq \frac{\lfloor N f(N, \alpha) \rfloor}{N}. \tag{4}$$

Figure 2 plots the functions $f(N, \alpha)$ and its floored version. By inspecting the figure, it is apparent that the best guarantee upper-bound on $\mathbb{P}(p_{LTO} \leq \alpha)$ for $\alpha < 1/N$ is $1/N$, which makes the Type I error inflation explicit. It is worth making this bound explicit, especially when we try to do inference in the $\alpha < 1/N$ setting.

Corollary 2.1. *When $\alpha < 1/N, \mathbb{P}(p_{LTO} \leq \alpha) \leq \frac{1}{N}$.*

Proof Sketch. The proof of the coverage guarantee relies on analyzing combinatorial structures related to tournaments. We present the proof here in order to introduce the main ideas of the analysis; it is an interesting extension of the proof of coverage in the Jackknife+ [BCRT21].

Throughout this section, let us introduce the notation

$$\mathbb{I}(k \succ i, j) := \mathbf{1}\left\{R_{i,j,k;k} > \max(R_{i,j,k;i}, R_{i,j,k;j})\right\}.$$

Call a unit k *strange* if

$$\sum_{\substack{i,j \in [N] \setminus k \\ i \neq j}} \mathbb{I}(k \succ i, j) \geq (1-\alpha)(N-1)(N-2).$$

Recall the intuition of the p-value as a tournament where matches are held between every triple of distinct units i, j, k. The winner of the match is the unit with the largest residual. We call a unit *strange* if it wins too many matches, quantified above. In a tournament, not everyone can win too often; following this intuition, we upper bound the number of strange units.

Let $\mathcal{S}$ denote the set of strange units, and let $s := |\mathcal{S}|$. Summing the definition of strangeness on both sides, we see that

$$\sum_{k \in \mathcal{S}} \sum_{i,j \in [N]} \mathbb{I}(k \succ i, j) \geq (1-\alpha) s (N-1)(N-2).$$

We will assume that all sums are over distinct indices, and will omit this notation from the sum indexing for clarity. For the first term, the left hand side may be split into three terms:

$$\overbrace{\sum_{i,j,k \in \mathcal{S}} \mathbb{I}(k \succ i, j)}^{\text{(I)}} + \overbrace{2 \sum_{k \in \mathcal{S}} \sum_{i \in \mathcal{S}, j \in \mathcal{S}^c} \mathbb{I}(k \succ i, j)}^{\text{(II)}} + \overbrace{\sum_{k \in \mathcal{S}} \sum_{i,j \in \mathcal{S}^c} \mathbb{I}(k \succ i, j)}^{\text{(III)}}.$$

For the first sum, note by renaming the labels that $\sum_{i,j,k \in \mathcal{S}} \mathbb{I}(k \succ i, j) = \sum_{i,j,k \in \mathcal{S}} \mathbb{I}(i \succ j, k) = \sum_{i,j,k \in \mathcal{S}} \mathbb{I}(j \succ k, i)$. Thus

$$\text{(I)} \leq \sum_{i,j,k \in \mathcal{S}} \frac{1}{3} \left(\mathbb{I}(k \succ i, j) + \mathbb{I}(i \succ j, k) + \mathbb{I}(j \succ k, i)\right) \leq s(s-1)(s-2)/3.$$

Crucially, we use the fact that there can be at most one winner in a triple comparison: $\mathbb{I}(k \succ i, j) + \mathbb{I}(i \succ j, k) + \mathbb{I}(j \succ k, i) \leq 1$.[1]

[1] We allow for zero winners in the triple comparison if there are ties in the residuals.

For the second sum, swap the naming of the i, k labels and use the bound $\mathbb{I}(k \succ i, k) + \mathbb{I}(i \succ k, j) \leq 1$ to see that

$$\text{(II)} \leq 2 \sum_{k \in \mathcal{S}} \sum_{i \in \mathcal{S}, j \in \mathcal{S}^c} \frac{1}{2} \left(\mathbb{I}(k \succ i, k) + \mathbb{I}(i \succ k, j) \right) \leq s(s-1)(N-s).$$

Finally, in the last sum, we use the naive bound $\text{(III)} \leq s(N-s)(N-1-s)$.

Combining these bounds and working through algebra, we arrive at a quadratic inequality in the variable $\beta := s/N$:

$$\frac{\beta^2}{3} - \beta(1 - \frac{1}{N}) - \frac{4}{3N^2} + \frac{1}{N} + \alpha(1 - \frac{1}{N})(1 - \frac{2}{N}) \geq 0.$$

We can solve this by quadratic formula, yielding

$$\beta \leq \frac{3 - \frac{3}{N} - \sqrt{9(1 - \frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \alpha(1 - \frac{1}{N})(1 - \frac{2}{N})\right)}}{2}. \tag{5}$$

To conclude, we note that $\mathbb{P}(p_{LTO} \leq \alpha) = \mathbb{P}(I \in \mathcal{S}) = \beta$. The bound in equation (5) shows that p_{LTO} is an approximate p-value. A full proof with further computational details may be found in the appendix.

Notes on the Type I error guarantee. Looking at the Type I error bound as conditional on the potential outcomes gives another perspective for the LTO guarantee. Letting $\mathcal{Y} = \{(Y_{it}(1), Y_{it}(0)) : i \in [N], t \in [T]\}$ denote the potential outcomes, the argument in the introduction says the following. Let $p(I)$ be conditionally valid p-value , i.e. $\mathbb{P}(p(I) \leq |\mathcal{Y}) \leq \alpha, \forall \alpha \in (0, 1)$, which is nonrandomized and only depends on the treatment I. Because I is uniform on $[N]$, the conditional probability $\mathbb{P}(p(I) \leq \alpha|\mathcal{Y})$ belongs to the set $\{0/N, 1/N, \ldots, N/N\}$. Thus if $\alpha < 1/N, \mathbb{P}(p(I) \leq \alpha|\mathcal{Y}) = 0$. Thus when $\alpha < 1/N$, any conditionally finite-sample valid p-value has non-zero power.

This does not imply unconditional impossibility. One way to get useful inference is to reject if the p-value $p(I) \leq 1/N$, and hope that $\mathbb{P}(p(I) \leq 1/N|\mathcal{Y})$ for some $\mathcal{Y}$. The placebo test does not have this property however, as the unconditional Type I error is always $\lfloor N\alpha \rfloor / N$. The LTO p-value, on the other hand, seems to have this property. The unconditional Type I error upper bound is theoretically $1/N$ in this case, and often much less than $1/N$ empirically.

Notice we can also create exact p-values in the $\alpha < 1/N$ regime by randomizing the placebo p-value: whenever the placebo p-value is $1/N$, flip a coin to

randomize it in order to make the Type I error exactly α. However, we are wary of the use of external randomization in this form. Although randomization may increase power, it allows the user to cheaply re-run the analysis with the possibility of obtaining statistical significance. In the worst case, external randomization of this form may encourage p-hacking for significance. The LTO Jackknife+ p-value, which is non-randomized, guards against adverse research incentives.

2.2 Constructing confidence intervals

The leave-two-out method may be adapted to yield a confidence interval for the counterfactual $Y_{(I,t)}(0)$. In this section, we create the residuals $R_{I,i,j;I}$ not by the RMSPE, but by taking the absolute difference between the synthetic control of unit I and the actual outcome of unit I at a fixed time t. This is equivalent to finding a confidence interval for the effect size at a fixed time t. If a strong null is assumed so that the effect size is fixed over time, a confidence interval can be created using a simple extension of the procedure in this section.

A direct inversion based confidence interval for the counterfactual is given by

$$\mathrm{CS}(1-\alpha) = \{y : p^y_{LTO} \geq \alpha\},$$

where p^y_{LTO} is the p-value obtained from the data replacing $Y_{I,t}(1)$ by y. From this definition, the direct inversion confidence interval has a miscoverage guarantee of $\frac{\lfloor Nf(N,\alpha) \rfloor}{N}$.

We can also construct the confidence interval indirectly, much like the construction of the original Jackknife+ confidence intervals [BCRT21], via

$$\mathrm{CS}(1-\alpha) := [\hat{q}^-_\alpha \left\{\hat{Y}^{i,j,I}_{I,t} - R^{LTO}_{i,j}\right\}, \hat{q}^+_\alpha \left\{\hat{Y}^{i,j,I}_{I,t} + R^{LTO}_{i,j}\right\}],$$

where the quantiles are taken over the sets indexed by indices i, j such that $i \neq j$ and each not equal to I. Note this is not the same as the direct inversion based confidence interval for the p-value; in general it will be larger than the inversion based set, but it still satisfies the following miscoverage guarantee.

Theorem 2.2. *The indirect LTO Jackknife+ interval* $\mathrm{CS}(1-\alpha)$ *satisfies*

$$\mathbb{P}(Y_{I,t}(0) \notin \mathrm{CS}(1-\alpha)) \leq f(N,\alpha).$$

Proof. Let us relate the p-value to the coverage guarantees on the leave-two-out confidence interval. Suppose that $Y_{I,t}(0) \notin \mathrm{CS}(1-\alpha)$. Then either $Y_{I,t}(0) - \hat{Y}_{I,t}^{i,j,I} \leq -R_{i,j,I}^{LTO}$ for at least $[(1-\alpha)\binom{N-1}{2}]$ many pairs of indices (i,j), or $Y_{I,t}(0) - \hat{Y}_{I,t}^{i,j,I} \geq R_{i,j,I}^{LTO}$, for at least $[(1-\alpha)\binom{N-1}{2}]$ many pairs of indices (i,j). Thus, for at least $[(1-\alpha)\binom{N-1}{2}]$ index pairs,

$$|Y_{I,t}(0) - \hat{Y}_{I,t}^{i,j,I}| > R_{i,j}^{LTO}.$$

Hence

$$\begin{aligned}\mathbb{P}(Y_{I,t}(0) \notin \mathrm{CS}_N(1-\alpha)) &\leq \mathbb{P}\left(\sum_{\substack{i,j \in [N] \setminus I \\ i>j}} \mathbf{1}\left\{R_{i,j,I;I} > R_{i,j}^{LTO}\right\} \geq (1-\alpha)\binom{N-1}{2}\right) \\ &\leq f(N,\alpha).\end{aligned}$$

by the previous part of the proof. □

2.3 Improving power

Notice that in defining the notion of strange unit, we have great flexibility in choosing the threshold. Call a unit k *c-strange* if

$$\sum_{\substack{i,j \in [N] \setminus k \\ i \neq j}} \mathbb{I}(k > i,j) \geq (1-\alpha)(N-1)(N-2) - c(N-2).$$

This results in a new p-value

$$p_{LTO,c} := p_{LTO} - \frac{c}{N-1}$$

Using the same proof technique as before, we can derive a similar quadratic Type I error guarantee as before, depending now on c.

Theorem 2.3. *Let all notation be as above. Then*

$$\mathbb{P}(p_{LTO,c} \leq \alpha) \leq \frac{\lfloor N f(N,\alpha,c) \rfloor}{N}$$

where

$$f(N,\alpha,c) := \frac{3 - \frac{3}{N} - \sqrt{9(1-\frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \alpha(1-\frac{1}{N})(1-\frac{2}{N}) + \frac{c(N-2)}{N^2}\right)}}{2}.$$

Proof of Thm. 2.3. Define unit k to be *c-strange* if

$$\sum_{\substack{i,j\in[N]\setminus k \\ i\neq j}} \mathbb{I}(k > i, j) \geq (1-\alpha)(N-1)(N-2) - c(N-2).$$

Let $\mathcal{S}_c$ be the set of c-strange units and $s = |\mathcal{S}_c|$. Imitating the proof of Thm. **??**, the number of strange points satisfies the inequality

$$\frac{s^2}{3} - \frac{4}{3} - s(N-1) + N + \alpha(N-1)(N-2) + c(N-2) \geq 0$$

Solving the quadratic inequality in the same fashion, we find that for $\beta = s/N$,

$$\beta \leq \frac{3 - \frac{3}{N} - \sqrt{9(1-\frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \alpha(1-\frac{1}{N})(1-\frac{2}{N}) + \frac{c(N-2)}{N^2}\right)}}{2}. \tag{6}$$

Thus, $\mathbb{P}(p_{LTO,c} \leq \alpha) = \mathbb{P}(I \in \mathcal{S}_c) \leq \beta$. Finally, because $p_{LTO,c}$ is a function of a uniform random variable, the probability $\mathbb{P}(p_{LTO,c} \leq \alpha)$ must be of the form $1/N$; from this we deduce that $\mathbb{P}(p_{LTO,c}) \leq \frac{\lfloor Nf(N,\alpha,c)\rfloor}{N}$. □

The advantage of this procedure is that $p_{LTO,c}$ is strictly more powerful than p_{LTO}. For a fixed budget α, we can choose $c > 0$ as large as possible depending on α such that the type I error guarantee is still the same as for p_{LTO}. That is, pick c as large as possible such that the equality

$$\frac{\lfloor Nf(N,\alpha,c)\rfloor}{N} = \frac{\lfloor Nf(N,\alpha)\rfloor}{N}$$

is retained. For example, when $\alpha < 1/N$, it can be shown that c may be taken as large as $\frac{1}{N-1} - \alpha$. The adjustment can be fairly large. For example, when $N = 15, \alpha = 0.05, c \approx 0.02$, which may increase power significantly in practice.

Thus whenever we refer to $p_{LTO,c}$ henceforth, the value of c will be chosen as large as possible depending on α to ensure the same Type I error upper bound above. We can denote this dependence by writing $p_{LTO,c}(\alpha)$. In Section 4.1 on power simulations of our method, we implement the optimized p-value. The power gains are significant, while the empirical Type I error grows slightly. In practice, we would recommend using this p-value. The only disadvantage is a lack of a sensitivity analysis procedure for $p_{LTO,c}$.

Rejection Sets. One undesirable feature of the powered LTO p-value is that the procedure is not decision-monotonic in α. That is, if the test based on $p_{LTO,c}(\alpha)$ rejects at some level α, the test using $p_{LTO,c}(\alpha')$ may not reject at an α' with $\alpha' > \alpha$. This is due to the step-like behavior of the Type I error upper bound.

