Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls

Nick Doudchenko[*]
Google Research
New York, NY 10011
nikolayd@google.com

Khashayar Khosravi[*]
Google Research
New York, NY 10011
khosravi@google.com

Jean Pouget-Abadie
Google Research
New York, NY 10011
jeanpa@google.com

Sebastien Lahaie
Google Research
New York, NY 10011
slahaie@google.com

Miles Lubin
Google Research
New York, NY 10011
mlubin@google.com

Vahab Mirrokni
Google Research
New York, NY 10011
mirrokni@google.com

Jann Spiess
Stanford GSB
Stanford, CA 94305
jspiess@stanford.edu

Guido Imbens
Stanford GSB
Stanford, CA 94305
imbens@stanford.edu

Abstract

We investigate the optimal design of experimental studies that have pre-treatment outcome data available. The average treatment effect is estimated as the difference between the weighted average outcomes of the treated and control units. A number of commonly used approaches fit this formulation, including the difference-in-means estimator and a variety of synthetic-control techniques. We propose several methods for choosing the set of treated units in conjunction with the weights. Observing the NP-hardness of the problem, we introduce a mixed-integer programming formulation which selects both the treatment and control sets and unit weightings. We prove that these proposed approaches lead to qualitatively different experimental units being selected for treatment. We use simulations based on publicly available data from the US Bureau of Labor Statistics that show improvements in terms of mean squared error and statistical power when compared to simple and commonly used alternatives such as randomized trials.

1 Introduction

Randomized experiments have long been a staple of applied causal inference. In his seminal paper, Rubin (1974) suggests that "given a choice between the data from a randomized experiment and an equivalent nonrandomized study, one should choose the data from the experiment, especially in the social sciences where much of the variability is often unassigned to particular causes." Using the language of Rubin's *potential-outcomes* framework, randomization guarantees that the treatment status is independent of the potential outcomes and that a simple and intuitive estimator that compares the average outcomes of the treatment and control units is an unbiased estimator of the *average treatment effect* (ATE). If both the treatment and control samples are sufficiently large, the hope is that this difference-in-means estimate is close to the population mean of the treatment effect.

[*]Equal contributions.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Another crucial property of randomized experimental designs is their robustness to alternative assumptions about the data generating process—a completely randomized experiment does not take into account any features of the observed data. Perhaps not surprisingly, when the researchers are willing to incorporate additional probabilistic assumptions in their design decisions, they can improve on the statistical properties of the average treatment effect estimators (see, for example, Kasy, 2016). These improvements, however, do not come for free and the performance of the estimators may suffer if the incorporated assumptions are violated (Banerjee et al., 2020).

For these reasons randomized experiments are widely used in academic and clinical settings, as well as industrial applications. However, not every practically important question can be easily answered using an experiment on a large sample of experimental units. For instance, the evaluation of major policies targeting large geographic areas has long been of interest in social and political sciences. One of the traditional approaches is to compare the affected unit—such as a metropolitan area or a state—to the average across a sample of carefully picked control units which are deemed to be suitable comparisons (Card, 1990). A more recent alternative, called *synthetic control*, first introduced by Abadie and Gardeazabal (2003) and later developed in Abadie et al. (2010) and Abadie et al. (2015), compares the unit of interest to a weighted average of the units unaffected by the treatment, where the weights are selected in a way that achieves a good fit on pre-treatment outcome variables as well as potentially other observed covariates.

While originally developed by academics for evaluating the effects of policies, approaches similar to the synthetic-control methodology have recently gained popularity in industry as well in cases when applied researchers decide to run experiments targeting larger units often representing geographic areas. This decision may be justified when more granular experiments are either unavailable (for example, television advertising can only be targeted at the media-market level) or are unlikely to capture the relevant effects due to interference or equilibrium concerns (Sobel, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008). For instance, a company like Uber may want to evaluate some of the possible treatments at the market level rather than at the driver level if the treatment in question is likely to affect the driver supply. Moreover, launching an experiment targeting even a single unit may be so expensive that the researchers try to minimize the required number of treated units. Privacy and fairness concerns may also make treatment assignment at a more granular level problematic.

Synthetic control and similar approaches may be attractive as estimation procedures in those cases, but they fail to address the equally if not more important aspect of the optimal choice of experimental units (see, for example, Rubin et al., 2008). We attempt to narrow this gap in the current paper. We consider a panel-data setting in which the researcher observes the outcome metric of interest for a number of units in a number of time periods and has to decide: (i) which units to experiment on and (ii) how to estimate the treatment effects after collecting the outcome data in the experimental time periods. The main difference between this setting compared to a typical synthetic-control study is that the treated units are not fixed, but rather chosen by the researcher. Proving that the underlying optimization problem is NP-hard, we rule out the possibility of designing polynomial-time algorithms for the problem under P$\neq$NP. Therefore, we formulate this combined design-and-analysis problem as a *mixed-integer program* (MIP). Depending on the particular estimands of interest, we propose one of the several formulations and discuss their advantages and drawbacks. We motivate the choice of the optimization objectives and discuss the selection of experimental units each of the objectives leads to. The MIP formulation allows for an easy inclusion of additional constraints as long as those are linear. For instance, it is easy to restrict the overall number of treated units, exclude specific units from treatment, or enforce a budget constraint if there is a varying cost to treat different units.

Using publicly available state-level unemployment data from the US Bureau of Labor Statistics, we compare the proposed methodology to a randomized design that utilizes either the conventional difference-in-means estimator or the synthetic-control approach. We estimate the average as well as individual state-level treatment effects in a simulated experiment and find that our approach substantially reduces the *root mean squared error* (RMSE) of the estimates. We show that our MIP-based design-and-analysis procedures consistently outperform the more traditional baselines regardless of whether the treatment effects are homogeneous or heterogeneous and whether the number of units selected for treatment is small or large relative to the total number of units in the sample. We also suggest a permutation-based inference procedure that follows Chernozhukov et al. (2021). We verify in our simulations that this procedure leads to correct test sizes and improved statistical power for testing the sharp null hypothesis of zero treatment affects across all treated units,

when used in conjunction with the proposed estimators. We provide theoretical guarantees—albeit under rather strong assumptions—which ensure that the proposed tests have proper sizes.

To our knowledge, this is one of the first papers to study experimental design in the context of synthetic control and adjacent estimation techniques. Doudchenko et al. (2019) consider the case when the underlying effects are homogeneous and only a single unit can be treated; they suggest searching for the unit that delivers the highest statistical power when testing the hypothesis of zero treatment effect with an artificially applied treatment. In an independent work, Abadie and Zhao (2021) also consider optimal design in settings where synthetic-control estimators are used for estimation. However, there are a few important aspects that differentiate the current paper from theirs. First, Abadie and Zhao (2021) use both pre-treatment unit-level covariates as well as the outcomes from some—but not necessarily all—of the pre-treatment periods whereas we focus exclusively on the outcome variable and utilize all of the pre-treatment outcome data. As a result, the inference procedure proposed by Abadie and Zhao (2021) does not apply to our approach. Second, we discuss formal hardness results. Finally—and, perhaps, most importantly—Abadie and Zhao (2021) choose an objective function that fixes a priori the target average treatment effect, as either the population average treatment effect or a weighted version thereof. In contrast, the objective function we use focuses on the average treatment effect only for the units we choose for the treatment, making the target estimand stochastic, but potentially easier to estimate and requiring different assumptions.

The rest of the paper is organized as follows. Section 2 introduces the setting and the proposed estimation approaches. Section 3 introduces the mixed-integer formulation of the suggested estimators. Section 4 presents some of the theoretical unit-selection results and the intuition behind them. Section 5 reports the empirical results obtained through simulations. Section 6 discusses some of the practical consideration that should accompany applied work that uses the proposed methodology. Section 7 concludes and outlines the directions of future research.

2 Setting

Let the researcher observe the outcome metric of interest, Y, for N units during T time periods, such that the observed data can be represented as an $N \times T$ matrix of values Y_{it}. At time $t = T$ the researcher decides—based on the data observed up until that point—which units should be treated and assigns a binary treatment described by variables $D_i \in \{0, 1\}$, $i = 1, \ldots, N$. The outcomes are then observed for an additional $S - T$ time periods $t = T + 1, \ldots, S$ and the treatment-effect estimates are constructed. Each unit $i = 1, \ldots, N$ in each time period $t = T + 1, \ldots, S$ is associated with two potential outcomes $(Y_{it}(0), Y_{it}(1))$ which are considered random. The potential outcome $Y_{it}(0)$ is realized if $D_i = 0$ and $Y_{it}(1)$ is realized if $D_i = 1$ so that the observed outcome is $Y_{it} = Y_{it}(D_i) = Y_{it}(0)(1 - D_i) + Y_{it}(1)D_i$.

Recall that, given a setting where a single unit $i = N$ has received the treatment in a single time period $t = T + 1 = S$, the synthetic-control literature (Abadie et al., 2010, among others) suggests constructing a counterfactual estimate for unit N as a weighted average of the other units' observed outcomes: $\hat{Y}_{N,T+1}(0) = \sum_{i=1}^{N-1} w_i Y_{i,T+1}$. In the previous equation, the w_i's are weights learned from the data observed during the pre-treatment periods $t = 1, 2, \ldots, T$, often by minimizing $\sum_{t=1}^{T}(Y_{Nt} - \sum_{i=1}^{N-1} w_i Y_{it})^2$ under some constraints on the weights. Assuming that the treatment effect, τ_N, for this unit is additive, it can then be estimated as $\hat{\tau}_N = Y_{N,T+1} - \hat{Y}_{N,T+1}(0)$.

