ClusterSC: Advancing Synthetic Control with Donor Selection

Saeyoung Rho Columbia University
Andrew Tang Columbia University
Noah Bergam Columbia University
Rachel Cummings Columbia University
Vishal Misra Columbia University

Abstract

In causal inference with observational studies, synthetic control (SC) has emerged as a prominent tool. SC has traditionally been applied to aggregate-level datasets, but more recent work has extended its use to individual-level data. As they contain a greater number of observed units, this shift introduces the curse of dimensionality to SC. To address this, we propose Cluster Synthetic Control (ClusterSC), based on the idea that groups of individuals may exist where behavior aligns internally but diverges between groups. ClusterSC incorporates a clustering step to select only the relevant donors for the target. We provide theoretical guarantees on the improvements induced by ClusterSC, supported by empirical demonstrations on synthetic and real-world datasets. The results indicate that ClusterSC consistently outperforms classical SC approaches.

1 Introduction

Synthetic control (SC) has emerged in the econometrics community as a natural extension of the Difference-in-Differences method (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).

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Recently, there has been increasing attention to employing SC on disaggregated data, observed in contexts like clinical trials with individual health records (Thorlund et al., 2020) and economic analyses using individual income data (Abadie and L'Hour, 2021). In disaggregated 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 SC 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 SC, 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 ClusterSC, a novel approach to disaggregated 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 and real-world datasets, demonstrating the improved prediction accuracy achieved by our method.

Section 2 introduces the SC family of methods and defines relevant notation. 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 5. Finally, Section 6 empirically evaluates the performance of our approach on synthetic and real-world datasets.

2 Synthetic Control (SC) Methods

First, we introduce some key notation. We denote a target unit as a vector x (usually indexed 0) and a donor pool as a matrix $X \in \mathbb{R}^{n \times T}$ with n donor units (index ranging from 1 to n) and T observations. Given a matrix X, let X_i be the i-th row , $X_{i:j}$ be the submatrix constructed by choosing the rows between i-th and j-th rows, and $X_{i,t}$ be the element in the i-th row and the t-th column of X. When donors are represented as a set of points, we use x_i to denote the point corresponding to the i-th donor unit. 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 $X^+ \in \mathbb{R}^{n \times (T-T_0)}$, and 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. We use $\|\cdot\|$ to denote the spectral norm for a matrix and ℓ_2 norm for a vector.

SC Family of Methods. Imagine that a new property tax policy was implemented in New York, but not in other states in the US. The time series data would include T observations of quarterly housing price index $x_{i,t} \in \mathbb{R}$ for all cities (units) $i \in V$ and for all time points $t \in [T]$. At time $T_0 < T$, only the cities in New York (treated units) $W \subset V$ adopt a new policy (intervention), while other cities outside of New York are not affected (control units, potential donors). Hence, for each treated unit $i \in W$, we have a pre-intervention time series $x_i^- \in \mathbb{R}^{T_0}$ without intervention and post-intervention time series $x_i^+ \in \mathbb{R}^{T-T_0}$ under the new policy. For a 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.

SC 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 ($i \in W$) 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 SC family of methods.

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 SC, 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 SC 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).

Algorithm 1: SC 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 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 disaggregated datasets.

Algorithm 2: Learn Step of SC, $\mathcal{M}(X, x^-; r)$

Input: donor data X, pre-intervention target data x^-, 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 $\hat{M} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top := HSVT(X; r)$.
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 data matrix. The pre-intervention donor matrix X^- is the regressor and the pre-intervention target time series x_0^- is the regressand, so the j-th element of the weight vector 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 Disaggregated Data When applying SC to disaggregated 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 is usually done for aggregate-level datasets (Abadie and Gardeazabal, 2003; Abadie et al., 2015). However, with disaggregated 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 (i.e., learning sparse weights). 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 disaggregated 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.

To 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. To more easily articulate the accuracy of SC methods, we focus on these placebo studies in the remainder of the paper.

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 disaggregated data, we assume $n \gg T$ (i.e., X is a tall matrix). We assume that the observation X is a noisy version of the true signal. That is, $X = M + E$, where $M_{i,t}$ is the true (deterministic) signal with entries bounded $-1 \le M_{i,t} \le 1$, 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 SC 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 5.1). Then, M is known to be well-approximated by a low-rank matrix (Chatterjee, 2015) with $\text{rank}(M) = O(\log T)$ (Udell and Townsend, 2019). We denote $\text{rank}(M) = r$ and assume $r < T$. Finally, we assume that there exists a vector f^* with $\|f^*\| \le \mu$ for some $\mu > 0$, satisfying $M_{0,t} = M_{1:n,t}^\top f^*$.

3.2 Existence of Subgroups

Our motivation comes from the idea that the donors may have some relevant subgroups in the population or underlying cluster structure, and the target unit belongs to one of these clusters. We formalize this by assuming a centroid-based separation structure on the row latent variables $\Theta = \{\theta_i : i \in [n]\}$. Let $P = \{P_j\}_{j\in[k]}$ be a Voronoi partition of Θ (i.e., $P_1 \sqcup ... \sqcup P_k = \Theta$), with induced centers $\{c_j\}_{j\in[k]} = \{\frac{1}{|P_j|}\sum_{i\in P_j}\theta_i\}_{j\in[k]}$. Then, we define the k-means cost $\Delta_k^2(\Theta; P) = \sum_{i=1}^n \min_{j\in[k]} \|\theta_i - c_j\|^2$, which captures the average distance from the cluster center to members of the cluster. We denote $\Delta_k^2(X) = \min_P \Delta_k^2(X; P)$ as the cost of an optimal k-means solution of input X. Finally, we assume the ε-separation condition on Θ, as in Definition 3.1.

Definition 3.1 (ε-separation)**.** *We say that $\Theta = (\theta_1, ..., \theta_n) \subset \mathbb{R}^d$ is ε-separated with k clusters if for some integer $k \ge 2$ and $\varepsilon \in (0, 1)$,*

$$\Delta_k^2(\Theta) \le \varepsilon^2 \Delta_{k-1}^2(\Theta). \tag{1}$$

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

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

Input: Donor matrix X, approximate rank r
1. Perform SVD
$X = U\Sigma V^\top$
$\tilde{M} = U\Sigma_r V^\top := HSVT_r(X)$ # Hard Singlular Value Thresholding
$\tilde{U} = U\Sigma_r$ # Features used for clustering
2. Perform k-means clustering $n^{O(1)}$ steps of Lloyd's method on the rows of $\tilde{U}$.
3. Return cluster centers and V

This 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 well-separated mixture of k distributions. This modeling assumption on the existence of subgroups provides a formal setup to show the ability of ClusterSC to approximate the clusters in the latent variable space.

4 Cluster Synthetic Control

In this section, we present Cluster Synthetic Control (ClusterSC, Algorithm 4), which integrates a donor selection step into the SC framework. The clustering subroutine is designed to identify structural patterns within the donor pool, ensuring that units within the same cluster exhibit similar behavior while differing across clusters. Given a target unit, the algorithm selects the most relevant cluster, after which SC is applied using only the chosen donors.

The core intuition behind our ClusterSC algorithm is that using more donor units corresponds to higher-dimensional inputs in the linear regression step of SC, 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:

1. Algorithm 3 is a clustering subroutine that partitions the donor units using k-means clustering.
2. Algorithm 4 finds a matching subgroup for a given target and performs SC using only this subgroup as donor. This donor selection leads to a more accurate predictions with lower computational cost.

Algorithm 3 uses the assumption that the signal matrix M is low-rank with rank r, hence the noisy version X is approximately low-rank. It first performs a singular value decomposition (SVD) $X = U\Sigma V^\top = \sum_{i=1}^{r} s_i u_i v_i^\top$, where the v_i's represent the r basis row vectors. Define $\tilde{U} = U\Sigma$. Then, for the j-th *row* of $\tilde{U}$, $\tilde{U}_{j,i}$ can be interpreted as the magnitude of basis vectors v_i that are used to describe the row X_j.

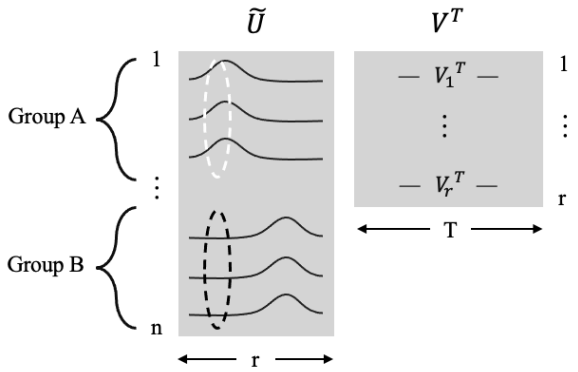

Figure 1: Visualization of the distribution of rows in $\tilde{U}$ with two different subgroups in the donor units. Each row $\tilde{U}_i$ can be interpreted as an embedding of the unit i, representing the composition of right singular vectors for that row.

Now, we pose a question to our readers. What if there are two groups with sufficiently different distributions of U_j's? Figure 1 visualizes this difference with the dashed circles in white and black, illustrating the differences in groups A and B in a single column: 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.

ClusterSC (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. A donor matrix A is then constructed using data from the selected cluster. Note that only pre-intervention data are used for these steps, based on the motivating use-case of SC, where post-intervention data are assumed to be missing for the target unit.

For the Learn step of SC within Algorithm 4, we adopt Algorithm 2 since de-noising with HSVT before the regression has shown to be more robust to noise (Amjad et al., 2018), although other preferred methods could be used in this step instead. This subroutine takes the selected donor matrix A as an input, instead of the whole donor set X.

5 Theoretical Guarantees

In this section, we provide theoretical guarantees for the performance of ClusterSC (Algorithm 4) by showing the accuracy of identifying subgroups (Section 5.1) and the impact of donor selection via ClusterSC on the

Algorithm 4: ClusterSC $(X, x_0; r)$

Input: Donor matrix X, target data x_0, 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
$\tilde{u} = V^{-\top} x_0^-$
$t = \arg\min_{t'} \|c_{t'} - \tilde{u}\|_2$ (t is target's cluster label)
3. Construct donor matrix A **and denoise**
$A = X_{C_t}$ (C_t is the set of units in cluster t)
$\hat{M}_{C_t} = HSVT(A; r)$ (Denoise selected donor)
4. SC: Learn $\hat{f} \leftarrow \mathcal{M}(A, x_0^-; r)$ (Algorithm 2)
5. SC: Project $\hat{m}_0^+ \leftarrow \hat{f}(\hat{M}_{C_t}^+)$
6. SC: Infer the estimated causal effect of the intervention for the target is $x_0^+ - \hat{m}_0^+$

prediction accuracy (Section 5.2).

Following the notation introduced in 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 based on the ClusterSC output. 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 a matrix X as $\sigma_X^* = \sigma_{r+1}(X)$.

The pre-intervention mean squared error (MSE) of SC estimator is given by: $\text{MSE}(\hat{m}^-; X) = \mathbb{E}[\frac{1}{T_0}\|m^- - \hat{M}^{-\top}\hat{f}\|^2]$, where $\hat{f} \leftarrow \mathcal{M}(X, x_0^-)$. Likewise, the post-intervention error is $\text{MSE}(\hat{m}^+; X) = \mathbb{E}[\frac{1}{T-T_0}\|m^+ - \hat{M}^{+\top}\hat{f}\|^2]$. RMSE is defined by taking a squared root inside the expectation of either expression. We are interested in the change in MSE (or RMSE) when replacing $X \in \mathbb{R}^{n \times T}$ with its subset $A \in \mathbb{R}^{n_A \times T}$.

5.1 Accuracy of Subgroup Identification

We show that existing subgroups in Θ-space are well-approximated by Algorithm 3. To do so, we first show that this structure is well-preserved in the signal M via L-bilipschitz mapping g. Then, we show $\tilde{M}$ — hard singular value thresholding (HSVT) applied to $X = M + E$ — is close to the signal matrix M. Finally, we show that clustering with $\tilde{U}$ features as in Algorithm 3 can well-approximate clusters in $\tilde{M}$. All omitted proofs from this section are presented in Appendix D. For our key lemmas used in the accuracy analysis (Lemmas 5.3, 5.6, and 5.7), proof sketches showing the intuition behind the arguments are also presented in Appendix C.

Some new notation is needed to discuss the clustering results. To measure the accuracy of approximation, we say partition P^A is ϵ-approximated by partition P^B if the two partitions agree with each other for all but ϵ fraction of the points. We use $A \ominus B = (A \setminus B) \cup (B \setminus A)$ to denote the symmetric difference between sets A and B. For a set of points $A = a_1, \ldots, a_n$, we define the k-mean optimal cluster centers $C^A = \{c_i^A\}_{i=1}^k$ and the induced Voronoi partition $P^A = \{P_i^A\}_{i=1}^k$. The optimal k-means objective is defined as $\Delta_k^2(A) = \sum_{i\in[n]} \min_{j\in[k]} ||a_i - c_j^A||^2 = \sum_{l\in[k]} \frac{1}{2|P_l^A|} \sum_{i,j\in P_l^A} ||a_i - a_j||^2$. When a partition $\hat{P}$ is specified, the optimal k-means objective can be written as $\Delta_k^2(A; \hat{P}) = \sum_{l\in[k]} \frac{1}{2|\hat{P}_l|} \sum_{i,j\in\hat{P}_l} ||a_i - a_j||^2$, with the centers recalculated as a mean of each partition.

The main result of this subsection is Theorem 5.8, which combines all three main steps (Lemmas 5.3, 5.6, and 5.7) to show that Algorithm 3 well-approximates the optimal k-means partition P^Θ. Each lemma bounds the symmetric difference between the partitions, and we can guarantee that the initial partition P^Θ and the algorithmic output will disagree by at most half of the sum of the symmetric difference in each step.

From Θ to M. To show $P^\Theta \approx P^M$, we consider P', a Voronoi partition induced by centers $c_i' = \frac{1}{|P_i^\Theta|} \sum_{l\in P_i^\Theta} m_l$, as an intermediate step. To show $P^\Theta \approx P'$, we use two lemmas stating that M is also $L^2\varepsilon$-separated and that with this separation structure, most of the points are close to their centers.

Lemma 5.1. *$M = g(\Theta)$ is $L^2\varepsilon$-separated with k clusters, and $\Delta_k^2(M; P^\Theta) \le L^4\varepsilon^2\Delta_{k-1}^2(M)$.*

Lemma 5.2 (Lemma 4.1 of (Ostrovsky et al., 2013))**.** *Let $r_i^2(\Theta) := \frac{1}{|P_i^\Theta|} \sum_{l\in P_i^\Theta} \|\theta_l - c_i^\Theta\|^2$. If Θ is ε-separated, $r_i^2(\Theta) \le \frac{\varepsilon^2}{1-\varepsilon^2} \min_{j\neq i} \|c_i^\Theta - c_j^\Theta\|^2$.*

Hence, the core of P_i^Θ will always belong to P_i'. Finally, we invoke Theorem B.1 to bound the difference between P' and P^M, and show the final bound between P^Θ and P^M.