We can enforce decision monotonicity and also improve power of the procedure by considering instead

$$I_\alpha = \max_{\alpha' \le \alpha} \mathbf{1}\left\{p_{LTO,c}(\alpha') \le \alpha'\right\},$$

rejecting at level α if $I_\alpha = 1$. Clearly this procedure is more powerful than rejecting only when $p_{LTO,c}(\alpha) \le \alpha$. The procedure also controls Type I error. To see this,

$$\begin{aligned}\mathbb{P}(I_\alpha = 1) &= \mathbb{P}(\exists \alpha' : p_{LTO,c}(\alpha') \le \alpha') \\ &= \mathbb{P}(\exists \alpha' : p_{LTO} \le \alpha' - f(\alpha')),\end{aligned}$$

where $f(\alpha')$ is the adjustment $c/(N-1)$ which is added to the LTO p-value. The last line may be written

$$\begin{aligned}&\mathbb{P}(p_{LTO} \le \max_{\alpha' \le \alpha} \alpha' - f(\alpha')) \\ =&\mathbb{P}(p_{LTO} \le g(\alpha) - f(g(\alpha))),\end{aligned}$$

where $g(\alpha) = \operatorname{argmax}_{\alpha' \le \alpha} \alpha' - f(\alpha')$. The latter is equal to $\mathbb{P}(p_{LTO,c}(g(\alpha)) \le g(\alpha)) \le \frac{\lfloor Nf(N,g(\alpha))\rfloor}{N} \le \frac{\lfloor Nf(N,\alpha)\rfloor}{N}$, which is the same Type I error bound as using just $p_{LTO,c}(\alpha)$. We do not use this procedure in our simulations, leaving this method for theoretical interest.

2.4 Inference for $\alpha \ge 1/N$

By normalizing p_{LTO} differently, we can create a finite-sample valid p-value. This is useful in standard settings where $\alpha \ge 1/N$. In this setting we are interested in comparing the finite sample valid LTO p-value with the ordinary placebo p-value. In Section 4, whenever we conduct inference in a setting where $\alpha \ge 1/N$, we will compare the finite sample valid LTO p-value with the usual placebo p-value.

Definition 2.1. *Recalling the notation* $\mathbb{I}(k \succ i,j) := \mathbf{1}\left\{R_{i,j,k;k} > \max(R_{i,j,k;i}, R_{i,j,k;j})\right\}$, *define the finite-sample valid LTO p-value* $p_{LTO,V}$ *by*

$$p_{LTO,V} = \frac{1}{(N-1)(N-1)} \sum_{j,k \ne I, j \ne k} \mathbb{I}(I \succ j,k) + \frac{1}{N-1}. \tag{7}$$

Proposition 2.1.

$$\mathbb{P}(p_{LTO,V} \leq \alpha) \leq \frac{3 - \frac{3}{N} - \sqrt{9(1-\frac{1}{N})^2 - 12[\alpha(1-\frac{1}{N})^2 - \frac{1}{3N^2}]}}{2}.$$

A quick inspection shows that the right hand side above is smaller than α, for α in some range $[0, c]$ for some c. In this sense, the p-value is finite sample valid. We can also define a more powerful version of this p-value in exactly the same way as in Section 2.3.

2.5 Related Work

The existing literature on synthetic controls is vast: see the survey article [Aba21] for an introduction. Accordingly, there has been much work addressing inference for synthetic controls. Among the first inferential-type ideas for synthetic controls was the placebo test, first described in [AG03], and extended by Firpo and Possebom [FP18] into a permutation type inference procedure. They work in the same design-based framework where the probability of treatment is uniform. Several extensions are made, including to a non-equally weighted inference procedure which is useful for sensitivity analysis.

Other works have also considered the design-based setup and randomization or permutation based inference. See [BISW21] for a design-based analysis of the synthetic controls procedure under a similar uniform treatment assumption, and also a random treatment time assumption. [ST21] gives a related randomization inference procedure in a staggered adoption setting where there are multiple treated units, under the additional assumption that the time at which each unit adopts treatment follows a Cox proportional hazards model.

Inference procedures have also been developed assuming underlying factor model structure and related time series assumptions on the data. [CWZ21] develop an inference procedure assuming a regularity condition on the residual process in time. The paper contains an extensive discussion on the various choices for creating the counterfactual or synthetic control. Another line of research [CFT21, CFPT22] provides prediction intervals via non-asymptotic concentration bounds, under stationarity and weak dependence time series assumptions on the data. See also [HS17, AAH+21, ASS18] for related analyses of the synthetic-control type estimators.

For many of these results, consistency of the estimator and the resulting inference procedure is shown in the limit as the number of controls and units goes to

infinity. Although many of these works offer finite sample bounds, it is difficult to contextualize these results when the applications of synthetic controls are usually small N, small T settings. Furthermore, many of these asymptotic results require simplifications on the synthetic control procedure to carry out analysis. The LTO Jackknife+ procedure can use other counterfactual models beyond the synthetic control method, and we require no conceptual simplifications to do inference. As a result, the LTO procedure may be used with generalizations of the original synthetic control procedure such as those outlined in [DI16, CWZ21, AAH+21, ASS18].

Connection to conformal inference. The LTO proposal extends the ideas of the Jackknife+ from the conformal inference literature [BCRT21], which creates confidence intervals for predictive inference by leaving out one data point at a time. Our method can be seen as a leave-two-out version of the Jackknife+, which is a leave-one-out (LOO) type proposal. The proof strategies are also similar. As with the original Jackknife+, the LTO Jackknife+ procedure described here can be generalized beyond synthetic controls and applies to the creation of prediction intervals.

Theoretically, the LTO method has a few advantages over the Jackknife+. Firstly, the Type I error bound is much closer to α, namely $\frac{3-\sqrt{9-12\alpha}}{2}$, when α is small and N is large. On the other hand, the primary guarantee of the LOO Jackknife+ method is a Type I error upper bound of 2α. The LTO method further gives more refined inference, giving $O(N^2)$ comparisons as opposed to $O(N)$ for the LOO procedure.

A detailed comparison to a leave-one-out version of the inference procedure is left for future work. It would be particularly interesting to empirically analyze the power of both the LTO and LOO procedures and identify the settings in which one dominates the other. Inspecting the looseness of Type I error upper bounds for both procedures is also an important direction.

There are further connections to the conformal inference literature. Construction of confidence intervals for the placebo p-value, based on inversion, can be seen as an analog of the *full conformal* procedure [VGS05]. Aside from full conformal, there are related procedures such as split conformal or cross conformal [Vov15], which may have useful implications for synthetic control inference.

3 Sensitivity Analysis

Outside of experimental settings, the assumption of uniform treatment is questionable. It is useful to have procedures to analyze sensitivity of our approach to this assumption. The goal of these procedures is to answer the question: by how much would our statistical conclusions change if the treatment were non-uniform? Phrased another way, by how much would the treatment probabilities have to change to overturn a conclusion of statistical significance? Firpo and Possebom [FP18] give an approach to do sensitivity analysis for their p-values under non-uniform treatments.

To answer this question, we first show that weighted analogs of the leave-two-out p-values may be derived, assuming that treatment probabilities were non-uniform. Suppose that the probability of treating unit k is given by $\pi_k, k = 1, \ldots, N$. Then the following weighted approximate p-value generalizes the LTO p-value of Section 2.

Definition 3.1 (Refined weighted p-values, $\alpha \leq 1/N$)**.**

$$p_{\pi,product} := \sum_{\substack{j \neq k \\ j,k \neq I}} \frac{\pi_j \pi_k}{(1-\pi_I)^2 - \sum_{l \neq I} \pi_l^2} \mathbf{I}\{I \not\succ j,k\}$$

In principle, these inexact p-values can be used for sensitivity analysis. The standard template is as follows:

1. Compute $p_{\pi,\text{product}}$ for the equal weights benchmark π_0 defined by $\pi_{0,k} = 1/N, \forall k$. Assume the result is significant at some pre-fixed level α_0.

2. Fix a $\Gamma \in [1, N]$. Solve the optimization problem

$$\max_{\pi} p_{\pi,\text{product}} \tag{8}$$

$$\text{subject to } \pi_i \in [\Gamma^{-1}\pi_{0,i}, \Gamma\pi_{0,i}] \forall i, \sum_i \pi_i = 1. \tag{9}$$

3. Do a grid search to find the smallest value of Γ such that the solution to the above optimization problem is greater than α_0.

If the result is insignificant for level α_0, we may also assess sensitivity of the this conclusion by finding the optimizing for the minimum of the above.

For Γ close to one and $\pi_i \approx 1/N$, we obtain the same p-value in Section 2, with similar Type I error guarantees:

Proposition 3.1. *Assume that $\alpha \le 1/N$ and $\pi_i \in [\frac{1}{\Gamma N}, \frac{\Gamma}{N}]$ for all i, with $\Gamma < N$. Then*

$$\mathbb{P}(p_{\pi,product} \le \alpha) \le \frac{3}{2} - \frac{1}{2}\sqrt{9 - 12\left(\alpha(1 - \sum_l \pi_l^2) + \sum_{\mathcal{S}} \pi_i^2 - 2(1-\alpha)(\frac{1}{N\Gamma} - \frac{1}{N^2\Gamma^2})\right)}$$

For the Type I error bound of $p_{\pi,\text{product}}$, we have an approximate bound of $\frac{3-\sqrt{9-12\alpha}}{2}$ as obtained for the $\Gamma = 1$ setting.

Weighted analogs of valid LTO p-values. If inference is conducted using the $\alpha > 1/N$ p-values, a similar sensitivity analysis procedure may be used, using the following weighted analogs of the p-values in Section 2.4

Definition 3.2 (Refined weighted p-values, $\alpha > 1/N$)**.**

$$p_{\pi,V,product} := \sum_{\substack{j\neq k \\ j,k\neq I}} \frac{\pi_j \pi_k}{(1-\pi_I)^2} \mathbf{I}\{I \not\asymp j,k\} + \frac{\sum_{l\neq I} \pi_l^2}{(1-\pi_I)^2}$$

$$p_{\pi,V,sum} := \sum_{\substack{j\neq k \\ j,k\neq I}} \frac{\pi_j + \pi_k}{2(N-1)(1-\pi_I)} \mathbf{I}\{I \not\asymp j,k\} + \frac{1}{N-1}$$

When $\pi_i = 1/N$ for all i, the p-values above reduce to those introduced in Section 2.4 respectively. The Type I error guarantees are also similar to those above.

Proposition 3.2. *We have the following results for the refined $\alpha \ge 1/N$ p-values, assuming $\pi_i \in [\frac{1}{\Gamma N}, \frac{\Gamma}{N}]$ for all i.*

$$\mathbb{P}(p_{\pi,V,product} \le \alpha) \le \frac{3}{2} - \frac{1}{2}\sqrt{9 - 12\left(\sum_{i\in\mathcal{S}} \pi_i^2 - \frac{1}{3N^2} - (1-\alpha)(1 - \sum_l \pi_l^2)\right)}$$

$$\mathbb{P}(p_{\pi,V,sum} \le \alpha) \le \alpha + \sqrt{\alpha^2 + (1-2\alpha)\sum_{k\in\mathcal{S}} \pi_k^2}$$

Computational Details. The optimization problem in step 2) may pose challenges. To optimize this problem, we use the following trick. Fixing a Γ, let B_Γ

be the set of π satisfying the constraints in Equation (8). We only need to check if $\max_{\pi \in B_\Gamma} p_{\pi,\text{product}}$ is greater than α_0. This is equivalent to checking that

$$\max_{\pi \in B_\Gamma} \sum_{\substack{j,k \neq I \\ j \neq k}} \pi_j \pi_k \mathbb{I}\{I \not\asymp j,k\} - \alpha_0 \left((1-\pi_I)^2 - \sum_{l \neq I} \pi_l^2 \right) > 0$$

The optimization problem can be written as a simple quadratic program. Namely, we must check whether

$$\max_{\pi \in B_\Gamma} \pi^\top A \pi + 2\alpha_0 \pi^\top e_I > \alpha_0$$

where $A := (G + \alpha_0 I_N - 2\alpha_0 e_I e_I^\top)$, with $G \in \mathbb{R}^{N \times N}$ is a zero-one matrix with entries $G_{jk} = \mathbb{I}\{I \not\asymp j,k\}$, whenever j and k do not equal I and zero otherwise. Here e_I is the standard basis vector for index I. If the goal is to calculate $\max_{\pi \in B_\Gamma} p_{\pi,\text{product}}$, we can do a binary search over $c \in [0,1]$ that satisfy $\max_{\pi \in B_\Gamma} p_{\pi,\text{product}} > c$, and apply the same trick as above.

The objective is not necessarily convex as G is not guaranteed to be positive semidefinite. As it turns out, the resulting optimization problem is a fundamental NP hard problem known as *quadratic nonconvex programming* with box constraints. In practice, non-convex quadratic programming problems are solved using heuristics such as branch-and-bound techniques. As the dimensionality of the problems that we are solving is rather small, we expect off the shelf solvers should do well on this problem. Moreover, whenever the computed LTO p-value is small, the matrix $G + I_N - 2e_{II}$ is usually rather sparse, which may help with computation.

4 Simulation Results

Thus far we have explored theoretical properties of the LTO Jackknife+ inference procedure. We compare empirical properties of the LTO Jackknife+ procedure on semisynthetic examples based on the California Proposition 99 dataset on smoking [ADH10] and the Basque Terrorism dataset on Spanish regional GDP. [AG03]. All simulation results rely on the R package Synth, and may be found on Github.

Firstly, we subsample both datasets in order to mimic an inference setting where $\alpha < 1/N$. In this setting, we compare the Type I error and power of the LTO p-value in Eq. (1) to an inexact variant of the placebo test. The inexact placebo test simply rejects if the residual $R_I = |Y_{I,t} - \hat{Y}_{I,t}|$ is the smallest among all the other

placebos. Secondly, we consider a setting in which $\alpha \geq 1/N$, and compare inference based on the valid LTO p-value in Eq. (7) and the usual placebo p-value. In both settings, we boost the power of the LTO p-value while keeping the same Type I error constraint as in Section 2.3.

Empirically, the LTO procedure demonstrates high power in large signal to noise ratios, compared against the placebo p-value. In the setting where we try to do inference when $\alpha < 1/N$, we observe that the empirical Type I error of the LTO p-value may be zero, whereas the Type I error of the placebo p-value is always $1/N$. This suggests that the upper bound on the Type I error establish in Theorem 2.1 is not tight, and in practice, we may use the method while still satisfying Type I error constraints.

4.1 Semisynthetic Power Simulations

California Proposition 99. In this section, we consider a semisynthetic simulation using the California proposition 99 example of [ADH10], one of the flagship applications of the synthetic control method. The dataset contains data on cigarette sales, prices, and various other regional attributes in 39 US states, over the years 1970 to 2000. In [ADH10], the data are used in the synthetic control procedure to analyze the effect of Proposition 99, a 1988 California state law increasing cigarette taxes, on smoking. In this example, $N = 38, T = 30, T_0 = 18$.