We now consider a more general setting where K units have received the treatment, with outcomes given by

$$Y_{it}(0) = \mu_{it} + \varepsilon_{it} \quad \text{and} \quad Y_{it} = Y_{it}(0) + D_i \tau_i$$

with homoscedastic noise ε_{it} that has mean zero and variance σ^2 and additive treatment effects τ_i. In order to estimate the treatment effect in this more general setting, we can apply a separate synthetic control method to each treated unit i, learning an appropriate set of weights $\{w_j^i\}$ for each treated unit i individually: $\hat{Y}_{i,T+1}(0) = \sum_{j:\, D_j = 0} w_j^i Y_{j,T+1}$. The treatment effect for unit i is then estimated as $\hat{\tau}_i = Y_{i,T+1} - \hat{Y}_{i,T+1}(0)$.

Rather than considering each weight-fitting optimization problem separately, we can express the (conditional) *mean squared error* (MSE) of the resulting estimator of the treatment effect as:

$$\mathbb{E}\left[(\hat{\tau}_i - \tau_i)^2 \,\middle|\, \{D_j, w_j^i\}_{j=1}^N\right] = \left(\mu_{i,T+1} - \sum_{j\colon D_j=0} w_j^i \mu_{j,T+1}\right)^2 + \sigma^2 \left(1 + \sum_{j\colon D_j=0} (w_j^i)^2\right).$$

A proof is included in the supplementary materials. The synthetic-control literature often operates in settings where the treated units are given. Here, we allow the experimenter to select which units should receive the treatment. The mean-squared-error formula above leads us to consider the following optimization problem over the weights $\{w_j^i\}_{i,j=1}^N$ and the treatment variables $\{D_i\}_{i=1}^N$ under some appropriate constraints on the weights $\{w_j^i\}_{i,j=1}^N$. The objective below can be seen as the empirical analog of the right-hand side of the population equation above in the pre-treatment period averaged across time and across the treated units.

$$\min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad \frac{1}{KT} \sum_{i=1}^N \sum_{t=1}^T D_i \left(Y_{it} - \sum_{j=1}^N w_j^i (1 - D_j) Y_{jt}\right)^2 + \frac{\sigma^2}{K} \sum_{i=1}^N \sum_{j=1}^N D_i \left(w_j^i\right)^2 \tag{per-unit}$$

Note that we can safely sum across all units, not just the control ones, in the second term since the optimal values w_j^i for units j with $D_j = 1$ will be equal to zero in an optimal solution.

In essence, this optimization problem attempts to minimize the discrepancy between the pre-treatment outcomes of the units chosen for treatment and the weighted averages of the outcomes of the units left as controls. At the same time, due to the second term, the objective attempts to balance the unit weights themselves.

The term σ^2 is unlikely to be known to the researcher and should be chosen based on the observed data. One possible way is to set it equal to the sample variance of the observed outcomes. We further discuss the selection of the penalty parameter in Section 6. Penalizing the weights is not uncommon in the synthetic-control literature. For example, Doudchenko and Imbens (2016) use the elastic-net penalty on the weights and Abadie and L'Hour (2021) introduce a lasso-style penalty in which the nonnegative weight, w_j^i, is multiplied by the squared distance between the vectors of the covariates used for matching the units. This way, depending on the magnitude of the penalty hyperparameter, they can balance between the synthetic-control fit and the nearest-neighbor fit that puts all the weight on unit j closest to i in terms of the observed covariates.

In many applications, the researcher may be interested in estimating some weighted average of the unit-level treatment effects on the treated units. Rather than considering each treated unit separately and then computing the weighted average of the estimated individual treatment effects, practitioners may wish to construct a synthetic-control-type estimate for the weighed average of the treated units directly.

Consider a set of treatment assignments $\{D_i\}_{i=1}^N$ and weights $\{w_i\}_{i=1}^N$ on outcomes at time $T+1$, and assume that we wish to estimate the *weighted average treatment effect on the treated* (wATET), $\tau = \sum_{i\colon D_i=1} w_i \tau_i = \sum_{i=1}^N D_i w_i \tau_i$, as a difference in weighted means: $\hat{\tau} = \sum_{i\colon D_i=1} w_i Y_{i,T+1} - \sum_{i\colon D_i=0} w_i Y_{i,T+1}$. Then, under the same outcomes model as presented above, the (conditional) mean squared error of the difference-in-(weighted)-means estimator is

$$\mathbb{E}\left[(\hat{\tau} - \tau)^2 \,\middle|\, \{D_i, w_i\}_{i=1}^N\right] = \left(\sum_{i\colon D_i=1} w_i \mu_{i,T+1} - \sum_{i\colon D_i=0} w_i \mu_{i,T+1}\right)^2 + \sigma^2 \sum_{i=1}^N w_i^2.$$

As before, our setting allows the experimenter to optimally select which units should receive the treatment as well as which particular weighting scheme should be used. This is especially appropriate when the treatment effects are homogeneous and $\tau_i = \tau$ for all $i = 1, \ldots, N$. In that case, any weighted average of the unit-level treatment effects is equal to τ and the weights can be chosen in a way that minimizes the mean squared error.

The population equation above suggests solving the following optimization problem on the weights $\{w_i\}_{i=1}^N$ and the treatment variables $\{D_i\}_{i=1}^N$ based on the data observed in periods $t = 1, \ldots, T$:

$$\min_{\{D_i, w_i\}_{i=1}^N} \quad \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i:\, D_i=1} w_i Y_{it} - \sum_{i:\, D_i=0} w_i Y_{it} \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2. \qquad \text{(two-way global)}$$

So far, we have not considered specific constraints on the weights, $\{w_i\}_{i=1}^N$. If $\sigma^2 > 0$, $w_i = 0$ for all $i = 1, \ldots, N$ is the unique optimal solution to the two-way global problem above if the weights are not constrained in any way. In order to avoid this clearly undesirable solution, we assume that the weights $\{w_i\}_{i=1}^N$ are normalized: $\sum_{i:\, D_i=1} w_i = \sum_{i:\, D_i=0} w_i = 1$. While not strictly required, another reasonable set of constraints motivated by estimating a proper weighted average is requiring all weights to be nonnegative, $w_i \geq 0$ for all $i = 1, \ldots, N$.

Similar constraints can be imposed in the context of the per-unit problem: $w_j^i \geq 0$ for all $i, j = 1, \ldots, N$ and $\sum_{j:\, D_j=0} w_j^i = 1$ for all $i = 1, \ldots, N$ such that $D_i = 1$.

Finally, the weights on the treated units may be fixed if a specific weighted average—for example, a simple average with equal weighting—needs to be estimated. This constraint is particularly justified when the treatment effects are heterogeneous and different weighting schemes lead to different estimands. For that reason, we formulate another variation of the global problem:

$$\min_{\{D_i, w_i\}_{i=1}^N} \quad \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i:\, D_i=1} w_i Y_{it} - \sum_{i:\, D_i=0} w_i Y_{it} \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2 \qquad \text{(one-way global)}$$

subject to an additional set of constraint that require $w_i = w_j$ for all $i, j = 1, \ldots, N$ such that $D_i = D_j = 1$.

It is possible to obtain other design-and-estimation approaches as special cases of these problems—something that we utilize later in the paper. For instance, the standard difference-in-means approach under randomized design can be viewed as the special case of the one-way global problem with the treatment indicators, D_i, set randomly and the weights on the control units restricted to be equal, $w_i = w_j$ for all $i, j = 1, \ldots, N$ such that $D_i = D_j = 0$. Likewise, the synthetic-control estimator can be viewed as the per-unit problem with the treatment indicators set according to the experimental design of choice—for example, randomly.

3 Mixed-integer formulation

The three optimization problems introduced in the first part of Section 2: the per-unit, the two-way global, and the one-way global problems can all be formulated as mixed-integer programs. We now describe the specific optimization problems we use in the empirical section of the paper.

The per-unit problem is formulated as:

$$\begin{aligned}
\min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad & \frac{1}{KT} \sum_{i=1}^{N} \sum_{t=1}^{T} D_i \left(Y_{it} - \sum_{j=1}^{N} w_j^i (1 - D_j) Y_{jt} \right)^2 + \frac{\lambda}{K} \sum_{i=1}^{N} \sum_{j=1}^{N} D_i \left(w_j^i \right)^2 \\
\text{s.t.} \quad & w_j^i \geq 0, \quad D_i \in \{0, 1\} \ \text{ for } i, j = 1, \ldots, N, \\
& \sum_{i=1}^{N} D_i = K, \quad \sum_{i=1}^{N} w_j^i (1 - D_j) = 1 \ \text{ for } i = 1, \ldots, N \text{ such that } D_i = 1.
\end{aligned}$$

The two-way global problem can be formulated as:

$$\begin{aligned}
\min_{\{D_i, w_i\}_{i=1}^N} \quad & \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{N} w_i D_i Y_{it} - \sum_{i=1}^{N} w_i (1 - D_i) Y_{it} \right)^2 + \lambda \sum_{i=1}^{N} w_i^2 \\
\text{s.t.} \quad & w_i \geq 0, \quad D_i \in \{0, 1\} \ \text{ for } i = 1, \ldots, N, \\
& \sum_{i=1}^{N} D_i = K, \quad \sum_{i=1}^{N} w_i D_i = 1, \quad \sum_{i=1}^{N} w_i (1 - D_i) = 1
\end{aligned}$$

and the one-way global problem is simply the two-way global problem with an additional set of constraints: $w_i = w_j$ for all $i, j = 1, \ldots, N$ such that $D_i = D_j = 1$.

All three problems require additional auxiliary variables representing the products of the weights and the treatment indicators as well as additional constraints in order to have a representation with a quadratic objective and only linear constraints so that they can be easier to solve by one of the academic or commercial MIP solvers.[2] See the supplementary materials for the exact formulations.

