clusterSC: Advancing Synthetic Control with Donor Clustering for Disaggregate-level Data

A Preprint

Saeyoung Rho
Department of Computer Science
Columbia University
New York, NY 10027

Andrew Tang
Department of Computer Science
Columbia University
New York, NY 10027

Noah Bergam
Department of Computer Science
Columbia University
New York, NY 10027

Rachel Cummings
IEOR Department
Columbia University
New York, NY 10027

Vishal Misra
Department of Computer Science
Columbia University
New York, NY 10027

October 27, 2024

Abstract

In the domain of causal inference with observational datasets, synthetic control (SC) has emerged as a prominent tool. Traditionally, SC has been applied to aggregate-level datasets, such as those measured at the city or state level. However, recent advancements have extended its application to more granular datasets (e.g., individual-level), which comes with a greater number of observed units. This change introduces the curse of dimensionality to the SC setup, making the model more challenging to learn. Motivated by the idea that groups of individuals may exist where behavior aligns internally but diverges between groups, we propose a novel approach to selecting an appropriate subset of donors when constructing an SC instance. We provide theoretical guarantees on the improvements induced by our method over existing approach, accompanied by empirical demonstration on synthetic datasets. Our method almost always increases the performance, particularly in highly noisy settings.

Keywords synthetic control · causal inference · clustering

1 Introduction

Synthetic control (SC) has emerged in the econometrics community as a natural extension of the Difference-in-Differences (D-in-D, Card and Krueger (1993)). By leveraging time-series data from both pre- and post-intervention periods, SC evaluates the impact of an intervention on a *target unit* by constructing a synthetic counterfactual using a weighted combination of *donor units*, rather than selecting the nearest neighbor as in D-in-D. Much of the practical usage of SC has been with aggregate-level data, such as assessing the economic impact of government policies or political events at the state or regional level (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015; Kreif et al., 2016).

Recently, there has been increasing attention to employing SC on disaggregate-level data, observed in contexts like clinical trials with individual health records (Thorlund et al., 2020) and economic analyses using data with individual incomes (Abadie and L'Hour, 2021). In disaggregate-level datasets, the number of observed donor units can increase dramatically, easily exceeding the number of time-series measurements. Although more data typically means more information, the dimension of the synthetic control weights is determined by the number of units in the donor data. Hence, increased number of donors may introduce the *curse of dimensionality*, where learning happens in a high-dimensional space with only a few time-series measurements.

In light of this, we revisit the core motivation of synthetic control, which is to construct a *similar* counterpart to the target unit. What if there is a group of donors that behaves most similarly to the given target unit? We hypothesize an underlying group-based structure where the latent variables have a certain *structural separation*. Specifically, we focus on the distribution of the right singular vectors in each unit, and suggest clustering the donor pool before learning SC weights. We then analyze the impact of selecting a subgroup of donors, rather than the entire donor pool, within the SC framework.

Our contribution is twofold. First, we introduce a novel approach to disaggregate-level SC to mitigate high noise and dimensionality issues by incorporating a donor clustering step. Second, we provide a theoretical analysis of our algorithm's guarantees, based on the structural assumptions in the latent variable space. We also validate our approach empirically on synthetic datasets, demonstrating the improved prediction accuracy achieved by our method.

Section 2 introduces the synthetic control family of methods and defines relevant notations. Based on this, we formalize the problem setup and introduce structural assumptions in Section 3. Our main algorithm is introduced in Section 4, with theoretical analyses in Section 4.2. Finally, Section 5 empirically evaluates the performance of our approach on synthetic datasets.

2 Synthetic Control (SC) Methods

Before introducing SC methods, we introduce some key notation. Let X_i be the i-th row and $X_{i,t}$ be the element in the i-th row and the t-th column of X. Usually, a target unit index is 0 and the data is vector x_0. A donor matrix $X \in \mathbb{R}^{n \times T}$ is formed with n donor units and T observations. Assuming an intervention at time $T_0 < T$, X can be split into pre-intervention portion $X^- \in \mathbb{R}^{n \times T_0}$ and post-intervention portion $X^+ \in \mathbb{R}^{n \times T - T_0}$. Similarly for a vector, $x = [x^-, x^+]$. We denote the i-th singular value of a matrix X by $\sigma_i(X)$ and the i-th eigenvalue of a square matrix X by $\lambda_i(X)$. If needed, we denote the left and right singular vectors of a matrix X as $u_i(X)$ and $v_i(X)$, respectively.

Synthetic Control Family of Methods. Imagine that you are a coffee shop owner. You have T observations of daily coffee consumption $x_{i,t} \in \mathbb{R}$ for all customers (units) $i \in V$ and for all time points $t \in [T]$. At time $T_0 < T$, you decide to run a new promotion (intervention) on a subset of the customers (treated units) $W \subset V$, while keeping others the same (control units, potential donors). Hence, for each treated unit $i \in W$, we have a pre-intervention time series $x_i^- \in \mathbb{R}^{T_0}$ and post-intervention time series $x_i^+ \in \mathbb{R}^{T - T_0}$. For control unit $j \in V \setminus W$, we can use the same notation but the post-intervention time series x_j^+ was not affected by the intervention.

Synthetic Control estimates the effect of an intervention on treated units in W by constructing the counterfactual for the post-intervention period. It is important to note that SC constructs a separate model for each treated unit, allowing the causal estimand to be calculated on a per-unit basis. The SC family of methods learns the relationship between a target unit ($i = 0$ from W) and donor units ($j = 1, \ldots, n$ from $V \setminus W$) using pre-intervention time series data. Assuming this relationship remains stable over time $t \in [T]$, the counterfactual post-intervention time series for the target unit is inferred using donor data from the post-intervention period. Algorithm 1 formally defines the synthetic control family of methods.

Algorithm 1: Synthetic Control Family of Methods

Data: Target time series vector $x_i \in \mathbb{R}^T$ for each treated unit $i \in W$. Donor data $X \in \mathbb{R}^{n \times T}$ containing all control units $j \in V \setminus W$.
for $i \in W$ **do**
  1. Learn $f = \mathcal{M}(X, x_i^-)$
  2. Project $\hat{m}_i^+ = f(X^+)$
  3. Infer the estimated causal effect of the intervention for target i is $x_i^+ - \hat{m}_i^+$
end for

In the first step of Algorithm 1, $\mathcal{M}$ learns weights f to represent the target unit as a linear combination of the donor units. In the original work on synthetic control, Abadie and Gardeazabal (2003) use linear regression with a simplex constraint on the weights (i.e., the regression coefficients should be non-negative and sum to one). They used data on per capita GDP in $n = 17$ Spanish regions (aggregate level) to measure the effect of terrorism on Basque Country's per capita GDP. Later, more advanced variations of synthetic control have been proposed to deal with multiple treated units (Dube and Zipperer, 2015; Abadie and L'Hour, 2021), to correct bias (Ben-Michael et al., 2021; Abadie and L'Hour, 2021), to remove simplex constraints (Doudchenko and Imbens, 2016; Amjad et al., 2018), to ensure differential privacy (Rho et al., 2023), to incorporate matrix completion techniques (Athey et al., 2021; Amjad et al., 2019), and to consider temporal order (Brodersen et al., 2015).

In this paper, we will use Algorithm 2 as our learning method $\mathcal{M}$, which is based on the method proposed by Amjad et al. (2018). It denoises the donor matrix by retaining only the top r singular values through hard singular value thresholding (HSVT) (Cai et al., 2010; Chatterjee, 2015), followed by ordinary least squares to obtain the weight vector f. This approach is known for its robustness to noisy data, making it well-suited to our objectives.

Algorithm 2: Synthetic Control Core Algorithm $\mathcal{M}(X, x^-; r)$

Hyperparameter: the number of singular values to keep r
1. Perform SVD
$X = \sum_{i=1}^{T} \sigma_i u_i v_i^\top$, σ_i in decreasing order.
2. Denoise (HSVT) $\hat{M} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$.
3. Return SC weights
$\hat{f} = \arg\min_{f \in \mathbb{R}^n} \|\hat{M}^\top f - x^-\|$ (SC weights)

An intuitive way to view SC is a linear regression vertically performed on the dataset. With the pre-intervention donor matrix X^- as a regressor and the pre-intervention target time series x_0^- as a regressand, the j-th element of the weight f represents the importance of the j-th donor unit in explaining the target unit 0. Since a column of the matrix X^- becomes one sample for learning, we call this a *vertical regression*.

Another way to view SC is as a matrix completion problem with post-intervention target data as missing values. Athey et al. (2021) formalizes SC as a matrix completion method by setting an objective function based on the Frobenius norm of the difference between the latent and the observed matrix. The core modeling assumption of this approach is that the matrix is approximately low-rank. This is achieved by assuming a Lipschitz-continuous latent variable model with bounded latent variables (Candes and Plan, 2010; Candes and Recht, 2012; Nguyen et al., 2019).

SC on Disaggreagate-level Data When applying SC to disaggregate-level data, meeting these assumptions becomes more challenging. For example, there might be a certain *type* of units (such as patients with a certain phenotype) that can be well-approximated by a low-rank matrix, but not when mixed with other units in different types. When the number of potential donors is small, it may be possible to hand-pick a suitable donor set based on background knowledge, which usually is the case for aggregate-level datasets (Abadie and Gardeazabal, 2003; Abadie et al., 2015).

However, with disaggregate-level data, researchers must devise more data-driven approaches to select the appropriate donor units for a given target. Abadie and L'Hour (2021) used a penalty term to keep the *active units* in the donor pool small. Other works suggest using LASSO (Chernozhukov et al., 2021) or elastic net (Doudchenko and Imbens, 2016) regularizers to achieve similar results. Still, these SC configurations operate in n-dimensional spaces, which is less feasible when n is large.

3 Problem Setup

In this paper, we focus on applying SC to disaggregate-level data. Given the abundance of donor units, our objective is to develop a pre-processing step for SC that selects the optimal set of donors for a given target unit. In the following subsections, we present a detailed model tailored to this setting.