We construct a semisynthetic version of the data, where a subsample of N units is randomly chosen from the full set of 38 units. California is removed from the dataset. In the smaller dataset, a state is chosen uniformly at random to be the 'treated' unit. For the treated unit I, we posit a uniform treatment effect τ, and we update the observed outcomes $Y_{I,t}$ by adding τ. A range of values of τ are tested, corresponding roughly to multiples of $(0, -0.5, -1, -2, -3)$ of the standard deviation of the outcome variable (cigarette sales). When $\tau = 0$ we are calculating the Type I error of the procedures. Similar semisynthetic experiments are provided in Section A.

In the first setting $(\alpha < 1/N)$, we take $\alpha = 0.02$ and subsample a dataset of size $N = 30$. We run the placebo inference procedure and the leave-two-out JK procedure in order to compare their Type I error and power properties. In the simulations, Type I error and power are calculated exactly by iterating over all units in the subsampled datasets. Further, we repeat the experiment over 20 Monte Carlo runs; for each run, a different subsampled dataset is chosen. For the first set of results, we use the RMSPE statistic when constructing all p-values. We compare four different procedures. The first is the standard placebo test, which should

have zero power and Type I error in this setting. The second is an inexact version of the placebo test, described above. As a result its Type I error is exactly $1/N$. The third and fourth are the LTO procedure for $\alpha \leq 1/N$ and its powered version described in Section 2.3. When constructing the synthetic controls, we matched on several covariates: average retail price of cigarettes, log per capita income, the proportion of the population age 15–24, and per capita beer consumption. We use the mean of these variables over the 1980–1988 period, along with lagged smoking consumption in 1975, 1980, and 1988. This replicates the specification of [ADH10]. The results of this comparison are shown in Figure 3.

Several observations can be made regarding empirical performance. Firstly, the Type I errors of the LTO p-values are on average smaller than $1/N$. This means that for a majority of the Monte Carlo resamples of the dataset, the empirical Type I error was zero. Thus, the LTO procedure may be empirically valid even if the theoretical upper bounds indicate that a Type I error correction is needed. On the other hand, the Type I error of the inexact placebo is always fixed. Simultaneously, the power of the two LTO procedures seems to be better than that of the inexact placebo test when the effect size is large, with the powered LTO dominating the standard LTO procedure. When the effect size is intermediate however, the inexact placebo seems to have the highest power, but not by much.

Figure 4 shows results for the same setup, using instead the absolute difference of the synthetic control and outcome at the year 2000 to construct the p-values of interest. The empirical benefits of the LTO p-values are less pronounced in this setting: the powered LTO procedure has about the same power as the inexact placebo at all signal to noise ratios except the intermediate case; the power of the LTO procedure is just slightly worse. The Type I error is only slightly smaller than $1/N$ on average, signaling that in most resampled datasets, the LTO procedures achieved the Type I error upper bound of Equation (2). We speculate that the choice of statistic used in constructing the p-values leads to this difference in power and Type I error. It would be interesting to establish theoretical results explaining how the choice of statistic affects the power of the procedure or the tightness of the Type I error upper bound.

In the second setting where $\alpha \geq 1/N$ we take $\alpha = 0.05, N = 30$. In this setting, we compare the standard placebo test with the LTO p-value described in Section 2.4 and its associated more powerful version. We can largely draw the same conclusions as with the $\alpha < 1/N$ setting. Notice that $\alpha < 2/N$, so the Type I error of the standard placebo test will also be exactly $1/N$. In this setting, the Type I error upper bound for the LTO procedure is also $1/N$. In the simulation, we observe

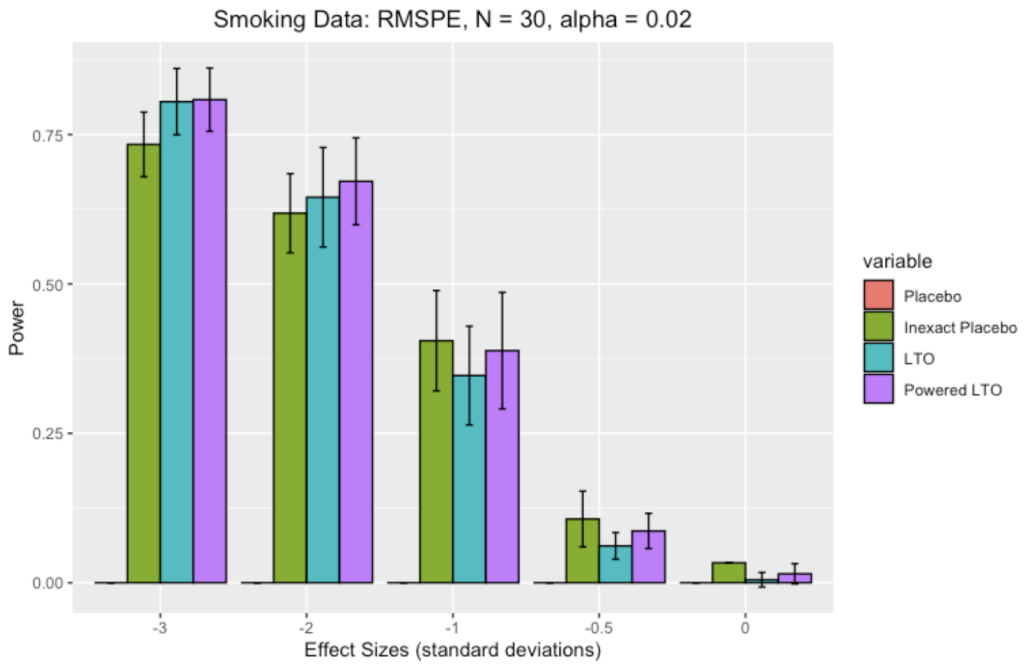

Figure 3: Power of several different p-values versus effect size, on size 30 subsamples of the California Proposition 99 dataset. Inference is done with $\alpha = 0.02$, which falls in the $\alpha < 1/N$ regime. The x-axis is the effect size τ scaled in terms of multiples of the standard deviation of the outcome variable of the dataset. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of the size 30 subsample. Four different methods are compared, all constructed using the RMSPE statistic. The rightmost column with $\tau = 0$ is just the Type I error of the procedures.

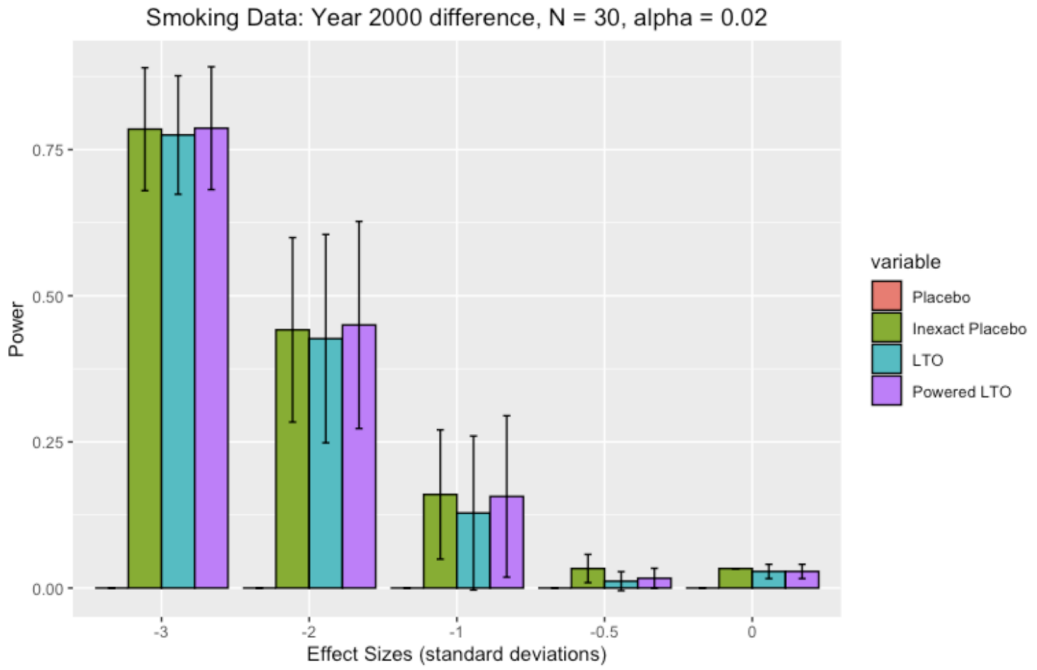

Figure 4: Power of several different p-values versus effect size, on size 30 subsamples of the California Proposition 99 dataset. Inference is done with $\alpha = 0.02$. The x-axis is the effect size τ scaled in terms of multiples of the standard deviation of cigarette sales. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of the size 30 subsample. Four different methods are compared, using the absolute difference at the year 2000.

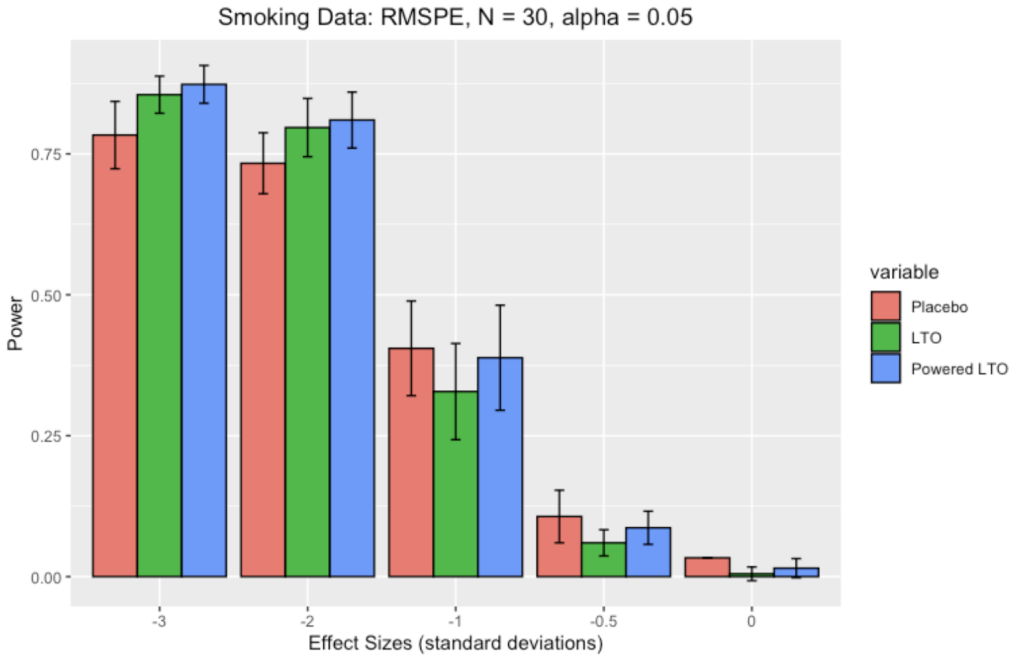

Figure 5: Power of three different p-values versus effect size, on size 30 subsamples of the California Proposition 99 dataset. Inference is done with $\alpha = 0.05$, and so the LTO procedures are constructed with the $\alpha > 1/N$ setting. All procedures are constructed with the RMSPE statistic. The x-axis is the effect size τ scaled in terms of multiples of the standard deviation of cigarette sales. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of the size 30 subsample.

that the empirical Type I error is far below this theoretical upper bound; in most Monte Carlo resamples, the Type I error is in fact zero. Moreover, the power of the powered LTO is on average larger than or equal to that of the standard placebo test.

The economic impacts of Basque conflict. Next, we consider a similar set up using a dataset of Spanish regions and GDP studied in [AG03]. In this setting, we take $\alpha = 0.05, N = 15$ as an example of inference in the $\alpha < 1/N$ setting, and compare the placebo, inexact placebo, and two LTO procedures. When constructing all synthetic controls, we imitate the choice of covariates found in the running Basque data example of [ADH11]. We repeat this comparison over 20 resamples of the original dataset. The results are found in Figure 6.

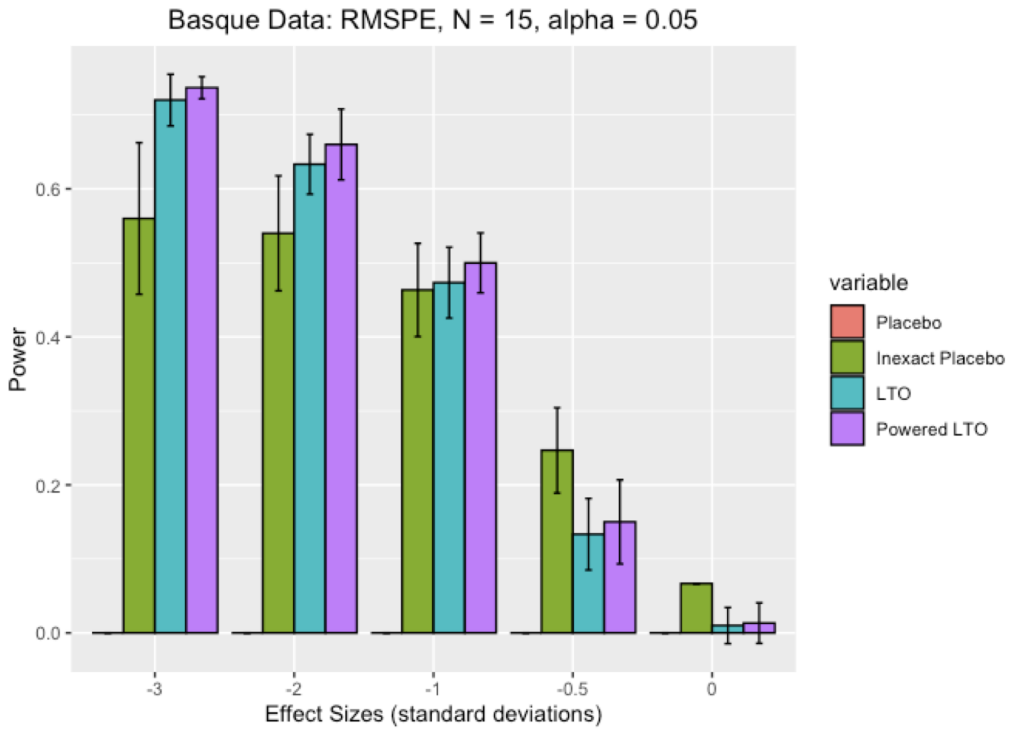

Figure 6: Power of several different p-values versus effect size, on size 15 subsamples of the Spanish GDP dataset. Inference is done with $\alpha = 0.05$, which falls in the $\alpha < 1/N$ regime. The x-axis is the effect size τ scaled in terms of multiples of the standard deviation of the outcome variable of the dataset. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of the size 15 subsample. Four different methods are compared, all constructed using the RMSPE statistic. The rightmost column with $\tau = 0$ is just the Type I error of the procedures.