The term λ that is used in all three objectives is a nonnegative penalty factor. Its selection is discussed in Section 6.

Note that the global versions of the problem can be solved without the constraint on the number of treated units, $\sum_{i=1}^{N} D_i = K$, but the per-unit problem requires it—without the constraint the per-unit problem will tend to select fewer treatment units unless the objective is divided by $\sum_{i=1}^{N} D_i$ which introduces a nonlinearity. See the supplementary materials where we introduce an alternative formulation that allows to circumvent this by imposing an additional quadratic constraint.

4 Design

The three optimization problems introduced in Section 2—the one-way and two-way global and the per-unit problems—tend to select certain treatment units in terms of their location within the distribution of the observed outcome data. In this section we illustrate that behavior using a simple example motivated by the objectives from Section 2. To simplify the analysis, we let $T = 1$ and denote $a_i = Y_{i1}$. We also do not restrict the unit weights to be nonnegative to allow for simpler closed form solutions.[3] Specifically, we consider the following per-unit problem:

$$\begin{aligned}
\min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad & \frac{1}{K}\sum_{i=1}^{N} D_i \left(a_i - \sum_{j=1}^{N} w_j^i (1 - D_j) a_j \right)^2 + \frac{\sigma^2}{K} \sum_{i=1}^{N}\sum_{j=1}^{N} D_i (w_j^i)^2 \\
\text{s.t} \quad & D_i \in \{0, 1\} \ \text{ for } i, j = 1, \ldots, N, \\
& \sum_{i=1}^{N} D_i = K, \quad \sum_{j=1}^{N} w_j^i (1 - D_j) = 1 \ \text{ for } i = 1, \ldots, N.
\end{aligned}$$

A similarly simplified two-way global problem can be written as:

$$\begin{aligned}
\min_{\{D_i, w_i\}_{i=1}^N} \quad & \left(\sum_{i=1}^{N} w_i D_i a_i - \sum_{i=1}^{N} w_i (1 - D_i) a_i \right)^2 + \sigma^2 \sum_{i=1}^{N} w_i^2 \\
\text{s.t} \quad & D_i \in \{0, 1\} \ \text{ for } i, j = 1, \ldots, N, \\
& \sum_{i=1}^{N} D_i = K, \quad \sum_{i=1}^{N} w_i D_i = 1, \quad \sum_{i=1}^{N} w_i (1 - D_i) = 1.
\end{aligned}$$

This becomes a one-way global problem if we impose an additional set of constraints requiring that $w_i = 1/K$ for all $i = 1, \ldots, N$ such that $D_i = 1$. When the unit weights are optimized (see the supplementary materials for the derivation), the optimal values of the objectives can be written in closed-form as functions of the set of treated units, I.

Theorem 1. *Let I denote the set of treated units and $\bar{I} = \{1, \ldots, N\} \setminus I$ denote the set of control units. Let $\overline{a}_I = \sum_{i \in I} a_i / |I|$ be the average outcome within the treatment group, $V_I^2 = \sum_{i \in I} (a_i - \overline{a}_I)^2$ a quantity proportional to the sample variance of the outcomes within the treatment group, and the corresponding quantities for set $\bar{I}$ defined similarly. After the unit weights are optimized away, the per-unit objective can be written as:*

$$J_{\textit{per-unit}}(I) = \sigma^2 \left(\frac{1}{N - K} + \frac{(\overline{a}_I - \overline{a}_{\bar{I}})^2 + K^{-1} V_I^2}{\sigma^2 + V_{\bar{I}}^2} \right),$$

[2] We use SCIP (Gamrath et al., 2020) when generating the empirical results in Sections 4 and 5.

[3] In some cases the nonnegativity constraint will not be binding, while in other cases the optimal weights without the constraint may actually turn out to be negative. This implies that only a subset of all units will have strictly positive weights in the optimal constrained solution.

the two-way global objective can be written as:

$$J_{\textit{2-way}}(I) = \sigma^2 \left(\frac{1}{K} + \frac{1}{N-K} + \frac{(\overline{a}_I - \overline{a}_{\bar{I}})^2}{\sigma^2 + V_I^2 + V_{\bar{I}}^2} \right),$$

and the one-way global objective can be written as:

$$J_{\textit{1-way}}(I) = \sigma^2 \left(\frac{1}{K} + \frac{1}{N-K} + \frac{(\overline{a}_I - \overline{a}_{\bar{I}})^2}{\sigma^2 + V_{\bar{I}}^2} \right).$$

It is evident from the objectives that all problems aim to select units in such a way that averages of the observed outcomes for the treatment and control groups are similar—the squared difference between the average outcomes within each group appears in the numerators of all three objectives. The problems differ in terms of how they approach the sample variances of the two groups. The per-unit problem maximizes the sample variance of the control units while minimizing that of the treated units—the term $V_{\bar{I}}^2$ appears in the denominator while the term V_I^2 appears in the numerator. The intuition for the latter is that the per-unit objective tries to simultaneously model each treated unit with a combination of control units, while keeping weight variance small, so it is best to keep treated units as homogeneous as possible. The two-way problem attempts to maximize both (taking into account that they are, of course, interdependent), the sample variances of the outcomes of the control and treated groups—both quantities appear in the denominator. The one-way problem maximizes the sample variance of the outcomes of the control units only, as the weights for the treated units are fixed—only the term $V_{\bar{I}}^2$ appears in the denominator.

To illustrate the same patterns visually, we solve a per-unit problem and a two-way global problem for a simulated dataset with $N = 25$ units, $T = 2$ pre-treatment periods and the outcomes Y_{it} drawn from a standard normal distribution independently across i and t. This allows 2-d plotting of the units in the space of observed pre-treatment outcomes (Y_{i1}, Y_{i2}). We select either $K = 3$ or $K = 25 - 3 = 22$ treated units and we use $\lambda = \sum_{i=1}^{N} (Y_{i1} - Y_{i2})^2/(4N)$ (which is the average two-period sample variance across all units). Figure 1 shows the units selected by each of the problems.

The results align with the intuition presented using the simple model with $T = 1$ above. Specifically, the per-unit problem maximizes the spread of the control units which is particularly apparent from the plot corresponding to $K = 22$ while keeping the treatment units as close to each other as possible (see the plot for $K = 3$). Since the control and the treatment units have symmetric roles in the two-way global objective, the three treatment units selected by the problem when $K = 3$ are the same as the three control units selected when $K = 22$. The control and treatment groups have similar average outcomes and both groups are relatively spread out.

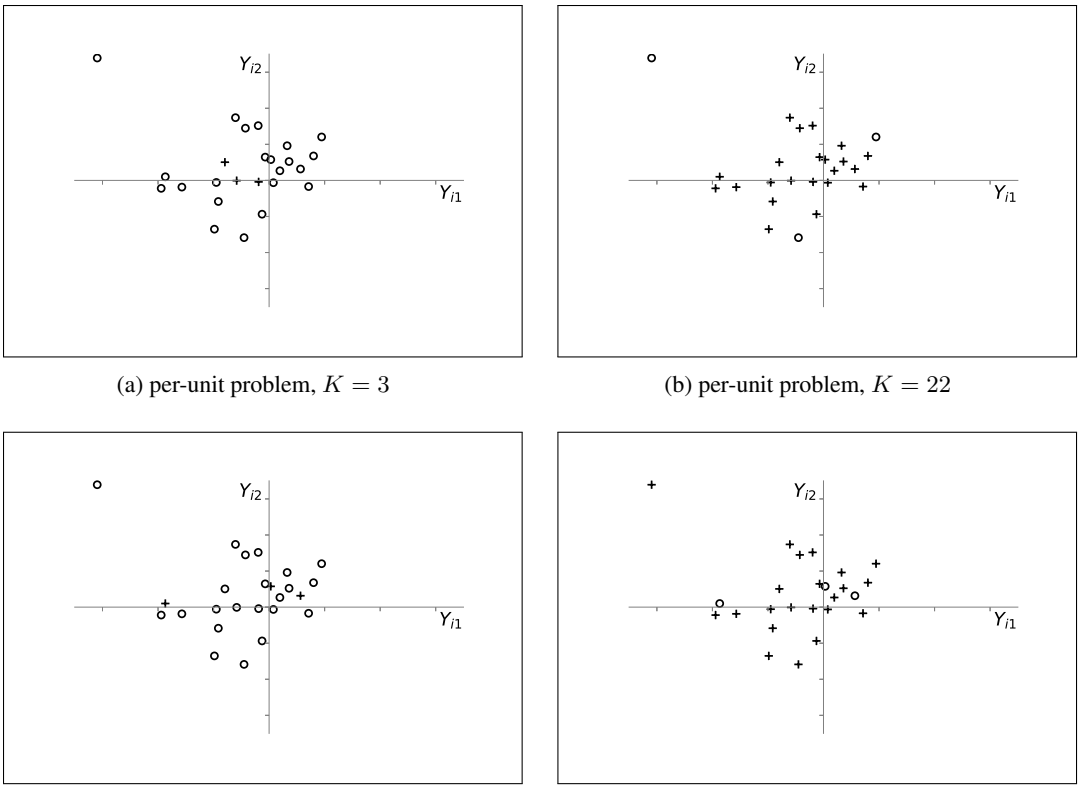
5 Empirical results

To evaluate the performance of the methods proposed in Section 2 we compare several design-and-estimation procedures: (i) the per-unit problem, (ii) the two-way global problem, (iii) the one-way global problem, (iv) the per-unit problem with randomly chosen treatment units, and (v) the standard randomized experiment which randomly assigns the treatment and estimates the average treatment effect on the treated as the difference in means between the two groups. It is important to note that approach (iv) is equivalent to using the synthetic-control method[4] for each randomly chosen treatment unit separately and then either averaging the unit-level treatment effect estimates or using the individual estimates directly. Taking this into account, comparing (i) to (iv) amounts to evaluating the role of optimal design in a synthetic-control study, while comparing (iv) to (v) evaluates the synthetic-control approach relative to the difference-in-means estimator.

