Synthetic Principal Component Design: Fast Covariate Balancing with Synthetic Controls

Yiping lu
ICME
Stanford University
yplu@stanford.edu

Jiajin Li
Management Science and Engineering (MS&E)
Stanford University
jiajinli@stanford.edu

Lexing Ying
Department of Mathematics
Stanford University
lexing@stanford.edu

Jose Blanchet
Management Science and Engineering (MS&E)
Stanford University
jose.blanchet@stanford.edu

Abstract

In this paper, we target at developing a globally convergent and yet practically tractable optimization algorithm for the optimal experimental design problem with synthetic controls. Specifically, we consider a setting when the pre-treatment outcome data is available. the average treatment effect is estimated via the difference between the weighted average outcomes of the treated and control units, where the weights are learned from the data observed during the pre-treatment periods. We find that if the experimenter has the ability to select an optimal set of non-negative weights, the optimal experimental design problem is identical to to a so-called *phase synchronization* problem. We solve this problem via a normalized variate of the generalized power method with spectral initialization. On the theoretical side, we establish the first global optimality guarantee for experiment design under a realizable assumption with linear fixed-effect models (also referred to an "interactive fixed-effect model"). These results are surprising, given that the optimal design of experiments, especially involving covariate matching, typically involves solving an NP-hard combinatorial optimization problem. Empirically, we apply our algorithm on US Bureau of Labor Statistics and the Abadie-Diemond-Hainmueller California Smoking Data. The experiments demonstrate that our algorithm surpasses the random design with a large margin in terms of the root mean square error.

1 Introduction

Estimating the average effects of a binary treatment is one of the main goals of empirical economic and political studies. Randomization in controlled trials is one of the golden rules for estimating average treatments effects (ATE). If the treatment assignment procedure guarantees that the potential outcomes are independent of the treatment status, then a simple difference-in-mean (i.e., average outcomes of the treated and control units) estimator becomes an unbiased estimator of ATE. Nevertheless, a completely randomized experiment may be affected by a significantly high variance in the final estimation. Such variance can be reduced by taking advantage of features in the observed data. In this paper, we focus on the following question *can the observed covariates improve the statistical properties of the ATE estimators via experimental designing?* [1, 2]. This problem is referred to as *covariate balancing* which restricts the randomization to achieve covariate balance between treatment groups [3, 4].

Covariate balancing has been substantially explored in the literature. However, an NP-hard combinatorial optimization problem [5, 6, 7] is always needed to be solved when balancing the covariance.

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

In this paper, following [6, 7], we consider the experiment design problem when a synthetic control [8, 9, 10] esitmaotr is implemented. That is, the experiment designer can observe the pre-treatment panel outcome data for a number of units in a number of time periods. Synthetic control compares treated units with a weighted average of untreated units. The weights are determined via empirical fit on the observed pre-treatment outcome. [6, 7] propose an optimization approach to select the control group based on the observed pre-treatment outcome. Namely, the choice of the treated units aims to balance the weighted average of treated and untreated covariates. As such, the designer can choose the best non-negative weights. [6] prove that the underlying optimization problem is still NP-hard and [7] relax the optimization problem into a canonical Quadratic Constraint Quadratic Program (QCQP). Nevertheless, the resulting QCQP is rather computationally demanding and applicable algorithms are not guaranteed to reach a global optimum.

In this paper, we remove the constraints of the number of units being treated in [6] and surprisingly find out the optimization problem can be recast as a phase synchronization problem [11]. Although phase synchronization is still an NP-hard problem [12], recent works [13, 14, 15, 16] have figured out that the problem is polynomial-time solvable under certain data generating process. Motivated by this line of research, we propose *Synthetic principal component Design* (SPCD), which optimizes the treatment decision via (a normalized variate of) the generalized power method with spectral initialization [17]. Under the realizable assumption in [7], we provide a *global* optimization guarantee of a normalized variate of the generalized power method and statistical estimation guarantee of the synthetic control procedure for the linear fixed-effect model studied [9, 18, 19, 20]. To the best of our knowledge, this is the first formulation of combinatorial optimization-based experiment design which enjoys a global optimization guarantee.