Remarkably, the conclusions identified for the Proposition 99 data carry over to this dataset. As before, the power of the placebo method is zero in all settings. The LTO procedures demonstrate much lower Type I error than the upper bound advertises, while at the same time, they exhibit higher power than the inexact placebo test in higher signal to noise ratios. It is also interesting to note less variability in the power of the Powered LTO than with the inexact placebo test. When the effect size is roughly less than one, the placebo test dominates.

4.2 LTO Sensitivity Analyses

It is worth revisiting the classic case studies of the synthetic control method in [AG03, ADH10, ADH15] using the LTO and placebo inference procedures. For each use of the LTO p-value, we analyze the sensitivity of the conclusion to the equal weights assumption. For all sensitivity analyses, we use Gurobi 10.0 to solve the nonconvex quadratic optimization problem for optimizing $p_{\pi,\text{product}}$. A similar sensitivity analysis for the placebo test may be done using the results of [FP18], which we do not replicate here.

In the introduction, we outlined the application of the LTO and placebo inference methods for the German Reunification example of [ADH15], along with the sensitivity analysis procedure for the LTO method. We can repeat the exercise with the California Proposition 99 dataset on smoking, where we analyze the impact of Proposition 99 on cigarette sales in California. Here, $N = 38$ with one treated unit. We conduct inference with $\alpha = 0.05$. This is above $1/N$, but for the sake of understanding the sensitivity analysis procedure, we use the $\alpha < 1/N$ LTO p-value. The LTO p-value obtained was 0.024 while the placebo p-value was 0.026, which exactly equals $1/N$. In this setting, we can also inspect how sensitive the conclusion of non-significance is with respect to the equal weights assumption. In this setting, the sensitivity analysis tries to minimize the p-value for each fixed Γ.

For the Spanish region GDP dataset of [AG03], we repeat the exercise to analyze the effect of terrorism on GDP of the Basque country. Here, $N = 17$ with the onset of the treatment being $T_0 = 1970$. We conduct inference with $\alpha = 0.05$, which is less than $1/N$. Surprisingly in this example, we obtained a LTO p-value of 0.67; the placebo p-value gave a similar non-significant result at 0.41. We expected much smaller p-values given the placebo test results displayed in [AG03, ADH11]. One possible explanation is that in the aforementioned analyses of the Basque dataset, several regions were removed if they had poor fit for the treatment period. We do not incorporate these adjustments in the LTO inference procedure. Because the LTO p-value is non-significant, in practice a sensitivity analysis procedure would not be run. It is still interesting to see how the curve of maximum weighted p-values grows with Γ. Figure 7 shows the output for this example and the previous one.

Table 1: The table shows a summary of LTO and Placebo inference procedures applied to the case studies in [AG03, ADH10, ADH15]. The row entitled Γ_{LTO} shows the Γ at which the sensitivity analysis for the LTO p-value overturns the conclusion of p_{LTO}. The value for Basque Country is blank, as the p-value was not significant at level α.

	Proposition 99	Basque Country	German Reunification
N	38	17	17
α	0.05	0.05	0.05
p_{placebo}	0.026	0.41	0.059
p_{LTO}	0.024	0.67	0.042
Γ_{LTO}	1.55	NA	1.42

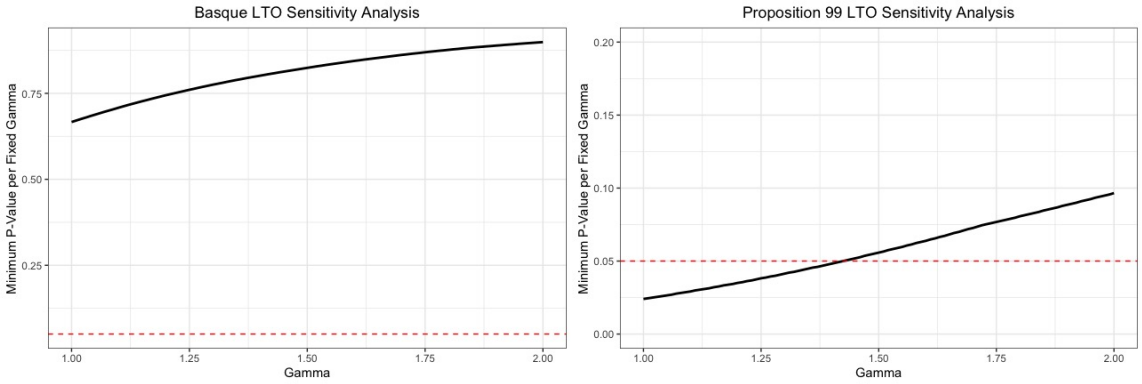

Figure 7: Left: curve of maximum weighted LTO p-value for the Basque country in the Spanish GDP dataset of [AG03]. The plot displays the maximum weighted LTO p-value over weights in B_Γ as a function of Γ. The red dashed line signifies the level $\alpha = 0.05$. As the unweighted LTO p-value is not significant, in practice this curve would not be computed. Right: output of a sensitivity analysis for the Proposition 99 smoking dataset of [ADH11]. The red dashed line signifies the level $\alpha = 0.05$. In both analyses, the RMSPE statistic was used in the construction of the synthetic control.

5 Discussion

The LTO procedure proposed in this paper was first motivated in an attempt to obtain more refined p-values, by which we mean it takes more than $O(n)$ values. Ideally, the p-value used for design-based SC inference or in more general randomization tests would take on and have minimum value less than $1/N$, following for useful inference in the $\alpha \leq 1/N$ regime where the placebo test has no power.

The issue of granularity of p-values for certain randomization based tests was observed a long time ago [Ber00, HG21]. Hemerik and Goeman [HG21] re-raise this issue in the context of randomized trials, where cohort of n units is randomly assigned to two treatment groups. Most commonly the randomization is chosen to enforce balance between the sizes of the two groups. As an alternative, the authors argue for the use of complete randomization by picking the treatment uniformly from the set $\{0,1\}^n$. This leads to p-values on the scale 2^{-n} as opposed to $\frac{1}{\binom{n}{n/2}}$, and consequently leads to greater power when the level α is very small. Situations where α is very small arise in multiple testing correction, which is increasingly applied in practice.

The impossibility result mentioned in the Introduction shows that it is not possible to obtain a useful p-value without a Type I error correction, but there are other useful potential ways to construct more refined p-values. For example, randomizing p-values can also solve this problem: whenever the p-value is $1/N$, flip a coin to randomize it. This allows for the p-value to be $1/2N$. As mentioned previously, we caution against the use of external randomization of this form. Although randomization may increase power, it allows the user to cheaply re-run the analysis with the possibility of obtaining statistical significance.

One could also combine several p-values obtained from randomization tests with slightly different statistics. In this example, instead of using the usual SC estimator we may use related techniques such as those proposed in [BISW21, DI16, CWZ21], to obtain p-values $p_1, \ldots, p_m$ for the null hypothesis, which are coarse in the sense that they take values in $\{1/N, 2/N, \ldots\}$. Then a more refined p-value might be obtained by considering $2(\sum_i p_i)/m$. This is a folklore result discovered and re-discovered by several authors, including Ruschendorf and Meng [Rüs82, Men94].

Extensions of the folklore result are systematically studied in the p-merging literature, where generalized averages of p-values, functionals of order statistics, and other proposals are shown to be p-values under arbitrary dependence. For further details, see [VW20, VWW22] and the references therein. Ideas from the p-

merging literature may let us combine both the LTO procedure and placebo procedure in fruitful ways. The power simulations of Section 4.1 suggest that the placebo p-value is more powerful in low signal to noise ratio regimes, while the LTO procedure has higher power in higher SNR regimes. It would be interesting to see if p-merging techniques such as [LX20] could combine LTO and placebo p-values to create ones with uniformly higher power and the empirical Type I error properties of the LTO procedure.

5.1 Extensions of the Leave-Two-Out procedure

We discuss several extensions of the Leave-two-out procedure. The first defines the notion of strange point in a different way, while the second notion leaves r points out at a time.

5.1.1 Rank Leave-Two-Out Jackknife+

Instead of labeling a unit as strange if it wins too many matches, we can consider a unit to be strange if its rank across all matches is too large. This leads to a variant of the procedure which has similar coverage guarantees as the original Jackknife+ procedure; see Theorem 1 of [BCRT21]. Computing the confidence set from this p-value in general requires inversion. Define $\mathrm{rank}\{I; I, i, j\} := \mathbb{I}\{I \succ j\} + \mathbb{I}\{I \succ i\}$, where in this context, $\{I \succ j\} := \{R_{I,i,j;I} > R_{I,i,j;j}\}$ and similarly for $\{I \succ i\}$. We can define a p-value

$$p_{LTO,\mathrm{rank}} := \frac{1}{(N-1)(N-2)} \sum_{\substack{j,k \in [N] \setminus I \\ j \neq k}} (2 - \mathrm{rank}\{I; I, i, j\}),$$

Similarly to p_{LTO}, there is a variant of $p_{LTO,\mathrm{rank}}$ which is a finite-sample valid p-value but has no power below $O(1/N)$, but we will not discuss this extension. This valid p-value may be created by normalizing by $1/N^2$ in the definition of $p_{LTO,\mathrm{rank}}$ above.

We may also define the related confidence set. For any given $y \in \mathbb{R}$, define $R^y_{I,j,k;I} = |y - \hat{Y}^{-\{I,j,k\}}_{I,t}|$ for any j, k. Using these residual quantities, we may define the analogous quantity $\mathrm{rank}^y\{I; I, i, j\}$ as above. Then the confidence set is given by

$$\mathrm{CS}_{LTO,\mathrm{rank}}(1-\alpha) := \left\{ y : \sum_{\substack{i,j \in [N] \setminus I \\ i \neq j}} \frac{1}{2} \mathrm{rank}^y\{I; I, i, j\} \leq (1 - \frac{\alpha}{2})(N-1)(N-2) \right\}$$

Notice that this is the same set as $\left\{y : p^y_{LTO,\text{rank}} \geq \alpha\right\}$, where $p^y_{LTO,\text{rank}}$ is the LTO p-value constructed with $Y_{I,t} = y$.

Theorem 5.1 (LTO Rank-Sum Jackknife+)**.** *$p_{LTO,\Sigma}$ is an approximate p-value:*

$$\mathbb{P}(p_{LTO,\Sigma} \leq \alpha) \leq \alpha + \frac{1-\alpha}{N}.$$

As an immediate corollary,

$$\mathbb{P}(Y_{I,t}(0) \notin \text{CS}_{LTO,\text{rank}}) \leq \alpha + \frac{1-\alpha}{N}.$$

A natural question to ask is which LTO procedure is more conservative. The next result provided shows that the p-values for the Rank Sum LTO procedure and Rank Threshold LTO procedure are sandwiched between one another, for different coverage levels.

Proposition 5.1 (Interleaving of Jackknife+ Confidence Intervals)**.**

$$\text{CS}_{LTO,Rank}(1-\alpha) \subseteq \text{CS}_{LTO}(1-\alpha) \subseteq \text{CS}_{LTO,Rank}(1-\frac{\alpha}{2})$$

Proof. Firstly, notice that $\mathbb{I}\{I \succ j\} + \mathbb{I}\{I \succ i\} \leq 2\mathbb{I}\{I \succ i,j\}$. This immediately implies

$$\sum_{\substack{i,j \in [N] \setminus I \\ i \neq j}} \frac{1}{2} \text{rank}\{I; I, i, j\} \leq \sum_{\substack{i,j \in [N] \setminus I \\ i \neq j}} \mathbf{1}\left\{R_{i,j,I;I} \geq R^{LTO}_{i,j}\right\},$$

and so $\text{CS}_{LTO}(1-\alpha) \supseteq \text{CS}_{LTO,\text{Rank}}(1-\alpha)$. On the other hand, suppose that for exactly $\gamma(N-1)(N-2)$ pairs of indices, the treated unit I wins the comparison against the pair:

$$\sum_{i,j} \mathbb{I}\{I \succ i,j\} = \gamma(N-1)(N-2).$$

Then $\sum_{i,j} \frac{1}{2}\mathbf{1}\{\text{rank}\{I; I, i, j\} = 1\} \leq \frac{(1-\gamma)}{2}(N-1)(N-2)$. Adding these inequalities, we find that

$$\sum_{i,j} \frac{1}{2} \text{rank}\{I; I, i, j\} \leq (\frac{1}{2} + \frac{\gamma}{2})(N-1)(N-2).$$

Thus if $\sum_{i,j} \mathbb{I}\{I \succ i,j\} \leq (1-\alpha)(N-1)(N-2)$, then $\sum_{i,j} \frac{1}{2} \text{rank}\{I; I, i, j\} \leq (1-\frac{\alpha}{2})(N-1)(N-2)$. □

5.1.2 Leave-r-out Jackknife+

We may consider an analogous procedure for Jackknife+ where r-tuples of data points are left out. The analysis of the method in this setting clarifies the essential features of the LTO case. Leaving out $r \geq 2$ points at a time allows for p-values with even smaller minimum values, which may be a boon over the leave-two-out method in some situations. For clarity, the results for LRO case will be phrased in terms of confidence sets.