To run a number of simulated experiments, we take publicly available data from the US Bureau of Labor Statistics (BLS) which contain unemployment rates of 50 states in 40 consecutive months.[5] We run 500 simulations such that each simulation utilizes a 10-by-10 matrix sampled from the original

[4]The only difference compared to the traditional synthetic-control methodology used, for example, in Abadie et al. (2010) is that no additional covariates are used and the weights are obtained using the outcome data alone—the approach similar to the one taken in, for instance, Doudchenko and Imbens (2016).

[5]The data are available from the BLS website, but the specific dataset we use is taken from `https://github.com/synth-inference/synthdid/blob/master/experiments/bdm/data/urate_cps.csv`.



(a) per-unit problem, $K = 3$

(b) per-unit problem, $K = 22$

(c) two-way global problem, $K = 3$

(d) two-way global problem, $K = 22$

Note: The units are plotted in the space of observed pre-treatment outcome data, (Y_{i1}, Y_{i2}); 'o' denote the selected control units and '+' denote the selected treatment units.

Figure 1: Units selected by the two-way global and per-unit problems.

50-by-40 dataset. Specifically, we randomly select 10 units and the first time period. The remaining 9 time periods are the consecutive months that follow. In each simulation we treat K units (equal to 3 in one set of simulations and 7 in another) which are chosen based on the data in the first 7 periods—or chosen randomly in cases (iv) and (v)—and the treatment is applied in the last 3 of the 10 periods. We either assign each treated unit the additive treatment effect of 0.05 (the homogeneous treatment case) or assume that the treatment effects increase linearly from 0 to 0.1 from the first unit in the (randomly selected) 10-by-10 matrix to the last one (the heterogeneous treatment case). This implies that the true value of the ATET changes depending on the identity of the units selected for treatment. However, the (overall) ATE remains 0.05.

We estimate the average treatment effect on the treated as well as the unit-level treatment effects. Only the per-unit problem, (i), and the synthetic control, (iv), allow for nontrivial estimation of heterogeneous treatment effects while the remaining approaches estimate all unit-level effects as being equal to the estimate of the average treatment effect on the treated.

We then compare approaches (i)–(v) in terms of the *root-mean-square error* (RMSE), where the squared differences between the true values of the treatment effects and the respective estimates are computed for each treatment period (and each treatment unit in case of the unit-level effects) and averaged. The square roots of these quantities are the RMSE's in question. Table 1 reports the RMSE's averaged across all simulations.

The specific improvements in terms of the RMSE over the baselines (iv) and (v)—the randomized synthetic control and the randomized difference-in-means—depend on the data and the true treatment effects. The main takeaways, however, are more general. The per-unit problem consistently outperforms the other methods when the underlying treatment effects are heterogeneous. For homogeneous treatment effects the difference between the per-unit and the global problems either vanishes, or the

Table 1: Root-mean-square errors of the average and unit-level treatment effect estimates

Homogeneous treatment				
	$K = 3$		$K = 7$	
	ATET RMSE	Unit-level RMSE	ATET RMSE	Unit-level RMSE
(i) Per-unit	8.5	13.9	**8.3**	16.0
(ii) Two-way global	**8.4**	**8.4**	8.4	**8.4**
(iii) One-way global	8.5	8.5	8.5	8.5
(iv) Synthetic control (random treat.)	9.7	15.9	10.3	19.0
(v) Diff-in-means (random treat.)	12.1	12.1	11.5	11.5
Heterogeneous treatment				
	$K = 3$		$K = 7$	
	ATET RMSE	Unit-level RMSE	ATET RMSE	Unit-level RMSE
(i) Per-unit	**8.5**	**13.9**	**8.3**	**16.0**
(ii) Two-way global	8.6	27.6	8.9	32.5
(iii) One-way global	8.5	27.6	8.5	32.5
(iv) Synthetic control (random treat.)	9.7	15.9	10.3	19.0
(v) Diff-in-means (random treat.)	12.1	29.7	11.5	33.6

Note: The reported RMSE's are multiplied by 10^3 for readability. The values in bold are the lowest in the respective columns and correspond to the methods that perform best.

two-way approach starts outperforming the alternatives since any weighting scheme leads to the same value of the ATET. Moreover, in the homogeneous treatment case the global problems always outperform the baselines when estimating the average treatment effect on the treated.

For the particular simulations we run, the one-way and two-way global objectives provide improvements over the baselines that vary from 12% to 31% in the homogeneous treatment case and from 11% to 30% in the heterogeneous treatment case when estimating the average treatment effect on the treated. The per-unit approach, (i), performs well across the board while being particularly effective when estimating the unit-level effects in the heterogeneous treatment case and providing an improvement of over 13% relative to the synthetic control approach, (iv), when the number of treated units is small and 16% when the number of treated units is large. It is not surprising that the per-unit problem provides a smaller improvement over the synthetic control when $K = 3$ because in that case the donor pool of units that are used for comparison with the treated units is large relative to the overall number of units and the probability that we will not be able to find a good synthetic comparison is relatively low. The situation is different when we only have 3 control units that are used for constructing synthetic outcomes for the remaining 7 treatment units. Section 4 provides an additional discussion of the optimal design when the number of treatment units changes. Neither is surprising that the per-unit approach performs poorly relative to the global problems and even the randomized difference-in-means estimator (but not the synthetic control) when estimating unit-level effects in the homogeneous treatment case. Since all treatment effects are the same, it is more efficient to pool all the data together and estimate the constant effect on the full sample rather than estimating the same quantity for each unit individually. What is important though, is the robustness of the per-unit approach which provides similar performance in the case of either homogeneous or heterogeneous treatment effects while the global problems and the randomized difference-in-means perform poorly when estimating unit-level effects in the heterogeneous treatment case.

6 Practice

There are a number of practical considerations that need to be addressed when using the proposed design and analysis approaches.

Formulating the mixed-integer programs. Both the per-unit problem and the global problems can be formulated as mixed-integer programs with quadratic objectives and linear constraints. As discussed in Section 2, the per-unit problem requires either an additional linear constraint that fixes the number of treated units, K, or an additional quadratic constraint that allows optimizing over the number of treated units. In addition to that, the per-unit problem uses more variables that need to be optimized since sets of weights vary across treatment units allowing separate estimation of every unit-level treatment effect. This implies that the per-unit problem is generally harder to solve and is only tractable for a smaller number of experimental units, N, compared to the global problems.

Choosing the penalty factor. The penalty factor, λ, used by all of the optimization problems can be chosen using cross-validation. Specifically, the pre-treatment time periods can be split into the consecutive training and validation time periods and the value of λ can be chosen by minimizing the RMSE over the validation period in a simulated experiment that is similar to the one we conduct in Section 5. An alternative approach motivated by the setting in Section 2 uses an estimate of the variance of the outcome variable. For example, the approach we take in Sections 4 and 5 computes the sample variances for every unit i across pre-treatment time periods $t = 1, \ldots, T$ and then uses the average of those quantities across all units as the penalty factor, λ.

Quantifying the uncertainty. Most applied settings require evaluating the uncertainty in the obtained estimates of the treatment effects. We suggest the permutation-based approach for testing the sharp null hypothesis of zero treatment effects across all treated units proposed by Chernozhukov et al. (2021). See the supplementary materials for the detailed description of the proposed inference procedure as well as a theoretical result that guarantees its validity—albeit under rather strong assumptions—and the power curves constructed using a simulated setting similar to that from Section 5. When the proposed procedure is used in conjunction with the per-unit or global problems it provides the correct (or conservative) test sizes and improves the power relative to the synthetic-control and difference-in-means approaches.

Computational complexity. Solving the global and the per-unit problems in their mixed-integer formulations becomes computationally burdensome as the total number of units increases, especially if the exact optimal solutions are required. In our simulations we were able to solve problems for $N = 50$ units—which is a meaningful threshold corresponding to the number of states, a typical experimental unit in synthetic-control-type studies—on a single machine within hours. However, we prove that the underlying optimization problem is NP-hard (by providing a reduction to the partitioning problem; for the exact proof see the supplementary materials), and therefore exact solutions to substantially larger problems are unlikely.

7 Conclusion

In this paper we evaluate the role of optimal experimental design in panel-data settings where traditionally the average treatment effect on the treated might be estimated using randomized design and the difference-in-means estimator. We propose several design-and-analysis procedures that can be solved as mixed-integer programs. Our empirical evaluations show that these procedures lead to a substantial improvement in terms of the root-mean-square error relative to the randomized difference-in-means as well as the randomized synthetic-control approaches.

We discuss the roles that underlying assumptions about the nature of the treatment effects, the estimands of interest, and the computational considerations play when deciding which approach should be used. We propose a permutation-based inference procedure that is shown to deliver the correct test sizes in simulations. We also discuss practical considerations when applying this methodology as well as its current limitations.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). 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.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review 93*(1), 113–132.

Abadie, A. and J. L'Hour (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association* (just-accepted), 1–34.

Abadie, A. and J. Zhao (2021). Synthetic controls for experimental design. *arXiv preprint arXiv:2108.02196*.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–30.

Bottmer, L., G. Imbens, J. Spiess, and M. Warnick (2021). A design-based perspective on synthetic control methods. *arXiv preprint arXiv:2101.09398*.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review 43*(2), 245–257.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 1–16.

Cieliebak, M., S. J. Eidenbenz, A. Pagourtzis, and K. Schlude (2008). On the complexity of variations of equal sum subsets. *Nord. J. Comput. 14*(3), 151–172.