The assess the performance of SC methods, researchers often construct a *placebo* test Abadie and Gardeazabal (2003), where SC is used to predict post-intervention data in the absence of an intervention, or equivalently, to predict the post-intervention time series of a control unit using other control units as the donor pool. In these settings, since the target is drawn from the same distribution as the donors, the estimated causal effect (from Algorithm 1) should be zero. In the remainder of the paper, we focus on these placebo studies.

3.1 Model

Let x_0 be the target unit and let $X \in \mathbb{R}^{n \times T}$ be the donor data matrix, where each row x_i is a T-length time-series measurement. In light of disaggregate-level data, we assume $n \gg T$ (i.e., X is a tall matrix). We assume that the true data generation model comes from a latent variable model, plus some observation noise. That is, $X = M + E$, where $M_{i,t}$ is the true (deterministic) signal with entries bounded $-1 \leq M_{i,t} \leq 1$, and $E_{i,t}$ is mean-zero noise with finite variance, for all $i \in \{0, \cdots, n\}$ and $t \in [T]$. Similarly, we assume the target $x_0 = m_0 + \epsilon_0$ with zero-mean finite-variance noise ϵ_0.

Consistent with the synthetic control literature, we assume the entries of M are generated by a latent variable model, i.e., $M_{i,t} = g(\theta_i, \rho_t)$ where θ_i and ρ_t are finite-dimensional latent vectors (Ben-Michael et al., 2021; Arkhangelsky

et al., 2021; Abadie, 2021; Amjad et al., 2018, 2019; Athey et al., 2021). Since we focus on placebo studies, we assume this holds for the target unit as well: $m_{0,t} = g(\theta_0, \rho_t)\ \forall t \in [T]$.[1] We assume g is L-bilipschitz continuous, so that cluster structure in the latent variables is recoverable by our algorithm (see Section 4.2.1). Then, M is known to be well-approximated by a low-rank matrix (Chatterjee, 2015) with $\mathrm{rank}(M) = O(\log T)$ (Udell and Townsend, 2019). We denote $\mathrm{rank}(M) = r$ and assume $r < T$. Finally, we assume that there exists a vector f^* with $\|f^*\| \leq \mu$ for some $\mu > 0$, such that $M_{0,t} = M_{1:n,t}^\top f^*$.

3.2 Existence of Subgroups

Our motivation comes from the idea that the donors may have some relevant groups or cluster structure, and the target unit belongs to one of these clusters. We formalize this by assuming a centroid-based (specifically, k-means based) separation structure on the row latent variables $\Theta = \{\theta_i : i \in [n]\}$. Let $P = (P_j)_{j\in[k]}$ be a k-partition of Θ (i.e., $P_1 \sqcup ... \sqcup P_k = \Theta$), with induced centers $\{\overline{P}_{j\in[k]}\} = \{\frac{1}{|P_j|}\sum_{\theta\in P_j}\theta\}_{j\in[k]}$. Then, we define the k-means cost $\Delta_k^2(\Theta; P) = \sum_{i=1}^n \min_{j\in[k]} \|\theta_i - \overline{P_j}\|^2$, and denote $\Delta_k^2(\Theta) = \min_{P'} \Delta_k^2(\Theta; P')$ as the optimal cost. We assume the following ϵ-separation condition on Θ from Ostrovsky et al. (2013): for some k and $\epsilon \in (0, 1)$:

$$\Delta_k(\Theta) \leq \epsilon \Delta_{k-1}(\Theta), \tag{1}$$

which captures the idea that k clusters fit the data significantly better than $k-1$ clusters (in the spirit of the "elbow method" heuristic). For example, this condition would be satisfied if Θ were generated by a sufficiently separated mixture of k distributions.

This modeling assumption on the existence of subgroups provides us with a formal setup to show the ability of our algorithm to approximate the clusters in the latent variable space.

4 clusterSC Algorithm

In this section, we present clusterSC (Algorithm 4), which uses Algorithm 3 as a pre-processing step to partition the donor pool and find the most appropriate subset of donors for a given target.

4.1 Intuition and Algorithms

The basic intuition behind our algorithm is that more donor units corresponds to higher-dimensional inputs in the linear regression step of synthetic control, which in turn leads to higher-dimensional noise and more instability. Therefore, we want to restrict to only the most relevant donors (via clustering) and thereby lower the dimension of the regression. We accomplish this in a two-step approach:

- Algorithm 3 is a clustering step that partitions the donor units using k-means clustering. This enables the donor selection algorithm to later identify the correct donor cluster for the target unit.
- With the identified clusters, Algorithm 4 finds a matching subgroup for a given target. This subgroup specialization leads to a more accurate SC predictions with lower computational cost.

Algorithm 3 uses the assumption that the signal matrix M is low-rank with rank r. It first performs a singular value decomposition (SVD) $M = U\Sigma V^\top = \sum_{i=1}^r s_i u_i v_i^\top$, where the v_i's represent the r basis row vectors; that is, any rows in M can be expressed as a linear combination of the v_i's. Define $\tilde{U} = U\Sigma$. Then, for the j-th *row* of $\tilde{U}$, $\tilde{U}_{j,i}$ can be interpreted as how many v_i's are used to describe the row M_j.

What if there are two groups with very different distributions of U_j's? Figure 1 visualizes this difference with the white dashed circle and the black dashed circle: for each column of $\tilde{U}$, we expect the distribution to be different across groups but similar within each group. Performing a clustering algorithm (e.g., k-means in our case) on $\tilde{U}$, we are separating units based on the use of right singular vectors v_i.

Algorithm 4 incorporates Algorithm 3 as a pre-processing step on the donor pool. The second step of Algorithm 4 computes $\tilde{u}$, a counterpart of $\tilde{U}$ for the target, and finds the matching cluster for the target. Note that we only use pre-intervention data for this step, based on the original use-case of SC, where post-intervention data is assumed to be missing for the target unit. A donor matrix A is constructed using data from the selected cluster. For the the core synthetic control algorithm within Algorithm 4, we adopt Algorithm 2 since de-noising with HSVT before the regression has been shown to be robust to noise (Amjad et al., 2018), although other preferred methods could be used in

[1] In the case of an intervention, this identity would not necessarily hold for $t > T_0$.

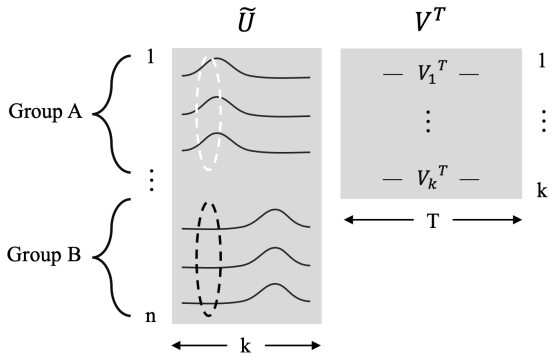

Figure 1: Visualization of the distribution of rows in $\tilde{U}$. Each row $\tilde{U}_i$ can be interpreted as an embedding of the unit i representing the composition of right singular vectors in that row.

Algorithm 3: Clustering Algorithm $\mathcal{C}(X; r)$

Input: Donor matrix X, approximate rank r
1. Perform SVD
$X = U\Sigma V^\top$ (If desired, replace singular values in Σ with zero, except for the top r values.)
$\tilde{U} = U\Sigma$
2. Perform K-means $n^{O(1)}$ steps of Lloyd's method on the rows of $\tilde{U}$.
3. Return cluster centers and V

this step instead. Note that the SC step takes the selected sub-group donor matrix A as an input, instead of the whole donor set X.

Algorithm 4: clusterSC $(X, x^-; r)$

Input: Donor matrix X, Pre-intervention target data x^-, approximate rank r
1. Learn clusters
$c_1, \cdots, c_k, V \leftarrow \mathcal{C}(X; r)$ (Algorithm 3, c_t are cluster centers)
2. Find target's matching donor cluster t
$\tilde{u} = V^{-\top} x^-$
$t = \arg\min_t \|c_t - \tilde{u}\|_2$ (t is target's cluster label)
3. Construct donor matrix A
$A = X_{C_t}$ (C_t is the set of units in cluster t)
4. Perform SC $\hat{f} \leftarrow \mathcal{M}(A, x^-; r)$ (Algorithm 2)
return $\hat{f}$ (SC weights)

4.2 Theoretical Guarantees

In this section, we provide theoretical guarantees on the performance of Algorithm 4 in terms of identifying subgroups (Section 4.2.1), the impact of using A instead of X (Section 4.2.2), and the RMSE of SC (Section 4.2.3). All omitted proofs are in Appendix C.

Following the notation from Section 3.1, let $X = M + E_M$ be a $n \times T$ donor pool matrix and $A = S + E_S$ be a sub-matrix constructed by taking n_A rows of X. Let the low-rank signal matrices have $\text{rank}(M) = r$ and $\text{rank}(S) = r_S$. Then, we say the approximate-rank of X is r, and we define the $(r+1)$-th singular value of the noisy matrix X as $\sigma_X^* = \sigma_{r+1}(X)$.

The pre-intervention error of the standard robust synthetic control algorithm is given by: $\text{MSE}(\hat{m}^-; X) = \mathbb{E}[\frac{1}{T_0}\|m^- - \hat{M}^{-\top}\hat{f}\|^2]$, where $\hat{f} \leftarrow \mathcal{M}(X, x^-)$. The post-intervention error is the same quantity as in (2) with $-$ replaced by $+$ and T_0 replaced by $T - T_0$. RMSE is defined by taking a squared root inside the expectation. We are interested in the change in MSE (or RMSE) when replacing $X \in \mathbb{R}^{n \times T}$ with its subset $A \in \mathbb{R}^{n_A \times T}$.

4.2.1 Accuracy of Subgroup Identification

First we show that existing subgroups in Θ-space are well-approximated by Algorithm 3. Consider the SVD of the true signal matrix $M = U\Sigma V^\top = \tilde{U}V^\top$ (here, $\tilde{U}$ denotes left singular vectors multiplied by corresponding singular values of M; in all other sections, $\tilde{U}$ denotes the noisy version obtained by decomposing X).

Lemma 4.1. *$\tilde{U}$ can be expressed as $\tilde{U} = h(\Theta)$ for some L-bilipschitz function h.*

Using the invariance of bilipschitz mapping, we show that Algorithm 3 identifies the cluster structure of the Θ, misclassifying all but a small fraction of points. To show this, we first require the following lemma.