Lemma 5.3. *For small $\varepsilon \le 0.1$, if $L^2 \le \min(\frac{1}{\sqrt{801}\varepsilon}, \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}})$, then $\sum_{i=1}^k |P_i^\Theta \ominus P_{\sigma(i)}^M| \le 8L^2\varepsilon n$ for some bijection $\sigma(i)$, where $n = |\Theta|$.*

From M to $\tilde{M}$. Next, we show the difference between the points represented by M and $\tilde{M}$, where $\tilde{M} = HSVT(X; r) = \sum_{i=1}^r \sigma_i u_i v_i^\top$ where σ_i, u_i, and v_i are respectively the i-th singular value, left singular vector, and right singular vector of $X = M + E$. To quantify the difference between M and $\tilde{M}$, we define $\eta := \max_{i\in[n]} \|m_i - \tilde{m}_i\|$. Then, define G as a set of *good* events where the noise is small enough that $\eta \le \frac{2s(\sqrt{n}+\sqrt{T})}{\delta}$ with high probability (at least $1 - \delta$).

Lemma 5.4. *With probability* $1-\delta$, $\eta \le \frac{2s(\sqrt{n}+\sqrt{T})}{\delta}$.

For the remainder of the proof, we only focus on the events in G. Then, Lemma 5.5 shows that a separation structure is preserved in $\tilde{M}$ with a scaled factor.

Lemma 5.5. *Choose* $\delta \in (0,1)$. *If* $s < \frac{\delta L^2 \varepsilon \Delta_{k-1}(M)}{2\sqrt{n}(\sqrt{n}+\sqrt{T})}$ *and* $L^2\varepsilon < \frac{1}{2}$, *then in the event of* G, $\tilde{M}$ *is* $4L^2\varepsilon$*-separated.*

Next, we define an intermediate step $\tilde{P}$ to connect P^M and $P^{\tilde{M}}$. Let $c_i^M = \frac{1}{|P_i^M|}\sum_{l\in P_i^M} m_l$ be the k-means optimal centers of the points represented by M. Define $\tilde{P}$ as the partition generated by $\{c_i^M\}_{i\in[k]}$ on $\tilde{M}$. It is possible both for the membership of points in $\tilde{P}$ to change in P^M, and for the re-calculated centers $c^{\tilde{M}}$ to additionally introduce a difference between $\tilde{P}$ and $P^{\tilde{M}}$. By considering both factors, Lemma 5.6 shows that adding observation noise E and then using HSVT to denoise does not substantially change the optimal k-means partition of M.

Lemma 5.6. *For* $\varepsilon < 0.1$ *and* $\delta \in (0,1)$, *if* $L^2\epsilon < 1/\sqrt{801}$, $\min_i r_i(M) \ge 1/64$, *and* $s < O(\frac{\delta\sqrt{T}}{\sqrt{n}+\sqrt{T}})$, *then, in the event of* G, $\sum_{i=1}^k |P_i^M \ominus P_{\sigma(i)}^{\tilde{M}}| \le 25L^2\varepsilon^2 n$ *for some bijection* $\sigma(i)$.

From $\hat{M}$ to algorithmic output. The last step is to show the approximation error of the output of Algorithm 3 with respect to $P^{\tilde{M}}$ is small. Note that $\tilde{M}$ is an approximation of M via $HSVT(X)$. We continue to condition on the events in the good set G, where $4L^2\varepsilon$-separation is guaranteed in $\tilde{M}$ (Lemma 5.5). Then, we translate the ε-separation condition to the proximity condition (Definition C.1) introduced in Kumar and Kannan (2010) so that we can invoke Theorem B.4.

Lemma 5.7. *Let* $\hat{P}$ *be the partition learned by Algorithm 3. In the event of* G, $\sum_{i=1}^k |P_i^{\tilde{M}} \ominus \hat{P}_{\sigma(i)}| = O(k^2L^4\varepsilon^2 n)$ *for some bijection* $\sigma(i)$.

Finally, we can prove our main result, Theorem 5.8, which shows that the k-mean optimal partition in Θ is well-approximated by the output of Algorithm 3.

Theorem 5.8. *For* $\varepsilon < 0.1$ *and* $\delta \in (0,1)$, *if* $L^2 \le \min(\frac{1}{\sqrt{801}\varepsilon}, \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}})$, $\min_i r_i(M) \ge 1/64$, *and* $s < O(\frac{\delta\sqrt{T}}{\sqrt{n}+\sqrt{T}})$, *then with probability* $1-\delta$, *the output of Algorithm 3* $O(k^2L^2\varepsilon n)$*-approximates* P^Θ.

Proof. To invoke Lemmas 5.3, 5.6, and 5.7, we require the following mild conditions: (i) the separation parameter $\varepsilon < 0.1$, (ii) the bilipschitz parameter$L^2 \le \min(\frac{1}{\sqrt{801}\varepsilon}, \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}})$ (Note that this upper bound on L^2 goes to infinity as ε becomes smaller, thus virtually removing the bound on L when ε is sufficiently small. When ε is bigger than ~ 0.011, the first term dominates.), (iii) the smallest cluster's $r_i(M) \ge 1/64$ (see Lemma 5.2 for the definition), (iv) the standard deviation of noise $s = O(\frac{\delta\sqrt{T}}{\sqrt{n}+\sqrt{T}})$, (v) an event in G occurs (the good set in Lemma 5.4).

With these reasonable conditions on parameters, we can combine the triangle inequality and Lemmas 5.3, 5.6, and 5.7 to bound $\sum_{i=1}^k |P_i^\Theta \ominus \hat{P}_{\sigma(i)}|$ by the sum of $\sum_{i=1}^k |P_i^\Theta \ominus P_{\sigma_1(i)}^M|$, $\sum_{i=1}^k |P_{\sigma_1(i)}^M \ominus P_{\sigma_2(i)}^{\tilde{M}}|$, and $\sum_{i=1}^k |P_{\sigma_2(i)}^{\tilde{M}} \ominus \hat{P}_{\sigma_3(i)}|$:

$$\begin{aligned}\sum_{i=1}^k |P_i^\Theta \ominus \hat{P}_{\sigma(i)}| &\le 8L^2\varepsilon n + 25L^2\varepsilon^2 n + O(k^2L^4\varepsilon^2 n)\\ &\le O(k^2L^2\varepsilon n)\end{aligned}$$

□

5.2 Effects of Donor Selection

The previous section shows that the clustering subroutine (Algorithm 3) in ClusterSC well-approximates the subgroup structure in Θ. Next, we analyze the changes in the SC pipeline when ClusterSC is used, which selects only donors from the relevant cluster A, instead of the whole donor pool X as in classical SC.

With $k \ge 2$ clusters, we expect three changes that will affect the prediction performance guarantees:

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

While the first two are trivial to see, the third one is not. We provide analyses of the gap $\sigma_X^* - \sigma_A^*$ under Gaussian noise setting, and additional analyses on sub-Gaussian and heavy-tailed noise settings are provided in Appendix E.1.

Theorem 5.9 presents the lower bound on the singular value gap, 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 the scale of noise s.

Theorem 5.9 (Singular Value Concentration with Gaussian Noise)**.** *Let the noise terms be sampled* $E_{i,t} \sim \mathcal{N}(0, s^2)$. *If* $r < T$ *and* $n_A < n + 4T - 4\sqrt{nT}$, *then* $\mathbb{E}[\sigma_X^* - \sigma_A^*] \ge s(\sqrt{n} - \sqrt{n_A} - 2\sqrt{T})$.

Finally, we translate the effect of donor selection induced by ClusterSC into an improvement in the upper bound of the prediction error. We compare the performance of ClusterSC, which uses only the selected

donors A as an input for $\mathcal{M}$, against the performance of SC using the whole donor pool X, and analyze the improvement in pre-intervention (Theorem 5.11) and post-intervention regime (Theorem 5.13).

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 5.11, which shows the improvement in pre-intervention MSE bound from using the selected donor pool A instead of the full donor pool X. To do this, Lemma 5.10 first gives an upper bound on the pre-intervention MSE of standard SC without clustering (i.e., Algorithm 2).

Lemma 5.10 (Pre-intervention MSE of SC)**.** *Given donor matrix $X \in \mathbb{R}^{n\times T}$, target unit $x_0 = m_0 + e_0$, 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, then,*

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

Combining Lemma 5.10 with the bounds on singular values in Theorem 5.9 allows us to show that the upper bound on pre-intervention MSE decreases when using selected donors A instead of the full donor pool X.

Theorem 5.11. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on pre-intervention MSE of ClusterSC (Algorithm 4) is strictly smaller than that of classical SC, and the difference in the upper bounds is $\Omega(s^2 n)$.*

Next, we analyze the post-intervention root mean squared error (RMSE), and show similar improvements when changing from X to A (Theorem 5.13). First, Lemma 5.12 gives an upper bound on the post-intervention error of SC without clustering (Algorithm 2), under the standard assumption that the SC weights $\hat{f} \leftarrow \mathcal{M}(X, x_0^-; r)$ satisfy $\|\hat{f}\|_2 \leq \eta$ for some $\eta \geq 0$ (Amjad et al., 2018).

Lemma 5.12 (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,* $\text{RMSE}(\hat{m}_0^+; X)$ *is upper bounded by*

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

This upper bound in Lemma 5.12 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 post-intervention error strictly decreases when using Algorithm 4. Theorem 5.13 gives a lower bound on the improvement of the post-intervention error bound by combining Lemma 5.12 and Theorem 5.9.

Theorem 5.13. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on post-intervention RMSE of ClusterSC (Algorithm 4) is strictly smaller than that of classical SC, and the difference in the upper bounds is $\Omega(s\sqrt{n})$.*

6 Empirical Evaluations

In this section, we test various design choices of ClusterSC on simulated datasets (Section 6.1), and demonstrate ClusterSC on a real-world dataset (Section 6.2).

6.1 Evaluation on Synthetic Datasets

To estimate a realistic size for the synthetic dataset, we turn to the literature that has applied SC on disaggregated datasets. Abadie and L'Hour (2021) adopted SC to measure the effect of participation in a government program on an individual's yearly income. They constructed SC instances out of $n = 2490$ individuals as a donor pool and with 10 covariates (equivalent to T_0). Robbins et al. (2017) examined 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) showed that having a child reduces womens' earnings by constructing SC with $n = 630$ women as donors. T_0 varied depending on the woman's first childbirth year, and was at most 7.

Based on this, we choose $T = 10$ and $n \in \{1000, 2000\}$ and set $T_0 = 8$. We construct a dataset X with two subgroups $A = S + E_A$ and $B = S' + E_B$, with even split ($n_A/n_B = 1$). The signal S (or S') is made by sum of multiple sinusoidal time series. Let the rank of S be r_S. Then, we sample three parameters—α_i (magnitude), ω_i (frequency), and ϕ_i (delay)—to generate a sine wave signal $v_{i,t} = \alpha_i \sin(2\pi\omega_i t + \phi_i)$, $\forall i \in [r_S]$. These parameters were independently sampled from the following distributions: $\alpha_i \sim \text{Beta}(2, 2)$, $\omega_i \sim \text{Unif}(1, 3)$, $\phi_i \sim \mathcal{N}(0, 1)$ for A, and $\alpha_i \sim \text{Beta}(2, 5)$, $\omega_i \sim \text{Unif}(3, 6)$, $\phi_i \sim \mathcal{N}(0, 1)$ for B. 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 observational noise to yield $A_{i,t} = S_{i,t} + E_{i,t}$, where $E_{i,t} \sim \mathcal{N}(0, s^2)$ for varying levels of s from 0.1 to 0.4 with 0.05 interval. We repeat the same process for B, and concatenate the two matrices to make $X = [A^\top, B^\top]^\top$. 500 datasets were generated for the experiment, and the `sklearn` implementation of Lloyd's k-means algorithm with `k-means++` initialization and silhouette scores[2] to find k were used.

[2] `https://scikit-learn.org/1.5/modules/generated`

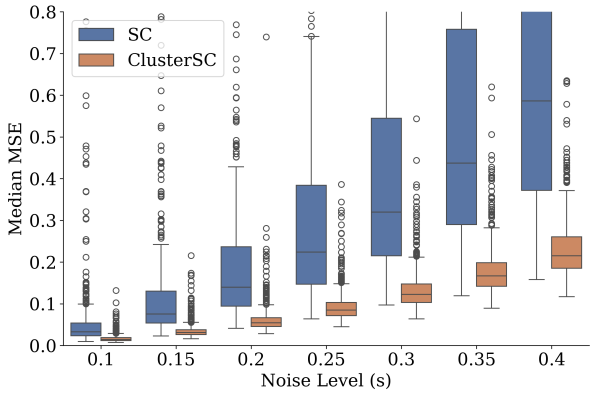

Figure 2: Median Post-intervention MSE using the classical SC without our clustering step (blue) and ClusterSC (orange) for varying levels of noise.

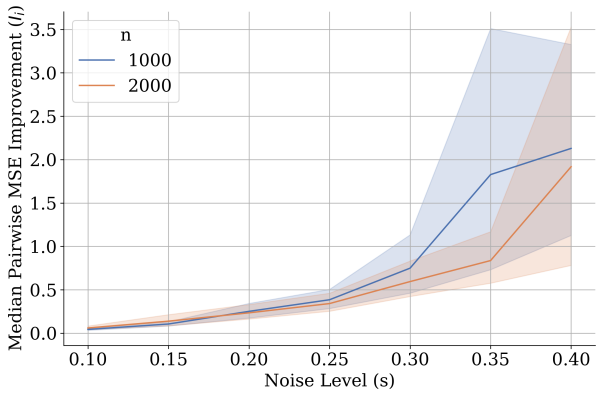

Figure 3: Median of the pairwise improvement I_i, measured for each dataset, for different noise levels (s). Shades represent 95% confidence interval.

For each dataset, we perform a leave-one-out placebo test on 30% of A, by choosing one unit in A as a target and the rest in X as a potential donor pool. For each target, we test two methods: (i) ClusterSC, using a subset of the donors A selected via Algorithm 4, and (ii) SC without clustering, using the whole donor pool X. Our method can flexibly adopt different versions of SC methods, and we present the results using Ridge regression in this section. We provide additional empirical evaluations with different choice of regression methods (OLS, Ridge, and Lasso) in Appendix F.

Figure 2 shows the distribution of median MSE for 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 SC, across all noise levels. This aligns with our Theorem 5.13, which promises a tighter error bound.

Next, we define the pairwise improvement for a target i as the difference in post-intervention MSE scores between SC and ClusterSC: $I_i = \text{MSE}(\hat{m}_i^+; X) - \text{MSE}(\hat{m}_i^+; A)$. Then, we take median$(I_i)$ as a metric to assess the overall improvement measured from the SC instances constructed from one dataset (under a leave-one-out placebo test). Figure 3 shows the median pairwise improvement, median(I_i), induced by ClusterSC at varying noise level. We observe that the median improvement is almost always positive, meaning that more than half of the individuals benefit from using ClusterSC instead of classical SC without clustering. The improvement grows as noise increases, aligning with our Theorem 5.13.

6.2 Evaluation on Real-world Dataset

Next, we evaluate ClusterSC using housing price index (HPI) data from the U.S. Federal Housing Finance Agency.[3] To avoid the effects of the subprime mortgage crisis ($2007-2010$), we use ten years of quarterly HPI data from 1997 to 2006, yielding a total of $T = 40$ time points. The dataset was preprocessed to retain only metropolitan areas without missing data, resulting in $n = 400$ units.