Doudchenko, N., D. Gilinson, S. Taylor, and N. Wernerfelt (2019). Designing experiments with synthetic controls. *Working paper*.

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

Gamrath, G., D. Anderson, K. Bestuzheva, W.-K. Chen, L. Eifler, M. Gasse, P. Gemander, A. Gleixner, L. Gottwald, K. Halbig, G. Hendel, C. Hojny, T. Koch, P. Le Bodic, S. J. Maher, F. Matter, M. Miltenberger, E. Mühmer, B. Müller, M. E. Pfetsch, F. Schlösser, F. Serrano, Y. Shinano, C. Tawfik, S. Vigerske, F. Wegscheider, D. Weninger, and J. Witzig (2020, March). The SCIP Optimization Suite 7.0. Technical report, Optimization Online.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis 24*, 324–338.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*(477), 191–200.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688.

Rubin, D. B. et al. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics 2*(3), 808–840.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

A Supplementary materials

A.1 Conditional MSE of the treatment effect estimator

The expression for the conditional mean squared error used in Section 2 can be derived as follows.

First, for a treated unit i,

$$\mathbb{E}\left[(\hat{\tau}_i - \tau_i)^2 \left|\{D_j, w_j^i\}_{j=1}^N\right.\right] = \text{var}\left(\hat{\tau}_i \middle| \{D_j, w_j^i\}_{j=1}^N\right) + \left(\mathbb{E}\left[\hat{\tau}_i - \tau_i \middle| \{D_j, w_j^i\}_{j=1}^N\right]\right)^2.$$

Next,

$$\begin{aligned}
\hat{\tau}_i - \tau_i &= Y_{i,T+1} - \sum_{j:\, D_j=0} w_j^i Y_{j,T+1} - \tau_i \\
&= \mu_{i,T+1} + \tau_i + \varepsilon_{i,T+1} - \sum_{j:\, D_j=0} w_j^i (\mu_{j,T+1} + \varepsilon_{j,T+1}) - \tau_i \\
&= \left(\mu_{i,T+1} - \sum_{j:\, D_j=0} w_j^i \mu_{j,T+1}\right) + \left(\varepsilon_{i,T+1} - \sum_{j:\, D_j=0} w_j^i \varepsilon_{j,T+1}\right).
\end{aligned}$$

Treating ε's as the only source of randomness in the above expression and assuming that they are independent across units,

$$\text{var}\left(\hat{\tau}_i \middle| \{D_j, w_j^i\}_{j=1}^N\right) = \sigma^2 + \sum_{j:\, D_j=0} \left(w_j^i\right)^2 \sigma^2$$

and

$$\left(\mathbb{E}\left[\hat{\tau}_i - \tau_i \middle| \{D_j, w_j^i\}_{j=1}^N\right]\right)^2 = \left(\mu_{i,T+1} - \sum_{j:\, D_j=0} w_j^i \mu_{j,T+1}\right)^2$$

which lead to the equation in Section 2.

Similarly, for another ATET estimator $\hat{\tau} = \sum_{i:\, D_i=1} w_i Y_{i,T+1} - \sum_{i:\, D_i=0} w_i Y_{i,T+1}$:

$$\begin{aligned}
\hat{\tau} - \tau &= \sum_{i:\, D_i=1} w_i Y_{i,T+1} - \sum_{i:\, D_i=0} w_i Y_{i,T+1} - \tau \\
&= \left(\sum_{i:\, D_i=1} w_i \mu_{i,T+1} - \sum_{i:\, D_i=0} w_i \mu_{i,T+1}\right) + \left(\sum_{i:\, D_i=1} w_i \varepsilon_{i,T+1} - \sum_{i:\, D_i=0} w_i \varepsilon_{i,T+1}\right) \\
&\quad + \sum_{i:\, D_i=1} w_i \tau_i - \tau
\end{aligned}$$

and, assuming that we are estimating $\tau = \sum_{i:\, D_i=1} w_i \tau_i$, we arrive at the conditional mean squared error formula from Section 2 using the same logic as we used for the unit-level treatment effects.

It is important to point out that unless this assumption about the average treatment effect of interest is made, another term,

$$\left(\sum_{i:\, D_i=1} w_i \tau_i - \tau\right)^2,$$

remains as part of the MSE. In the current paper we make no claims about this term apart from the case when the treatment effects are homogeneous and the term vanishes.

Empirical analogs. The empirical analogs of the population-level equations above were presented in Section 2. While we do not provide any theoretical guarantees for the estimators obtained as solutions to the corresponding optimization problems, there are two types of assumptions that are likely necessary for a formal result. First, those that guarantee that the weights obtained based on the

pre-treatment data provide good approximations for the counterfactual outcomes in the treatment periods—these types of assumptions are typically used in the synthetic-control literature, such as the assumption that the underlying data generating process is described by a latent factor model (e.g. Abadie et al., 2010), or the assumption that treatment periods are themselves chosen at random and are thus comparable to the control periods (Bottmer et al., 2021, e.g.). Second, those that guarantee that the average is a good approximation of the corresponding conditional expectation which are likely satisfied when T goes to infinity.

A.2 Exact mixed-integer formulations

In this section we present the exact mixed-integer programming formulations that can be used for solving the proposed models in one of the available academic or commercial solvers. We use SCIP (Gamrath et al., 2020) which can handle mixed-integer nonlinear programs (MINLP's) with constraints that can be written as "expressions" with certain operations. It is preferable, however, that the constraints are quadratic or linear.[6]

Note that a general nonlinear objective $f(x)$ can be replaced by a linear objective y with an auxiliary variable y and an additional constraint $y \geq f(x)$ (for a minimization problem).

Per-unit problem. The per-unit problem was formulated as

$$\begin{aligned}
\min_{\{D_i, \{w_j^i\}_{j=1}^N\}_{i=1}^N} \quad & \frac{1}{K} \sum_{i=1}^N D_i \left[\frac{1}{T} \sum_{t=1}^T \left(Y_{it} - \sum_{j=1}^N w_j^i (1 - D_j) Y_{jt} \right)^2 + \lambda \sum_{j=1}^N \left(w_j^i \right)^2 \right] \\
\text{s.t.} \quad & w_j^i \geq 0, \quad D_i \in \{0, 1\} \text{ for } i, j = 1, \ldots, N, \\
& \sum_{i=1}^N D_i = K, \\
& \sum_{i=1}^N w_j^i (1 - D_j) = 1 \text{ for } i = 1, \ldots, N \text{ such that } D_i = 1
\end{aligned}$$

which can be rewritten by introducing auxiliary variables $q_{ij} = w_j^i(1 - D_j)$ for $i, j = 1, \ldots, N$ with a few additional constraints. Specifically, we impose $q_{ij} \geq 0$, $q_{ij} \leq 1 - D_j$, $q_{ij} \leq w_j^i$, and $q_{ij} \geq w_j^i - D_j$. Indeed, when $D_j = 1$, q_{ij} has to be equal to 0 given the constraints $q_{ij} \geq 0$ and $q_{ij} \leq 1 - D_j = 0$ while the constraints $q_{ij} \leq w_j^i$ and $q_{ij} \geq w_j^i - D_j = w_j^i - 1$ are non-binding. When $D_j = 0$, on the other hand, q_{ij} has to be equal to w_j^i since $q_{ij} \leq w_j^i$ and $q_{ij} \geq w_j^i - D_j = w_j^i$ while $q_{ij} \leq 1 - D_j = 1$ and $q_{ij} \geq 0$ are non-binding.

We need two additional observations to formulate the problem as a quadratic objective with linear constraints. First, since $D_i \in \{0, 1\}$, $D_i^2 = D_i$ and therefore D_i's can be carried inside the parentheses. Second, w_j^i need to be able to take nonzero values only for such i's that have $D_i = 1$. By imposing additional constraints $w_j^i \leq D_i$ for $i, j = 1, \ldots, N$ we can use w_j^i instead of $w_j^i D_i$.

[6] See `https://www.scipopt.org/doc/html/FAQ.php#minlptypes` for more details.

By utilizing an additional set of auxiliary variables, z_{it}, the per-unit problem can be written as

$$\begin{aligned}
\min_{\{D_i,\{w_j^i,q_{ij}\}_{j=1}^N,\{z_{it}\}_{t=1}^T\}_{i=1}^N} \quad & \frac{1}{KT}\sum_{i=1}^N\sum_{t=1}^T z_{it}^2 + \frac{\lambda}{K}\sum_{i=1}^N\sum_{i=1}^N \left(w_j^i\right)^2 \\
\text{s.t.} \quad & w_j^i \geq 0, \quad q_{ij} \geq 0, \quad D_i \in \{0,1\} \text{ for } i,j = 1,\ldots,N, \\
& \sum_{i=1}^N D_i = K, \\
& \sum_{j=1}^N q_{ij} = D_i \text{ for } i = 1,\ldots,N, \\
& q_{ij} \leq 1 - D_j, \quad q_{ij} \leq w_j^i, \quad q_{ij} \geq w_j^i - D_j \text{ for } i,j = 1,\ldots,N, \\
& w_j^i \leq D_i \text{ for } i,j = 1,\ldots,N, \\
& z_{it} = Y_{it}D_i - \sum_{j=1}^N q_{ij}Y_{jt} \text{ for } i = 1,\ldots,N \text{ and } t = 1,\ldots,T
\end{aligned}$$

which has a quadratic objective with a positive semi-definite Hessian and linear constraints.