Lemma 4.2. *If h is L-bilipschitz, then $(1/L)\Delta_k(\Theta) \leq \Delta_k(h(\Theta)) \leq L\Delta_k(\Theta)$, and for all k-partitions P, $(1/L)\Delta_k(\Theta; P) \leq \Delta_k(h(\Theta); P) \leq L\Delta_k(\Theta; P)$.*

Then, we can combine guarantees from Kumar and Kannan (2010) and Ostrovsky et al. (2013) to show that Algorithm 3 (effectively, Lloyd's method on the left singular vectors) detects the planted cluster structure in poly-time.

Theorem 4.3. *Take $\epsilon \leq (400L^4 + 401)^{-1/2}$. If $\Theta = \{\theta_i : i \in [n]\}$ satisfies $\Delta_k(\Theta) \leq \epsilon\Delta_{k-1}(\Theta)$ with optimal k-partition P^*, then Algorithm 3 outputs a partition that matches P^* for all but $O((k^2 + \epsilon^2)n)$ of the points.*

Proof sketch. Observe that $h(\Theta)$ is (ϵL^2)-separated by Lemma 4.2. Hence, by a guarantee from Kumar and Kannan (2010), Algorithm 3 can well-approximate the k-means optimal partition on $h(\Theta)$, call it P', with only a $(k^2\epsilon L^2)$-fraction of points misclassified. From there, it suffices to show P' and P differ on at most $O(\epsilon^2)$-fraction of points. To do this, we invoke Theorem 5.1 of (Ostrovsky et al., 2013), which restricts us to choosing $\epsilon = O(1/L^2)$. □

Theorem 4.3 implies that Algorithm 3 can well-approximate the subgroups in Θ using $\tilde{U}$.

4.2.2 Effects of Subgroup Specialization

Next, we analyze the effect of performing Algorithm 2 on a selected donor pool A (i.e., $\mathcal{M}(A, x^-; r_A)$) instead of the entire donor set X (i.e., $\mathcal{M}(X, x^-; r)$). With $k \geq 2$, we expect the following changes:

1. The number of donor units $n_A < n$.
2. The rank of the signal matrix $r_S \leq r$.
3. The largest singular value suppressed in the HSVT step $\sigma_A^* < \sigma_X^*$. (Recall $\sigma_X^* = \sigma_{r+1}(X)$)

While the first two are trivial, the third one is not. We provide analyses of the gap $\sigma_X^* - \sigma_A^*$ under three different noise settings: Gaussian, sub-Gaussian, and heavy-tailed. Theorem 4.4 presents our first result on the singular values, under Gaussian noise $E_{i,t} \sim \mathcal{N}(0, s^2)$. This result shows that the gap between σ_X^* and σ_A^* will grow with larger scale of noise (larger s).

Theorem 4.4 (Singular Value Concentration with Gaussian Noise)**.** *Let noise terms $E_{i,t} \sim \mathcal{N}(0, s^2)$. If $r < T$ and $n_A < n + 4T - 4\sqrt{nT}$, then*

$$\mathbb{E}[\sigma_X^* - \sigma_A^*] \geq s(\sqrt{n} - \sqrt{n_A} - 2\sqrt{T}).$$

Proof. First, we show that $\mathbb{E}[\sigma_A^*] \leq s(\sqrt{n_A} + \sqrt{T})$.

$$\begin{aligned}\mathbb{E}[\sigma_A^*] &= \mathbb{E}[\sigma_{r_S+1}(S + E_S)] \\ &\leq \sigma_{r_S+1}(S) + \mathbb{E}[\sigma_1(E_S)] \\ &\leq s(\sqrt{n_A} + \sqrt{T}).\end{aligned}$$

The first inequality is from Weyl's inequality on singular values, and $\sigma_{r_S+1}(S) = 0$ by construction. The second inequality is from Gordon's theorem: $\mathbb{E}[\sigma_1(E_s)] \leq s(\sqrt{n_A} + \sqrt{T})$ (Vershynin, 2010).

Next, we show $\sigma_X^* \geq s(\sqrt{n} - \sqrt{T})$ by analyzing the eigenvalues of $X^\top X$:

$$\begin{aligned}\lambda_{r+1}(X^\top X) &= \lambda_{r+1}(M^\top M + 2M^\top E + E^\top E) \\ &\geq \lambda_{r+1}(M^\top M) + \lambda_T(2M^\top E) + \lambda_T(E^\top E) \\ &\geq \lambda_T(E^\top E) \\ \therefore \mathbb{E}[\lambda_{r+1}(X^\top X)] &\geq \left(s(\sqrt{n} - \sqrt{T})\right)^2.\end{aligned}$$

The first inequality is from Weyl's Inequality and $\lambda_{r+1}(M^\top M) = \lambda_T(2M^\top E) = 0$. The second inequality is from Gordon's theorem: $\mathbb{E}[\sigma_T(E)] \geq s(\sqrt{n} - \sqrt{T})$. Finally, we obtain $\mathbb{E}[\sigma_X^*] = \mathbb{E}[\sqrt{\lambda_{r+1}(X^\top X)}] \geq s(\sqrt{n} - \sqrt{T})$.

Combining these two bounds and rearranging terms, we see that the lower bound on σ_X^* is greater than the upper bound on σ_A^* when $n_A < n + 4T - 4\sqrt{nT}$. □

Next we consider the the sub-gaussian noise setting (Definition 4.1), where $\|E_{i,t}\|_{\psi_2} = K$. Our result in this setting, Corollary 4.5, follows from Theorem 4.4 by instantiating Theorem 5.39 from Vershynin (2010), instead of Gordon's.

Definition 4.1 (Sub-gaussian norm)**.** *The sub-gaussian norm of X, denoted by $||X||_{\psi_2}$ is defined as*

$$||X||_{\psi_2} = \sup_{p \geq 1} \frac{1}{\sqrt{p}} \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$

Corollary 4.5 (Singular Value Concentration with Sub-gaussian Noise)**.** *Let the noise terms satisfy $\|E_{i,t}\|_{\psi_2} = K$ For every $t \geq 0$, if $r < T$ and $n_A < \left(\sqrt{n} - CK^2\sqrt{T} - 2t\right)^2$, then with probability at least $1 - 2e^{-ct^2}$,*

$$\sigma_X^* - \sigma_A^* \geq \sqrt{n} - \sqrt{n_A} - CK^2\sqrt{T} - 2t,$$

where $C > 0$ and $c > 0$ are constants, and only $c > 0$ depends on the sub-gaussian norm $K = \|E_{i,t}\|_{\psi_2}$.

Finally, we consider settings where noise comes from a heavy-tailed distribution. This is the most challenging of the three settings considered because the random noise terms will be less concentrated around zero, and thus learning from the noisy data will be more difficult.

Corollary 4.6 (Singular Value Concentration with Heavy-tail Noise)**.** *Let the noise terms $E_{i,t}$ follow a heavy-tailed distribution. If $r < T$ and $n_A < n + 4Tt^2 - 4t\sqrt{nT}$, then for every $t \geq 0$, with probability at least $1 - 2Te^{-ct^2}$,*

$$\mathbb{E}[\sigma_X^* - \sigma_A^*] \geq \sqrt{n} - \sqrt{n_A} - 2t\sqrt{T}.$$

The bound of Theorem 4.4 is used in Section 4.2.3 to argue the improvement in SC performance under Gaussian noise; if needed, one could instead adopt Corollary 4.5 and Corollary 4.6 depending on practical assumptions about noise distributions.

4.2.3 Improvement in SC Performance

Finally, we translate the effect of subgroup specialization into an improvement in the upper bound of the SC prediction error. We compare SC performance using only the selected donors A as an input for $\mathcal{M}$, against performance using the whole donor pool X, and give results for pre-intervention (Theorem 4.8) and post-intervention error (Theorem 4.10).

Let $x_0 = m_0 + e_0$ be the placebo target unit that did not receive the intervention. Then, our goal is to construct an SC prediction $\hat{m}_0$ that approximate m_0 as accurately as possible.

Our first main result in this section is Theorem 4.8, which bounds the improvement in pre-intervention MSE from using the selected donor pool A instead of the full donor pool X. To do this, Lemma 4.7 first gives an upper bound on the pre-intervention MSE of standard synthetic control without clustering (i.e., Algorithm 2).

Lemma 4.7 (Pre-intervention MSE of SC)**.** *Given donor matrix $X \in \mathbb{R}^{n \times T}$, target $x_0 = m_0 + e_0$, rank parameter r, noise distribution $E_{i,t} \sim \mathcal{N}(0, s^2)$, and SC weights SC weights $\hat{f} \leftarrow \mathcal{M}(X, x_0^-; r)$ learned using Algorithm 2, then,*

$$\text{MSE}(\hat{m}_0^-; X) \leq \frac{\mu^2}{T_0}\mathbb{E}[(\sigma_X^* + 2s(\sqrt{n} + \sqrt{T}))^2] + \frac{2s^2 r}{T_0}.$$

Proof. From Lemma 25 of Amjad et al. (2018),

$$\mathbb{E}[\|m_0^- - \hat{m}_0^-\|^2] \leq \mathbb{E}[\|(M^- - \hat{M}^-)^\top f^*\|^2] + 2s^2 r.$$

We bound the first term inside the expectation by

$$\|(M^- - \hat{M}^-)^\top f^*\|^2 \leq \|M^- - \hat{M}^-\|^2 \|f^*\|^2, \tag{2}$$

using the property of the operator norm: $\|Ax\| \leq \|A\| \cdot \|x\|$ for any matrix A and vector x. We bound the first term of (2) by

$$\begin{aligned}\|M^- - \hat{M}^-\| \leq \|M - \hat{M}\| &\leq \sigma_X^* + 2\|X - M\| \\ &\leq \sigma_X^* + 2\|E\|.\end{aligned}$$

Combing these bounds and the assumption $\|f^*\| \leq \mu$, we obtain

$$\text{MSE}(\hat{m}_0^-; X) \leq \frac{1}{T_0}\mathbb{E}\left[(\sigma_X^* + 2\|E\|)^2\right]\mu^2 + \frac{2s^2 r}{T_0}.$$

Using the fact that $\mathbb{E}[\|E\|] \leq s(\sqrt{n} + \sqrt{T})$ completes the proof. □

Combining Lemma 4.7 with the bounds on singular values in Theorem 4.4 in Section 4.2.2 allows us to show that the upper bound on pre-intervention MSE decreases when using selected donor pool A instead of the full donor pool X.