To evaluate ClusterSC, we conduct a placebo test to assess the model's ability to accurately predict *observations* that would serve as *counterfactual* outcomes in the presence of an intervention. We formulate 100 iterations of test, each involving a random split of the units into a donor set (80%) and a target set (20%). In each iteration, a unique SC model is fitted for each target using the corresponding donor set. The accuracy of an iteration is measured by the median MSE of the post-intervention predictions.

Figure 4 presents boxplots of the median MSE measured over 100 iterations. We fix the post-intervention period to year 2006 (four quarterly data points) and use nine years prior to 2006 as pre-intervention time points ($T_0 = 36$). The performance of ClusterSC (orange) is compared against two benchmarks. The first benchmark applies SC using the entire donor pool (blue), while the second benchmark selects a randomly subsampled donor set of the same size as the ClusterSC-selected donor set (green). For example, if ClusterSC selects a cluster of 50 donors for a given target, the second benchmark randomly selects 50 donors from the full donor pool. The plot consists of nine boxplots with different choice of learning method for SC

`/sklearn.metrics.silhouette_score.html`

[3] https://www.fhfa.gov/data/hpi/datasets

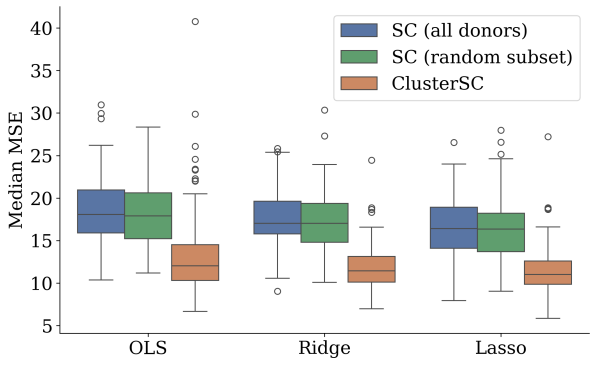

Figure 4: Comparison of ClusterSC and two SC benchmarks on different regression methods (OLS, Ridge, and Lasso). Each boxplot contains 100 points representing the median MSE of each iteration.

weights (e.g., Step 3 of Algorithm 2): OLS (first three), Ridge (middle three), and Lasso (last three). Regularization coefficients for Ridge and Lasso were set to 0.1 after testing values of 0.01, 0.05, 0.1 and 0.2, which resulted in minimal performance differences. For all SC instances, we used $k = 2$ clusters[4] and determined the rank cutoff for HSVT at the 95% threshold.

Compared to the first benchmark, SC with all donors (blue), the second benchmark, SC with a random subset of donors (green), does not show any meaningful change in Figure 4. However, ClusterSC (orange) consistently achieves lower median MSE compared to both benchmarks over all learning methods (OLS, Ridge, or Lasso). This indicates that the clustering approach in ClusterSC improves prediction accuracy by selecting a more relevant donor pool, not just by using fewer donors. The thinner orange boxes, which indicate lower variance in MSE, also suggest that the clustering approach not only improves accuracy but also enhances stability.

7 Discussion and Future Work

This paper presents Cluster Synthetic Control (ClusterSC), a synthetic control (SC) approach that incorporates a donor selection step. By only selecting the most relevant donors, ClusterSC addresses the challenges of higher noise and increased dimensionality in disaggregated datasets. To the author's knowledge, ClusterSC 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 (Abadie and L'Hour, 2021; Chernozhukov et al., 2021; Doudchenko and Imbens, 2016). ClusterSC advances SC methodology to be better suited for applications where individual-level conditional treatment effects are of interest, such as in drug trials or targeted marketing analyses.

ClusterSC is supported by two main theoretical guarantees. Theorem 5.8 shows the accuracy of our clustering step in identifying subgroups in the donor latent variables Θ. Theorems 5.11 and 5.13 establish a tighter upper bound on prediction error induced by ClusterSC, which is empirically validated in Section 6.1 with simulation data and Section 6.2 with real-world data. In both experiments, ClusterSC consistently shows significant improvement across the choice of the learning algorithm for SC weights.

The improved performance of ClusterSC, achieved by partitioning the donors, relies on the observation that each *cluster* may exhibit a more pronounced low-rank structure than the combined matrix. This aligns with prior findings suggesting that, in models such as Gaussian mixtures, the middle components of singular value decomposition can carry more informative signals than the principal component (Nadakuditi, 2013). By incorporating a clustering step, ClusterSC effectively isolates these mixtures, ensuring that the principal components remain the most informative. A similar approach has been explored in matrix completion, where rows are iteratively partitioned based on their projections onto the principal component (Ruchansky et al., 2017). In the same spirit, ClusterSC leverages the top few principal components to identify clusters.

Conceptually, ClusterSC shares similarities with Lasso in that it selects a small subset of donors. While Lasso is good at learning sparse weights, ClusterSC has demonstrated further refinement. The impact of the clustering step on different variations of SC requires further analysis. However, the fundamental principle of concentrating on the most informative signals remains valid under the common assumption that the data is generated from a latent variable model, resulting in an approximately low-rank matrix.

Lastly, we acknowledge a potential fairness issue in our approach. As empirically shown in Section 6, our method guarantees improved performance overall. 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 analyses. Investigating the potential disproportionate effects on minority groups presents an avenue for future research.

[4]Based on silhouette scores.

Acknowledgements

S.R. and R.C. were supported in part by NSF grant CNS-2138834 (CAREER). S.R., A.T., and V.M. were supported in part by Novartis AG. We thank Professor Daniel Hsu for his constructive feedback.

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.

Nadakuditi, R. R. (2013). When are the most informative components for inference also the principal components? *arXiv preprint arXiv:1302.1232*.

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.

Ruchansky, N., Crovella, M., and Terzi, E. (2017). Targeted matrix completion. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 255–263. SIAM.

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 from Prior Work

B.1 Related to Theorem 5.8

Theorem B.1 (Theorem 5.1 (i) of Ostrovsky et al. (2013))**.** *Suppose that X is ε-separated with k clusters. If there is a Voronoi partition $P = \{P_1, \ldots, P_k\}$ such that*

$$\Delta_k^2(X; P) \leq \alpha \Delta_{k-1}^2(X)$$

for some $\alpha \in (0, \frac{1-401\varepsilon^2}{400}]$, then for each cluster P_i, there is a cluster P_i' induced by a distinct optimal center, such that:

$$|P_i \ominus P_i'| \leq 161\varepsilon^2 |P_i'|$$

where $A \ominus B$ denotes the symmetric difference between sets A and B.

Theorem B.2 (Theorem 5.1 (ii) of Ostrovsky et al. (2013))**.** *Let $X = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ be ε-separated and let $X' = \{x_1', ..., x_n'\}$ such that $\|x_i - x_i'\| \leq \frac{\epsilon \Delta_{k-1}(X)}{\sqrt{n}}$. Then $\Delta_k^2(X') \leq \frac{8\varepsilon^2}{1-2\varepsilon^2} \Delta_{k-1}^2(X')$.*

Theorem B.3 (Claim 6 of Kumar and Kannan (2010))**.** *If X is ε-separated with k clusters, then all but ε^2 fraction of points in X satisfy the proximity condition.*

Theorem B.4 (Theorem 2.2 of Kumar and Kannan (2010))**.** *If all but ε fraction of points satisfy the proximity condition, then there exists an algorithm running in polynomial time which correctly partitions all but $O(k^2\varepsilon n)$ points.*

B.2 Related to Theorem 5.9

First of all, we introduce two version sof Weyl's inequality used in the proof of Theorem 5.9.

Theorem B.5 (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.6 (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 5.9.

Theorem B.7 (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.3 Related to Theorem 5.11 and Theorem 5.13

In this section, we introduce theorems used in the proof of Theorem 5.11 and Theorem 5.13. The first one is from Chatterjee (2015), and the proof can be found in the cited paper.

Theorem B.8 (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,*

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

Using Theorem B.8, 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.9 (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 $\mu \ge 0$, let $S = \{i : \sigma_i \ge \mu\}$. 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

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

Proof. By Theorem B.8, we have that $\sigma_i \le \sigma_i(B) + ||A - B||$ for all i. Applying triangle inequality, we obtain

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

□

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.10 (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 $Var(\epsilon_{0,j}) \le s^2$ 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,*

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

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:

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

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

$$\mathbb{E}\|\hat{m}_0^- - m_0^-\|^2 \leq \mathbb{E}\|(Q - \hat{Q})f^*\|^2 + 2\mathbb{E}\|\epsilon_0^-\|^2 + 2\mathbb{E}[\langle \epsilon_0^-, (Q - \hat{Q})f^*\rangle] + 2\mathbb{E}[\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle]. \tag{3}$$

We will now deal with the two inner products on the right hand side of equation (3).

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

Also,

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

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 $\hat{f} = \hat{Q}^\dagger x_0^-$,

$$\begin{aligned}
\mathbb{E}[\langle -\epsilon_0^-, x_0^- - \hat{Q}\hat{f}\rangle] &= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{f}] - \mathbb{E}[(\epsilon_0^-)^Tx_0^-] \\
&= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger x_0^-] - \mathbb{E}[(\epsilon_0^-)^T]m_0^- - \mathbb{E}[(\epsilon_0^-)^T\epsilon_0^-] \\
&= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger]m_0^- + \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] - \mathbb{E}[(\epsilon_0^-)^T\epsilon_0^-] \\
&\overset{(a)}{=} \mathbb{E}[(\epsilon_0^-)^T]\mathbb{E}[\hat{Q}\hat{Q}^\dagger]m_0^- + \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] - \mathbb{E}[(\epsilon_0^-)^T\epsilon_0^-] \\
&= \mathbb{E}[(\epsilon_0^-)^T\hat{Q}\hat{Q}^\dagger\epsilon_0^-] - \mathbb{E}\|\epsilon_0^-\|^2 \\
&\leq s^2r - \mathbb{E}\|\epsilon_0^-\|^2,
\end{aligned}$$

where (a) follows from the independence of noise.

Finally, we get

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

□

C Proof Sketches of Main Lemmas in Section 5.1

In this appendix, we present proof sketches for the three main lemmas in in Section 5.1: Lemmas 5.3, 5.6, and 5.7. All omitted proofs, including complete proofs of these lemmas are presented in Appendix D.

First, we define some notation frequently used in these proofs. In addition to the k-means objective defined in Section 3.2, we additionally define more variations of the k-means objective. For a set of points $A = a_1, \ldots, a_n$, we define the k-mean optimal cluster centers $C^A = \{c_i^A\}_{i=1}^k$ and the induced Voronoi partition $P^A = \{P_i^A\}_{i=1}^k$. Note that the optimal k-means objective can be defined in two equivalent ways,

$$\Delta_k^2(A) = \sum_{i\in[n]} \min_{j\in[k]} ||a_i - c_j^A||^2 = \sum_{l\in[k]} \frac{1}{2|P_l^A|} \sum_{i,j\in P_l^A} ||a_i - a_j||^2.$$

In the analysis, we often use non-optimal k-means cost by introducing artificial partitions $\hat{P}$ and cluster centers $\{\hat{c}_i\}_{i=1}^k$. When the new cluster centers are the mean of all points belonging to each cluster, we omit the center and denote,

$$\Delta_k^2(A; \hat{P}) = \sum_{l\in[k]} \frac{1}{2|\hat{P}_l|} \sum_{i,j\in\hat{P}_l} ||a_i - a_j||^2.$$

Hence, $\Delta_k^2(A; P^A) = \Delta_k^2(A)$ by definition. When the new cluster centers are not the mean of points in each cluster, we explicitly denote,

$$\Delta_k^2(A; \hat{P}, \{\hat{c}_i\}_{i=1}^k) = \sum_{j\in[k]} \sum_{i\in\hat{P}_j} ||a_i - \hat{c}_j||^2.$$

C.1 Proof sketch for Lemma 5.3

To show that $P^\Theta \approx P^M$, we consider an intermediate step P'. Take the labeling under P^Θ and consider the distribution of points in M space. Let $c' = (c'_1, ..., c'_k)$ be the centers induced by P^Θ on M, i.e., $c'_i = \frac{1}{|P_i^\Theta|}\sum_{l\in P_i^\Theta} m_l$, and P' be its induced Voronoi partition. We will use P' as an intermediate step to show that P^Θ and P' are similar by using the property coming from separation structure, and then show P' and P^M are similar by invoking Theorem B.1. Note that P' will be the same as P^Θ if P^Θ on M forms a Voronoi partition.

To invoke Theorem B.1 later, we first require Lemma C.1, which shows the relationship between the optimal k-means objective with input $M = g(\Theta)$ and with input Θ.

Lemma C.1. *For any L-bilipschitz function g,* $(1/L^2)\Delta_k^2(\Theta) \le \Delta_k^2(g(\Theta)) \le L^2\Delta_k^2(\Theta)$.

Based on this, we can obtain the conditions for Theorem B.1—M is $L^2\varepsilon$-separated and the k-means cost of partition P^Θ calculated on M is bounded by $L^4\varepsilon^2\Delta_{k-1}^2(M)$—presented in Lemma 5.1.

Lemma 5.1. *$M = g(\Theta)$ is $L^2\varepsilon$-separated with k clusters, and $\Delta_k^2(M; P^\Theta) \le L^4\varepsilon^2\Delta_{k-1}^2(M)$.*

Proof. We first show the $L^2\varepsilon$-separation of M by comparing the the cost of clustering $M = g(\Theta)$ with k and $k-1$ clusters.

$$\begin{aligned}
\Delta_k^2(M) &\le L^2\Delta_k^2(\Theta) && \text{(Lemma C.1)}\\
&\le L^2\varepsilon^2\Delta_{k-1}^2(\Theta) && \text{(Modeling assumption (1))}\\
&\le L^4\varepsilon^2\Delta_{k-1}^2(M). && \text{(Lemma C.1)}
\end{aligned}$$

Note that the first line decomposes to

$$\Delta_k^2(M) \le \Delta_k^2(M; P^\Theta) \le L^2\Delta_k^2(\Theta),$$

as shown in Equation (5), (6), and (7). Taking the second term, we obtain $\Delta_k^2(M; P^\Theta) \le L^4\varepsilon^2\Delta_{k-1}^2(M)$. □

With this in mind, we now focus on bounding the difference between P^Θ and P'. First, we define $r_i^2(A) := \frac{1}{|P_i^A|}\sum_{l\in P_i^A}\|A_l - c_i^A\|^2$, the mean squared error of cluster P_i^A, for any set of points A. Then, Ostrovsky et al. (2013) shows a useful lemma.

Lemma 5.2 (Lemma 4.1 of (Ostrovsky et al., 2013)). *Let $r_i^2(\Theta) := \frac{1}{|P_i^\Theta|}\sum_{l\in P_i^\Theta}\|\theta_l - c_i^\Theta\|^2$. If Θ is ε-separated, $r_i^2(\Theta) \le \frac{\varepsilon^2}{1-\varepsilon^2}\min_{j\neq i}\|c_i^\Theta - c_j^\Theta\|^2$.*

Using this, we can bound the distance between the centers c' and the bilipschitz-map of centers in Θ, $g(c^\Theta)$.