Define the residual

$$R_{(i_1,\ldots,i_r);i} = |Y_{i,t} - \hat{\mu}_{-(i_1,\ldots,i_r)}(Y_{i,\bullet})|$$

Define $\operatorname{rank}\{i_0; i_1, \ldots, i_r\} = \sum_{t=1}^{r} \mathbf{1}\left\{R_{(i_1,\ldots,i_r);i_0} > R_{(i_1,\ldots,i_r);i_t}\right\}$. Then, we can create the confidence sets

$$\mathrm{CS}_{\mathrm{Win}}(\alpha) = \left\{ Y_{I,t} : \sum_{\{i_1,\ldots,i_r\}} \mathbb{I}\{I \succ i_1, \ldots, i_r\} \leq \frac{1}{\lfloor \alpha N \rfloor}\left[\binom{N}{r+1} - \binom{N - \lfloor \alpha N \rfloor}{r+1}\right]\right\}$$

$$\mathrm{CS}_{\mathrm{Rank}}(\alpha) = \left\{ Y_{I,t} : \sum_{\{i_1,\ldots,i_r\}} \frac{1}{r} \operatorname{rank}\{I; i_1, \ldots, i_r\} \leq (1-\alpha)\binom{N}{r}\right\}.$$

Creating these confidence sets for both of these procedures requires grid search in general. Practically, however, running this method is computationally intensive for synthetic controls. In math, the two quantities above satisfy the following guarantees:

Theorem 5.2 (Leave-r-out Jackknife+ guarantees)**.**

$$\mathbb{P}(Y_{I,t}(0) \notin \mathrm{CS}_{\mathit{Win}}(\alpha)) \leq 2\alpha + \frac{1-2\alpha}{N}$$

$$\mathbb{P}\left(Y_{I,t}(0) \notin \mathrm{CS}_{\mathit{Rank}}(\alpha)\right) \leq \alpha$$

Heuristically the threshold $Q := \frac{1}{\lfloor \alpha N \rfloor}\left[\binom{N}{r+1} - \binom{N-\lfloor \alpha N \rfloor}{r+1}\right]$ can be given a nice interpretation. For large N and small α, Q is approximately equal to $\frac{1}{N\alpha}\left[\frac{N^{r+1}}{(r+1)!} - \frac{((1-\alpha)N)^{r+1}}{(r+1)!}\right]$, which is equal to

$$\frac{1-(1-\alpha)^{r+1}}{\alpha}\frac{N^r}{(r+1)!}. \tag{10}$$

When α is small, $(1-\alpha)^{r+1} \approx 1 - (r+1)\alpha + \binom{r+1}{2}\alpha^2$. Using this approximation, equation (10) simplifies to $(1 - \frac{r}{2}\alpha)\frac{N^r}{r!}$, which is effectively

$$(1 - \frac{r}{2}\alpha)\binom{N}{r}. \tag{11}$$

Under this heuristic, an approximation of the confidence sets for the value of the counterfactual $Y_{I,t}(0)$ can be written as

$$\widehat{\mathrm{CS}}_{\mathrm{Win}}(\alpha) = \left\{ y_{I,t} : \sum_{\{i_1,\ldots,i_r\}} \mathbb{I}\{I \succ i_1,\ldots,i_r\} \leq (1-\alpha)\binom{N}{r} \right\}$$
$$\widehat{\mathrm{CS}}_{\mathrm{Rank}}(\alpha) = \left\{ y_{I,t} : \sum_{\{i_1,\ldots,i_r\}} \frac{1}{r}\,\mathrm{rank}\,\{I; i_1,\ldots,i_r\} \leq (1-\alpha)\binom{N}{r} \right\}.$$

By the results of Theorem 5.2, the miscoverage rate of $\widehat{\mathrm{CS}}_{\mathrm{Win}}(\alpha)$ should be roughly $\frac{2}{r}\alpha$, while the miscoverage of $\widehat{\mathrm{CS}}_{\mathrm{Rank}}(\alpha)$ should be about 2α. This can likely be made precise in the asymptotic limit as $\alpha \to 0, n \to \infty$.

There is an interleaving between the approximate confidence sets of these procedures. Firstly, notice that $\sum_{\{i_1,\ldots,i_r\}} \mathbb{I}\{I \succ i_1,\ldots,i_r\} \leq \sum_{\{i_1,\ldots,i_r\}} \frac{1}{r}\,\mathrm{rank}\,\{I; i_1,\ldots,i_r\}$. This immediately implies that $\widehat{\mathrm{CS}}_{\mathrm{Rank}}(\alpha) \subseteq \widehat{\mathrm{CS}}_{\mathrm{Win}}(\alpha)$. Further, suppose that

$$\sum_{\{i_1,\ldots,i_r\}} \mathbb{I}\{I \succ i_1,\ldots,i_r\} = \gamma\binom{N}{r},$$

for some $\gamma \in [0,1]$. Then

$$\sum_{k=0}^{r-1} \sum_{\{i_1,\ldots,i_r\}} \frac{k}{r}\mathbf{1}\{\mathrm{rank}\,\{I; i_1,\ldots,i_r\} = k\} \leq \frac{r-1}{r}(1-\gamma)\binom{N}{r}.$$

Adding these two bounds, we find $\sum_{\{i_1,\ldots,i_r\}} \frac{1}{r}\,\mathrm{rank}\,\{I; i_1,\ldots,i_r\} \leq \left(\frac{r-1}{r} + \frac{1}{r}\gamma\right)\binom{N}{r}$. These two bounds show that if $\gamma \leq (1 - \frac{r}{2}\alpha)$ then $\sum_{\{i_1,\ldots,i_r\}} \frac{1}{r}\,\mathrm{rank}\,\{I; i_1,\ldots,i_r\} \leq (1 - \frac{\alpha}{2})$. This proves the inclusion $\widehat{\mathrm{CS}}_{\mathrm{Win}}(\frac{r}{2}\alpha) \subseteq \widehat{\mathrm{CS}}_{\mathrm{Rank}}(1 - \frac{\alpha}{2})$. The miscoverages of these two intervals are both theoretically bounded by α according to Theorem 5.2. This suggests that the confidence set produced by the Leave-r-out rank-jackknife+ is shorter than that of the LRO max-jackknife+, for the same coverage level.

5.2 Alternative Sensitivity Analysis Procedures

Definition 5.1.

$$p_{\pi,sum} := \sum_{\substack{j\neq k \\ j,k\neq I}} \frac{\pi_j + \pi_k}{2(N-2)(1-\pi_I)} \mathbf{I}\{I \not\succ j,k\}$$

Proposition 5.2. *We have the following results for the refined $\alpha \le 1/N$ p-values, assuming $\pi_i \in [\frac{1}{\Gamma N}, \frac{\Gamma}{N}]$ for all i. Let $\beta = \alpha\frac{N-2}{N-1} + \frac{1}{N-2}$. Then*

$$\mathbb{P}(p_{\pi,sum} \le \alpha) \le \beta + \sqrt{\beta^2 + [(1-2\alpha)\frac{N-1}{N-2} - \frac{1}{N-1}]\sum_{k\in\mathcal{S}} \pi_k^2}.$$

From the formula of the Type I error of $p_{\pi,\text{sum}}$, when $\Gamma \approx 1$ and N is large, the Type I error when is approximately 2α.

Weighted analogs of valid LTO p-values. If inference is conducted using the $\alpha > 1/N$ p-values, a similar sensitivity analysis procedure may be used, using the following weighted analogs of the p-values in Section 2.4

Definition 5.2.

$$p_{\pi,V,sum} := \sum_{\substack{j\neq k \\ j,k\neq I}} \frac{\pi_j + \pi_k}{2(N-1)(1-\pi_I)} \mathbf{I}\{I \not\in j,k\} + \frac{1}{N-1}$$

When $\pi_i = 1/N$ for all i, the p-value reduces to those introduced in Section 2.4 respectively. The Type I error guarantees are also similar to those above.

Proposition 5.3. *We have the following results for the refined $\alpha \ge 1/N$ p-value, assuming $\pi_i \in [\frac{1}{\Gamma N}, \frac{\Gamma}{N}]$ for all i.*

$$\mathbb{P}(p_{\pi,V,sum} \le \alpha) \le \alpha + \sqrt{\alpha^2 + (1-2\alpha)\sum_{k\in\mathcal{S}} \pi_k^2}$$

Computational Details. The optimization procedure for $p_{\pi,\text{sum}}$ is comparatively more straightforward than that of $p_{\pi,\text{product}}$. Using the same approach, for any fixed Γ we need only check whether $\max_{\pi\in B_\Gamma} p_{\pi,\text{sum}} \ge \alpha_0$. Because $2(N-2)(1-\pi_I) > 0$, it suffices to check whether

$$\max_{\pi\in B_\Gamma} \sum_{j\neq k, j,k\neq I} (\pi_j + \pi_k)G_{jk} - 2\alpha_0(N-2)(1-\pi_I) > 0.$$

The resulting optimization problem is a linear program. We may rewrite the problem as verifying whether

$$\max_{\pi\in B_\Gamma} w^\top \pi > 2\alpha_0(N-2),$$

for the vector $w := \text{diag}((G+G^\top)E)+2\alpha_0(N-2)e_I$, where G is the matrix described above and E is another zero one matrix, such that $E_{jk} = 0$ whenever $j = k$ or $j = I$. This is computationally tractable. In numerical simulations of this method, we used the R package lpSolve [B+15] to solve the linear programming problem for $p_{\pi,\text{sum}}$.

For $p_{\pi,V,\text{sum}}$, the same approach works. We only need to check if $\max_{\pi\in B_\Gamma} p_{\pi,V,\text{sum}}$ is greater than α_0, which reduces to checking that

$$\max_{\pi\in B_\Gamma} \sum_{j\neq k, j,k\neq I} (\pi_j + \pi_k)G_{jk} + 2(1-\pi_I) > 2\alpha_0(N-1)(1-\pi_I).$$

This is equivalent to verifying whether the maximal objective of the linear program $\max_{\pi\in B_\Gamma} \tilde{w}^\top\pi$ with $\tilde{w} := \text{diag}((G+G^\top)E) + (2\alpha_0(N-1)-2)e_I$ is greater than $2\alpha_0(N-1)-2$.

6 Acknowledgements

TS acknowledges support from the NSF Graduate Research Fellowship Program under Grant DGE-1656518.

References

[AAH+21] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021. 14, 15

[Aba21] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021. 14

[ABD+21] 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*, 116(536):1716–1730, 2021. 2

[ADH10] 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(490):493–505, 2010. 2, 3, 18, 19, 20, 25, 26

[ADH11] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13), 2011. 2, 23, 25, 26

[ADH15] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015. 3, 4, 5, 6, 25, 26

[AG03] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003. 2, 3, 14, 18, 23, 25, 26, 37

[ASS18] Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018. 2, 14, 15

[B+15] Michel Berkelaar et al. Package 'lpsolve', 2015. 33

[BCRT21] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. 7, 10, 15, 28

[Ber00] Vance W Berger. Pros and cons of permutation tests in clinical trials. *Statistics in medicine*, 19(10):1319–1328, 2000. 27

[BISW21] Lea Bottmer, Guido Imbens, Jann Spiess, and Merrill Warnick. A design-based perspective on synthetic control methods. *arXiv preprint arXiv:2101.09398*, 2021. 14, 27

[CFPT22] Matias D Cattaneo, Yingjie Feng, Filippo Palomba, and Rocio Titiunik. Uncertainty quantification in synthetic controls with staggered treatment adoption. *arXiv preprint arXiv:2210.05026*, 2022. 14

[CFT21] Matias D Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021. 14

[CWZ21] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021. 14, 15, 27

[DI16] 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. 15, 27

[FP18] Sergio Firpo and Vitor Possebom. Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2), 2018. 2, 3, 14, 16, 25

[HG21] Jesse Hemerik and Jelle J Goeman. Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*, 89(2):367–381, 2021. 27

[HS17] Jinyong Hahn and Ruoyao Shi. Synthetic control and inference. *Econometrics*, 5(4):52, 2017. 14

[LX20] Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020. 28

[Men94] Xiao-Li Meng. Posterior predictive p-values. *The annals of statistics*, 22(3):1142–1160, 1994. 27

[Rüs82] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982. 27

[ST21] Azeem M Shaikh and Panos Toulis. Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848, 2021. 14

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005. 15

[Vov15] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015. 15

[VW20] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020. 27

[VWW22] Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022. 27

A Additional Simulation Results

In addition to the results provided in Section 4.1, we provide a few more results for completeness on the Proposition 99 dataset. The first additional result is a power comparison between the standard placebo, LTO, and the powered LTO p-value in the $\alpha > 1/N$ regime, when evaluating the synthetic controls using the absolute difference at year 2000. Figure 8 displays these results. The advantages of the LTO procedure are not as prnounced here. The powered LTO has slightly lower average Type I error and comparable or slightly greater power than the placebo. This setting reinforces the observation that construction of the synthetic control using the RMSPE statistic highlights the benefits of the LTO procedure.

Next, it is interesting to consider how these conclusions may change when the synthetic controls are evaluated at a year other than year 2000. Figures 9 and 10 show the Type I error and power for large effect size when evaluating the synthetic controls at different years in the range 1996 - 2000. We draw the same conclusions on the same benefits of the LTO procedure as before, but dependent on the year, the empirical Type I error is larger than for other years.

In addition, some results are provided in Figures 11, 12 for the Basque country dataset of [AG03]. It is interesting to note that when constructing the placebo test and LTO statistic using the difference at a fixed year, the Type I errors are almost comparable for several years. See the left panel of Figure 12.