The constraint on the number of treated units, $\sum_{i=1}^N D_i = K$, can be removed from the per-unit problem too if we recall the technique used for formulating general nonlinear objectives. Suppose that the objective written above is denoted as $f\left(\{\{w_j^i\}_{j=1}^N,\{z_i\}_{t=1}^T\}_{i=1}^N\right)/K$. Let us introduce an auxiliary variable y and two additional constraints: $y\sum_{i=1}^N D_i \geq f\left(\{\{w_j^i\}_{j=1}^N,\{z_i\}_{t=1}^T\}_{i=1}^N\right)$ (which is quadratic as long as f is quadratic) and $\sum_{i=1}^N D_i \geq 1$ which ensures that we are not normalizing by zero in the objective that we are actually trying to minimize.

Two-way global problem. The two-way global problem was formulated in the following way in the main part of the paper:

$$\begin{aligned}
\min_{\{D_i,w_i\}_{i=1}^N} \quad & \frac{1}{T}\sum_{t=1}^T\left(\sum_{i=1}^N w_iD_iY_{it} - \sum_{i=1}^N w_i(1-D_i)Y_{it}\right)^2 + \lambda\sum_{i=1}^N w_i^2 \\
\text{s.t.} \quad & w_i \geq 0, \quad D_i \in \{0,1\} \text{ for } i = 1,\ldots,N, \\
& \sum_{i=1}^N D_i = K, \\
& \sum_{i=1}^N w_iD_i = 1, \quad \sum_{i=1}^N w_i(1-D_i) = 1
\end{aligned}$$

This can be rewritten in a slightly different way that utilizes auxiliary variables z_t, $q_i = w_iD_i$ and the constraints similar to those used for the per-unit problem:

$$\begin{aligned}
\min_{\{D_i,w_i,q_i\}_{i=1}^N,\{z_t\}_{t=1}^T} \quad & \frac{1}{T}\sum_{t=1}^T z_t^2 + \lambda\sum_{i=1}^N w_i^2 \\
\text{s.t.} \quad & w_i \geq 0, \quad q_i \geq 0, \quad D_i \in \{0,1\} \text{ for } i = 1,\ldots,N, \\
& \sum_{i=1}^N D_i = K, \\
& \sum_{i=1}^N q_i = 1, \quad \sum_{i=1}^N w_i = 2, \\
& q_i \leq D_i, \quad q_i \leq w_i, \quad q_i \geq w_i - (1-D_i) \text{ for } i = 1,\ldots,N, \\
& z_t = \sum_{i=1}^N (2q_i - 1)Y_{it} \text{ for } t = 1,\ldots,T,
\end{aligned}$$

where the inequality constraints on variables q_i enforce the nonlinear equality constraints $q_i = w_i D_i$.

The problem above has a quadratic objective with a positive semi-definite (diagonal, in fact) Hessian and linear constraints.

Note that the constraint on the number of treated units, $\sum_{i=1}^{N} D_i = K$, can be safely removed without complicating any of the other constraints.

One-way global problem. If the constraint on the number of treated units, $\sum_{i=1}^{N} D_i = K$, is imposed, the problem above becomes the one-way global problem as soon as we impose additional constraints

$$q_i = \frac{D_i}{K} \quad \text{for } i = 1, \ldots, N.$$

Indeed, both sides are equal to zero when $D_i = 0$ and when $D_i = 1$ the constraint is equivalent to $w_i = 1/K$.

The problem becomes more complicated when there is no constraint on the number of treated units. We want to impose

$$q_i = \frac{D_i}{\sum_{j=1}^{N} D_j} \quad \text{for } i = 1, \ldots, N$$

which are nonlinear.[7] These constraints can be rewritten as linear by multiplying both sides by the denominator of the right-had side and introducing additional variables $r_{ij} = q_i D_j$ for $i, j = 1, \ldots, N$ and enforcing this equality in the same way that we used for $q_i = w_i D_i$. This, however, substantially increases the number of required variables from $O(N)$ to $O(N^2)$.

A.3 Proof of Theorem 1

In this section we provide a proof of Theorem 1. To derive the formulas presented in this theorem, we solve for the optimal set of weights in each optimization problem. We will start with the proofs for the two- and one-way global problems which are then utilized in the proof for the per-unit problem.

Two-way global problem. In this case, we can define the Lagrangian of the relaxed objective—recall that we allowed weights to be negative—as follows:

$$\mathcal{L}(w, \lambda_1, \lambda_2) = \left(\sum_{i \in I} a_i w_i - \sum_{j \in \bar{I}} a_j w_j \right)^2 + \sigma^2 \sum_{i=1}^{n} w_i^2 - \lambda_1 \left(\sum_{i \in I} w_i - 1 \right) - \lambda_2 \left(\sum_{j \in \bar{I}} w_j - 1 \right).$$

Taking a derivative with respect to w_l where $l \in I$ and equating it to zero gives us

$$\frac{\partial \mathcal{L}(w, \lambda_1, \lambda_2)}{\partial w_l} = 2a_l \left(\sum_{i \in I} a_i w_i - \sum_{j \in \bar{I}} a_j w_j \right) + 2\sigma^2 w_l - \lambda_1 = 0. \tag{1}$$

Similarly, for $l \in \bar{I}$ we have:

$$\frac{\partial \mathcal{L}(w, \lambda_1, \lambda_2)}{\partial w_l} = 2a_l \left(\sum_{j \in \bar{I}} a_j w_j - \sum_{i \in I} a_i w_i \right) + 2\sigma^2 w_l - \lambda_2 = 0. \tag{2}$$

These equations together with $\sum_{i \in I} w_i = 1$ and $\sum_{j \in \bar{I}} w_j = 1$ lead to a (linear) system of equations with $N + 2$ variables which can be solved. We claim that the solution to this system is given by:

$$w_l^* = \frac{1}{K} + \frac{(\bar{a}_I - \bar{a}_{\bar{I}})(\bar{a}_I - a_l)}{\sigma^2 + V_I^2 + V_{\bar{I}}^2} \quad \text{for } l \in I,$$

$$w_l^* = \frac{1}{N - K} - \frac{(\bar{a}_I - \bar{a}_{\bar{I}})(\bar{a}_{\bar{I}} - a_l)}{\sigma^2 + V_I^2 + V_{\bar{I}}^2} \quad \text{for } l \in \bar{I}\,.$$

[7] We do not need to worry about $\sum_{j=1}^{N} D_j = 0$ which is ruled out by other constraints.

It is easy to show that $\sum_{l\in I} w_l^* = \sum_{l\in \bar{I}} w_l^* = 1$. Now it remains to show that for a suitable choice of λ_1 and λ_2, Eq. (1) and (2) are satisfied. For any $l \in I$, we can write

$$\begin{aligned}
&2a_l\left(\sum_{i\in I} a_i w_i - \sum_{j\in \bar{I}} a_j w_j\right) + 2\sigma^2 w_l \\
&= 2a_l\left(\bar{a}_I - \bar{a}_{\bar{I}}\right)\left(1 + \frac{1}{\sigma^2 + V_I^2 + V_{\bar{I}}^2}\left[\sum_{i\in I}(\bar{a}_I - a_i)a_i + \sum_{j\in \bar{I}}(\bar{a}_{\bar{I}} - a_j)a_j\right]\right) \\
&\quad + 2\sigma^2\left(\frac{1}{K} + \frac{(\bar{a}_I - \bar{a}_{\bar{I}})(\bar{a}_I - a_l)}{\sigma^2 + V_I^2 + V_{\bar{I}}^2}\right).
\end{aligned}$$

For simplicity, denote $\bar{a}_I - \bar{a}_{\bar{I}} = A$ and $\sigma^2 + V_I^2 + V_{\bar{I}}^2 = B$. Then, noting that

$$\begin{aligned}
\sum_{i\in I}(\bar{a}_I - a_i)a_i &= -V_I^2, \\
\sum_{j\in \bar{I}}(\bar{a}_{\bar{I}} - a_j)a_j &= -V_{\bar{I}}^2,
\end{aligned}$$

the above is equal to:

$$\begin{aligned}
2a_l A\left(1 - \frac{V_I^2 + V_{\bar{I}}^2}{B}\right) + \frac{2\sigma^2}{K} + \frac{(2\sigma^2 A)(\bar{a}_I - a_l)}{B} &= 2a_l A(\frac{\sigma^2}{B}) + \frac{2\sigma^2}{K} + \frac{2\sigma^2 A}{B}\bar{a}_I - 2a_l A\frac{\sigma^2}{B} \\
&= 2\sigma^2\left(\frac{1}{K} + \bar{a}_I\frac{A}{B}\right)
\end{aligned}$$

which is independent of the index l. In particular, if we let

$$\lambda_1 = 2\sigma^2\left(\frac{1}{K} + \bar{a}_I\frac{A}{B}\right),$$

then for all $l \in I$, the condition in Eq. (1) is satisfied. Similarly, it can be shown that for

$$\lambda_2 = 2\sigma^2\left(\frac{1}{N-K} - \bar{a}_{\bar{I}}\frac{A}{B}\right),$$

Eq. (2) is satisfied for all $l \in \bar{I}$. Given the computed sets of weights, we can now calculate the optimal objective value as follows:

$$\begin{aligned}
J_{\text{2-way}}(I) &= \left(\sum_{i\in I} a_i w_i^* - \sum_{j\in \bar{I}} a_j w_j^*\right)^2 + \sigma^2\sum_{i=1}^{n} w_i^{*2} \\
&= (\bar{a}_I - \bar{a}_{\bar{I}})^2\left(1 + \frac{1}{\sigma^2 + V_I^2 + V_{\bar{I}}^2}\left[\sum_{i\in I}(\bar{a}_I - a_i)a_i + \sum_{j\in \bar{I}}(\bar{a}_I - a_j)a_j\right]\right)^2 \\
&\quad + \sigma^2\left(\frac{1}{K} + \sum_{i\in I}\frac{(\bar{a}_I - \bar{a}_{\bar{I}})^2(\bar{a}_I - a_i)^2}{(\sigma^2 + V_I^2 + V_{\bar{I}}^2)^2}\right) \\
&\quad + \sigma^2\left(\frac{1}{N-K} + \sum_{j\in \bar{I}}\frac{(\bar{a}_I - \bar{a}_{\bar{I}})^2(\bar{a}_{\bar{I}} - a_j)^2}{(\sigma^2 + V_I^2 + V_{\bar{I}}^2)^2}\right) \\
&= \sigma^2\left(\frac{1}{K} + \frac{1}{N-K}\right) + A^2\left[\left(\frac{\sigma^2}{B}\right)^2 + \sigma^2\left(\frac{V_I^2 + V_{\bar{I}}^2}{B^2}\right)\right] \\
&= \sigma^2\left(\frac{1}{K} + \frac{1}{N-K} + \frac{A^2}{B}\right) \\
&= \sigma^2\left(\frac{1}{K} + \frac{1}{N-K} + \frac{(\bar{a}_I - \bar{a}_{\bar{I}})^2}{\sigma^2 + V_I^2 + V_{\bar{I}}^2}\right)
\end{aligned}$$

where we used $A = \bar{a}_I - \bar{a}_{\bar{I}}$ and $B = \sigma^2 + V_I^2 + V_{\bar{I}}^2$. This completes the proof for the two-way problem.