Lemma C.2. *For all $i \in [k]$, $\|g(c_i^\Theta) - c_i'\| \le L \cdot r_i(\Theta)$.*

Now, we define $core(P_i^\Theta)$, a core set that contains at least a $1-\varepsilon$ fraction of the points in partition P_i^Θ.

Lemma C.3. *Let $core(P_i^\Theta) := \{l \in P_i^\Theta : \|\theta_l - c_i^\Theta\| \le \sqrt{\frac{\epsilon}{1-\epsilon^2}}\min_{j\neq i}\|c_i^\Theta - c_j^\Theta\|\}$. Then, for all $i \in [k]$, $|core(P_i^\Theta)| \ge (1-\varepsilon)|P_i^\Theta|$.*

Lastly, we show a useful characteristic of the set $core(P_i^\Theta)$.

Lemma C.4. *Choose two distinct partitions P_i^Θ and P_j^Θ. Then, for all $l \in core(P_i^\Theta)$,*

$$\|\theta_l - c_j^\Theta\| - \|\theta_l - c_i^\Theta\| \ge \left(1 - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right)\|c_j^\Theta - c_i^\Theta\|.$$

Combining Lemmas C.2, C.3, and C.4, we show that P' and P^Θ are similar.

Lemma C.5. *For small $\varepsilon \le 0.1$, if $L^2 < \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}}$, then $\sum_{i=1}^k |P_i' \ominus P_i^\Theta| \le 2\varepsilon n$.*

Finally, we show the bound on the difference between k-means optimal clusters in Θ and M. Lemma 5.3 uses Lemma C.5 to show $P^\Theta \approx P'$ and instantiates Theorem B.1 to show $P' \approx P^M$. Again, if P^Θ on M were Voronoi partition, $P^\Theta = P'$ and the difference coming from $P^\Theta \approx P'$ would be zero. Note that the bound on L depends on ε—we need approximately $\varepsilon < 0.1$ to ensure $L > 1$, and the upper bound on L increases as we have smaller ε.

Lemma 5.3. *For small $\varepsilon \le 0.1$, if $L^2 \le \min(\frac{1}{\sqrt{801\varepsilon}}, \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}})$, then $\sum_{i=1}^k |P_i^\Theta \ominus P_{\sigma(i)}^M| \le 8L^2\varepsilon n$ for some bijection $\sigma(i)$, where $n = |\Theta|$.*

C.2 Proof sketch for Lemma 5.6

The goal is to show the difference between the points represented by M and $\tilde{M}$, where $\tilde{M} = HSVT(X;r) = \sum_{i=1}^r \sigma_i u_i v_i^\top$ where σ_i, u_i, and v_i are respectively the i-th singular value, left singular vector, and right singular vector of $X = M + E$.

To quantify the difference between M and $\tilde{M}$, we define $\eta := \max_{i\in[n]}\|m_i - \tilde{m}_i\|$. Then, define G as a set of *good* events where the noise is small enough that $\eta \le \frac{2s(\sqrt{n}+\sqrt{T})}{\delta}$ for some $\delta > 0$, which happens with high probability (at least $1-\delta$).

Lemma 5.4. *With probability $1-\delta$, $\eta \le \frac{2s(\sqrt{n}+\sqrt{T})}{\delta}$.*

For the remainder of the proof, we only focus on the events in G. Then, Lemma 5.5 shows that a separation structure is preserved in $\tilde{M}$ with a scaled factor.

Lemma 5.5. *Choose $\delta \in (0,1)$. If $s < \frac{\delta L^2\varepsilon\Delta_{k-1}(M)}{2\sqrt{n}(\sqrt{n}+\sqrt{T})}$ and $L^2\varepsilon < \frac{1}{2}$, then in the event of G, $\tilde{M}$ is $4L^2\varepsilon$-separated.*

Next, we define an intermediate step $\tilde{P}$ to connect P^M and $P^{\tilde{M}}$. Let $c_i^M = \frac{1}{|P_i^M|}\sum_{l\in P_i^M} m_l$ be the k-means optimal centers of the points represented by M. Define $\tilde{P}$ as the partition generated by $\{c_i^M\}_{i\in[k]}$ on $\tilde{M}$. It is possible both for the membership of points in $\tilde{P}$ to change in P^M, and for the re-calculated centers $c^{\tilde{M}}$ to additionally introduce a difference between $\tilde{P}$ and $P^{\tilde{M}}$. The next two lemmas address these changes.

Lemma C.6. *If $s < \frac{\delta\left(-\sqrt{16T}+\sqrt{16T+\frac{6\varepsilon^2L^4\Delta_{k-1}^2(\tilde{M})}{n}}\right)}{12(\sqrt{n}+\sqrt{T})}$ and $L^2\varepsilon < \frac{1}{2}$, then in the event of G, $|\tilde{P}\ominus P^{\tilde{M}}| \le 644L^4\varepsilon^2 n$.*

Lemma C.7. *If* $s \leq \frac{\delta\left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})}$*, then in the event of* G*,* $|\tilde{P} \ominus P^M| \leq 2L^2\varepsilon n$

Finally, we combine Lemmas C.6 and C.7 to bound the difference between $P^{\tilde{M}}$ and P^M. Specifically, Lemma 5.6 shows that adding observation noise E and then using HSVT to denoise does not substantially change the optimal k-means partition of M.

Lemma 5.6. *For* $\varepsilon < 0.1$ *and* $\delta \in (0,1)$*, if* $L^2\epsilon < 1/\sqrt{801}$*,* $\min_i r_i(M) \geq 1/64$*, and* $s < O(\frac{\delta\sqrt{T}}{\sqrt{n}+\sqrt{T}})$*, then, in the event of* G*,* $\sum_{i=1}^k |P_i^M \ominus P_{\sigma(i)}^{\tilde{M}}| \leq 25L^2\varepsilon^2 n$ *for some bijection* $\sigma(i)$*.*

C.3 Proof sketch for Lemma 5.7

We show the approximation error of the output of Algorithm 3 with respect to $P^{\tilde{M}}$ is small. Note that $\tilde{M}$ is an approximation of M via $HSVT(X)$. We continue to condition on the events in the good set G, where $4L^2\varepsilon$-separation is guaranteed in $\tilde{M}$ (Lemma 5.5). Then, we will translate the ε-separation condition to the proximity condition introduced in Kumar and Kannan (2010). This proximity condition is defined as below.

Definition C.1 (Proximity Condition (Kumar and Kannan, 2010))**.** *Let* $X \in \mathbb{R}^{n\times T}$ *be the data matrix where the rows* X_i *are divided into* k *clusters* $P_1, \ldots, P_k$ *with corresponding cluster centers* $c_1, \ldots, c_k$*. Let* C *be the* n *by* T *matrix with each row* C_i *as the cluster center of which* X_i *belongs to (i.e.,* $c_{\pi(X_i)}$ *where* π *denotes a function that coutputs the cluster of* X_i*.) Define*

$$\Delta_{r,s} = \left(\frac{ck}{\sqrt{|P_r|}} + \frac{ck}{\sqrt{|P_s|}}\right)||X - C||,$$

where c *is a large enough constant. We say a point* $X_i \in P_r$ *satisfies the proximity condition if for any* $s \neq r$*, the projection of* X_i *onto the line connecting* c_r *to* c_s *is at least* $\Delta_{r,s}$ *closer to* c_r *than to* c_s*.*

Theorem B.3 (from Kumar and Kannan (2010)) shows that the ε-separation condition implies that at least a $1-\varepsilon^2$ fraction of points satisfy the proximity condition. Then, we can instantiate Theorem B.4 to show that our Algorithm 3 well-approximates $P^{\tilde{M}}$.

Lemma 5.7. *Let* $\hat{P}$ *be the partition learned by Algorithm 3. In the event of* G*,* $\sum_{i=1}^k |P_i^{\tilde{M}} \ominus \hat{P}_{\sigma(i)}| = O(k^2L^4\varepsilon^2 n)$ *for some bijection* $\sigma(i)$*.*

Proof. In the event of G, $\tilde{M}$ is $4L^2\varepsilon$-separated. hence we have all but $16L^4\varepsilon^2$ fraction of points satisfying the proximity condition. By instantiating Theorem B.4, Algorithm 3 can correctly classify all but $16k^2L^4\varepsilon^2 n = O(k^2L^4\varepsilon^2 n)$ points with respect to $P^{\tilde{M}}$ in polynomial time. □

D Omitted proofs from Section 5.1

In this appendix, all omitted proofs from Section 5.1 are provided.

D.1 Proof of Lemma 5.1

Lemma 5.1. $M = g(\Theta)$ *is* $L^2\varepsilon$*-separated with* k *clusters, and* $\Delta_k^2(M; P^\Theta) \leq L^4\varepsilon^2\Delta_{k-1}^2(M)$*.*

Proof. We first show the $L^2\varepsilon$-separation of M by comparing the the cost of clustering $M = g(\Theta)$ with k and $k-1$ clusters.

$$\begin{aligned}
\Delta_k^2(M) &\leq L^2\Delta_k^2(\Theta) && \text{(Lemma C.1)}\\
&\leq L^2\varepsilon^2\Delta_{k-1}^2(\Theta) && \text{(Modeling assumption (1))}\\
&\leq L^4\varepsilon^2\Delta_{k-1}^2(M). && \text{(Lemma C.1)}
\end{aligned}$$

Note that the first line decomposes to

$$\Delta_k^2(M) \leq \Delta_k^2(M; P^\Theta) \leq L^2\Delta_k^2(\Theta),$$

as shown in Equation (5), (6), and (7). Taking the second term, we obtain $\Delta_k^2(M; P^\Theta) \leq L^4\varepsilon^2\Delta_{k-1}^2(M)$. □

D.2 Proof of Lemma C.1

Lemma C.1. *For any L-bilipschitz function g,* $(1/L^2)\Delta_k^2(\Theta) \le \Delta_k^2(g(\Theta)) \le L^2\Delta_k^2(\Theta)$.

Proof. We begin with the optimal k-means objective in Θ

$$\begin{aligned}
\Delta_k^2(\Theta) = \Delta_k^2(\Theta; P^\Theta) &\le \Delta_k^2(\Theta; P^M) && (4)\\
&= \sum_{l=1}^{k} \frac{1}{2|P_l^M|} \sum_{i,j \in P_l^M} \|\theta_i - \theta_j\|^2 \\
&\le L^2 \sum_{l=1}^{k} \frac{1}{2|P_l^M|} \sum_{i,j \in P_l^M} ||g(\theta_i) - g(\theta_j)||^2 \\
&= L^2\Delta_k^2(g(\Theta); P^M) = L^2\Delta_k^2(g(\Theta)) && (5)\\
&\le L^2\Delta_k^2(g(\Theta); P^\Theta) = L^2 \sum_{l=1}^{k} \frac{1}{2|P_l^\Theta|} \sum_{i,j \in P_l^\Theta} ||g(\theta_i) - g(\theta_j)||^2 && (6)\\
&\le L^4 \sum_{l=1}^{k} \frac{1}{2|P_l^\Theta|} \sum_{i,j \in P_l^\Theta} \|\theta_i - \theta_j\|^2 = L^4\Delta_k^2(\Theta; P^\Theta) = L^4\Delta_k^2(\Theta) && (7)
\end{aligned}$$

The first step is because of the optimality of P, the second step is from the definition of $\Delta_k^2(\Theta; P^M)$, the third step is because g is L-bilipschitz, the fourth step is by definition, the fifth step is because P^M is optimal for $g(\Theta)$, and the fifth step is again because g is L-bilipschitz.

Combining Equations (4), (5), and (7) and dividing by L^2 yields:

$$\frac{1}{L^2}\Delta_k^2(\Theta) \le \Delta_k^2(g(\Theta)) \le L^2\Delta_k^2(\Theta).$$

□

D.3 Proof of Lemma C.2

Lemma C.2. *For all* $i \in [k]$, $\|g(c_i^\Theta) - c_i'\| \le L \cdot r_i(\Theta)$.

Proof. Expand the definition of c_i' and rearrange.

$$\begin{aligned}
\|g(c_i^\Theta) - c_i'\| &= \Big\|g(c_i^\Theta) - \frac{1}{|P_i^\Theta|} \sum_{j \in P_i^\Theta} g(\theta_j)\Big\| \\
&= \frac{1}{|P_i^\Theta|}\Big\| \sum_{j \in P_i^\Theta} (g(c_i^\Theta) - g(\theta_j))\Big\| \\
&\le \frac{1}{|P_i^\Theta|} \sum_{j \in P_i^\Theta} \|g(c_i^\Theta) - g(\theta_j)\| && \text{(triangle inequality)} \\
&\le \frac{L}{|P_i^\Theta|} \sum_{j \in P_i^\Theta} \|c_i^\Theta - \theta_j\| && \text{(Lipschitzness of } g\text{)} \\
&\le L\sqrt{\frac{1}{|P_i^\Theta|} \sum_{j \in P_i^\Theta} \|c_i^\Theta - \theta_j\|^2} = L \cdot r_i(\Theta) && \text{(Jensen's inequality)}
\end{aligned}$$

□

D.4 Proof of Lemma C.3

Lemma C.3. *Let* $core(P_i^\Theta) := \{l \in P_i^\Theta : \|\theta_l - c_i^\Theta\| \leq \sqrt{\frac{\epsilon}{1-\epsilon^2}} \min_{j\neq i} \|c_i^\Theta - c_j^\Theta\|\}$. *Then, for all* $i \in [k]$, $|core(P_i^\Theta)| \geq (1-\varepsilon)|P_i^\Theta|$.