B Proofs

B.1 Proofs for Section 2

Full Proof of Thm. 2.1. Let us start with the decomposition of the sum outlined in Section 2:

$$\overbrace{\sum_{i,j,k\in\mathcal{S}} \mathbb{I}(k \succ i,j)}^{\text{(I)}} + \overbrace{2\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S},j\in\mathcal{S}^c} \mathbb{I}(k \succ i,j)}^{\text{(II)}} + \overbrace{\sum_{k\in\mathcal{S}}\sum_{i,j\in\mathcal{S}^c} \mathbb{I}(k \succ i,j)}^{\text{(III)}}.$$

For the first sum, note by renaming the labels that $\sum_{i,j,k\in\mathcal{S}} \mathbb{I}(k \succ i,j) = \sum_{i,j,k\in\mathcal{S}} \mathbb{I}(i \succ j,k) = \sum_{i,j,k\in\mathcal{S}} \mathbb{I}(j \succ k,i)$. Thus

$$\text{(I)} \leq \sum_{i,j,k\in\mathcal{S}} \frac{1}{3}\left(\mathbb{I}(k \succ i,j) + \mathbb{I}(i \succ j,k) + \mathbb{I}(j \succ k,i)\right) \leq s(s-1)(s-2)/3.$$

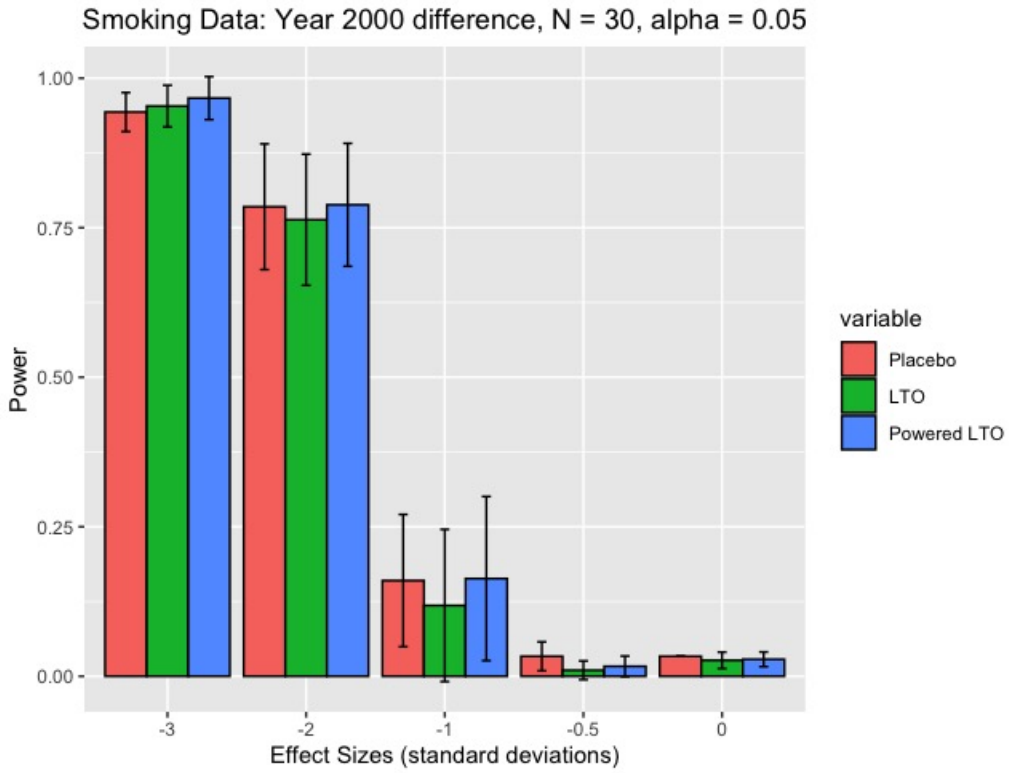

Figure 8: Power of several different p-values versus effect size, on size 30 subsamples of the California Proposition 99 dataset. The rightmost column indicates Type I error of each method. Inference is done with $\alpha = 0.05$. The x-axis is the effect size τ scaled in terms of multiples of the standard deviation of cigarette sales. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of the size 30 subsample. Three different methods are compared, using the absolute difference at the year 2000. As $\alpha > 1/N$ in this setting, the LTO procedure is constructed to be finite sample valid.

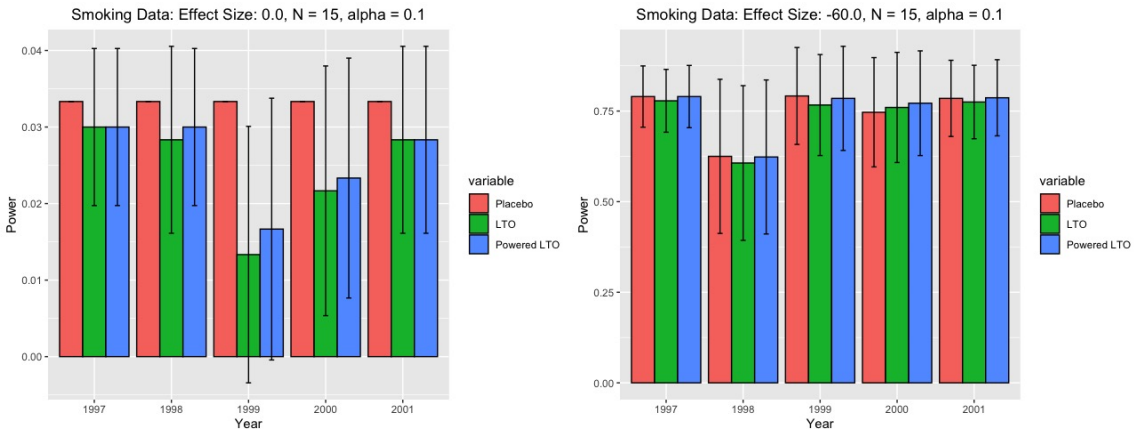

Figure 9: Power of several different p-values, grouped by year. Left figure denotes Type I error; right figure displays power with respect to roughly 2 standard deviations of the outcome variable. The x-axis signifies the year with which the synthetic controls were compared on. The rightmost column indicates Type I error of each method. Inference is done with $\alpha = 0.05$. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of size 30 subsample of the original dataset. Three different methods are compared, using the absolute difference at the year indicated on the x-axis.

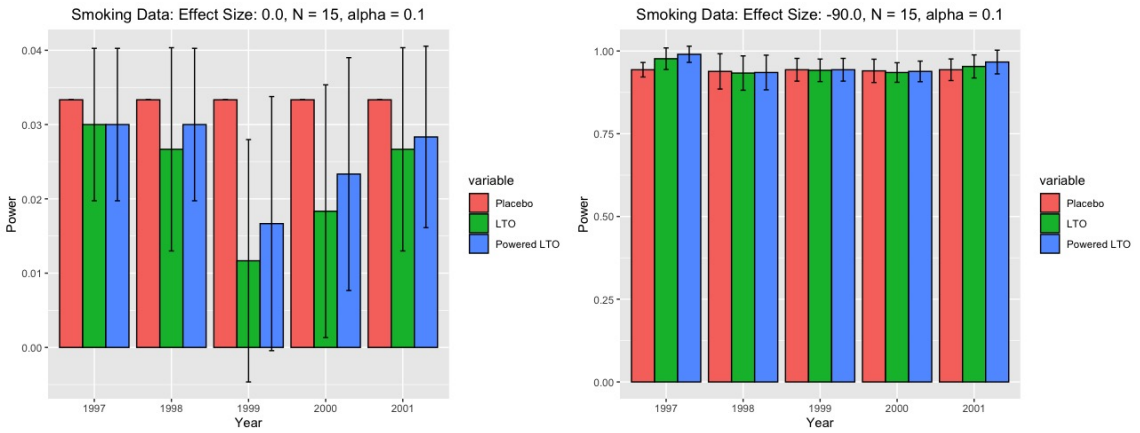

Figure 10: Power of several different p-values, grouped by year. Left figure denotes Type I error; right figure displays power with respect to roughly 2 standard deviations of the outcome variable. The x-axis signifies the year with which the synthetic controls were compared on. The rightmost column indicates Type I error of each method. Inference is done with $\alpha = 0.05$. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of size 30 subsample of the original Prop. 99 smoking dataset. Three different methods are compared, using the absolute difference at the year indicated on the x-axis. As $\alpha > 1/N$ in this setting, the LTO procedure is constructed to be finite sample valid.

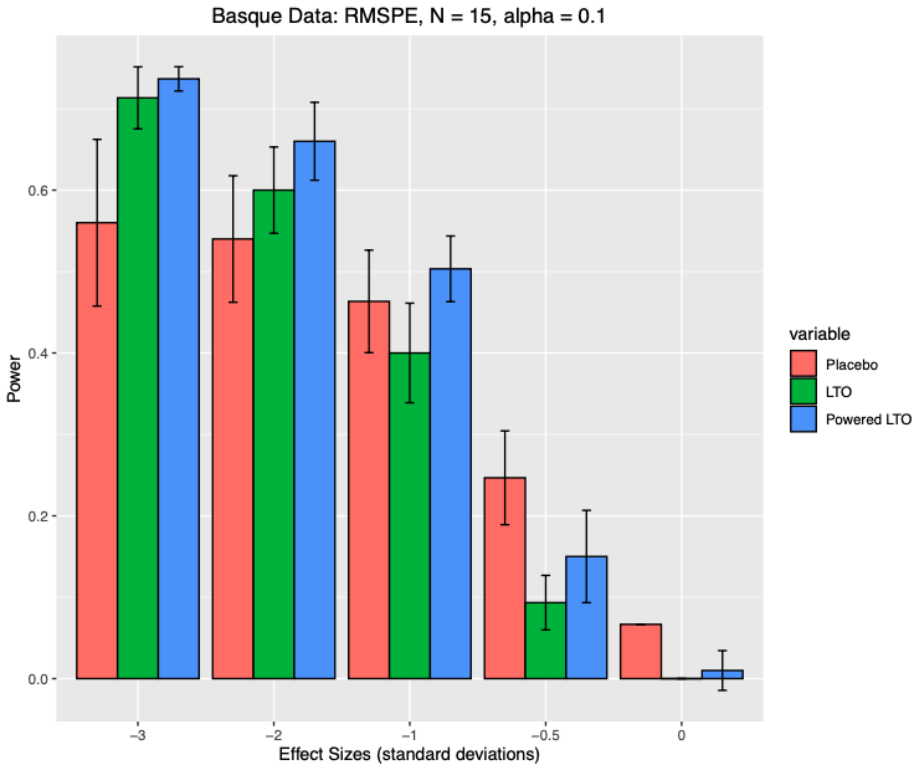

Figure 11: Power of several different p-values versus effect size, on size 15 subsamples of the Basque country dataset. The rightmost column indicates Type I error of each method. Inference is done with $\alpha = 0.05$. The x-axis is the effect size τ scaled in terms of multiples of the standard deviation of cigarette sales. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of the size 15 subsample. Three different methods are compared, using the RMSPE statistic. As $\alpha > 1/N$ in this setting, the LTO procedure is constructed to be finite sample valid.

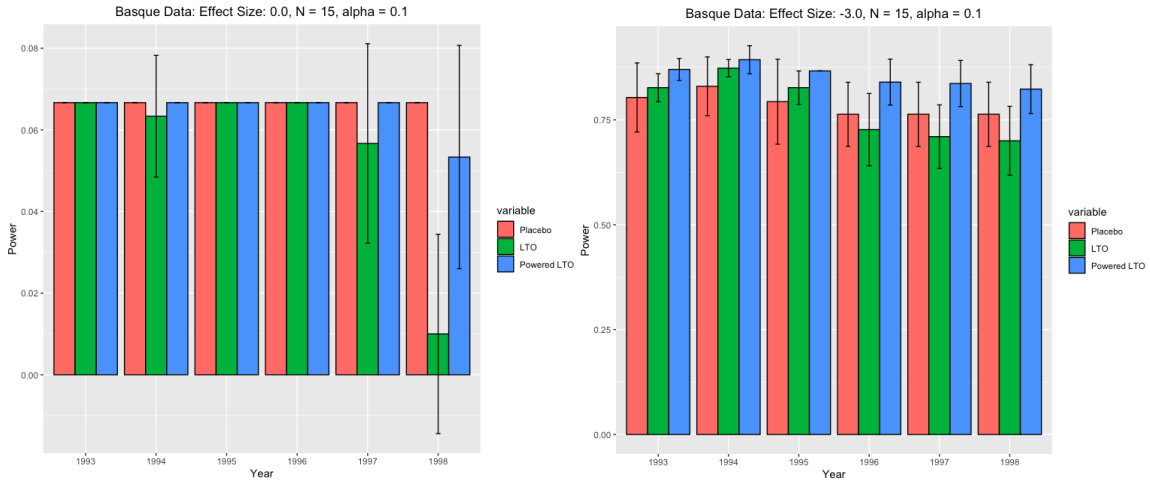

Figure 12: Power of several different p-values, grouped by year. Left figure denotes Type I error; right figure displays power with respect to roughly 2 standard deviations of the outcome variable. The x-axis signifies the year with which the synthetic controls were compared on. The rightmost column indicates Type I error of each method. Inference is done with $\alpha = 0.1$. Error bars indicate one standard deviation variability of the power over 20 Monte Carlo samples of size 15 subsample of the original Basque country dataset. Three different methods are compared, using the absolute difference at the year indicated on the x-axis. As $\alpha > 1/N$ in this setting, the LTO procedure is constructed to be finite sample valid.

For the second sum, swap the naming of the i, k labels and use the bound $\mathbb{I}(k > i, k) + \mathbb{I}(i > k, j) \leq 1$ to see that

$$\text{(II)} \leq 2 \sum_{k \in \mathcal{S}} \sum_{i \in \mathcal{S}, j \in \mathcal{S}^c} \frac{1}{2} \left(\mathbb{I}(k > i, k) + \mathbb{I}(i > k, j) \right) \leq s(s-1)(N-s).$$

Finally, in the last sum, we use the naive bound $\text{(III)} \leq s(N-s)(N-1-s)$.

Combining these bounds, canceling a factor of s, we find that

$$(1-\alpha)(N-1)(N-2) \leq (s-1)(s-2)/3 + (s-1)(N-s) + (N-s)(N-1-s).$$

The right hand side reduces to

$$\begin{aligned} &\frac{1}{3}s^2 - s + \frac{2}{3} + sN - N - s^2 + s + (N^2 - N - Ns - sN + s + s^2) \\ &= \frac{1}{3}s^2 + \frac{2}{3} - s(N-1) - 2N + N^2. \end{aligned}$$

Thus

$$\frac{1}{3}s^2 + \frac{2}{3} - s(N-1) - 2N + N^2 - (N^2 - 3N + 2) + \alpha(N-1)(N-2) \geq 0,$$

which reduces to

$$\frac{s^2}{3} - \frac{4}{3} - s(N-1) + N + \alpha(N-1)(N-2) \geq 0$$

Introducing $\beta = s/N$, we may reduce to the quadratic inequality

$$\frac{\beta^2}{3} - \beta(1 - \frac{1}{N}) - \frac{4}{3N^2} + \frac{1}{N} + \alpha(1 - \frac{1}{N})(1 - \frac{2}{N}) \geq 0.$$

We can solve this by quadratic formula. The right endpoint is greater than 1, so we conclude that

$$\beta \leq \frac{3 - \frac{3}{N} - \sqrt{9(1-\frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \alpha(1-\frac{1}{N})(1-\frac{2}{N})\right)}}{2}. \tag{12}$$

To conclude, we note that $\mathbb{P}(p_{LTO} \leq \alpha) = \mathbb{P}(I \in \mathcal{S}) = \beta$. The bound in equation (5) shows that p_{LTO} is an approximate p-value.

□

Proof of Corollary 2.1. We will show that $f(N, \frac{1}{N-1}) < \frac{2}{N}$. Noting that $f(N, \alpha)$ is continuous and increasing in α, this shows that $f(N, \alpha) < \frac{2}{N}, \forall \alpha < 1/N$. As a consequence of Theorem 2.1, it follows that $\mathbb{P}(p_{LTO} \leq \alpha) \leq \frac{1}{N}$ for all $\alpha < 1/N$.