Theorem 4.8. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on pre-intervention MSE of Algorithm 4 is strictly smaller than that of Algorithm 2, and the difference in the upper bounds is $\Omega(s^2 n)$.*

The change in MSE can be seen from the bound stated in Lemma 4.7, which has three elements that change with the donor matrix: σ_X^* to σ_A^*, n to n_A, and r to r_S. We know that all three changes will reduce this upper bound.

Next, we analyze the post-intervention root mean squared error (RMSE), and show similar improvements when changing from X to A (Theorem 4.10). First, Lemma 4.9 gives an upper bound on the post-intervention RMSE of SC without clustering (Algorithm 2), under the standard assumption that the SC weights $\hat{f} \leftarrow \mathcal{M}(X, x_0^-; r)$ have $\|\hat{f}\|_2 \leq \eta$ for some $\eta \geq 0$ (Amjad et al., 2018).

Lemma 4.9 (Post-intervention RMSE of SC)**.** *Given a donor matrix $X \in \mathbb{R}^{n\times T}$, a target x_0, rank parameter r, noise distribution $E_{i,t} \sim \mathcal{N}(0, s^2)$, and SC weights $\hat{f} \leftarrow \mathcal{M}(X, x_0^-; r)$ learned using Algorithm 2,*

$$\begin{aligned}\text{RMSE}(\hat{m}_0^+; X) \leq &\frac{\eta}{\sqrt{T - T_0}}\mathbb{E}[\sigma_X^* + 2s(\sqrt{n} + \sqrt{T})] \\ &+ \sqrt{n}(\mu + \eta).\end{aligned}$$

Proof. We use triangle inequality and the property of induced norm to upper bound the following quantity:

$$\begin{aligned}\|m_0^+ - \hat{m}_0^+\| &= \|(M^+)^\top f^* - (\hat{M}^+)^\top \hat{f}\| \\ &\leq \|(M^+ - \hat{M}^+)^\top \hat{f}\| + \|(M^+)^\top (f^* - \hat{f})\| \\ &\leq \|M^+ - \hat{M}^+\| \cdot \|\hat{f}\| + \|M^+\| \cdot \|f^* - \hat{f}\| \\ &\leq \|M^+ - \hat{M}^+\|\eta + \|M^+\|_F(\|f^*\| + \|\hat{f}\|).\end{aligned}$$

Taking the expectation of both sides and using the fact that $\mathbb{E}[\|M^+ - \hat{M}^+\|] \leq \mathbb{E}[\sigma_X^* + 2\|E\|_2]$ from Lemma B.5 gives,

$$\mathbb{E}[\|m_0^+ - \hat{m}_0^+\|] \leq \mathbb{E}[\sigma_X^* + 2\|E\|_2]\eta + \|M^+\|_F(\mu + \eta).$$

Since $\|M^+\|_F \leq \sqrt{n(T - T_0)}$, we obtain,

$$\text{RMSE}(\hat{m}_0^+; X) \leq \frac{\eta}{\sqrt{T - T_0}}\mathbb{E}[\sigma_X^* + 2\|E\|_2] + \sqrt{n}(\mu + \eta).$$ □

This lemma can be combined with the bound on the difference in singular values from Theorem 4.4 to bound the difference in post-intervention RMSE from using A versus X.

Theorem 4.10. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on post-intervention RMSE of Algorithm 4 is strictly smaller than that of Algorithm 2, and the difference in the upper bounds is $\Omega(s\sqrt{n})$.*

Again, the upper bound stated in Lemma 4.9 has three elements that changes when the donor matrix becomes A instead of X: σ_X^* to σ_A^*, n to n_A, and r to r_S. All three changes reduce the bound, and hence the upper bound on RMSE strictly decreases when using Algorithm 4.

5 Empirical Evaluations

In this section, we empirically demonstrate the performance improvements of our method using simulated datasets. To get a sense of a plausible size for the dataset, we review the literature that applied synthetic control on disaggregate-level datasets. Abadie and L'Hour (2021) adopt synthetic control to measure the effect of participation in a government program on an individual's yearly income. They construct SC instances out of $n = 2490$ individuals as a donor pool

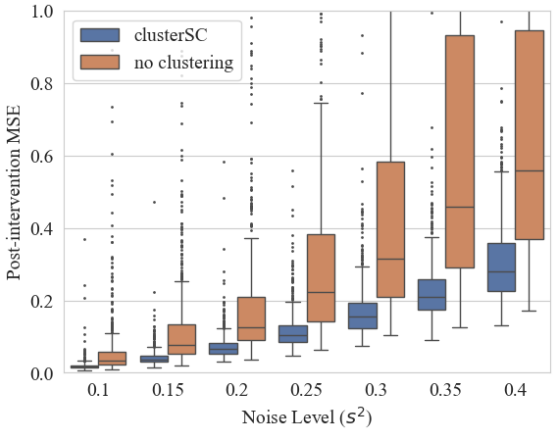

Figure 2: Post-intervention MSE using clusterSC (blue) and using the same SC algorithm without our clustering step (orange), for varying levels of noise.

and with 10 covariates (equivalent to T_0). Robbins et al. (2017) examine the effect of a crime intervention on crime levels measured at the census block level. With 3535 donor units, SC was constructed with $T_0 = 12$ pre-intervention time-series measurements along with auxiliary variables. Vagni and Breen (2021) show that having a child reduces womens' earnings by about 45 per cent by constructing SC with 630 women as donor units. T_0 varies depending on the woman's first childbirth year, and was at most 7.

Based on this, we choose $T = 10$ and $n \in \{1000, 2000\}$ with $n_A/n_B = 1$ (even split), and set $T_0 = 8$ in all experiments. We construct a dataset X with two subgroups, $A = S + E_A$ and $B = S' + E_B$, using multiple sinusoidal time series. Let the rank of S be r_S. For each base time series, we sample three parameters—α_i (magnitude), ω_i (frequency), and ϕ_i (delay)—to generate a sine wave signal $v_{i,t} = \alpha_i \sin(2\pi\omega_i t + \phi_i),\ \forall i \in [r_S]$. The observation matrix is then constructed with elements $S_{i,t} = \sum_{i=1}^{k} w_i \cdot v_{i,t}$, where $w_i \sim \text{Unif}([0,1]),\ \forall i \in [r_S]$. Finally, we introduce observation noise, yielding $A_{i,t} = S_{i,t} + E_{i,t}$, where $E_{i,t} \sim \mathcal{N}(0, s^2)$ for varying levels of $s \geq 0$. We repeat the same process with different parameters to construct B, and concatenate the two matrices to make $X = [A^\top, B^\top]^\top$. To simulate the latent variable model, we use $\alpha \sim \text{Beta}(2,2)$, $\omega \sim \text{Unif}(1,3)$, $\phi \sim \mathcal{N}(0,1)$ for A, and $\alpha \sim \text{Beta}(2,5)$, $\omega \sim \text{Unif}(3,6)$, $\phi \sim \mathcal{N}(0,1)$ for B. Observation noise following $\mathcal{N}(0, s^2)$ was added with $s^2 \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. We use the `sklearn` implementation of Lloyd's k-means algorithm with `k-means++` initialization, and automated search for k with silhouette scores.[2]

For each generated dataset, we perform a leave-one-out placebo test by choosing one unit as a target and the rest as a potential donor pool. For each target, we run synthetic control using two methods: i) clusterSC (using a subset of the donors A selected via Algorithm 4) and ii) robust synthetic control with no clustering (using the whole donor pool X). We iterate this process for 500 datasets in each setting.

Figure 2 shows the distribution of post-intervention MSE of the two algorithms, when $n_A = n_B = 500$. The boxplot shows the quartiles of MSE, the whiskers extend to the furthest datapoint within 1.5 times the interquartile range, and the rest are shown as small dots. We observe that clusterSC consistently outperforms no clustering setup, across all noise levels and other choices of n. This aligns with our Theorem 4.10, which promises a tighter error bound.

Next, we define the pair-wise improvement as the difference in post-intervention MSE scores; $I_i := \text{MSE}(\hat{m}_i^+; X) - \text{MSE}(\hat{m}_i^+; A)$. Then, we take median$(I_i)$ as a metric to assess the overall improvement of one dataset. For computational efficiency, we randomly select 30% of the units in A as target units for the leave-one-out test.

Figure 3 shows the median pair-wise improvements (i.e., median(I_i)) for varying levels of noise. We observe that the median improvement is almost always non-negative, meaning that more than half of the individuals benefits from our method in a pair-wise comparison with high probability. The improvement grows as noise increases, especially with large $n = 2000$, corroborating our Theorem 4.10. We provide additional empirical evaluations in Appendix D.

[2] `https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.silhouette_score.html`

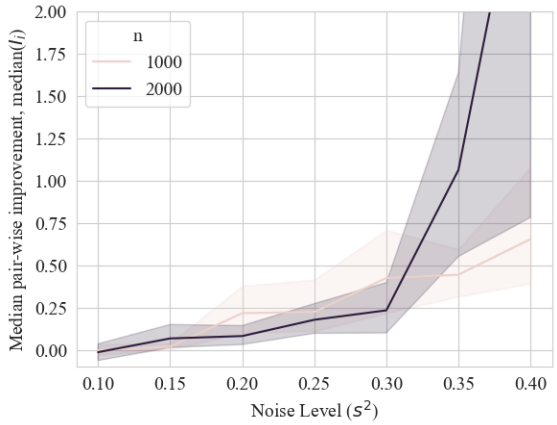

Figure 3: Median of the pair-wise improvement I_i, measured for each dataset, for different noise levels (s^2). Shades represent 95% confidence interval.

6 Discussion and Future Work

This paper presents a novel approach to SC on disaggregate-level datasets, addressing the challenges of higher noise and increased dimensionality by incorporating a principled clustering step. To the best of our knowledge, this is the first method to directly reduce the dimension of regression weights, in contrast to approaches that rely on regularization to suppress the number of active donors. Our approach advances synthetic control methodology, making it better suited for applications where individual-level conditional treatment effects are of interest, such as in drug trials or targeted marketing analysis.