Proof. Define $d_l = \|\theta_l - c_i^\Theta\|^2$ with probability mass distributed uniformly over $\forall l \in P_i^\Theta$. Then,

$$\mathbb{E}[d_l] = \frac{1}{|P_i^\Theta|} \sum_{l \in P_i^\Theta} \|\theta_l - c_i^\Theta\|^2 = r_i^2(\Theta)$$

by construction. Using Markov's inequality, for all $t > 0$, we have

$$\mathbb{P}(d_l \geq t) \leq \frac{\mathbb{E}[d_l]}{t} = \frac{r_i^2(\Theta)}{t} \leq \frac{\frac{\epsilon^2}{1-\epsilon^2} \min_{j\neq i} \|c_i^\Theta - c_j^\Theta\|^2}{t},$$

where the last inequality is from Lemma 5.2. Take $t = \frac{r_i^2(\Theta)}{\epsilon} = \frac{\epsilon}{1-\epsilon^2} \min_{j\neq i} \|c_i^\Theta - c_j^\Theta\|^2$. Then,

$$\mathbb{P}\left(\|\theta_l - c_i^\Theta\|^2 \geq \frac{r_i^2(\Theta)}{\epsilon}\right) = \mathbb{P}\left(\|\theta_l - c_i^\Theta\| \geq \frac{r_i(\Theta)}{\sqrt{\epsilon}}\right) = \mathbb{P}\left(\|\theta_l - c_i^\Theta\| \geq \sqrt{\frac{\epsilon}{1-\epsilon^2}} \min_{j\neq i} \|c_i^\Theta - c_j^\Theta\|\}\right) \leq \epsilon$$

Hence, we have at most ε faction of points l in P_i^Θ that will not belong to $core(P_i^\Theta)$. Finally, we have $|core(P_i^\Theta)| \geq (1-\varepsilon)|P_i^\Theta|$. □

D.5 Proof of Lemma C.4

Lemma C.4. *Choose two distinct partitions* P_i^Θ *and* P_j^Θ. *Then, for all* $l \in core(P_i^\Theta)$,

$$\|\theta_l - c_j^\Theta\| - \|\theta_l - c_i^\Theta\| \geq \left(1 - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right) \|c_j^\Theta - c_i^\Theta\|.$$

Proof. Fix $i \in [k]$. For all $j \in [k] \setminus \{i\}$ and all $l \in core(P_i^\Theta)$, we have

$$\|c_j^\Theta - \theta_l\| + \|\theta_l - c_i^\Theta\| \geq \|c_j^\Theta - c_i^\Theta\| \tag{8}$$

by triangle inequality. By subtracting $2\|\theta_l - c_i^\Theta\|$ from each side, we obtain

$$\begin{aligned}
\|c_j^\Theta - \theta_l\| - \|\theta_l - c_i^\Theta\| &\geq \|c_j^\Theta - c_i^\Theta\| - 2\|\theta_i - c_i^\Theta\| \\
&\geq \|c_j^\Theta - c_i^\Theta\| - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}} \min_{i\neq j} \|c_j^\Theta - c_i^\Theta\| && \text{(Lemma C.3)} \\
&\geq \left(1 - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right) \|c_j^\Theta - c_i^\Theta\|
\end{aligned}$$

Note that the equality is achieved in the last transition if c_j^Θ is the closest neighboring center from c_i^Θ. □

D.6 Proof of Lemma C.5

Lemma C.5. *For small* $\varepsilon \leq 0.1$, *if* $L^2 < \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}}$, *then* $\sum_{i=1}^k |P_i' \ominus P_i^\Theta| \leq 2\varepsilon n$.

Proof. Take $core(P_i^\Theta)$ as in Lemma C.3 and imagine the points plotted in M space. Suppose towards contradiction that there exists $l \in core(P_i^\Theta)$ such that $l \in P_j'$ for some $j \neq i$. Hence, l satisfies the following condition:

$$\|m_l - c_i'\| \geq \|m_l - c_j'\| \tag{9}$$

For the left hand side, we can bound from above

$$\begin{aligned}
\|m_l - c_i'\| &\leq \|m_l - g(c_i^\Theta)\| + \|g(c_i^\Theta) - c_i'\| && \text{(Triangle inequality)} \\
&\leq \|m_l - g(c_i^\Theta)\| + Lr_i(\Theta). && \text{(Lemma C.2)}
\end{aligned}$$

To bound the right hand side from below, we start from the distance between m_l and $g(c_j^\Theta)$

$$\begin{aligned}
\|m_l - g(c_j^\Theta)\| &\le \|m_l - c_j'\| + \|c_j' - g(c_j^\Theta)\| && \text{(Triangle inequality)}\\
&\le \|m_l - c_j'\| + Lr_j(\Theta). && \text{(Lemma C.2)}\\
\therefore\ \|m_l - g(c_j^\Theta)\| - Lr_j(\Theta) &\le \|m_l - c_j'\|
\end{aligned}$$

Then, by the inequality in (9), we have

$$\begin{aligned}
\|m_l - g(c_i^\Theta)\| + Lr_i(\Theta) &\ge \|m_l - g(c_j^\Theta)\| - Lr_j(\Theta)\\
\|g(\theta_l) - g(c_i^\Theta)\| + Lr_i(\Theta) &\ge \|g(\theta_l) - g(c_j^\Theta)\| - Lr_j(\Theta) && \text{(by } m_i = g(\theta_l)\text{)}\\
L\|\theta_l - c_i^\Theta\| + Lr_i(\Theta) &\ge \frac{1}{L}\|\theta_l - c_j^\Theta\| - Lr_j(\Theta) && (L\text{-bilipschitzness of } g)\\
Lr_i(\Theta) + Lr_j(\Theta) &\ge \frac{1}{L}\|\theta_l - c_j^\Theta\| - L\|\theta_l - c_i^\Theta\|\\
r_i(\Theta) + r_j(\Theta) &\ge \frac{1}{L^2}\|\theta_l - c_j^\Theta\| - \|\theta_l - c_i^\Theta\|
\end{aligned}$$

The left hand side can be bounded by Lemma 5.2,

$$r_i(\Theta) + r_j(\Theta) \le \frac{2\varepsilon}{\sqrt{1-\varepsilon^2}}\|c_i^\Theta - c_j^\Theta\|.$$

The right hand side can be arranged and bounded below by Lemma C.4

$$\begin{aligned}
\frac{1}{L^2}\|\theta_l - c_j^\Theta\| - \|\theta_l - c_i^\Theta\| &= \left(\frac{1}{L^2} - 1\right)\|\theta_l - c_j^\Theta\| + \|\theta_l - c_j^\Theta\| - \|\theta_l - c_i^\Theta\|\\
&\ge \left(\frac{1}{L^2} - 1\right)\|\theta_l - c_j^\Theta\| + \left(1 - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right)\|c_j^\Theta - c_i^\Theta\|.
\end{aligned}$$

Combining the upper and the lower bounds and rearranging terms,

$$\begin{aligned}
\frac{2\varepsilon}{\sqrt{1-\varepsilon^2}}\|c_i^\Theta - c_j^\Theta\| &\ge \left(\frac{1}{L^2} - 1\right)\|\theta_l - c_j^\Theta\| + \left(1 - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right)\|c_j^\Theta - c_i^\Theta\|\\
\left(1 - \frac{1}{L^2}\right)\|\theta_l - c_j^\Theta\| &\ge \left(1 - 2\sqrt{\frac{\varepsilon}{1-\varepsilon^2}} - \frac{2\varepsilon}{\sqrt{1-\varepsilon^2}}\right)\|c_i^\Theta - c_j^\Theta\|\\
\left(1 - \frac{1}{L^2}\right)\|\theta_l - c_j^\Theta\| &\ge \left(1 - \frac{2\sqrt{\varepsilon} + 2\varepsilon}{\sqrt{1-\varepsilon^2}}\right)\|c_i^\Theta - c_j^\Theta\|.
\end{aligned}$$

Note that l is coming from the set $core(P_i^\Theta)$, and hence by Lemma C.3, we have

$$\|\theta_l - c_j^\Theta\| \le \|\theta_l - c_i^\Theta\| + \|c_i^\Theta - c_j^\Theta\| \le \left(1 + \sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right)\|c_i^\Theta - c_j^\Theta\|.$$

Using this, we will replace $\|\theta_l - c_j^\Theta\|$ in the previous inequality by a function of $\|c_i^\Theta - c_j^\Theta\|$

$$\left(1 - \frac{1}{L^2}\right)\left(1 + \sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right)\|c_i^\Theta - c_j^\Theta\| \ge \left(1 - \frac{2\varepsilon + 2\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}\right)\|c_i^\Theta - c_j^\Theta\|$$

Dividing both sides by $\|c_i^\Theta - c_j^\Theta\|$ yields

$$\begin{aligned}\left(1-\frac{1}{L^2}\right)\left(1+\sqrt{\frac{\varepsilon}{1-\varepsilon^2}}\right) &\geq \left(1-\frac{2\varepsilon+2\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}\right)\\ 1-\frac{1}{L^2} &\geq \frac{1-\frac{2\varepsilon+2\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}}{1+\frac{\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}}\\ \frac{1}{L^2} &\leq 1-\frac{1-\frac{2\varepsilon+2\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}}{1+\frac{\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}}\\ L &\geq \left(1-\frac{1-\frac{2\varepsilon+2\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}}{1+\frac{\sqrt{\varepsilon}}{\sqrt{1-\varepsilon^2}}}\right)^{-1/2} = \left(\frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}}\right)^{1/2}.\end{aligned}$$

Since we assumed that $L < \left(\frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}}\right)^{1/2}$, this contradicts our assumption. Hence, there is no point l such that $\theta_l \in core(P_i^\Theta)$ and $m_l \notin P_i'$. Finally, we conclude that $|P_i' \ominus P_i^\Theta| \leq \varepsilon |P_i^\Theta|$, yielding $\sum_{i=1}^k |P_i' \ominus P_i^\Theta| \leq 2\varepsilon n$.

□

D.7 Proof of Lemma 5.3

Lemma 5.3. *For small $\varepsilon \leq 0.1$, if $L^2 \leq \min(\frac{1}{\sqrt{801}\varepsilon}, \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}})$, then $\sum_{i=1}^k |P_i^\Theta \ominus P_{\sigma(i)}^M| \leq 8L^2\varepsilon n$ for some bijection $\sigma(i)$, where $n = |\Theta|$.*

Proof. First, we investigate the meaning of the second part of Lemma 5.1: $\Delta_k^2(M; P^\Theta) \leq L^2\Delta_k^2(\Theta)$. We define the center of the partition P^Θ in M space as $\{c_i'\}_{i=1}^k = \{\frac{1}{|P_j^\Theta|}\sum_{i\in P_j^\Theta} m_i\}_{j=1}^k$. Note that this partition may not be Voronoi partitions anymore in M space.

Remember P', the Voronoi partition induced by centers $\{c_i'\}_{i=1}^k$. Then,

$$\begin{aligned}\Delta_k^2(M; P') &= \Delta_k^2(M; P', \{c_i'\}_{i=1}^k)\\ &\leq \Delta_k^2(M; P^\Theta, \{c_i'\}_{i=1}^k)\\ &= \Delta_k^2(M; P^\Theta)\\ &\leq L^4\varepsilon^2\Delta_{k-1}^2(M)\end{aligned}$$

The first step is by definition, the second step is because P' is the induced Voronoi partition of $\{c_i'\}_{i=1}^k$, the third step is because $\{c_i'\}_{i=1}^k$ are included cluster means of P^Θ, and the last step is by the second part of Lemma 5.1. Taking the first, third, and last part of this inequalities, we obtain

$$\Delta_k^2(M; P') \leq \Delta_k^2(M; P^\Theta) \leq L^4\varepsilon^2\Delta_{k-1}^2(M).$$

If P^Θ represented in M space is still Voronoi partition, $P' = P^\Theta$. If not, the difference between P' and P^Θ is bounded by $O(\varepsilon n)$ by Lemma C.5. Hence, we will show the difference between P' and P^M first, and then use this fact to conclude $P^\Theta \approx P^M$.

Lemma 5.1 shows $L^2\varepsilon$-separation of M, and we have $\Delta_k^2(M; P') \leq L^4\varepsilon^2\Delta_{k-1}^2(M)$. With these two conditions, we can now instantiate the first part of Theorem B.1 (Theorem 5.1 of (Ostrovsky et al., 2013)) with $\alpha = L^4\varepsilon^2$. To satistfy the condition on α, we impose $\varepsilon \leq \frac{1}{\sqrt{801}L^2}$, so that $\alpha = L^4\varepsilon^2 \leq \frac{1-401L^4\varepsilon^2}{400}$. Then, Theorem B.1 tells us that for each cluster P_i', we can match it with one of the optimal clusters $P_{\sigma(i)}^M$, with small fraction of error:

$$|P_i' \ominus P_{\sigma(i)}^M| \leq 161L^4\varepsilon^2|P_{\sigma(i)}^M|,$$

where $\sigma(i)$ is some bijection. By summing this over all $i \in [k]$ and using the assumption $L^2\varepsilon < \frac{1}{\sqrt{801}}$, we get

$$\sum_{i=1}^{k} |P_i' \ominus P_{\sigma(i)}^M| \leq 161L^4\varepsilon^2 n \leq 6L^2\varepsilon n,$$

where $n = |\Theta|$.

Finally, we use Lemma C.5 showing that $\sum_{i=1}^{k} |P_i' \ominus P_i^\Theta| \leq 2\varepsilon n \leq 2L^2\varepsilon n$, and conclude

$$\sum_{i=1}^{k} |P_i^\Theta \ominus P_{\sigma(i)}^M| \leq 8L^2\varepsilon n = O(L^2\varepsilon n).$$

This requires assumptions on ε and L: $\varepsilon < 0.1$ and $L^2 \leq \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}}$. Combining this with the previous assumption $L^2\varepsilon \leq \frac{1}{\sqrt{801}}$, this theorem assumes $L^2 \leq \min(\frac{1}{\sqrt{801}\varepsilon}, \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}})$. Note that we always have $L^2 \leq \frac{1}{\sqrt{801}\varepsilon} \leq \frac{\sqrt{1-\varepsilon^2}+\sqrt{\varepsilon}}{2\varepsilon+3\sqrt{\varepsilon}}$ for $\varepsilon > 0.011$, and the two values do not differ too much for $\varepsilon \leq 0.011$.

□

D.8 Proof of Lemma 5.4

Lemma 5.4. *With probability* $1-\delta$, $\eta \leq \frac{2s(\sqrt{n}+\sqrt{T})}{\delta}$.

Proof. Observe that for any row of a matrix A, $\|A_i\| = \|U_i\Sigma V^\top\| = \|U_i\Sigma\|$, where U, Σ, V denote singular value decomposition of A. Then, $\max \|A_i\| \leq \sigma_1\sqrt{\sum_{j=1}^{r} U_{i,j}^2} = \sigma_1 = ||A||$. Using this, we bound η from above

$$\begin{aligned}
\eta &= \max_i \|m_i - \tilde{m}_i\| \\
&\leq \|M - \tilde{M}\| \\
&= \|M - (M+E) + (M+E) - \tilde{M}\| \\
&\leq \|M - (M+E)\| + \|(M+E) - \tilde{M}\| \\
&= \|E\| + \|(M+E) - \tilde{M}\|,
\end{aligned}$$

where the first inequality is by the previously shown fact and the second inequality is from a subadditivity property of a norm.