Checking the inequality $f(N, \frac{1}{N-1}) < \frac{2}{N}$ reduces to verifying

$$3 - \frac{3}{N} - \sqrt{9(1-\frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \frac{1}{N}(1-\frac{1}{N})(1-\frac{2}{N})\right)} < \frac{4}{N},$$

which is equivalent to

$$(3 - \frac{3}{N} - \frac{4}{N})^2 < 9(1-\frac{1}{N})^2 - 12\left(-\frac{4}{3N^2} + \frac{1}{N} + \frac{1}{N}(1-\frac{1}{N})(1-\frac{2}{N})\right).$$

Expanding and cancelling the $9(1-\frac{1}{N})^2$ terms, this is equivalent to

$$-\frac{24}{N}(1-\frac{1}{N}) + \frac{16}{N^2} < -\frac{12}{N} + \frac{16}{N^2} - \frac{12}{N}(1-\frac{1}{N})(1-\frac{2}{N}),$$

which is satisfied as long as $N > 2$.

□

Proof of 2.1. For this proof, call a unit k *strange* if

$$\sum_{\substack{i,j \in [N] \setminus k \\ i \neq j}} \mathbb{I}(k \succ i, j) \geq (1-\alpha)(N-1)(N-1).$$

Let $\mathcal{S}$ denote the set of strange units, and let $s := |\mathcal{S}|$. Summing the definition of strangeness on both sides, we see that

$$\sum_{k \in \mathcal{S}} \sum_{i,j \in [N]} \mathbb{I}(k \succ i, j) \geq (1-\alpha)s(N-1)(N-1).$$

Using the same decomposition as in the proof of 2.1, we find that

$$\frac{s(s-1)(s-2)}{3} + s(s-1)(N-s) + s(N-s)(N-1-s) \geq (1-\alpha)s(N-1)^2,$$

introducing $\beta = s/N$, this reduces to the quadratic inequality

$$\frac{\beta^2}{3} - \beta(N-1) - \frac{1}{3N^2} + \alpha(1-\frac{1}{N})^2$$

Solving this yields

$$\beta \leq \frac{3 - \frac{3}{N} - \sqrt{9(1-\frac{1}{N})^2 - 12[\alpha(1-\frac{1}{N})^2 - \frac{1}{3N^2}]}}{2}.$$

Finally, note that unit I is strange if and only if $p_{LTO,V} \leq \alpha$. □

B.2 Proofs for Section 3

Proof of Proposition 3.1. We first analyze the refined weighted p-value. Because $\sum_{\substack{j\neq k\\ j,k\neq I}} \pi_j\pi_k = (1-\pi_I)^2 - \sum_{l\neq I}\pi_l^2$, the event that the p-value is less than α is equal to the event that

$$\sum_{\substack{i,j\neq I\\ j\neq i}} \frac{\pi_i\pi_j}{(1-\pi_I)^2 - \sum_{l\neq I}\pi_l^2}\mathbb{I}\{I \succ i,j\} \geq (1-\alpha)$$

Define a unit k to be strange if it satisfies the above bound in place of I. Call $\mathcal{S}$ the set of strange points and $s = |\mathcal{S}|$. Then multiply on both sides by $\pi_k((1-\pi_k)^2 - \sum_{l\neq k}\pi_l^2)$ and sum over $k \in \mathcal{S}$ to obtain the inequality

$$\sum_{k\in\mathcal{S}}\sum_{i,j}\pi_i\pi_j\pi_k\mathbb{I}\{I \succ j,k\} \geq (1-\alpha)\left(\sum_{k\in\mathcal{S}}\pi_k(1-\pi_k)^2 - \sum_{k\in\mathcal{S},l\neq k}\pi_k\pi_l^2\right) \tag{13}$$

the left hand side of the above may be split into three terms. The sums are over distinct indices.

$$\overbrace{\sum_{i,j,k\in\mathcal{S}}\pi_k\pi_i\pi_j\mathbb{I}(k\succ i,j)}^{\text{(I)}} + \overbrace{2\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S},j\in\mathcal{S}^c}\pi_k\pi_i\pi_j\mathbb{I}(k\succ i,j)}^{\text{(II)}} + \overbrace{\sum_{k\in\mathcal{S}}\sum_{i,j\in\mathcal{S}^c}\pi_k\pi_i\pi_j\mathbb{I}(k\succ i,j)}^{\text{(III)}}.$$

For sum (I), split the sum into three copies of itself, and permute the labels cyclically. That is, by switching the indexing labels, we note $\sum_{i,j,k\in\mathcal{S}}\pi_k\pi_i\pi_j\mathbb{I}(k\succ i,j) = \sum_{i,j,k\in\mathcal{S}}\pi_k\pi_i\pi_j\mathbb{I}(i\succ j,k) = \sum_{i,j,k\in\mathcal{S}}\pi_k\pi_i\pi_j\mathbb{I}(j\succ k,i)$. Next, because at most one of the residuals $R_{i,j,k;\cdot}$ can be the largest, $\mathbb{I}(i\succ j,k)+\mathbb{I}(j\succ k,i)+\mathbb{I}(k\succ i,j)\leq 1$. This leads to the bound

$$\text{(I)} \leq \frac{1}{3}\sum_{i,j,k\in\mathcal{S}}\pi_i\pi_j\pi_k = \frac{1}{3}[p_*^3 - 3\sum_{i,j}\pi_i^2\pi_j + 2\sum_i\pi_i^3].$$

For the second sum (II), switching the labels of i,k gives the identity

$$\sum_{k\in\mathcal{S}}\sum_{i\in\mathcal{S},j\in\mathcal{S}^c}\pi_k\pi_i\pi_j\mathbb{I}(k\succ i,j) = \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S},j\in\mathcal{S}^c}\pi_k\pi_i\pi_j\mathbb{I}(i\succ k,j).$$

Bounding the indicator sum $\mathbb{I}(k\succ i,j)+\mathbb{I}(i\succ k,j)$ by one, we have the bound

$$\text{(II)} \leq \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S},j\in\mathcal{S}^c}\pi_k\pi_i\pi_j = [p_*^2 - \sum_{i\in\mathcal{S}}\pi_i^2](1-p_*).$$

Lastly, we bound (III) by the naive bound; replacing the indicator by the constant 1, the sum can be bounded by $p_*[(1-p_*)^2 - \sum_{i\in\mathcal{S}^c}\pi_i^2]$.

Summing these bounds, we find

$$\frac{1}{3}p_*^3 - p_*^2 + p_*(1-\sum_{i\in\mathcal{S}^c}\pi_i^2) + \frac{2}{3}\sum_{i\in\mathcal{S}}\pi^3 \geq (1-\alpha)\left(\sum_{k\in\mathcal{S}}\pi_k(1-\pi_k)^2 - \sum_{k\in\mathcal{S},l\neq k}\pi_k\pi_l^2\right)$$
$$= (1-\alpha)\left(p_* - 2\sum_{k\in\mathcal{S}}\pi_k^2 + 2\sum_{k\in\mathcal{S}}\pi_k^3 - p_*\sum_{l=1}^{n}\pi_l^2\right).$$

After some simplification, the resulting bound gives

$$\frac{1}{3}p_*^3 - p_*^2 + p_*\left(\sum_{i\in\mathcal{S}}\pi_i^2 + \alpha(1-\sum_l\pi_l^2)\right) \geq 2(1-\alpha)[\sum_{k\in\mathcal{S}}\pi_k^2 - \pi_k^3]$$

Using the fact that $\pi_k - \pi_k^2$ is minimized at $\pi_k = \frac{1}{N\Gamma}$ under the constraint that $\pi_k \in [\frac{1}{N\Gamma}, \frac{\Gamma}{N}]$, so long as $\Gamma << N$. we may lower bound the right hand side by $2(1-\alpha)p_*(\frac{1}{N\Gamma} - \frac{1}{N^2\Gamma^2})$. Now we may cancel a factor of p_*, which yields the quadratic inequality

$$\frac{1}{3}p_*^2 - p_* + \left[\sum_{\mathcal{S}}\pi_i^2 + \alpha(1-\sum_l\pi_l^2) - 2(1-\alpha)(\frac{1}{N\Gamma} - \frac{1}{N^2\Gamma^2})\right] \geq 0.$$

This may be solved to yield the following quadratic inequality:

$$p_* \leq \frac{3}{2} - \frac{1}{2}\sqrt{9 - 12\left(\alpha(1-\sum_l\pi_l^2) + \sum_{\mathcal{S}}\pi_i^2 - 2(1-\alpha)(\frac{1}{N\Gamma} - \frac{1}{N^2\Gamma^2})\right)}$$

□

Proof of Proposition 3.2. To analyze this expression, first notice that $\{p_{\pi,LTO} \leq \alpha\}$ is equivalent to the event

$$\sum_{\substack{j,k\neq I\\ j\neq k}}\frac{\pi_j\pi_k}{1-\sum_l\pi_l^2}\mathbb{I}\{I \succ j,k\} \geq (1-\alpha).$$

Let unit k be a strange point if it satisfies the inequality

$$\sum_{\substack{i,j\neq k\\ j\neq i}}\frac{\pi_i\pi_j}{1-\sum_l\pi_l^2}\mathbb{I}\{k \succ i,j\} \geq (1-\alpha).$$

Let $p_* = \sum_{i \in S} \pi_i$, which is the probability that I is a strange point. Let $\mathcal{S}$ be the set of strange points. Upon multiplying by $(1 - \sum_l \pi_l^2)\pi_k$ on both sides and summing over $k \in \mathcal{S}$, we have

$$\sum_{k \in \mathcal{S}} \sum_{i,j} \pi_i \pi_j \pi_k \mathbb{I}\{I \succ j, k\} \geq (1-\alpha) p_* (1 - \sum_l \pi_l^2).$$

Imitate the bound on the left hand side made in the proof of Prop. 3.1. The only modification we make is a bound on the first term of the three-part decomposition.

$$\frac{1}{3} \sum_{i,j,k \in \mathcal{S}} \pi_i \pi_j \pi_k \leq \frac{1}{3}[p_*^3 - 3p_* \sum_{i \in \mathcal{S}} \pi_i^2 + 2 \sum_i \pi_i^2 \frac{\Gamma}{N}].$$

Summing the bounds, we find

$$\frac{1}{3} p_*^3 + (\frac{2\Gamma}{3N} - 1) \sum_{i \in \mathcal{S}} \pi_i^2 - p_*^2 + p_* - p_* \sum_{i \in \mathcal{S}^c} \pi_i^2 \geq (1-\alpha) p_* (1 - \sum_l \pi_l^2)$$

We may reduce this further. Noting that $\frac{2\Gamma}{3N} - 1 \leq -\frac{\Gamma}{3N}, \sum \pi_i^2 \geq p_* \frac{1}{\Gamma N}$ we may upper bound $(\frac{2\Gamma}{3N} - 1) \sum_{i \in \mathcal{S}} \pi_i^2 \leq -\frac{p_*}{3N^2}$. Using this upper bound and dividing through by p_* we obtain

$$\frac{1}{3} p_*^2 - p_* + \sum_{i \in \mathcal{S}} \pi_i^2 - \frac{1}{3N^2} \geq (1-\alpha)(1 - \sum_l \pi_l^2).$$

Solving the resulting quadratic equation gives the following expression for p_*:

$$\frac{3}{2} - \frac{1}{2} \sqrt{9 - 12 \left(\sum_{i \in \mathcal{S}} \pi_i^2 - \frac{1}{3N^2} - (1-\alpha)(1 - \sum_l \pi_l^2) \right)}$$

□

B.3 Proofs for Section 5.1.1

Proof of Theorem 5.1. In this proof, call a unit i *strange* if

$$\sum_{\substack{j,k \in [N] \setminus I \\ j \neq k}} \frac{1}{2} (\mathbb{I}\{i \succ j\} + \mathbb{I}\{i \succ k\}) \geq (1-\alpha)(N-1)(N-2).$$

Let $\mathcal{S}$ be the set of strange points and $s := \mathcal{S}$. Summing this inequality over all strange points, we find

$$2(1-\alpha)(N-1)(N-2)s \leq \sum_{i \in \mathcal{S}} \sum_{j,k \in [N]} \mathbb{I}\{i \succ j\} + \mathbb{I}\{i \succ k\}.$$

Again, we will assume all indices in the summations are distinct, and omit this from notation. Let us split this sum into three cases. The first case sums over $i, j, k \in \mathcal{S}$; the second is twice the summation over $i, j \in \mathcal{S}, k \in \mathcal{S}^c$, and the third case sums over $i \in \mathcal{S}, j, k \in \mathcal{S}^c$.

The first sum is the quantity $\sum_{i,j,k\in\mathcal{S}}(\mathbb{I}\{i > j\} + \mathbb{I}\{i > k\})$. By swapping the naming of the indices i, j and i, k, observe that the first sum is equal to

$$\frac{1}{3}\sum_{i,j,k\in\mathcal{S}}(\mathbb{I}\{i > j\} + \mathbb{I}\{i > k\} + \mathbb{I}\{j > i\} + \mathbb{I}\{j > k\} + \mathbb{I}\{k > j\} + \mathbb{I}\{k > i\}).$$

The summand is bounded above by 3. We may thus bound the first sum above by s^3. The second sum is treated similarly. Swapping the naming of the indices i, j, the second sum can be rewritten

$$\frac{1}{2}\sum_{i,j\in\mathcal{S}}\sum_{k\in\mathcal{S}^c}(\mathbb{I}\{i > j\} + \mathbb{I}\{i > k\} + \mathbb{I}\{j > i\} + \mathbb{I}\{j > k\}).$$

The sum of the indicators on the inside is bounded above by 3, so the second sum can be bounded by $3s(s-1)(N-s)$. Finally, for the third sum, there is no gain in playing tricks with swapping indices. We will use the naive bound $2s(N-s)(N-s-1)$.

Combining these bounds yields the following inequality on s:

$$2(1-\alpha)(N-1)(N-2)s \le s(s-1)(s-2) + 3s(s-1)(N-s) + 2s(N-s)(N-s-1),$$

which reduces to $2(1-\alpha)(N-1)(N-2) \le 2N^2 - 5N - (N-2)s + 2$. Thus

$$\begin{aligned}(N-2)s &\le 2\alpha(N-1)(N-2) - 2(N^2-3N+2) + 2N^2 - 5N + 2\\ (N-2)s &\le 2\alpha(N-1)(N-2) + N - 2.\end{aligned}$$

Thus

$$s \le 2\alpha(N-1) + 1,$$

and so $s/N \le 2\alpha + \frac{1-2\alpha}{N}$.