Our algorithm is supported by two main theoretical guarantees. Theorem 4.3 demonstrates the accuracy of our clustering step in identifying intrinsic cluster structure among the donor latent variables Θ. Theorems 4.8 and 4.10 establish a tighter upper bound on prediction error induced by our algorithm, which is empirically validated in Section 5.

Conceptually, clusterSC is similar to LASSO in that it selects a small subset of donors. However, our clustering method offers an alternative solution for cases where LASSO regression may be less desirable, such as when dealing with missing data. While the effect of the clustering step on different SC versions requires further analysis, the core principle of focusing on the singular value concentration remains valid under the common assumption of a latent variable model that produces an approximately low-rank matrix.

Lastly, we acknowledge a potential fairness issue in our approach. As shown empirically in Section 5, our method guarantees improved overall performance of the SC algorithm. However, it does not ensure that the prediction error will decrease for every individual target unit—while the majority of units may benefit, some could experience worse outcomes. This uneven distribution of benefits raises concerns about fairness, especially in individual-level datasets. Investigating the potential disproportionate effects on minority groups presents an avenue for future research.

References

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

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

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

Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, pages 1–18.

Amjad, M., Misra, V., Shah, D., and Shen, D. (2019). mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–27.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, pages 1–15.

Ben-Michael, E., Feller, A., and Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, (just-accepted):1–34.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, pages 247–274.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Card, D. and Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864.

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

Dube, A. and Zipperer, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*. JHU press.

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., and Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health economics*, 25(12):1514–1528.

Kumar, A. and Kannan, R. (2010). Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE.

Nguyen, L. T., Kim, J., and Shim, B. (2019). Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. (2013). The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):1–22.

Rho, S., Cummings, R., and Misra, V. (2023). Differentially private synthetic control. In *International Conference on Artificial Intelligence and Statistics*, pages 1457–1491. PMLR.

Robbins, M. W., Saunders, J., and Kilmer, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *Journal of the American Statistical Association*, 112(517):109–126.

Roughgarden, T. and Valiant, G. (2015). Cs168: the modern algorithmic toolbox lecture 9: the singular value decomposition (svd) and low-rank matrix approximations. *Online, http://theory. stanford.edu/tim/s15/l/l9.pdf*.

Thorlund, K., Dron, L., Park, J. J., and Mills, E. J. (2020). Synthetic and external controls in clinical trials–a primer for researchers. *Clinical epidemiology*, pages 457–467.

Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160.

Vagni, G. and Breen, R. (2021). Earnings and income penalties for motherhood: estimates for british women using the individual synthetic control method. *European Sociological Review*, 37(5):834–848.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

A Technical Definitions

Throughout the paper, we use lower-case letters to denote a vector x and upper-case letters to denote a matrix X. The norms $\|X\|$ and $\|x\|$ refer to the spectral norm and ℓ_2 norm, respectively.

In this section, we summarize important definitions used in our paper. (Some are repeated in the main part too.)

Definition A.1 (Sub-gaussian norm)**.** *The sub-gaussian norm of X, denoted by $||X||_{\psi_2}$ is defined as*

$$||X||_{\psi_2} = \sup_{p \geq 1} \frac{1}{\sqrt{p}} \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$

Definition A.2 (Bilipschitz continuity)**.** *Let $(X, d), (Y, \rho)$ be metric spaces. A map $g : (X, d) \mapsto (Y, d')$ is L-bilipschitz, for $L > 0$, if, for all $x, x' \in X$*

$$\frac{1}{L}\, d(x, x') \leq \rho(g(x), g(x')) \leq L\, d(x, x')$$

B Useful Theorems and Lemmas

B.1 Related to Theorem 4.4

First of all, we introduce two version sof Weyl's inequality used in the proof of Theorem 4.4.

Theorem B.1 (Weyl's Inequality on Singular Values)**.** *For matrices A and B in $\mathbb{R}^{n\times m}$, let $k = \min(n, m)$. Then, the following holds for all $i, j \in [k]$, $i + j - 1 \leq k$.*

$$\sigma_{i+j-1}(A + B) \leq \sigma_i(A) + \sigma_j(B)$$

Theorem B.2 (Weyl's Inequality on Eigenvalues)**.** *For square matrices A and B in $\mathbb{R}^{n\times n}$, the following holds for all $i, j \in [n]$, $i + j - 1 \leq k$*

$$\lambda_{i+j-1}(A + B) \leq \lambda_i(A) + \lambda_j(B)$$

And for all $i \in [n]$,

$$\lambda_i(A) + \lambda_n(B) \leq \lambda_i(A + B) \leq \lambda_i(A) + \lambda_1(B).$$

Next, we introduce Gordon's theorem that bounds the singular values of Gaussian matrices, also used in the proof of Theorem 4.4.

Theorem B.3 (Gordon's theorem for Gaussian matrices)**.** *Let A be an $N \times n$ matrix whose entries are independent standard normal random variables. Then*

$$\sqrt{N} - \sqrt{n} \leq \mathbb{E}[\sigma_{min}(A)] \leq \mathbb{E}[\sigma_{max}(A)] \leq \sqrt{N} + \sqrt{n}.$$

B.2 Related to Theorem 4.8 and Theorem 4.10

In this section, we introduce theorems used in the proof of Theorem 4.8 and Theorem 4.10. The first one is from Chatterjee (2015), and the proof can be found in the cited paper.

Theorem B.4 (Perturbation of Singular Values, Chatterjee (2015))**.** *Let A and B be two $m \times n$ matrices. Let $k = \min\{m, n\}$. Let $\sigma_1(A), \ldots, \sigma_k(A)$ be the singular values of A in decreasing order and repeated by multiplicities. Similarly, we define $\sigma_1(B), \ldots, \sigma_k(B)$ for B and $\sigma_1(A - B), \ldots, \sigma_k(A - B)$ for matrix $A - B$. Then,*

$$\max_{1\leq i\leq k} |\sigma_i(A) - \sigma_i(B)| \leq \max_{1\leq i\leq k} |\sigma_i(A - B)|.$$

Using Theorem B.4, we derive the following lemma. We provide the proof for completeness, but the proof is available in Chatterjee (2015) and Amjad et al. (2018) as well.

Lemma B.5 (Approximation Bound Between Two Matrices, Lemma 20 of Amjad et al. (2018).)**.** *Let A and B be two matrices of the same size. Let $A = \sum_{i=1}^{m} \sigma_i(A) u_i v_i^T$ be the singular value decomposition of A with $\sigma_1(A), \ldots, \sigma_m(A)$ in decreasing order and with repeated multiplicities. For any choice of $\mu \geq 0$, let $S = \{i : \sigma_i \geq \mu\}$. Then, define*

$$\hat{B} = \sum_{i\in S} \sigma_i(A) u_i v_i^T.$$

Let $\sigma_i(B)$ be the singular values of B in decreasing order and repeated by multiplicities, with $\sigma_B^* = \max_{i\notin S} \sigma_i(B)$. Then

$$||\hat{B} - B|| \leq \sigma_B^* + 2||A - B||.$$

Proof. By Theorem B.4, we have that $\sigma_i \leq \sigma_i(B) + ||A - B||$ for all i. Applying triangle inequality, we obtain

$$\begin{aligned}\|\hat{B} - B\| &\leq \|\hat{B} - A\| + \|A - B\| \\ &= \max_{i \notin S} \sigma_i(A) + \|A - B\| \\ &\leq \max_{i \notin S} \Big(\sigma_i(B) + \|A - B\|\Big) + \|A - B\| \\ &= \sigma_B^* + 2\|A - B\|.\end{aligned}$$

□

The following Lemma is Lemma 25 of Amjad et al. (2018). Again, we provide a simplified version of the proof here using our notation for completeness.

Lemma B.6 (Universal Bound on Pre-intervention MSE of OLS, Lemma 25 of Amjad et al. (2018).)**.** *Suppose $x_0^- = m_0^- + \epsilon_0^-$ with $\mathbb{E}[\epsilon_{0,j}] = 0$ and Var$(\epsilon_{0,j}) \leq s^2$ for all $j \in [T_0]$. Let f^* be the true weights assumed in Section 3.1 and $\hat{f}$ be the output of Algorithm 2. Then,*

$$\mathbb{E}\|m_0^- - \hat{m}_0^-\|^2 \leq \mathbb{E}\|(M^- - \hat{M}^-)^\top f^*\|^2 + 2s^2 r, \tag{3}$$

where $r = rank(M)$.

Proof. For easier notation, we define the following:

$$Q := (M^-)^T, \;\; \hat{Q} := (\hat{M}^-)^T.$$

Then, the following is true:

$$m_0^- := Qf^*, \;\; \hat{m}_0^- := \hat{Q}\hat{f}.$$

Recall that for the target row decomposes to $x_0^- = m_0^- + \epsilon_0^-$. $m_0^- = Qf^*$. Since $\hat{f}$ minimizes $\|x_0^- - \hat{Q}f\|$ for any $f \in \mathbb{R}^n$, we have

$$\begin{aligned}\|m_0^- - \hat{m}_0^-\|^2 &= \|(x_0^- - \epsilon_0^-) - \hat{Q}\hat{f}\|^2 \\ &= \|(x_0^- - \hat{Q}\hat{f}) + (-\epsilon_0^-)\|^2 \\ &= \|x_0^- - \hat{Q}\hat{f}\|^2 + \|\epsilon_0^-\|^2 + 2\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle \\ &\leq \|x_0^- - \hat{Q}f^*\|^2 + \|\epsilon_0^-\|^2 + 2\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle \\ &= \|(Qf^* + \epsilon_0^-) - \hat{Q}f^*\|^2 + \|\epsilon_0^-\|^2 + 2\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle \\ &= \|(Q - \hat{Q})f^* + \epsilon_0^-\|^2 + \|\epsilon_0^-\|^2 + 2\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle \\ &= \|(Q - \hat{Q})f^*\|^2 + 2\|\epsilon_0^-\|^2 + 2\langle \epsilon_0^-, (Q - \hat{Q})f^*\rangle + 2\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle.\end{aligned}$$