Let the true rank of M be r. Then by Eckart-Young Theorem, we have

$$\begin{aligned}
\|(M+E) - \tilde{M}\| &= \min_{A:\, rank(A)=r} \|(M+E) - A\| && \text{(Eckart-Young Theorem)} \\
&\leq \|(M+E) - M\| \\
&= \|E\|.
\end{aligned}$$

Applying this to bound η, we obtain $\eta \leq 2\|E\|$. Also, by Gordon's theorem, we have $\mathbb{E}[\|E\|] \leq s(\sqrt{n}+\sqrt{T})$. Combining these yields

$$\begin{aligned}
\mathbb{E}[\eta] &\leq \mathbb{E}[2\|E\|] \\
&\leq 2s(\sqrt{n}+\sqrt{T}).
\end{aligned}$$

Instantiating Markov's inequality on η, we have

$$\eta \leq \frac{2s(\sqrt{n}+\sqrt{T})}{\delta},$$

with probability at least $1-\delta$. □

D.9 Proof of Lemma 5.5

Lemma 5.5. *Choose $\delta \in (0,1)$. If $s < \frac{\delta L^2 \varepsilon \Delta_{k-1}(M)}{2\sqrt{n}(\sqrt{n}+\sqrt{T})}$ and $L^2\varepsilon < \frac{1}{2}$, then in the event of G, $\tilde{M}$ is $4L^2\varepsilon$-separated.*

Proof. By Lemma 5.4 and the assumption on s, we have

$$\max_i \|\tilde{m}_i - m_i\| \leq \frac{2s(\sqrt{n}+\sqrt{T})}{\delta} \leq \frac{L^2\varepsilon\Delta_{k-1}(M)}{\sqrt{n}}$$

with probability $1-\delta$. Then, by Theorem B.2, $\tilde{M}$ is $\sqrt{\frac{8L^4\varepsilon^2}{1-2L^4\varepsilon^2}}$-separated. Using a mild assumption $L^2\varepsilon < \frac{1}{2}$, we can bound $\sqrt{\frac{8L^4\varepsilon^2}{1-2L^4\varepsilon^2}} \leq 4L^2\varepsilon$. We conclude that $\tilde{M}$ is $4L^2\varepsilon$-separated. □

D.10 Proof of Lemma C.6

Lemma C.6. *If $s < \frac{\delta\left(-\sqrt{16T}+\sqrt{16T+\frac{6\varepsilon^2L^4\Delta_{k-1}^2(\tilde{M})}{n}}\right)}{12(\sqrt{n}+\sqrt{T})}$ and $L^2\varepsilon < \frac{1}{2}$, then in the event of G, $|\tilde{P} \ominus P^{\tilde{M}}| \leq 644L^4\varepsilon^2 n$.*

Proof. We can establish the relationship between $\Delta_k^2(\tilde{M};\tilde{P})$ and $\leq \Delta_{k-1}^2(M)$ as follows.

$$\begin{aligned}
\Delta_k^2(\tilde{M};\tilde{P}) \leq \Delta_k^2(\tilde{M};P^M,\{c_i^M\}_{i=1}^k) &= \sum_{i=1}^k \sum_{l\in P_i^M} \|\tilde{m}_l - c_i^M\|^2 \\
&= \sum_{i=1}^k \sum_{l\in P_i^M} \|\tilde{m}_l - m_l + m_l - c_i^M\|^2 \\
&= \sum_{i=1}^k \sum_{l\in P_i^M} \|m_l - c_i^M\|^2 + \|\tilde{m}_l - m_l\|^2 + 2\langle \tilde{m}_l - m_l, m_l - c_i^M\rangle \\
&\leq \sum_{i=1}^k \sum_{l\in P_i^M} \|m_l - c_i^M\|^2 + \|\tilde{m}_l - m_l\|^2 + 2\|\tilde{m}_l - m_l\| \cdot \|m_l - c_i^M\| \\
&\leq \Delta_k^2(M) + \sum_{k=1}^n \sum_{l\in P_i^M} \left(\eta^2 + 4\eta\sqrt{T}\right) \\
&\leq \Delta_k^2(M) + n\eta^2 + 4n\eta\sqrt{T} \qquad (10)
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz and the second is by assuming the worst case bound $\|m_l - c_i^M\| \leq \max_{i\neq j}\|m_i - m_j\| \leq \|\mathbf{1}_T - (-\mathbf{1}_T)\| = 2\sqrt{T}$, where $\mathbf{1}_T = (1,1,1,\ldots,1,1)$ denotes a vector of length T with all elements being 1.

Now, we compare the bound in $k-1$ clusters. Let $P^{\tilde{M},k-1}$ be the optimal $(k-1)$-means partition for $\tilde{M}$ and

let $\{c_i^{\tilde{M},k-1}\}_{i=1}^{k-1}$ be the corresponding centers. Then,

$$\begin{aligned}
\Delta_{k-1}^2(M) &\le \Delta_{k-1}^2(M; P^{\tilde{M},k-1}) \\
&= \sum_{i\in[k-1]} \sum_{l\in P_i^{\tilde{M},k-1}} \|m_l - c_i^{\tilde{M},k-1}\|^2 \\
&= \sum_{i\in[k-1]} \sum_{l\in P_i^{\tilde{M},k-1}} \|m_l - \tilde{m}_l + \tilde{m}_l - c_i^{\tilde{M},k-1}\|^2 \\
&\le \sum_{i\in[k]} \sum_{l\in P_i^{\tilde{M},k-1}} \|\tilde{m}_l - c_i^{\tilde{M},k-1}\|^2 + \|m_l - \tilde{m}_l\|^2 + 2\|m_l - \tilde{m}_l\| \cdot \|\tilde{m}_l - c_i^{\tilde{M},k-1}\| \\
&\le \Delta_{k-1}^2(\tilde{M}) + n\eta^2 + 2n\eta \max_{i\neq j} \|\tilde{m}_i - \tilde{m}_j\| \\
&\le \Delta_{k-1}^2(\tilde{M}) + n\eta^2 + 2n\eta \left(\max_{i\neq j} \|m_i - m_j\| + 2\eta\right) \\
&\le \Delta_{k-1}^2(\tilde{M}) + n\eta^2 + 4n\eta(\sqrt{T} + \eta) \qquad (11)
\end{aligned}$$

the first inequality is by Cauchy-Schwarz, and the second inequality is by $\|\tilde{m}_l - c_i^{\tilde{M},k-1}\| \le \max_{i\neq j} \|\tilde{m}_i - \tilde{m}_j\|$, and the third inequality is by triangle inequality: $\|\tilde{m}_i - \tilde{m}_j\| \le \|\tilde{m}_i - m_i\| + \|m_i - m_j\| + \|m_j - \tilde{m}_j\| \le \|m_i - m_j\| + 2\eta$.

Combining inequalities in (10) and (11), we have

$$\begin{aligned}
\Delta_k^2(\tilde{M}; c_i^M) &\le \Delta_k^2(M) + n\eta^2 + 4n\eta\sqrt{T} \\
&\le \Delta_{k-1}^2(\tilde{M}) + n\eta^2 + 4n\eta(\sqrt{T} + \eta) + n\eta^2 + 4n\eta\sqrt{T} \\
&\le \Delta_{k-1}^2(\tilde{M}) + 2n\eta(3\eta + 4\sqrt{T}) \\
&\le \varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M}) + 2n\eta(3\eta + 4\sqrt{T}) \qquad \text{(Lemma 5.5)}
\end{aligned}$$

Now, we want to make this upper bound to $2\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})$, so that we conclude that $\tilde{M}$ is εL^2-separated. To do so, we want to bound $2n\eta(3\eta + 4\sqrt{T})$ by $\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})$. By Lemma 5.4, in the event of G, we have

$$2n\eta(3\eta + 4\sqrt{T}) \le \frac{4ns(\sqrt{n} + \sqrt{T})}{\delta} \left(\frac{6s(\sqrt{n} + \sqrt{T})}{\delta} + 4\sqrt{T}\right).$$

So the goal is to find a condition on s such that

$$\frac{4ns(\sqrt{n} + \sqrt{T})}{\delta} \left(\frac{6s(\sqrt{n} + \sqrt{T})}{\delta} + 4\sqrt{T}\right) \le \varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M}).$$

By rearranging terms to get a quadratic form in s,

$$\begin{aligned}
s\left(\frac{6s(\sqrt{n} + \sqrt{T})}{\delta} + 4\sqrt{T}\right) &\le \frac{\delta\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{4n(\sqrt{n} + \sqrt{T})} \\
\frac{6(\sqrt{n} + \sqrt{T})}{\delta} s^2 + 4\sqrt{T}s - \frac{\delta\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{4n(\sqrt{n} + \sqrt{T})} &\le 0
\end{aligned}$$

Define this quadratic formula as $f(s) := \frac{6(\sqrt{n}+\sqrt{T})}{\delta} s^2 + 4\sqrt{T}s - \frac{\delta\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{4n(\sqrt{n}+\sqrt{T})}$. Since $\frac{6(\sqrt{n}+\sqrt{T})}{\delta} > 0$ and $f(s = 0) < 0$, we can find the condition on s that makes $f(s) \le 0$ by finding the positive solution for $f(s) = 0$. Using

the quadratic formula, we get

$$\begin{aligned} s &= \frac{-4\sqrt{T} \pm \sqrt{16T + \frac{24(\sqrt{n}+\sqrt{T})}{\delta}\frac{\delta\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{4n(\sqrt{n}+\sqrt{T})}}}{12(\sqrt{n}+\sqrt{T})/\delta} \\ &= \frac{\delta\left(-4\sqrt{T} \pm \sqrt{16T + \frac{6\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{n}}\right)}{12(\sqrt{n}+\sqrt{T})} \\ &= \frac{\delta\left(-\sqrt{16T} \pm \sqrt{16T + \frac{6\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{n}}\right)}{12(\sqrt{n}+\sqrt{T})}. \end{aligned}$$

Therefore, $2n\eta(3\eta + 4\sqrt{T}) \leq \varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})$ is satisfied for all

$$s \in \left(0, \frac{\delta\left(-\sqrt{16T} + \sqrt{16T + \frac{6\varepsilon^2 L^4 \Delta_{k-1}^2(\tilde{M})}{n}}\right)}{12(\sqrt{n}+\sqrt{T})}\right).$$

Finally, using the fact that $\tilde{M}$ is $4L^2\varepsilon$-separated in the event of G, we can use Theorem B.1 to conclude that $|\tilde{P} \ominus P^{\tilde{M}}| \leq 644L^4\varepsilon^2 n$.

□

D.11 Proof of Lemma C.7

Lemma C.7. *If* $s \leq \frac{\delta\left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})}$, *then in the event of* G, $|\tilde{P} \ominus P^M| \leq 2L^2\varepsilon n$

Proof. Define

$$core(P_i^M) = \left\{l \in P_i^M : \|m_l - c_i^M\| \leq \sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\min_{i\neq j}\|c_i^M - c_j^M\|\right\}.$$

Similar to Lemma C.3, we construct a uniform probability distribution over $\forall l \in P_i^M$:

$$\mathbb{E}[\|m_l - c_i^M\|^2] = \frac{1}{|P_i^M|}\sum_{l\in P_i^M}\|m_l - c_i^M\|^2 = r_i^2(M)$$

Usking Markov's inequality with the fact that M is $L^2\varepsilon$-separated,

$$\mathbb{P}(\|m_l - c_i^M\|^2 \geq t) \leq \frac{r_i^2(M)}{t} \leq \frac{\frac{L^4\epsilon^2}{1-L^4\epsilon^2}\min_{j\neq i}\|c_i^M - c_j^M\|^2}{t}$$

Take $t = \frac{r_i^2(M)}{L^2\varepsilon} = \frac{L^2\varepsilon}{1-L^4\epsilon^2}\min_{j\neq i}\|c_i^M - c_j^M\|^2$. Then,

$$\mathbb{P}\left(\|m_l - c_i^M\| \geq \sqrt{t}\right) = \mathbb{P}\left(m_i \notin core(P_i^M)\right) \leq L^2\varepsilon$$

Hence, we have at most $L^2\varepsilon$ faction of points l in P_i^M that will not belong to $core(P_i^M)$.

Similar to Lemma C.4, for all $l \in core(P_i^M)$ and for all $j \neq i$, we have

$$\|m_l - c_j^M\| - \|m_l - c_i^M\| \geq \left(1 - 2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|. \tag{12}$$

We will assume that there exists j such that

$$\|\tilde{m}_l - c_j^M\| \leq \|\tilde{m}_l - c_i^M\|,$$

and show a contradiction to conclude that all $l \in core(P_i^M)$ should also belong to $\tilde{P}_i$. From this assumption, we apply triangle inequality twice to yield

$$\|m_l - c_j^M\| - \|\tilde{m}_l - m_l\| \leq \|m_l - c_i^M\| + \|\tilde{m}_l - m_l\|.$$

Rearranging, we have

$$\|m_l - c_j^M\| - \|m_l - c_i^M\| \leq 2\|\tilde{m}_l - m_l\| \leq 2\eta,$$

and in the event of G, we have the bound on η (Lemma 5.4):

$$\|m_l - c_j^M\| - \|m_l - c_i^M\| \leq \frac{4s(\sqrt{n}+\sqrt{T})}{\delta}.$$

With the assumption that $s < \frac{\delta\left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})}$, this bound becomes

$$\|m_l - c_j^M\| - \|m_l - c_i^M\| < \left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|,$$

which contradicts the inequality in (12). Hence we conclude $core(P_i^M) \subseteq \tilde{P}_i$. Recall that $|core(P_i^M)| \geq (1 - L^2\varepsilon)|P_i^M|$. Therefore, we have $|\tilde{P} \ominus P^M| = \sum_{i=1}^k |\tilde{P}_i \ominus P_i^M| \leq 2L^2\varepsilon n$. □

D.12 Proof of Lemma 5.6

Lemma 5.6. *For $\varepsilon < 0.1$ and $\delta \in (0,1)$, if $L^2\epsilon < 1/\sqrt{801}$, $\min_i r_i(M) \geq 1/64$, and $s < O(\frac{\delta\sqrt{T}}{\sqrt{n}+\sqrt{T}})$, then, in the event of G, $\sum_{i=1}^k |P_i^M \ominus P_{\sigma(i)}^{\tilde{M}}| \leq 25L^2\varepsilon^2 n$ for some bijection $\sigma(i)$.*

Proof. To instantiate Lemmas C.6 and C.7, we need to assume

$$s < \min\left\{\frac{\delta\left(-\sqrt{16T}+\sqrt{16T+\frac{6\varepsilon^2L^4\Delta_{k-1}^2(\tilde{M})}{n}}\right)}{12(\sqrt{n}+\sqrt{T})}, \frac{\delta\left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})}\right\}.$$

Since $\tilde{M}$ is $4L^2\varepsilon$-separated in the event G,

$$\Delta_k^2(\tilde{M}) \leq 4\varepsilon^2L^4\Delta_{k-1}^2(\tilde{M}).$$

Using this, we suggest a slightly stronger bound on s:

$$s < \min\left\{\frac{\delta\left(-\sqrt{16T}+\sqrt{16T+\frac{3\Delta_k^2(\tilde{M})}{2n}}\right)}{12(\sqrt{n}+\sqrt{T})}, \frac{\delta\left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})}\right\}.$$

Define $s_1 := \frac{\delta\left(-\sqrt{16T}+\sqrt{16T+\frac{3\Delta_k^2(\tilde{M})}{2n}}\right)}{12(\sqrt{n}+\sqrt{T})}$ and $s_2 := \frac{\delta\left(1-2\sqrt{\frac{L^2\varepsilon}{1-L^4\varepsilon^2}}\right)\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})}$. We will simplify this bound by showing that s_2 (the left element in minimum) is asymptotically smaller than s_2 (the right element).