One-way global problem. The derivation here is similar to the two-way problem. Indeed, the Lagrangian can be written as

$$\mathcal{L}(w, \lambda) = \left(\bar{a}_I - \sum_{j \in \bar{I}} a_j w_j \right)^2 + \frac{\sigma^2}{K} + \sigma^2 \sum_{j \in \bar{I}} w_j^2 - \lambda \left(\sum_{j \in \bar{I}} w_j - 1 \right).$$

For any $l \in \bar{I}$, the weights need to satisfy

$$\frac{\partial \mathcal{L}(w, \lambda)}{\partial w_l} = 2a_l \left(\sum_{j \in \bar{I}} a_j w_j - \sum_{i \in I} a_i w_i \right) + 2\sigma^2 w_l - \lambda = 0. \tag{3}$$

This, together with the condition that $\sum_{l \in \bar{I}} w_l = 1$, leads to a system of equations that can be solved. Similarly to the two-way problem, we claim that the following set of weights satisfy the first-order conditions:

$$w_l^* = \frac{1}{K} \quad \text{for } l \in I,$$
$$w_l^* = \frac{1}{N-K} - \frac{(\bar{a}_I - \bar{a}_{\bar{I}})(\bar{a}_{\bar{I}} - a_l)}{\sigma^2 + V_{\bar{I}}^2} \quad \text{for } l \in \bar{I}\,.$$

It is straightforward to verify that $\sum_{l \in \bar{I}} w_l^* = 1$. Also, for any $l \in \bar{I}$ we can write:

$$\begin{aligned}
&2a_l \left(\sum_{j \in \bar{I}} a_j w_j - \bar{a}_I \right) + 2\sigma^2 w_l \\
&= 2a_l \left(\bar{a}_{\bar{I}} - \bar{a}_I \right) \left(1 + \frac{1}{\sigma^2 + V_{\bar{I}}^2} \left[\sum_{j \in \bar{I}} (\bar{a}_{\bar{I}} - a_j) a_j \right] \right) + 2\sigma^2 \left(\frac{1}{N-K} + \frac{(\bar{a}_{\bar{I}} - \bar{a}_I)(\bar{a}_{\bar{I}} - a_l)}{\sigma^2 + V_{\bar{I}}^2} \right) \\
&= 2 \left(\bar{a}_{\bar{I}} - \bar{a}_I \right) \left[\left(\frac{\sigma^2}{\sigma^2 + V_{\bar{I}}^2} a_l \right) + \sigma^2 \frac{\bar{a}_{\bar{I}} - a_l}{\sigma^2 + V_{\bar{I}}^2} \right] + \frac{2\sigma^2}{N-K} \\
&= 2\sigma^2 \left[\frac{(\bar{a}_{\bar{I}} - \bar{a}_I)\bar{a}_{\bar{I}}}{\sigma^2 + V_{\bar{I}}^2} + \frac{1}{N-K} \right]
\end{aligned}$$

which is independent of l. Hence, for

$$\lambda = 2\sigma^2 \left[\frac{(\bar{a}_{\bar{I}} - \bar{a}_I)\bar{a}_{\bar{I}}}{\sigma^2 + V_{\bar{I}}^2} + \frac{1}{N-K} \right]$$

Eq. (3) is satisfied for all $l \in \bar{I}$. Substituting these optimal weights into the formula for the one-way objective yields

$$\begin{aligned}
J_{\text{1-way}}(I) &= \left(\sum_{i \in I} a_i w_i^* - \sum_{j \in \bar{I}} a_j w_j^* \right)^2 + \sigma^2 \sum_{i=1}^{n} {w_i^*}^2 \\
&= (\bar{a}_I - \bar{a}_{\bar{I}})^2 \left(1 + \frac{1}{\sigma^2 + V_{\bar{I}}^2} \left[\sum_{j \in \bar{I}} (\bar{a}_I - a_j) a_j \right] \right)^2 + \sigma^2 \left(\frac{1}{K} \right) \\
&\quad + \sigma^2 \left(\frac{1}{N-K} + \sum_{j \in \bar{I}} \frac{(\bar{a}_I - \bar{a}_{\bar{I}})^2 (\bar{a}_{\bar{I}} - a_j)^2}{(\sigma^2 + V_{\bar{I}}^2)^2} \right) \\
&= \sigma^2 \left(\frac{1}{K} + \frac{1}{N-K} \right) + (\bar{a}_I - \bar{a}_{\bar{I}})^2 \left[\left(\frac{\sigma^2}{\sigma^2 + V_{\bar{I}}^2} \right)^2 + \sigma^2 \frac{V_{\bar{I}}^2}{(\sigma^2 + V_{\bar{I}}^2)^2} \right] \\
&= \sigma^2 \left(\frac{1}{K} + \frac{1}{N-K} + \frac{(\bar{a}_I - \bar{a}_{\bar{I}})^2}{\sigma^2 + V_{\bar{I}}^2} \right)
\end{aligned}$$

which completes the proof for the one-way problem.

Per-unit problem. The per-unit problem can be thought of as solving K separate two-way (or one-way) global problems where in each sub-problem, a single treated unit is selected as set I and all $N-K$ control units are in the set $\bar{I}$. Hence, using our derivation for the one-way global problem, in each sub-problem with units in P and $\bar{P}$ where $P = \{i\}$ for $i \in I$ and $\bar{P} = \bar{I}$ the optimal weights are given by

$$w_j^i = \frac{1}{N-K} - \frac{(a_i - \bar{a}_{\bar{I}})(\bar{a}_{\bar{I}} - a_j)}{\sigma^2 + V_{\bar{I}}^2}.$$

Note that we can also use our derivation for the one-way global problem to calculate the optimal objective within each sub-problem, with a single change that in the per-unit objective we do not penalize the weight of the single treated unit in P. In other words, the term $\sigma^2/1$ of the objective will not show up in the calculations. Hence, denoting J_i^* as the optimal value of this sub-problem, we have

$$J_i^* = \sigma^2 \left(\frac{1}{N-K} + \frac{(a_i - \bar{a}_{\bar{I}})^2}{\sigma^2 + V_{\bar{I}}^2} \right).$$

Furthermore, for the per-unit objective we can write $J_{\text{per-unit}}(I) = \sum_{i \in I} J_i^*/K$ which implies

$$\begin{aligned}
J_{\text{per-unit}}(I) = \frac{1}{K} \sum_{i \in I} J_i^* &= \frac{1}{K} \sum_{i \in I} \sigma^2 \left(\frac{1}{N-K} + \frac{(a_i - \bar{a}_{\bar{I}})^2}{\sigma^2 + V_{\bar{I}}^2} \right) \\
&= \frac{\sigma^2}{N-K} + \frac{\sigma^2}{\sigma^2 + V_{\bar{I}}^2} \cdot \frac{\sum_{i \in I} (a_i - \bar{a}_{\bar{I}})^2}{K} \\
&= \frac{\sigma^2}{N-K} + \frac{\sigma^2}{\sigma^2 + V_{\bar{I}}^2} \cdot \frac{\sum_{i \in I} (a_i - \bar{a}_I + \bar{a}_I - \bar{a}_{\bar{I}})^2}{K} \\
&= \frac{\sigma^2}{N-K} + \frac{\sigma^2}{\sigma^2 + V_{\bar{I}}^2} \left[(\bar{a}_I - \bar{a}_{\bar{I}})^2 + \frac{V_I^2}{K} \right] \\
&= \sigma^2 \left(\frac{1}{N-K} + \frac{(\bar{a}_I - \bar{a}_{\bar{I}})^2 + K^{-1} V_I^2}{\sigma^2 + V_{\bar{I}}^2} \right).
\end{aligned}$$

A.4 Inference

A standard way of performing permutation-based inference for synthetic-control procedures suggested in Abadie et al. (2010) involves permuting which units receive the treatment—or choosing placebo treated units among the control units—to obtain a reference distribution of treatment-effect estimates under the sharp null hypothesis of no effect on any of the units in the treated periods. However, our method chooses the treated units themselves which makes this type of permutation infeasible. Instead we focus on inference methods that permute the treatment periods closely following the methodology suggested by Chernozhukov et al. (2021)—subsequently CWZ.