By taking expectations, we have

$$\mathbb{E}\|\hat{m}_0^- - m_0^-\|^2 \leq \mathbb{E}\|(Q - \hat{Q})f^*\|^2 + 2\mathbb{E}\|\epsilon_0^-\|^2 + 2\mathbb{E}[\langle \epsilon_0^-, (Q - \hat{Q})f^*\rangle] + 2\mathbb{E}[\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle]. \tag{4}$$

We will now deal with the two inner products on the right hand side of equation (4).

$$\begin{aligned}\mathbb{E}[\langle \epsilon_0^-, (Q - \hat{Q})f^*\rangle] &= \mathbb{E}[(\epsilon_0^-)^T]Qf^* - \mathbb{E}[(\epsilon_0^-)^T\hat{Q}]f^* \\ &= -\mathbb{E}[(\epsilon_0^-)^T]\mathbb{E}[\hat{Q}]f^* \\ &= 0.\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] &= \mathbb{E}[\text{tr}((\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-)] \\
&= \mathbb{E}[\text{tr}(\hat{Q}\hat{Q}^\dagger\epsilon_0^-(\epsilon_0^-)^T)] \\
&= \text{tr}\Big(\mathbb{E}[\hat{Q}\hat{Q}^\dagger\epsilon_0^-(\epsilon_0^-)^T]\Big) \\
&= \text{tr}\Big(\mathbb{E}[\hat{Q}\hat{Q}^\dagger]\mathbb{E}[\epsilon_0^-(\epsilon_0^-)^T]\Big) \\
&\leq \text{tr}\Big(\mathbb{E}[\hat{Q}\hat{Q}^\dagger]s^2 I\Big) \\
&= s^2\mathbb{E}[\text{tr}(\hat{Q}\hat{Q}^\dagger)] \\
&\overset{(a)}{=} s^2\mathbb{E}[\text{rank}(\hat{Q})] \\
&\leq s^2 r,
\end{aligned}$$

where (a) follows from the fact that $\hat{Q}\hat{Q}^\dagger$ is a projection matrix with rank r.

For the second inner product, using $\hat{f} = \hat{Q}^\dagger x_0^-$,

$$\begin{aligned}
\mathbb{E}[\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle] &= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{f}] - \mathbb{E}[(\epsilon_0^-)^T x_0^-] \\
&= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger x_0^-] - \mathbb{E}[(\epsilon_0^-)^T]m_0^- - \mathbb{E}[(\epsilon_0^-)^T\epsilon_0^-] \\
&= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger]m_0^- + \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] - \mathbb{E}[(\epsilon_0^-)^T\epsilon_0^-] \\
&\overset{(a)}{=} \mathbb{E}[(\epsilon_0^-)^T]\mathbb{E}[\hat{Q}\hat{Q}^\dagger]m_0^- + \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] - \mathbb{E}[(\epsilon_0^-)^T\epsilon_0^-] \\
&= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] - \mathbb{E}\|\epsilon_0^-\|^2 \\
&\leq s^2 r - \mathbb{E}\|\epsilon_0^-\|^2,
\end{aligned}$$

where (a) follows from the independence of noise.

Finally, we get

$$\begin{aligned}
\mathbb{E}\|\hat{m}_0^- - m_0^-\|^2 &\leq \mathbb{E}\|(Q-\hat{Q})f^*\|^2 + 2\mathbb{E}\|\epsilon_0^-\|^2 + 2(s^2 r - \mathbb{E}\|\epsilon_0^-\|^2) \\
&= \mathbb{E}\|(Q-\hat{Q})f^*\|^2 + 2s^2 r.
\end{aligned}$$

□

C Omitted Proofs

In this section, we provide omitted proofs from Section 4.2.

C.1 Proof of Lemma 4.1

Lemma 4.1. *$\tilde{U}$ can be expressed as $\tilde{U} = h(\Theta)$ for some L-bilipschitz function h.*

Proof. We can write the rows of M as $M_i = g(\theta_i, \rho) = (g(\theta_i, \rho_1), \cdots, g(\theta_i, \rho_T))$ with L-bilipschitz g.

$$\begin{aligned}
\|\theta_i - \theta_j\| &\leq L\|g(\theta_i,\rho) - g(\theta_j,\rho)\| \\
&= L\|M_i - M_j\| \\
&= L\|\tilde{U}_i V^\top - \tilde{U}_j V^\top\| \\
&\leq L\|V^T\| \cdot \|\tilde{U}_i - \tilde{U}_j\|
\end{aligned}$$

For the lower bound, we have

$$\begin{aligned}
\|\theta_i - \theta_j\| &\geq \frac{1}{L}\|M_i - M_j\| \\
\|\theta_i - \theta_j\|\|V\| &\geq \frac{1}{L}\|(\tilde{U}_i - \tilde{U}_j)V^\top\|\|V\| \\
&\geq \frac{1}{L}\|\tilde{U}_i - \tilde{U}_j\|,
\end{aligned}$$

where the last inequality is from the property of induced norm. Finally, since the columns of V are orthogonal, we have $\|V\| = \|V^\top\| = 1$. □

C.2 Proof of Lemma 4.2

Lemma 4.2. *If h is L-bilipschitz, then $(1/L)\Delta_k(\Theta) \leq \Delta_k(h(\Theta)) \leq L\Delta_k(\Theta)$, and for all k-partitions P, $(1/L)\Delta_k(\Theta; P) \leq \Delta_k(h(\Theta); P) \leq L\Delta_k(\Theta; P)$.*

Proof. Let P be the optimal partition for Θ and let P' be the optimal for $h(\Theta)$. Observe

$$\begin{aligned}\Delta_k^2(\Theta) \leq \Delta_k^2(\Theta; P') = \sum_{t=1}^{k} \frac{1}{|P'_t|} \sum_{i,j \in P'_t} \|\theta_j - \theta_j\|^2 &\leq L^2 \sum_{t=1}^{k} \frac{1}{|P'|} \sum_{i,j \in P'_t} ||h(\theta_i) - h(\theta_j)||^2 = L^2 \Delta_k^2(h(\Theta)) \\ &\leq L^2 \sum_{t=1}^{k} \frac{1}{|P_t|} \sum_{i,j \in P_t} ||h(\theta_i) - h(\theta_j)||^2 \\ &\leq L^4 \sum_{t=1}^{k} \frac{1}{|P_t|} \sum_{i,j \in C_t} \|\theta_i - \theta_j\|^2 = L^4 \Delta_k^2(\Theta)\end{aligned}$$

Dividing by L^2 and taking a square root returns claim (i). Claim (ii) follows immediately from the fact that $\frac{\|h(\theta_i) - h(\theta_j)\|}{\|\theta_i - \theta_j\|} \in [\frac{1}{L}, L]$. □

C.3 Proof of Theorem 4.3

In this section, we show the full proof of Theorem 4.3. We restate the theorem below for convenience.

Theorem 4.3. *Take $\epsilon \leq (400L^4 + 401)^{-1/2}$. If $\Theta = \{\theta_i : i \in [n]\}$ satisfies $\Delta_k(\Theta) \leq \epsilon\Delta_{k-1}(\Theta)$ with optimal k-partition P^*, then Algorithm 3 outputs a partition that matches P^* for all but $O((k^2 + \epsilon^2)n)$ of the points.*

Proof. Let $P^* = (P_1^*, ..., P_n^*)$ be the optimal k-means partition for Θ; let P' be the optimal k-means partition for $h(\Theta)$. Let $\Delta(\Theta; P)$ be the k-means cost of a partition P on points of Θ. Observe:

$$\begin{aligned}\Delta_k(h(\Theta); P) &\leq L\Delta_k(\Theta; P) && \text{(Lemma 4.2(ii))} \\ &\leq L\epsilon\Delta_{k-1}(\Theta) && \text{(Definition of } P\text{)} \\ &\leq L^2\epsilon\Delta_{k-1}(h(\Theta)) && \text{(Lemma 4.2(i))}\end{aligned}$$

Since $L^4\epsilon^2 \leq \frac{1-401\epsilon^2}{400}$, by the assumption, we can apply Theorem 5.1 from (Ostrovsky et al., 2013): there exists some matching σ of clusters in P to clusterings in P' such that for each $i \in [k]$, $|P_i \ominus P'_{\sigma(i)}| \leq 161\epsilon^2|P'_{\sigma(k)}|$ (where $\ominus$ denotes symmetric difference). This implies:

$$\#(\text{misclassified}) \leq \sum_{i=1}^{k} |P_i \ominus P'_{\sigma(i)}| \leq \sum_{i=1}^{k} 161\epsilon^2|P'_{\sigma(k)}| = 161\epsilon^2 n$$

By Lemma 4.2, $\Delta_k(h(\Theta)) \leq \epsilon L^2\Delta_{k-1}(h(\Theta))$. In other words, $h(\Theta)$ is (ϵL^2)-separated. In Claim 6.10, (Kumar and Kannan, 2010) show that this condition implies $(1 - \epsilon L^2)$-fraction of the points in $h(\Theta)$ are "good" points in the sense of satisfying a more general separation condition, which they define in their Definition 2.1. Their Theorem 2.2 implies that Lloyd's method on $\{\tilde{U}_i\}$ will yield a partition matching P' on all but $k^2(L^2\epsilon)n$ points. This means the output will deviate from P' on at most $O(k^2L^2\epsilon)n + 161\epsilon^2 n = O((k^2L^2\epsilon + \epsilon^2)n) = O((k^2 + \epsilon^2)n)$ (due to our choice of $\epsilon = O(1/L^2)$) of the points. □

C.4 Proof of Theorem 4.8

Theorem 4.8. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on pre-intervention MSE of Algorithm 4 is strictly smaller than that of Algorithm 2, and the difference in the upper bounds is $\Omega(s^2 n)$.*

Proof. Let $n_A = \alpha^2 n$ for some $\alpha \in (0,1)$. We want to investigate the dependence of the gap between the two upper bounds presented in Lemma 4.7 on n (the number of units) and s^2 (noise). Specifically, we focus on the terms inside the expectation in the upper bound since $\frac{\mu^2}{T_0}$ does not change and $\frac{2s^2 r}{T_0}$ can only decrease.