First, we show an upper bound of s_1. Note that $\sqrt{x+y}-\sqrt{x} \leq \frac{y}{2\sqrt{x}}$, $\forall x, y > 0$ because the second derivative is always negative. Hence,

$$\begin{aligned} s_1 &\leq \frac{\delta}{\sqrt{n}+\sqrt{T}}\frac{\Delta_k^2(\tilde{M})}{64n\sqrt{T}} \\ &\leq \frac{\delta}{\sqrt{n}+\sqrt{T}}\frac{2n\sqrt{T}}{64n\sqrt{T}} && (\Delta_k^2(\tilde{M}) \leq 2n\sqrt{T}) \\ &= \frac{\delta}{\sqrt{n}+\sqrt{T}}\frac{1}{32}. \end{aligned}$$

Then, we analyze a lower bound for the second element in the minimum.

$$\begin{aligned}
s_2 &\geq \frac{\delta(1-\frac{\sqrt{89}}{267})\min_{i\neq j}\|c_i^M - c_j^M\|}{4(\sqrt{n}+\sqrt{T})} && (L^2\varepsilon < 1/\sqrt{801}) \\
&\geq \frac{\delta(1-2\frac{\sqrt{89}}{267})\sqrt{\frac{1-\varepsilon^2}{\varepsilon^2}\min_i r_i^2(M)}}{4(\sqrt{n}+\sqrt{T})} && \text{(Lemma 5.2)} \\
&\geq \frac{\delta(1-2\frac{\sqrt{89}}{267})\sqrt{99}\min_i r_i(M)}{4(\sqrt{n}+\sqrt{T})} && (\varepsilon < 0.1) \\
&\geq \frac{2.311\cdot\delta\min_i r_i(M)}{(\sqrt{n}+\sqrt{T})} \\
&\geq \frac{\delta}{\sqrt{n}+\sqrt{T}}\,2\cdot\min_i r_i(M).
\end{aligned}$$

With our assumption $\min_i r_i(M) \geq 1/64$, we have

$$s_1 \leq 1/32 \leq 2\cdot\min_i r_i(M) \leq s_2$$

Hence, by assuming $s < s_1 \leq \frac{\delta}{\sqrt{n}+\sqrt{T}}\frac{\Delta_k^2(\tilde{M})}{64n\sqrt{T}} = O(\frac{\delta\sqrt{T}}{\sqrt{n}+\sqrt{T}})$, we satisfy the constraint for s.

Instantiating Lemmas C.6 and C.7 yeild

$$\begin{aligned}
|P^M \ominus P^{\tilde{M}}| &\leq |P^M \ominus \tilde{P}| + |\tilde{P} \ominus P^{\tilde{M}}| && \text{(Triangle inequality)} \\
&\leq 2L^2\epsilon n + 644L^4\epsilon^2 n && \text{(Lemmas C.6 and C.7)} \\
&\leq 2L^2\epsilon n + 23L^2\epsilon n && (L^2\varepsilon < 1/\sqrt{801}) \\
&= 25L^2\epsilon n.
\end{aligned}$$

□

E Omitted Proofs from Section 5.2

E.1 Different Noise Setting

In the main body of the paper, we present the proof assuming Gaussian noise. In this section, we present different noise assumptions (sub-gaussian and heavy-tailed distributions). In Section **??**, the bound of $\sigma_X^* - \sigma_A^*$ is used to show improvement in SC performance from using ClusterSC in the presence of Gaussian noise terms; if needed, one could instead adopt Corollary E.2 and Corollary E.4 depending on the relevant assumptions about the noise distributions (sug-Gaussian or heavy-tailed) in a given application.

E.1.1 Sub-Gaussian Noise Setting

We consider the the sub-gaussian noise setting (Definition E.1) where $\|E_{i,t}\|_{\psi_2} = K$. Our result in this setting, Corollary E.2, follows from Theorem 5.9 by instantiating Theorem E.1 from Vershynin (2010) instead of Gordon's theorem.

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

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

Theorem E.1 (Theorem 5.39 of Vershynin (2010))**.** *Let A be an $N\times n$ matrix whose rows A_i are independent sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for every $t\geq 0$, with probability at least $1-2\exp(-ct^2)$ one has*

$$\sqrt{N} - C\sqrt{n} - t \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{N} + C\sqrt{n} + t.$$

Here $C = C_K$, $c = c_K > 0$ depend only on the subgaussian norm $K = \max_i \|A_i\|_{\psi_2}$ of the rows.

Corollary E.2 (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* $K = \|E_{i,t}\|_{\psi_2}$*.*

E.1.2 Heavy-tailed Noise Settings

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. Using the bound in Theorem E.3 on maximum and minimum singular values of the noise matrix, we can draw a lower bound on the gap of singular values for heavy-tailed distributions. This is used in place of Theorem E.1 or Gordon's theorem to prove Corollary E.4.

Theorem E.3 (Theorem 5.41 of Vershynin (2010))**.** *Let* A *be an* $N \times n$ *matrix whose rows* A_i *are independent isotropic random vectors in* $\mathbb{R}^n$*. Let* m *be a number such that* $\|A_i\|_2 \leq \sqrt{m}$ *almost surely for all* i*. Then for every* $t \geq 0$*, one has*

$$\sqrt{N} - t\sqrt{m} \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{N} + t\sqrt{m}$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$*, where* $c > 0$ *is an absolute constant.*

Corollary E.4 (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* $1 - 2Te^{-ct^2}$*,*

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

E.2 Proof of Theorem 5.9

Theorem 5.9 (Singular Value Concentration with Gaussian Noise)**.** *Let the noise terms be sampled* $E_{i,t} \sim \mathcal{N}(0, s^2)$*. 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 (Theorem B.5), 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 $X^\top X$:

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

The first inequality is from Weyl's inequality on eigenvalues (Theorem B.6) and the second inequality is due to $\lambda_{r+1}(M^\top M) = \lambda_T(2M^\top E) = 0$. By taking the expectation on both sides, we reach the last inequality by Gordon's theorem, which says that $\mathbb{E}[\sigma_T(E)] \geq s(\sqrt{n} - \sqrt{T})$. 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 get the desired difference, and see that the lower bound on σ_X^* is greater than the upper bound on σ_A^* when $n_A < n + 4T - 4\sqrt{nT}$. □

E.3 Proof of Lemma 5.10

Lemma 5.10 (Pre-intervention MSE of SC)**.** *Given donor matrix $X \in \mathbb{R}^{n\times T}$, target unit $x_0 = m_0 + e_0$, 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, then,*

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

Proof. From Lemma B.10, presented in Appendix B.3,

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

We bound the first term inside the expectation by

$$\|(M^- - \hat{M}^-)^\top f^*\|^2 \le \|M^- - \hat{M}^-\|^2 \|f^*\|^2, \tag{13}$$

using the property of the operator norm: $\|Ax\| \le \|A\| \cdot \|x\|$ for any matrix A and vector x. We bound the first term of (13) by

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

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

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

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

E.4 Proof of Theorem 5.11

Theorem 5.11. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on pre-intervention MSE of ClusterSC (Algorithm 4) is strictly smaller than that of classical SC, 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 5.10 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

$$\begin{aligned}\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 &= \sigma_X^{*2} + 4s(\sqrt{n}+\sqrt{T})\sigma_X^* + 4s^2(\sqrt{n}+\sqrt{T})^2 \\ &= \sigma_X^{*2} + 4s(\sqrt{n}+\sqrt{T})\sigma_X^* + 4s^2 n + 8s^2\sqrt{nT} + 4s^2 T.\end{aligned}$$

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

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

Then, the difference between the two becomes

$$\begin{aligned}&\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 - \left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2 \\ &= \sigma_X^{*2} - \sigma_A^{*2} + 4s(\sqrt{n}+\sqrt{T})(\sigma_X^* - \sigma_A^*) - (1-\alpha)4s\sqrt{n}\sigma_A^* + (1-\alpha^2)4s^2 n + (1-\alpha)8s^2\sqrt{nT} \\ &= (\sigma_X^* + \sigma_A^* + 4s(\sqrt{n}+\sqrt{T}))(\sigma_X^* - \sigma_A^*) - (1-\alpha)4s\sqrt{n}\sigma_A^* + (1-\alpha^2)4s^2 n + (1-\alpha)8s^2\sqrt{nT}\end{aligned} \tag{14}$$

Let Δ denote the quantity in Equation (14), i.e., $\Delta \coloneqq \left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 - \left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2$. To lower bound the expectation of Δ, we apply Theorem 5.9:

$$\begin{aligned}
\mathbb{E}[\Delta] &\geq (2\sigma_A^* + s(5-\alpha)\sqrt{n} + 2s\sqrt{T})((1-\alpha)s\sqrt{n} - 2s\sqrt{T}) \\
&\quad - 4(1-\alpha)s\sqrt{n}\sigma_A^* + 4(1-\alpha^2)s^2 n + 8(1-\alpha)s^2\sqrt{nT} \\
&= 2(1-\alpha)s\sqrt{n}\sigma_A^* - 4(1-\alpha)s\sqrt{n}\sigma_A^* - 4s\sqrt{T}\sigma_A^* + s^2(5-\alpha)(1-\alpha)n + 2s^2(1-\alpha)\sqrt{nT} \\
&\quad - 2s^2(5-\alpha)\sqrt{nT} - 4s^2T + (1-\alpha^2)4s^2 n + (1-\alpha)8s^2\sqrt{nT} \\
&= -2(1-\alpha)s\sqrt{n}\sigma_A^* - 4s\sqrt{T}\sigma_A^* + s^2((9-6\alpha-3\alpha^2)n - 4T - 8\alpha^2\sqrt{nT}) \\
&= -s\left(2(1-\alpha)\sqrt{n} + 4\sqrt{T}\right)\sigma_A^* + \underbrace{3(\alpha+3)(1-\alpha)s^2 n}_{\text{dominating term}} - \left(4s^2T + 8s^2\alpha^2\sqrt{nT}\right).
\end{aligned}$$

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, $\mathbb{E}[\Delta] = \Omega(s^2 n)$.

Finally, The difference in the two upper bounds is

$$\begin{aligned}
&\frac{\mu^2}{T_0}\mathbb{E}[\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2] + \frac{2s^2 r}{T_0} - \frac{\mu^2}{T_0}\mathbb{E}[\left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2] - \frac{2s^2 r_S}{T_0} \\
&= \frac{\mu^2}{T_0}\mathbb{E}[\left(\sigma_X^* + 2s(\sqrt{n}+\sqrt{T})\right)^2 - \left(\sigma_A^* + 2s(\alpha\sqrt{n}+\sqrt{T})\right)^2] + \frac{2s^2}{T_0}(r - r_S) \\
&= \Omega(s^2 n),
\end{aligned}$$

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

E.5 Proof of Lemma 5.12

Lemma 5.12 (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,* $\mathrm{RMSE}(\hat{m}_0^+; X)$ *is upper bounded by*

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

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.9 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,

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

E.6 Proof of Theorem 5.13

Theorem 5.13. *If $n_A < n + 4T - 4\sqrt{nT}$, then the upper bound on post-intervention RMSE of ClusterSC (Algorithm 4) is strictly smaller than that of classical SC, 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 5.12. 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

$$\begin{aligned}
&\frac{\eta}{\sqrt{T-T_0}}\mathbb{E}[\sigma_X^* - \sigma_A^* + 2s(1-\alpha)\sqrt{n}] + (1-\alpha)\sqrt{n}(\mu+\eta) \\
&\geq \frac{\eta}{\sqrt{T-T_0}}\left(s((1-\alpha)\sqrt{n} - 2\sqrt{T}) + 2s(1-\alpha)\sqrt{n}\right) + (1-\alpha)\sqrt{n}(\mu+\eta) \\
&= \frac{\eta}{\sqrt{T-T_0}}\left(s(3(1-\alpha)\sqrt{n} - 2\sqrt{T})\right) + (1-\alpha)\sqrt{n}(\mu+\eta),
\end{aligned}$$

where the first inequality comes from Theorem 5.9. Since $n \gg T$ and μ and η are constants, we conclude that the difference is $\Omega(s\sqrt{n})$. □

F Additional Experiments with Synthetic Datasets

In this section, we share more detailed results with synthetic simulated datasets. Appendix F.1 shows the performance of ClusterSC with OLS and Ridge regression (via Robust Synthetic Control). Appendix F.2 presents an analysis of ClusterSC with Lasso regression.

F.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 against our ClusterSC. The number of distinct signals in each submatrices A and B is $r_A = r_B = 3$, and the ridge coefficient was fixed to 0.01.

Figure 5 shows the average MSE per dataset over varying noise levels (s), 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).

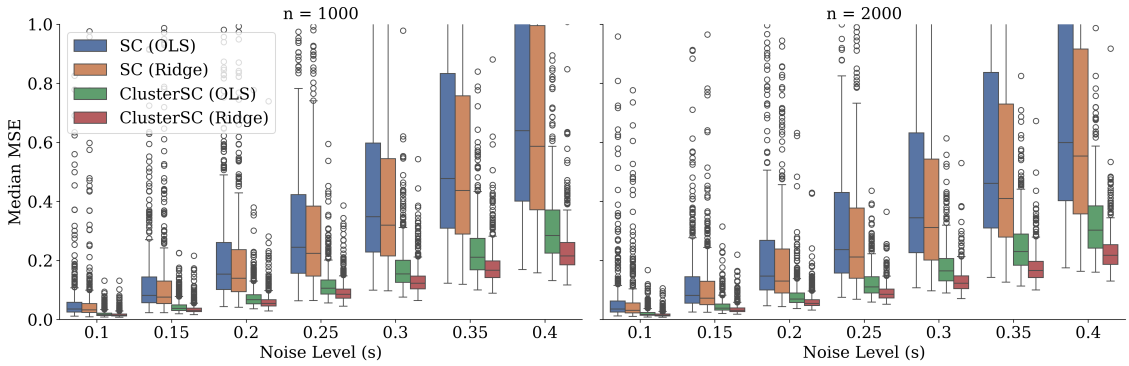

Figure 5: Median post-intervention MSE, measured per dataset. Each boxplot corresponds to ridge, OLS, cluster and then ridge, and cluster and then OLS, from left to right, plotted for each noise level. Left plot is with $n = 1000$ donor units in total and the right plot is with $n = 2000$.

Next, we define the pairwise improvement for a target i as the difference in post-intervention MSE scores between the two methods: $I_i = \text{MSE}(\hat{m}_i^+; X) - \text{MSE}(\hat{m}_i^+; A)$. Then, we take median$(I_i)$ as a metric to assess the overall

improvement measured from n SC instances constructed from one dataset (leave-one-out placebo test). Figure 6 shows the pairwise improvement (i.e., median(I_i)) induced by ClusterSC at varying noise level. We observe that the median improvement is almost always positive, meaning that more than half of the individuals benefit from using ClusterSC instead of RSC. In the $n = 2000$ case, the improvement grows as noise increases, corroborating our Theorem 5.13. On the other hand, when $n = 1000$, the improvement continues to increase until $s = 0.35$, after which it plateaus at $s = 0.4$ for Ridge and decreases for OLS. This may be attributed to the noise level $s = 0.4$ being sufficiently high that the donor $n = 1000$ is not enough to effectively capture the true signal. Nonetheless, a significant improvement is observed in median MSE for the same setting from Figure 5. The improvement is more stable (low variance) with higher n, and the improvement in OLS and ridge regression is not too different.