1.1 Contribution

- We show an equivalence between the experiment design with synthetic control [6, 7] and phase synchronization problem [11, 14, 15], where separating the experiment and control group can be transformed to finding the phase of a complex signal. Based on this observation, we linked covariate balancing with the smallest eigenvector of gram matrix and utilize a spectral method for fast experiment design.
- We proposed a novel normalized version of generalized power method which enjoys *global* convergence results under certain generative models. The normalization technique needs weaker generative models assumption for global results and consistently improves the empirical results.
- Our method surpass random design a large margin empirically on both synthetic and real world dataset. Our method even exceed 500000 times of rerandomization over the Abadie-Diamond-Hainmueller smoking legislation data.

2 Problem Setup

In this section, we follow the setting of synthetic control (SC) and corresponding experiment design introduced in [6]. We aim to estimate the effect of a binary treatment under the panel data setting. Researchers have access to the outcome metric of interest $Y \in \mathbb{R}^{N \times T}$ for N units during T time periods. At time T, researchers are required to execute an experiment by assigning a binary treatment described by $D_i \in \{-1, 1\}, i = 1, 2, \cdots, N$ based on the observed pre-treatment data.

If $D_i = 1$, then a treatment needs to be applied to unit i. After the treatment experiment, furthermore outcomes are observed for additional S time periods $t = T+1, \cdots, T+S$. During this period, every unit $i \in [N]$ in each time period t is associated with the following two random outcomes: $Y_{it}(-1) = \mu_{it} + e_{it},$ and $Y_{it}(1) = Y_{it}(-1) + \tau_i$, where μ_{it} is the base outcome, τ is the treatment effect aiming to estimate and e_{it} is the zero mean i.i.d idiosyncratic noise with variance $\text{Var}(\epsilon_{it}) = \sigma$. Once treatment D_i is applied, the experimenter is able to realize $Y_{it} = \frac{(D_i+1)}{2} Y_{it}(1) + \frac{(1-D_i)}{2} Y_{it}(-1)$.

Estimating the treatment effect τ is quite challenging because once we implement a treatment on unit j (*i.e.,* $D_j = 1$) and observe the outcome $Y_{j,T+1}(1)$, then counterfactual outcome $Y_{j,T+1}(-1)$ is not observable. With the pre-treatment observation Y_{iT}, synthetic control literature [9, 18] constructs the counterfactual estimate for a treated unit j (*i.e.,* $D_j = 1$) from a weighted average of other units' outcomes: $\hat{Y}_{j,T+1}(-1) = \sum_{i:D_i=-1} w_i Y_{i,T+1}$. The weights w_i are learned from the pre-treatment observed data via minimizing $\sum_{t=1}^{T}(Y_{jt} - \sum_{i:D_i=0} w_i Y_{it})^2$. Then the treatment effect of unit j we estimate can be written as $\tau_j = Y_{j,T+1} - \hat{Y}_{j,T+1}(-1)$.

2.1 Synthetic Design

In this section, we consider the synthetic design objective function proposed for two-way fixed effect in [6, 7] and showed an hidden connection with the phase synchronization problem. [6, 7] aims to design treatment assignments $\{D_i = \pm 1\}_{i=1}^N$ and weights $\{w_i \geq 0\}_{i=1}^N$ for outcome experiments at time $T+1$. For we aims to estimate a two-way fixed effect where the treatment effects are homogeneous, we can consider the *weighted average treatment effect on the treated (wATET)* $\tau = \sum_{i=1}^N \frac{D_i+1}{2} w_i \tau_i$ instead [21]. wATET can be estimated as a difference in weighted means estimator $\hat{\tau} = \sum_{i:D_i=1} w_i Y_{i,T+1} - \sum_{i:D_i=-1} w_i Y_{i,T+1}$ with $\sum_{i:D_i=1} w_i = \sum_{i:D_i=-1} w_i = 1$. Following [6], upon the outcome model, the mean squared error of the difference-in-weighted-means estimator admits the decomposition

$$\mathbb{E}\left[(\hat{\tau}-\tau)^2 | \{D_i, w_i\}_{i=1}^N\right] = \underbrace{\left(\sum_{i:D_i=1} w_i \mu_{i,T+1} - \sum_{i:D_i=-1} w_i \mu_{i,T+1}\right)^2}_{\text{weighted covariate balancing}} + \sigma \sum_{i=1}^N w_i^2.$$