By expanding the terms inside the expectation, we get

$$\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 = \sigma_X^{*2} + 4s(\sqrt{n}+\sqrt{T})\sigma_X^* + 4s^2(\sqrt{n}+\sqrt{T})^2 \tag{5}$$

$$= \sigma_X^{*2} + 4s(\sqrt{n}+\sqrt{T})\sigma_X^* + 4s^2 n + 8s^2\sqrt{nT} + 4s^2 T. \tag{6}$$

When we change the donor matrix to A instead of X, this changes to

$$\left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2 = \sigma_A^{*2} + 4s(\alpha\sqrt{n}+\sqrt{T})\sigma_A^* + 4s^2\alpha^2 n + 8s^2\alpha\sqrt{nT} + 4s^2 T. \tag{7}$$

Then, the difference between the two becomes

$$\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 - \left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2 \tag{8}$$

$$= \sigma_X^{*2} - \sigma_A^{*2} + 4s(\sqrt{n}+\sqrt{T})(\sigma_X^* - \sigma_A^*) - (1-\alpha)4s\sqrt{n}\sigma_A^* + (1-\alpha^2)4s^2 n + (1-\alpha)8s^2\sqrt{nT} \tag{9}$$

$$= (\sigma_X^* + \sigma_A^* + 4s(\sqrt{n}+\sqrt{T}))(\sigma_X^* - \sigma_A^*) - (1-\alpha)4s\sqrt{n}\sigma_A^* + (1-\alpha^2)4s^2 n + (1-\alpha)8s^2\sqrt{nT} \tag{10}$$

Now we take the expectation and apply Theorem 4.4 to lower bound this quantity

$$\mathbb{E}[(8)] \geq (2\sigma_A^* + s(5-\alpha)\sqrt{n} + 2s\sqrt{T})((1-\alpha)s\sqrt{n} - 2s\sqrt{T}) \tag{11}$$

$$- 4(1-\alpha)s\sqrt{n}\sigma_A^* + 4(1-\alpha^2)s^2 n + 8(1-\alpha)s^2\sqrt{nT} \tag{12}$$

$$= 2(1-\alpha)s\sqrt{n}\sigma_A^* - 4(1-\alpha)s\sqrt{n}\sigma_A^* - 4s\sqrt{T}\sigma_A^* + s^2(5-\alpha)(1-\alpha)n + 2s^2(1-\alpha)\sqrt{nT} \tag{13}$$

$$- 2s^2(5-\alpha)\sqrt{nT} - 4s^2 T + (1-\alpha^2)4s^2 n + (1-\alpha)8s^2\sqrt{nT} \tag{14}$$

$$= -2(1-\alpha)s\sqrt{n}\sigma_A^* - 4s\sqrt{T}\sigma_A^* + s^2((9-6\alpha-3\alpha^2)n - 4T - 8\alpha^2\sqrt{nT}) \tag{15}$$

$$= -s\left(2(1-\alpha)\sqrt{n} + 4\sqrt{T}\right)\sigma_A^* + \underbrace{3(\alpha+3)(1-\alpha)s^2 n}_{\text{dominating term}} - \left(4s^2 T + 8s^2\alpha^2\sqrt{nT}\right). \tag{16}$$

The first and the third terms are negative but they are relatively small numbers compared to the middle one (highlighted as dominating term). Hence, for sufficiently large n, we have Equation $\mathbb{E}[(8)] = \Omega(s^2 n)$.

Finally, The difference in the two upper bounds is

$$\frac{\mu^2}{T_0}\mathbb{E}[\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2] + \frac{2s^2 r}{T_0} - \frac{\mu^2}{T_0}\mathbb{E}[\left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2] - \frac{2s^2 r_S}{T_0} \tag{17}$$

$$= \frac{\mu^2}{T_0}\mathbb{E}[\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 - \left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2] + \frac{2s^2}{T_0}(r - r_S) \tag{18}$$

$$= \Omega(s^2 n), \tag{19}$$

since $n \gg T > T_0$, μ is a constant, and $r - r_S \geq 0$. □

C.5 Proof of Theorem 4.10

Note: there was a typo in the main text where Ω was replaced by O. The correct version is with Ω as presented here.

Theorem 4.10. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on post-intervention RMSE of Algorithm 4 is strictly smaller than that of Algorithm 2, and the difference in the upper bounds is $\Omega(s\sqrt{n})$.*

Proof. Let $n_A = \alpha^2 n$ for some $\alpha \in (0,1)$. Now we want to investigate the gap between post-intervention RMSE upper bounds presented in Lemma 4.9. Starting from the upper bound

$$\frac{\eta}{\sqrt{T-T_0}}\mathbb{E}[\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})] + \sqrt{n}(\mu+\eta),$$

we obtain the difference

$$\frac{\eta}{\sqrt{T-T_0}}\mathbb{E}[\sigma_X^* - \sigma_A^* + 2s(1-\alpha)\sqrt{n}] + (1-\alpha)\sqrt{n}(\mu+\eta) \tag{20}$$

$$\geq \frac{\eta}{\sqrt{T-T_0}}\left(s((1-\alpha)\sqrt{n} - 2\sqrt{T}) + 2s(1-\alpha)\sqrt{n}\right) + (1-\alpha)\sqrt{n}(\mu+\eta) \tag{21}$$

$$= \frac{\eta}{\sqrt{T-T_0}}\left(s(3(1-\alpha)\sqrt{n} - 2\sqrt{T})\right) + (1-\alpha)\sqrt{n}(\mu+\eta), \tag{22}$$

where the first inequality comes from Theorem 4.4. Since $n \gg T$ and μ and η are constants, we conclude that the difference is $\Omega(s\sqrt{n})$. □

D More Results on Simulation Datasets

In this section, we show more results from empirical evaluations of our method. Our method can flexibly adopt different versions of synthetic control methods. To be specific, we try different regression methods for step 3 of Algorithm 2. Appendix D.1 shows the performance of clusterSC with OLS and Ridge regression. Appendix D.2 presents an analysis of clusterSC with Lasso regression.

D.1 clusterSC with Robust Synthetic Control (OLS and Ridge regression)

Robust synthetic control (Amjad et al., 2018) first applies de-noising step (HSVT) and then learns weights using OLS or ridge regression. This is simply adopting OLS or ridge regression in step 3 of Algorithm 2. In this section, we show the results comparing the performance of robust synthetic control both with and without the clustering step. We use the same data generating method as in Section 5, with the same parameters $n_A = n_B \in \{500, 1000\}$ and $T = 10$. The number of distinct signals in each sub-matrices A and B are $r_A = r_B = 3$. The experiment was iterated for 100 randomly generated datasets, and the ridge coefficient was 0.01 for all cases (except for OLS where there is no regularizer).

Figure 4 shows the average MSE per dataset over varying noise levels (s^2), using 1) robust synthetic control with OLS (blue), 2) robust synthetic control with ridge (orange), 3) clusterSC with OLS (green), and 4) clusterSC with ridge (red). We observe that ridge regression performs better than OLS with or without the clustering step. With clusterSC, we reduce the expectation of MSE and the variance as well, regardless of the choice of regression method (OLS or ridge).

Figure 5 shows a similar result to Figure 3; the pair-wise improvement induced by the clustering step, for each SC instance over varying noise level. We observe that the improvement is almost always positive, and the magnitude is larger in a high-noise regime. The improvement is more stable (low variance) with higher n, and the improvement in OLS and ridge regression is not too different.

D.2 clusterSC with Lasso regression

In this section, we use Lasso regression for step 3 of Algorithm 2. Again, we use the same data generating method as in Section 5, with the same parameters $n_A = n_B \in \{500, 1000\}$ and $T = 10$. The number of distinct signals in each

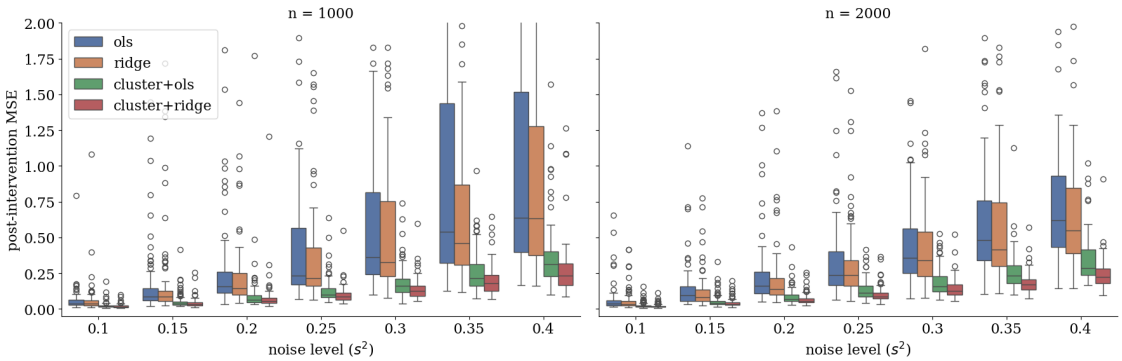

Figure 4: Median post-intervention MSE, measured per dataset. Each boxplot corresponds to ridge, OLS, cluster and then ridge, and cluster and then OLS, from left to right, plotted for each noise level.

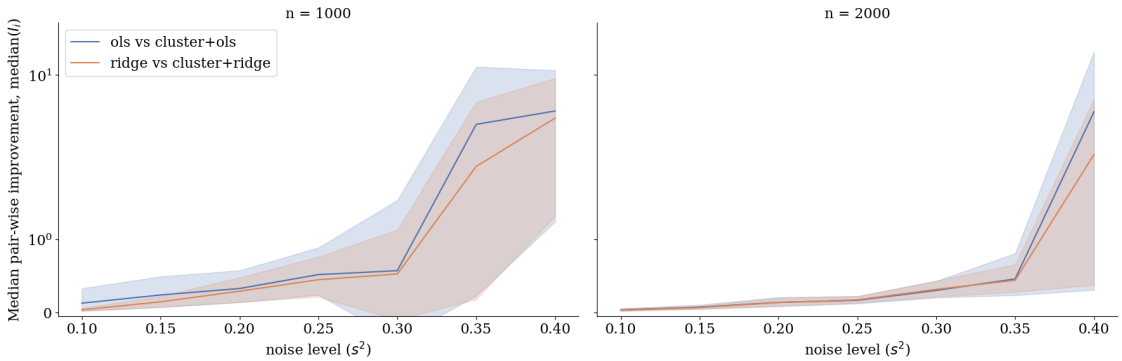

Figure 5: Median of the pair-wise improvements observed when using OLS (blue) and ridge regression (orange), over different noise levels.

sub-matrices A and B are $r_A = r_B = 3$. The experiment was iterated for 100 randomly generated datasets, and the Lasso coefficient was 0.01 for all cases. Due to the high computational cost of Lasso, we only test for noise levels $s^2 \in \{0.1, 0.2, 0.3, 0.4\}$.