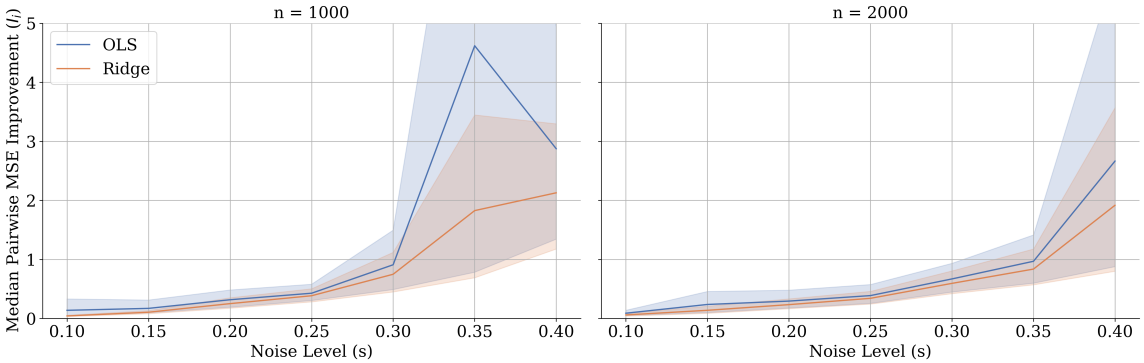

Figure 6: Median of the pair-wise improvements measured by comparing SC and ClusterSC, when using OLS (blue) and Ridge regression (orange), over different noise levels.

F.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 with the same parameters $n_A = n_B \in \{500, 1000\}$, $T = 10$, and $r_A = r_B = 3$. The Lasso coefficient was 0.01 for all experiments. Due to the high computational cost of Lasso, we only test for noise levels $s \in \{0.1, 0.2, 0.3, 0.4\}$.

Figure 7 shows the average post-intervention MSE. We observe more improvement with clustering as noise level increases. Compared to the results in Figure 5, the improvement induced by the clustering step when regression is performed with Lasso is not as large as compared to OLS or Ridge in absolute value. Still, the improvement is evident.

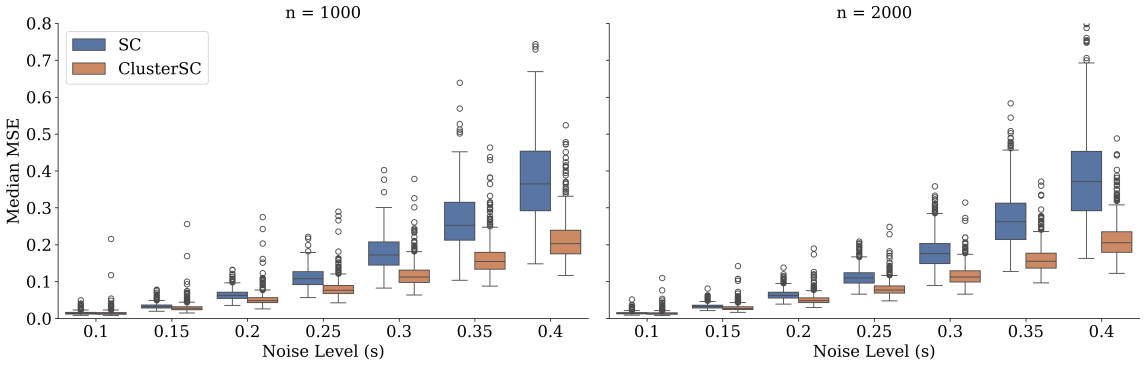

Figure 7: Median post-intervention MSE, measured per dataset. We compare using SC with Lasso (blue) and ClusterSC with Lasso (orange).

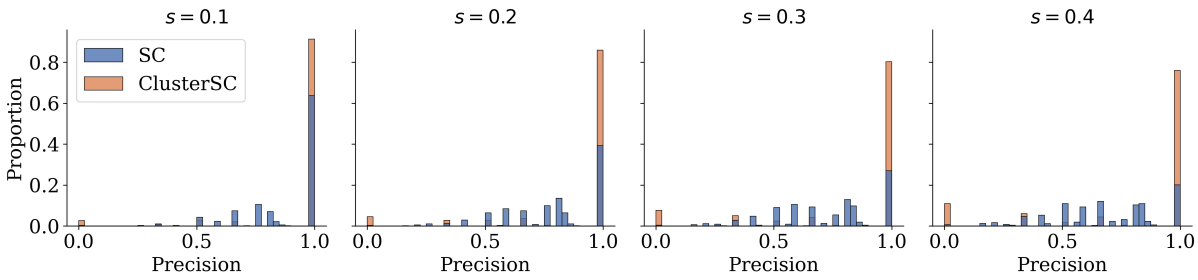

Figure 8: Histogram of the precision score of active donor units from using SC with Lasso (blue) and ClusterSC with Lasso (orange). 100 iterations are displayed for $n = 1000$, each with 150 leave-one-out scores.

Analysis on Active Donors. 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 correctly belong to group A.

Figure 8 illustrates the distribution of precision scores for active donor units selected by SC with Lasso (blue) and ClusterSC with Lasso (orange). For all noise regimes, ClusterSC shows more concentrated precision score around 1 compared to SC. As noise increases, this difference in concentration increases further, indicating that ClusterSC is more robust to noise in selecting relevant donors.

For SC, a small proportion of donors from group B are incorrectly included, with their precision scores spread relatively evenly across all score ranges. When ClusterSC fails to achieve high precision, the scores tend to be near 0, rather than being somewhere in between 0 and 1. This pattern suggests that ClusterSC selected an incorrect cluster (e.g., the selected donor set predominantly consists of group B), and hence it can only choose the donors from B. Nonetheless, incorrect selection of donor set does not occur frequently, which agrees with ClusterSC's decreased median MSE compared to SC (see Figure 7).

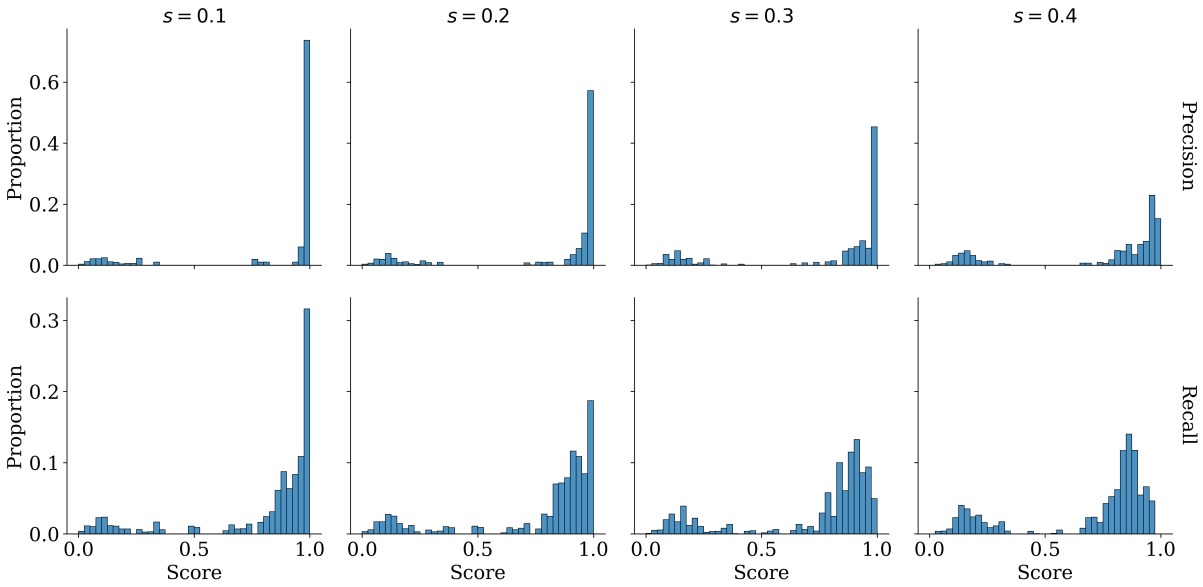

Figure 9: Histograms of precision and recall scores of donor units selected by ClusterSC compared to units in A, over varying noise levels. 100 iterations are displayed for $n = 1000$, each with 150 leave-one-out scores.

Analysis on Clusters. We can further analyze this improvement by investigating how effectively the clustering step (k-means) selects a relevant donor set by computing its precision and recall scores with units in group A (from Step 4 of Algorithm 4) as the true label. Figure 9 presents histograms of precision scores (top row) and recall scores (bottom row) for varying noise levels. A precision score close to 1 indicates that most of the donors selected by ClusterSC are already from the relevant group A, making it easier for Lasso to select the best fit among them. In the low noise regimes, the precision score is 1 for more than 70% of the cases, and the regression step does not need to filter it any further. However, the ability of ClusterSC to select only the relevant donors (high precision score) degrades as noise increases. ClusterSC, together with the power of Lasso to learn sparse weights, can significantly improve the precision scores in high noise regimes, from the first row of Figure 9 to the orange plots in Figure 8. In contrast, without clustering, Lasso must filter out irrelevant donors from group B solely through the power of regularization in the regression step, which is shown in blue bars in Figure 8. This shows that ClusterSC with Lasso has a synergistic effect for only selecting relevant donors, improving from using only Lasso (Figure 8) or clustering (Figure 9) individually.

Computational Efficiency. Another advantage that ClusterSC brings in with Lasso regression is computational efficiency. Like Lasso, ClusterSC seeks to isolate only the most important donors for target reconstruction. Unlike Lasso, ClusterSC avoids running a linear regression in the full n dimensions. ClusterSC comprises of three main parts: 1) SVD on the original matrix, 2) clustering, and 3) regression (on the subsampled donor). We recall the computational complexity of each of these steps.

Lemma F.1 (Theorem 4.16, (Ostrovsky et al., 2013)). *Fix any $\omega > 0$ and a dataset $X \in \mathbb{R}^{n\times d}$. Assuming $\Delta_k^2(X) \leq \varepsilon^2 \Delta_{k-1}^2(X)$ for ε small enough, there is an algorithm which, with constant probability, outputs a partition $\hat{P}$ that is $(1+\omega)$-optimal solution to k-means on X, meaning $\Delta_k^2(X;\hat{P}) \leq (1+\omega)\Delta_k^2(X)$. Furthermore, this algorithm runs in time $O(2^{O(k(1+\varepsilon^2)/\omega)}nd)$.*

Lemma F.2 (Golub and Van Loan (2013), Figure 5.5.1 in p. 263). *Computing the singular value decomposition for a dense $m \times n$ matrix takes time $O(mn\min\{m,n\})$.*

Lemma F.3 (Efron et al. (2004), Section 7). *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 for regression purposes is actually the number of donors n, as the goal is to predict the behavior of the target donor per-time-step. Suppose we use a constant-factor k-means approximation for ClusterSC (i.e., the algorithm will correctly identify clusters for all but ω fraction of points with $\omega = \Omega(1)$). We compare the runtime of synthetic control with Lasso versus ClusterSC with Lasso in terms of n (number of all donor units), n_A (number of donors selected by ClusterSC), and m (the number of targets we test). Note the we do not consider T as we assume a tall matrix ($n \gg T$).

For ClusterSC (Algorithm 4), the major computations will be:

- Step 1. Learn clusters: $O(n) + O(n)$ (Lemmas F.2 and F.1)
- Step 3. Construct donor matrix A and denoise: $O(n_A)$ (Lemma F.2)
- Step 4. (m rounds of) SC Learning: $m \cdot O(n_A^3)$ (Lemma F.3)

Note that we only n and n_A to grow (where $n > n_A$), but not T. Hence, considering runtime in terms of parameters n, n_A, and m, the time complexity of ClusterSC is $O(n + mn_A^3)$. On the other hand, the classical synthetic control with Lasso will have time complexity of $O(mn^3)$.

G Additional Details about the Housing Dataset

We provide more insights into the housing price index dataset used in Section 6.2. Figure 10 provides a graphical summary of this time series panel dataset.

To further understand the importance of the singular value cutoff, we plot the singular value spectrum of the entire dataset in Figure 11. Note that in every iteration with different train test split, we recompute SVD for the donor matrix, so the spectrum may not be exactly the same across iterations. However, this does still give us an understanding of the dataset, which we assume to be approximately low rank. On the left side, we plot

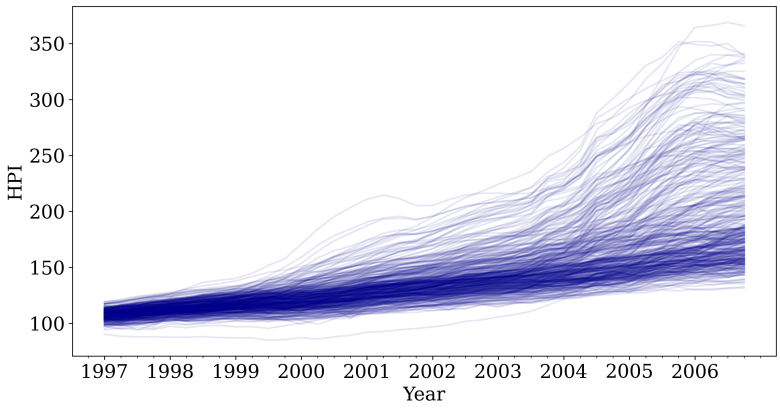

Figure 10: Full time series plot of cleaned housing price index (HPI) dataset with $n = 400$ metropolitan areas and $T = 40$ quarters from 1997 to 2006.

the cumulative singular value ratio, which shows that the top three or four singular values contain about 95% of the total singular values. On the right side, we can see the gap in singular values decreases as we increase the index, which shows that the dataset does satisfy the assumption of approximately low-rank structure.

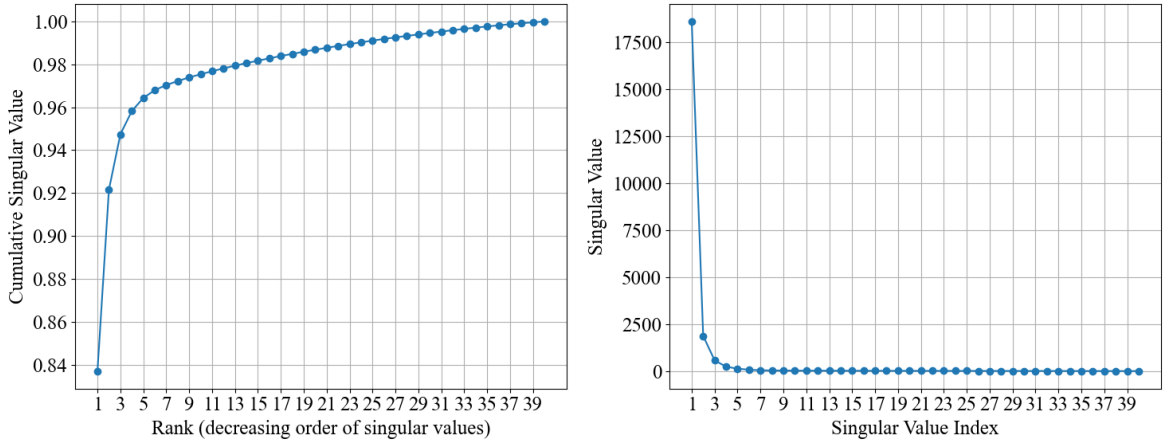

Figure 11: Cumulative singular value ratio (left) and the singular value spectrum (right), ordered by decreasing singular values.