Specifically, we draw bootstrap samples (without replacement) of the S time periods (including both the pre-treatment periods and the treatment periods). CWZ suggest using one of the two permutation regimes: (i) *iid permutations* which allow for an arbitrary order of the time periods in the bootstrap sample or (ii) *moving block permutations* in which every bootstrap sample is a cyclic shift of the original sample over the time periods. While the conditions that ensure the validity of the second approach are less strict than those required by the first approach, the second approach only generates at most S unique bootstrap samples, and therefore requires more data for convergence. While the overall size of our dataset, $N \times S = 50 \times 40$, is relatively small, we present the results obtained using both permutation schemes. For each bootstrap sample we re-estimate the average treatment effect on the treated (ATET) assuming that the same number of periods at the end of the bootstrap sample are treated as the number of treatment periods in the original sample. Following CWZ, we use the absolute value of the estimate divided by the square root of the number of treatment periods as the test statistic, denoted $U(Y)$, and construct its permutation distribution under the sharp null hypothesis of all individual treatment effects being equal to zero in all of the treatment time periods.

We reject the null hypothesis at a significance level $1-\alpha$ if the original test statistic is larger than the $1-\alpha$ fraction of the computed bootstrap values.

As a baseline, we note that both permutation procedures lead to a valid test for the sharp null of no treatment effects if time periods are exchangeable in the following sense:

Proposition 1. *Assume that across time periods $t = 1, \ldots, S$ the vector of potential outcomes $Y_t(0) = (Y_{1t}(0), \ldots, Y_{Nt}(0))$ of all units at time t in the absence of the treatment is drawn independently and identically, and that the test statistic $U(Y)$ permits a bounded density function. Then the test of the sharp null $H_0\colon Y_{it}(1) = Y_{it}(0)$ for all $i = 1, \ldots, N$ in all treatment periods $t = T+1, \ldots, S$ of no treatment effects is unbiased in size, in the sense that it rejects a true null hypothesis with probability α.*

Note that this proposition follows directly from permutation invariance of the outcome vectors $Y_t(0)$, and does not require that units themselves are exchangeable or that treatment is assigned randomly across units. Nevertheless, the assumption that the distribution of the full vector of potential control outcomes is independently drawn across time periods, including the treated periods, is unrealistic in many time series settings. A more realistic treatment would establish exchangeability only for the regression residuals. While a rigorous proof of a version of Proposition 1 under such weaker assumptions is beyond the scope of the current paper, CWZ state sufficient conditions for valid inference in a similar setting where treatment units are fixed. Specifically, they require that: (i) the estimator used for the construction of individual counterfactual outcomes in the absence of the treatment is unbiased for $Y_{it}(0)$, (ii) the treatment effects are fixed (nonrandom) and additive, $Y_{it}(1) = Y_{it}(0) + \tau_{it}$ (this would not be required to test the null hypothesis $Y_{it}(1) = Y_{it}(0)$, but allows for constructing of confidence sets and testing other sharp hypotheses), and (iii) the remaining noise variables, $\hat{\varepsilon}_{it}$, equal to the differences between the estimators of $Y_{it}(0)$ and the values themselves, are mean zero and either $i.i.d.$ across units and time (justifying the iid permutations) or $i.i.d.$ across units and following a stationary weakly dependent process across time (in which case moving block permutations should be used). CWZ also provide sufficient conditions on commonly used estimators satisfying assumption (i).

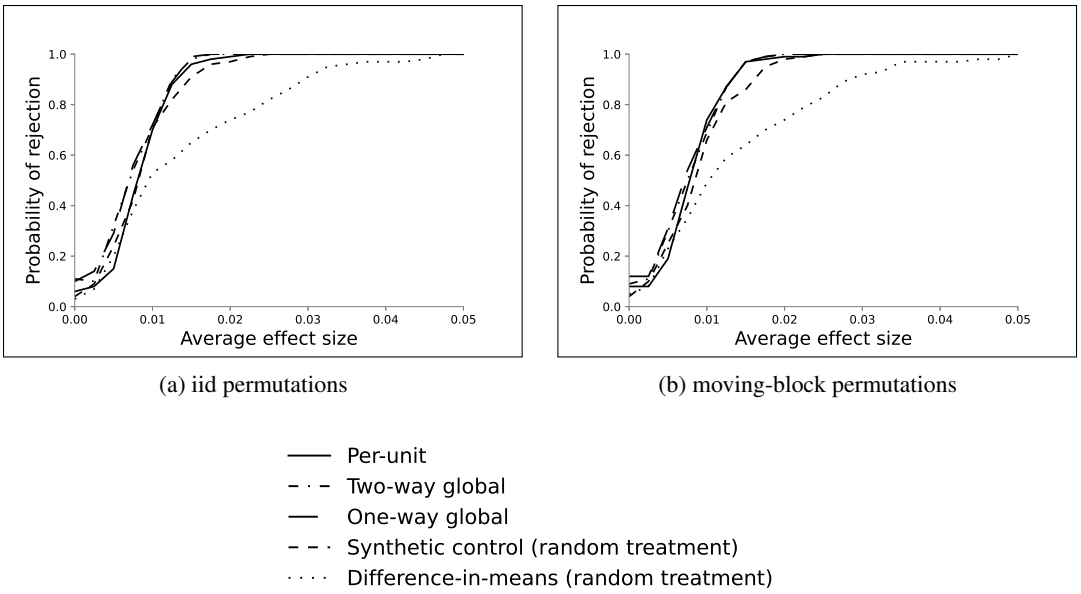
We construct power curves using simulated data and both types of permutations. Figure 2 plots the probability of rejecting the sharp null hypothesis of zero treatment effects as a function of the true value of the ATET. It presents two types of inference results. First, for the true treatment effects equal to zero across all units and all time periods the plots show the fraction of simulations that reject the sharp null hypothesis of zero treatment effects at the 90% significance level. This provides an estimate of the actual size of the nominally 10% size test. Second, the plots show how the rejection probabilities change as the true value of the average treatment effect on the treated increases.

Specifically, we run 100 simulations. The data for each simulation include 10 units that are chosen randomly (out of 50 available) and all 40 of the available time periods. We assume that the treatment is applied in the last 5 time periods to the 3 chosen treatment units. Depending on the method, the treatment units are chosen either randomly or using a mixed-integer program. The treatment effects are assumed to be heterogeneous and equal to 0, $\tau/2$, and τ (implying the average treatment effect on the treated of $\tau/2$) for the chosen treatment units in the order that corresponds to their order in the sampled data.[8] The average treatment effect on the treated is estimated as described in Section 5 of the paper. Within each simulation, we repeat this procedure for each of the 40 permuted samples and compute the 90% quantile of the bootstrap distribution of the test statistic.[9] If the test statistic computed on the originally sampled data (before the bootstrap) exceeds that quantile, we reject the sharp null hypothesis. Notably, the methods proposed in the paper—the per unit problem as well as the two- and one-way global problems—have test sizes at most as large and the power generally exceeding those of the randomized methods.[10]

A.5 Hardness

We prove that the presented optimization problems are indeed NP-Hard. We do so by providing a formal reduction from a variant of the partitioning problem that is known to be NP-hard. In an

[8] This setting is different from the one used in Section 5 where the treatment effects did not depend on the identity of the treated units. The setting we use here can still be formulated in terms of the potential outcomes framework. However, it will require the potential outcomes under treatment to depend on the total number of treated units as well as the order of the treated units introducing interference. We use this setting to guarantee



(a) iid permutations

(b) moving-block permutations

Figure 2: Power curves.

instance of the equal-size partitioning problem, we are given a set of numbers $B = \{b_1, b_2, \cdots, b_n\}$. Let $T = \sum_{i=1}^{n} b_i$. The equal-size partitioning problem is to decide if there exists a subset $S \subset B$ of $n/2$ items with total sum of elements equal to $T/2$. This decision problem is NP-complete (Cieliebak et al., 2008). We show that we can distinguish between a "YES" and "NO" instance of this problem using an optimal algorithm for our optimization problems, hence proving that our problem is also NP-hard.

Given an instance $(B, T/2)$ of the partitioning problem, consider an instance of the $J_{\text{2-way}}(I)$ or $J_{\text{1-way}}(I)$ problem where we set $N = \{a_i | 1 \leq i \leq n\}$ with $a_i = b_i + T$. We note that with this transformation, there exists a subset I with a total sum of $\frac{nT+T}{2}$ if and only if there exists a subset $S \subset B$ of $n/2$ items with total sum of $T/2$. Furthermore, for such a subset I, since $|I| = |N \backslash I|$, the average of items in I will be the same as the average of $\bar{I} = N \backslash I$ and equivalently, $(\overline{a}_I - \overline{a})^2 = 0$. We also observe that the optimal solution for both $J_{\text{2-way}}(I)$ and $J_{\text{1-way}}(I)$ is at least $\frac{1}{K} + \frac{1}{N-K}$. As a result, the optimal solution for problems $J_{\text{2-way}}(I)$ or $J_{\text{1-way}}(I)$ is $\frac{1}{K} + \frac{1}{N-K}$ if and only if we can find a subset of size $n/2$ of N with the total sum of $\frac{nT+T}{2}$ or, equivalently, if and only if there exists a subset $S \subset B$ of $n/2$ items of B with the total sum of $T/2$. In other words, determining if the optimal solution is $\frac{1}{K} + \frac{1}{N-K}$ corresponds to having a "YES" instance of the equal-size partitioning problem $(B, T/2)$. Therefore, finding such an optimal solution is NP-hard.

that the true value of the average treatment effect on the treated remains constant across all simulations and methods allowing us to minimize the noise while using a relatively small number of simulations.

[9]We limit the number of bootstrap samples to 40 to make the inference results obtained using either of the permutation schemes more easily comparable since the maximum number of distinct samples that the moving block permutations allow is equal to $S = 40$.

[10]Note that the effect size of 0.05 corresponds to roughly the average outcome value observed in the data.