Figure 6 shows the average post-intervention MSE. We observe more improvement with clustering as noise level increases. Compared to the results in Figure 4, the improvement induced by the clustering step when regression is performed with Lasso is not as large as compared to OLS or Ridge.

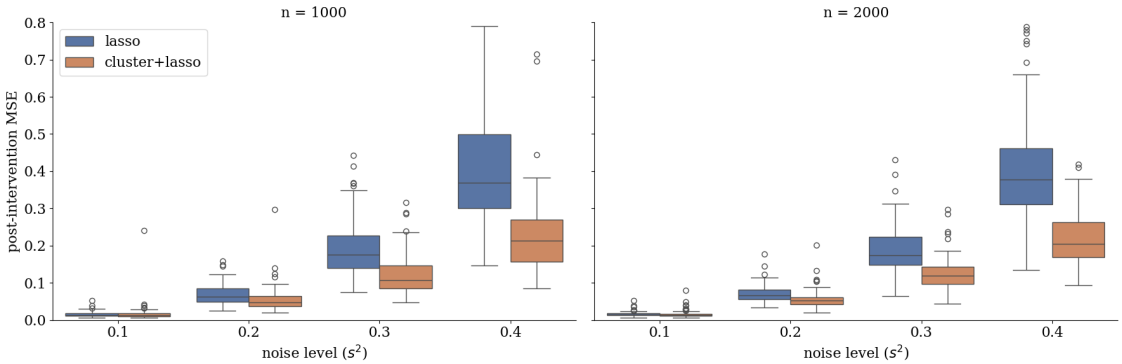

Figure 6: Median post-intervention MSE, measured per dataset. We compare using Lasso with HSVT (blue) and the same setting with cluster step (orange).

To further investigate, we analyze the *active donors* selected by Lasso regression, which correspond to donor units with non-zero SC weights. Since Lasso produces a sparse vector, it effectively selects a subset of relevant donors to reconstruct the target unit. In our experimental setup, units in group A share the same signals, and all target units are sampled from A. Ideally, the relevant donors should be chosen from A rather than B. To quantify this, we use precision scores to assess the proportion of selected donors that belong to group A.

Figure 7 presents the precision scores of the active donor units selected by Lasso regression, i.e., the fraction of active donors that correctly belong to the true group A. The precision scores for Lasso (top row) are right skewed with a high concentration at 1, meaning that Lasso by itself can accurately select a subset of relevant donors. This naturally leads to a lesser degree of improvement when applying our clustering step on top of Lasso, since Lasso captures part of the performance gain from clustering. The bottom row, which shows clusterSC with Lasso regression, is even more right skewed and concentrate at a precision score of 1. Additionally, we observe that although the distributions for both become less right skewed as noise increases, lasso with clustering better maintains its skew. This explains why the MSE improvement in Figure 6 increases for higher noise levels.

We can further analyze this improvement by investigating how effectively the clustering step (k-means) selects a good donor set by computing its F1 and precision scores with units in group A (from Step 4 of Algorithm 4). Figure 8 presents histograms of F1 scores (top row) and precision scores (bottom row) for varying noise levels. A precision score close to 1 indicates that most of the selected donors from the clustering step are already from the relevant group

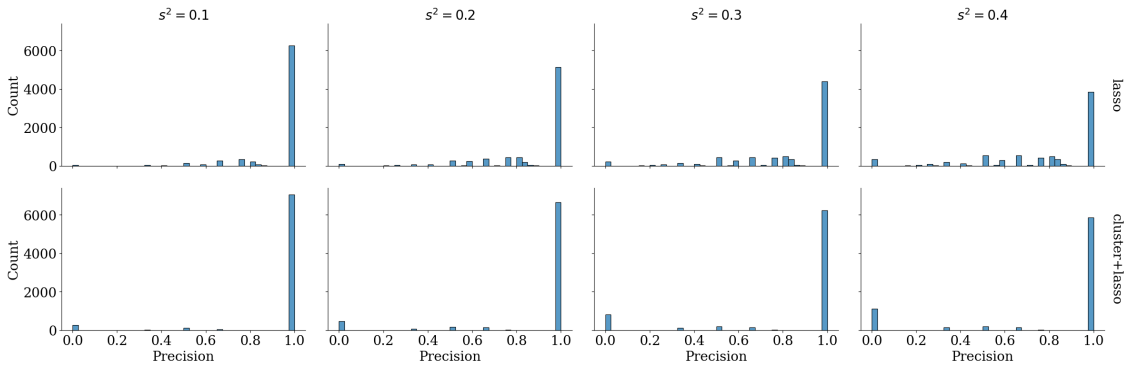

Figure 7: Histogram of the precision score of active donor units from using Lasso regression with HSVT. The first row is without clustering, and the second row is with the clusering step (our method, clusterSC with Lasso regression).

A, making it easier for Lasso to select the best fit among them. This allows the regression to more easily pick up the *relevant donors*. In contrast, without clustering, Lasso must filter out irrelevant donors from group B solely through the regression step.

Also in Figure 8, we observe that both F1 and precision decrease as noise increases, and the distributions are bimodal with increasingly larger left mode as noise increases, similar to Figure 7. The left mode suggests that our clustering algorithm is assigning a new target unit to the wrong cluster for some fraction (Step 2 of Algorithm 4), indicating that increased noise leads to lowered generalization capabilities of k-means. This agrees with the overall higher MSE scores for higher noise levels.

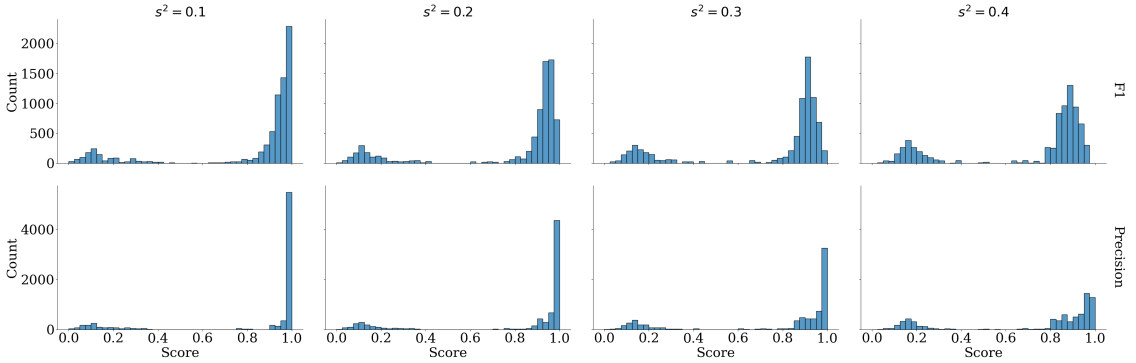

Figure 8: Histograms of Precision scores of donor units selected by our clustering step compared to units in A, over varying noise levels.

D.2.1 Runtime analysis

Like Lasso, the cluster pre-processing step seeks to isolate only the most important donors for target reconstruction. Unlike Lasso, the cluster pre-processing step avoids running a linear regression in the full n dimensions. This allows for significant runtime improvements when Lasso regression was used in step 4 of Algorithm 2, which we explore in this section.

The asymptotic runtime analysis of our algorithm decomposes into three main bottlenecks: (1) SVD on the original matrix, (2) clustering , and (3) regression (on the downsampled data). For each step, we present the following lemmas to show the aymptotic bound on the computational complexity.

Lemma D.1 (Ostrovsky et al. (2013))**.** *Algorithm* 3 *with* $O(2^{O(k/\epsilon)}rT)$ *steps of Lloyd's method outputs a* $(1+\epsilon)$*-approximate optimal* k*-means partitions for* $h(\Theta)$ *(which, per Theorem 4.3, is approximately optimal for* Θ*).*

Lemma D.2 (Golub and Van Loan (2013), Roughgarden and Valiant (2015))**.** *Computing the top* r *singular values and corresponding singular vectors of an* $n \times m$ *matrix takes* $O(nmr)$ *time.*

Lemma D.3 (Efron et al. (2004))**.** *Lasso regression on* v *variables (number of features) with sample size* s *(number of observations) each takes time* $O(v^3 + v^2 s)$.

Note that in synthetic control, the number of features in usual regression setting is actually the number of donors n, as we are predicting the behavior of the target donor per-time-step. Then, we have

- Runtime of clusterSC with Lasso: $O(nTr + 2^{O(k/\epsilon)}rT + n_A^3 + n_A^2T) \asymp n + n_A^3$
- Runtime of HSVT + Lasso without Clustering: $O(nTr + n^3 + n^2T) \asymp n^3$

For disaggregate-level data, we think of n as dominating the rest of the terms (since $n \gg T$, as explained in Section 3.1.), and we think of k as constant or near-constant. Note that if $k = \Theta(\log(\log n))$ and clusters are balanced, i.e. $n_A = n/k$, we see that the cluster algorithm achieves an asymptotic improvement over the pure lasso algorithm, by a factor of $(\log(\log n))^3$. In the case of imbalanced clusters and $k = O(1)$ (and under the assumption that our clustering step is robust to this imbalance[3]), then n_A can be arbitrarily smaller than n, yielding arbitrary runtime improvement.

Note that using the cluster pre-processing with more efficient regression methods like ridge regression—which more easily take advantage of the low-rank assumption—would yield even greater runtime improvements.

[3]Note that clustering condition used in this paper, from (Ostrovsky et al., 2013), is sensitive to cluster size imbalance. More general separation conditions, such as that of (Kumar and Kannan, 2010), are more robust.