□

B.4 Proofs for Section 5.1.2

Proof of Theorem 5.2. Throughout the proof, define the quantity $Q := r\binom{N-1}{r}(1-\alpha)$. Define a unit i_0 to be strange if $Q \le \sum_{\{i_1,\ldots,i_r\}\subset[N]}\sum_{j=1}^{r}\mathbb{I}\{i_0 > i_j\}$, where the notation

$\mathbb{I}\{i_0 > i_r\}$ is shorthand for the indicator $\mathbb{I}\left\{R_{(i_1,\ldots,i_r);i_0} > R_{(i_1,\ldots,i_r);i_j}\right\}$ that the residual of i_0 is greater than that of i_j. Let $\mathcal{S}$ denote the set of all strange units and $s := |\mathcal{S}|$. Summing over all $i_0 \in \mathcal{S}$,

$$Qs \le \sum_{i_0 \in \mathcal{S}} \sum_{\substack{\{i_1,\ldots,i_r\} \\ \subseteq [N] \setminus i_0}} \sum_{j=1}^{r} \mathbb{I}\{i_0 > i_j\}.$$

By considering the number of indices t in the tuple $\{i_1, \ldots, i_r\}$ which are in $\mathcal{S}$, let us break up the right hand side of the sum into

$$\sum_{t=0}^{r} \sum_{i_0 \in \mathcal{S}} \sum_{\substack{\{i_1,\ldots,i_t\} \\ \subseteq \mathcal{S}}} \sum_{\substack{\{i_{t+1},\ldots,i_r\} \\ \subseteq \mathcal{S}^c}} \sum_{j=1}^{r} \mathbb{I}\{i_0 > i_j\}. \tag{14}$$

Here we assumed without loss of generality that the first t indices are in $\mathcal{S}$, because the summand is invariant to permutations in $i_1, \ldots, i_r$. Introducing ordering of the labels, splitting the summand, we find that Eq. (14) equals a sum over t of the quantities

$$\frac{1}{t!(r-t)!} \sum_{\substack{(i_0,i_1,\ldots,i_t) \\ \subseteq \mathcal{S}}} \sum_{\substack{(i_{t+1},\ldots,i_r) \\ \subseteq \mathcal{S}^c}} \left(\overbrace{\sum_{j=1}^{t} \mathbb{I}\{i_0 > i_j\}}^{\text{(I)}} + \overbrace{\sum_{j'=t+1}^{r} \mathbb{I}\{i_0 > i_{j'}\}}^{\text{(II)}} \right)$$

Fixing t, we apply symmetry to improve on the naive upper bound for Eq. (14). Let Π_t denote the subgroup of permutations whose elements permute the labels of $i_0, \ldots, i_t$. We may then bound the quantity (I) as

$$\begin{aligned}
&\frac{1}{(t+1)!} \sum_{\pi \in \Pi_{t+1}} \sum_{j=1}^{t} \mathbb{I}\{\pi(i_0) > \pi(i_j)\} \\
&= \frac{1}{(t+1)!} \sum_{\pi \in \Pi_{t+1}} \operatorname{rank}(\pi(i_0)) \\
&= \frac{t!}{(t+1)!} \sum_{j=0}^{t} \operatorname{rank}(i_j) = \frac{1}{t+1} \sum_{j=0}^{t} j = \frac{t}{2}.
\end{aligned}$$

We note that the notion of rank used above ranges from 0 to t.

On the other hand, the sum over the quantity (II) can be bounded above naively by $(r-t)$. Combining these two bounds, we find that

$$\begin{aligned}
Qs &\leq \sum_{t=0}^{r} \frac{1}{t!(r-t)!} \frac{s!}{(s-t-1)!} \frac{(N-s)!}{((N-s)-(r-t))!}(r-\frac{t}{2}) \\
&= \sum_{t=0}^{r} s\binom{s-1}{t}\binom{N-s}{r-t}(r-\frac{t}{2}) \\
&= \sum_{t=0}^{r} sr\binom{s-1}{t}\binom{(N-1)-(s-1)}{r-t} - \sum_{t=1}^{r} \frac{s(s-1)}{2}\binom{s-2}{t-1}\binom{N-1-(s-1)}{(r-1)-(t-1)} \\
&= \sum_{t=0}^{r} sr\binom{s-1}{t}\binom{(N-1)-(s-1)}{r-t} - \sum_{t'=0}^{r-1} \frac{s(s-1)}{2}\binom{s-2}{t'}\binom{N-1-(s-1)}{(r-1)-t'}
\end{aligned}$$

Applying the Vandermonde Identity, the last line yields $Q \leq r\binom{N-1}{r} - \frac{(s-1)}{2}\binom{N-2}{r-1}$. Solving for s, we find

$$s \leq \frac{2r\binom{N-1}{r}}{\binom{N-2}{r-1}} - \frac{2Q}{\binom{N-2}{r-1}} + 1$$

Plugging in the value of Q, we find that

$$s \leq \frac{2r\binom{N-1}{r}}{\binom{N-2}{r-1}} - \frac{2r\binom{N-1}{r}(1-\alpha)}{\binom{N-2}{r-1}} + 1 = 2(N-1)\alpha + 1,$$

so that $\frac{s}{N} \leq 2\alpha + \frac{1-2\alpha}{N}$. □

Proof of LRO Rank Threshold Procedure Guarantee, Theorem 5.2 . In this setting, let us introduce the notation $\mathbb{I}(i_0 \succ i_1, \ldots, i_r)$ to denote the event where the residual $|Y_{i_0} - \mu_{-i_0,\ldots,i_r}(X_{i_0})|$ is greater than the all the other residuals $|Y_{i_k} - \mu_{-i_0,\ldots,i_r}(X_{i_k})|, k = 1, \ldots, r$. Define a unit i_0 to be strange if

$$Q \leq \sum_{i_1,\ldots,i_r} \mathbb{I}(i_0 \succ i_1, \ldots, i_r), \tag{15}$$

for the threshold $Q := \frac{1}{\lfloor \alpha N \rfloor}\left[\binom{N}{r+1} - \binom{N-\lfloor \alpha N \rfloor}{r+1}\right]$. Define $\mathcal{S}$ to be the set of strange points and $s := |\mathcal{S}|$. Summing both sides of the equation (15), we obtain the bound

$$Qs \leq \sum_{i_0 \in \mathcal{S}} \sum_{i_1,\ldots,i_r} \mathbb{I}\{i_0 \succ i_1, \ldots, i_r\}.$$

Much as before, let us consider the number of indices t in the tuple $\{t_1, \ldots, t_r\}$ which are in $\mathcal{S}$. This lets us write:

$$\begin{aligned} Qs &\leq \sum_{i_0 \in \mathcal{S}} \sum_{t=0}^{r} \sum_{\substack{\{i_1,\ldots,i_t\} \\ \subset \mathcal{S}}} \sum_{\substack{\{i_{t+1},\ldots,i_r\} \\ \subset \mathcal{S}^c}} \mathbb{I}\{i_0 > i_1, \ldots, i_r\} \\ &= \frac{1}{t!(r-t)!} \sum_{i_0 \in \mathcal{S}} \sum_{t=0}^{r} \sum_{\substack{i_1,\ldots,i_t \\ \subset \mathcal{S}}} \sum_{\substack{i_{t+1},\ldots,i_r \\ \subset \mathcal{S}^c}} \mathbb{I}\{i_0 > i_1, \ldots, i_r\}. \end{aligned}$$

Swapping the naming of i_0 with $i_1, i_2, \ldots, i_t$ we have the bound

$$\sum_{\substack{i_1,\ldots,i_t \\ \subset \mathcal{S}}} \mathbb{I}\{i_0 > i_1, \ldots, i_r\} = \sum_{\substack{i_1,\ldots,i_t \\ \subset \mathcal{S}}} \frac{1}{t+1} \sum_{j=0}^{t} \mathbb{I}\{i_j > i_1, \ldots, i_r\} \leq \sum_{\substack{i_1,\ldots,i_t \\ \subset \mathcal{S}}} \frac{1}{t+1}.$$

Plugging this back into the bound on Qs, we find

$$\begin{aligned} Qs &\leq \sum_{t=0}^{r} \frac{1}{t!} \frac{1}{(r-t)!} \left(\frac{s!}{(s-t-1)!} \frac{(N-s)!}{(N-s-(r-t))!} \frac{1}{t+1} \right) \\ &= \sum_{t=0}^{r} \binom{s}{t+1} \binom{N-s}{r-t} \\ &= \binom{N}{r+1} - \binom{N-s}{r+1} \end{aligned}$$

where the last line follows by the Vandermonde identity and some reindexing. By Lemma B.1, $Q \leq \frac{1}{s}\left[\binom{N}{r+1} - \binom{N-s}{r+1}\right]$ implies $s \leq \lfloor \alpha N \rfloor$. □

B.5 Proofs for Section 5.2

Proof of Proposition 5.2. we define a unit k to be strange if

$$\sum_{i \neq j, i, j \neq k} (\pi_i + \pi_j) \mathbb{I}(k > i, j) \geq 2(N-2)(1-\alpha)(1-\pi_k). \tag{16}$$

Label the set of strange points as $\mathcal{S}$ and let $s := |\mathcal{S}|$. As before, define $p_* = \sum_{k \in \mathcal{S}} \pi_k$. Multiply both sides of Equation (16) by π_k, and sum over $k \in \mathcal{S}$ to obtain

$$\begin{aligned} 2(N-2)(1-\alpha)(p_* - \sum_{k \in \mathcal{S}} \pi_k^2) \leq \overbrace{\sum_{i,j,k \in \mathcal{S}} \pi_k(\pi_i + \pi_j)\mathbb{I}(k > i, j)}^{\text{(I)}} + \overbrace{2 \sum_{i,k \in \mathcal{S}} \sum_{j \in \mathcal{S}^c} \pi_k(\pi_i + \pi_j)\mathbb{I}(k > i, j)}^{\text{(II)}} \\ + \overbrace{\sum_{k \in \mathcal{S}} \sum_{i,j \in \mathcal{S}^c} \pi_k(\pi_i + \pi_j)\mathbb{I}(k > i, j)}^{\text{(III)}}. \end{aligned}$$

For the first sum (I) in the right hand side, break the sum into two terms, $\sum_{i,j,k\in\mathcal{S}} \pi_k\pi_i\mathbb{I}(k \succ i,j)$ and $\sum_{i,j,k\in\mathcal{S}} \pi_k\pi_j\mathbb{I}(k \succ i,j)$. For the first term, swap indices i,k and use the bound $\mathbb{I}(k \succ i,j) + \mathbb{I}(i \succ k,j) \le 1$; this yields $\sum_{i,j,k\in\mathcal{S}} \pi_k\pi_i\mathbb{I}(k \succ i,j) \le \frac{1}{2}\sum_{i,j,k\in\mathcal{S}} \pi_k\pi_i \le \frac{1}{2}(p_*^2 - \sum_{k\in\mathcal{S}} \pi_k^2)s$. A similar logic on the second term, swapping j,k, yields the same upper bound of $\frac{1}{2}p_*^2 s$. Thus (I) $\le (p_*^2 - \sum_{k\in\mathcal{S}} \pi_k^2)s$.

For the second sum (II), split into the two terms $2\sum_{i,k\in\mathcal{S}}\sum_{j\in\mathcal{S}^c} \pi_k\pi_i\mathbb{I}(k \succ i,j)$ and $2\sum_{i,k\in\mathcal{S}}\sum_{j\in\mathcal{S}^c} \pi_k\pi_j\mathbb{I}(k \succ i,j)$. In the first term, swap the indices i,k; leave the second term alone. Swapping i,k in the first term leads to the upper bound $\sum_{i,k\in\mathcal{S}}\sum_{j\in\mathcal{S}^c} \pi_k\pi_i \le (p_*^2 - \sum_{k\in\mathcal{S}} \pi_k^2)(N-1-s)$. The naive upper bound on the second term yields $2p_*(1-p_*)s$. Finally, the third sum (III) is bounded naively by $2p_*(1-p_*)(N-1-s)$.

Combining all of these, we obtain

$$2(1-\alpha)(N-2)\left[p_* - \sum_{\mathcal{S}} \pi_k^2\right] \le \left[p_*^2 - \sum_{k\in\mathcal{S}} \pi_k^2\right](N-1) + 2p_*(N-1) - 2p_*^2(N-1),$$

which reduces to the quadratic inequality

$$(N-1)p_*^2 - [2\alpha(N-2)+2]\,p_* + [(2\alpha - 1 + \frac{1}{N-2})(N-2)]\sum_{k\in\mathcal{S}} \pi_k^2 \le 0$$

dividing by $(N-1)$ and solving, we find that

$$p_* \le \alpha\frac{N-2}{N-1} + \frac{1}{N-2} + \sqrt{(\alpha\frac{N-2}{N-1} + \frac{1}{N-2})^2 + [(1-2\alpha)\frac{N-1}{N-2} - \frac{1}{N-1}]\sum_{k\in\mathcal{S}} \pi_k^2}.$$

□

Proof of Proposition 5.3. In this setting, define a unit k to be strange if it satisfies

$$\sum_{i\neq j, i,j\neq k} (\pi_i + \pi_j)\mathbb{I}(k \succ i,j) \ge 2(N-1)(1-\alpha)(1-\pi_k). \tag{17}$$

Imitating the proof of Proposition 5.2, the factors of $(N-1)$ cancel nicely and we can arrive at the quadratic inequality

$$p_*^2 - 2\alpha p_* - (1-2\alpha)\sum_{k\in\mathcal{S}} \pi_k^2 \le 0,$$

which is easily solved to yield the desired bound on p_*. □

B.6 Minor lemmas

Lemma B.1. *The sequence*

$$s \mapsto A(s) := \frac{1}{s}\left[\binom{N}{r+1} - \binom{N-s}{r+1}\right]$$

is decreasing.

Proof. Observe the two following identities on binomial coefficients:

$$\binom{N-s}{r+1} = \sum_{k=r}^{N-1-s} \binom{k}{r}$$
$$s\binom{N-1-s}{r} \leq \sum_{k'=N-s}^{N-1} \binom{k'}{r}.$$

The first is true because of the hockeystick identity, and the second is by monotonicity. Summing these inequalities, we find

$$\binom{N-s}{r+1} + s\binom{N-1-s}{r} \leq \sum_{k=r}^{N-1} \binom{k}{r} = \binom{N}{r+1},$$

which is equivalent to

$$\binom{N-s}{r+1} + s\left[\binom{N-s}{r+1} - \binom{N-1-s}{r+1}\right] \leq (s+1)\binom{N}{r+1} - s\binom{N}{r+1}.$$

Rearranging, $A(s) \geq A(s+1)$. □