The designer aims to design the experiment with a lowest expected mean square error. Thus, [6] proposed the following mixed-integer programming for experimenting design with Synthetic Control:

$$\begin{aligned} \min_{\{D_i, w_i\}_{i=1}^n} \quad & \frac{1}{T}\sum_{t=1}^T \left(\sum_{i:D_i=1} w_i Y_{it} - \sum_{i:D_i=-1} w_i Y_{it}\right)^2 + \sigma \sum_{i=1}^N w_i^2 \\ \text{s.t.} \quad & w_i \geq 0, D_i \in \{-1,1\},\ \forall i \in [N], \qquad \sum_{i:D_i=1} w_i = \sum_{i:D_i=-1} w_i = 1. \end{aligned} \tag{1}$$

Remark 1. *We remove the constraint $\sum_{i:D_i=1} D_i = K$ for a given integer $K \in \mathbb{N}$ in the mixed-integer programming in [6] mainly as this constraints is empirically proved not critical in [7]. The NP-hard proof in [6] depends on the constraint $\sum_{i:D_i=1} D_i = K$. In the following discussion, we will show that the problem is also NP-hard even $\sum_{i:D_i=1} D_i = K$ is removed, as the resulting optimization problem can be reformulated as the ℓ_1-PCA [22, 23] and phase Synchronization [11, 14].*

By making a further simplification of the problem (1), we introduce a change of variable $W_i = w_i D_i$. For $w_i \geq 0$, then $D_i = \operatorname{sgn}(W_i)$ and $w_i = |W_i|$. At the same time, the constraint $\sum_{i:D_i=1} w_i = \sum_{i:D_i=-1} w_i = 1$ is equivalent to $\mathbb{1}^\top W = 0$ and the objective function

$$\frac{1}{T}\sum_{t=1}^T \left(\sum_{i:D_i=1} w_i Y_{it} - \sum_{i:D_i=-1} w_i Y_{it}\right)^2 + \sigma \sum_{i=1}^N w_i^2$$

can be reformulated as $W^\top(YY^\top + \lambda I)W$, where $W = [w_1, \cdots, w_N]^\top$ and $\mathbb{1} \in \mathbb{R}^N$ is the all one vector. Thus, (1) can be recast into

$$\min_{W\in\mathbb{R}^n, \mathbb{1}^\top W = 0, \|W\|_1 = 1} W^\top (YY^\top + \sigma I)W. \tag{2}$$

Although the reformulation (2) translates the problem into a compact matrix form, it is still a nonconvex problem due to the constraint $\|W\|_1 = 1$. To deal with the constraint $\mathbb{1}^\top W = 0$, we add an extra term $\lambda(\mathbb{1}^\top W)^2$ to the objective function, where λ is a pre-defined hyper-parameter. Although this penalty method cannot produce the exact global solution, we can still recover the sign of the global solution (see Theorem 1). Once the sign of the global solution is identified, the remaining effort of computing the magnitude reduces to solving a convex problem (5).

Theorem 1. *For large enough λ, the global solution W^* of (2) satisfies*

$$\operatorname{sgn}(W^*) = \operatorname{sgn}\left(\operatorname*{arg\,min}_{W\in\mathbb{R}^n, \|W\|_1=1} W^\top (YY^\top + \sigma I + \lambda \mathbb{1}\mathbb{1}^\top)W\right).$$

The following theorem states that the problem is equivalent to another well-known NP-hard non-convex problem — Phase Synchronization [11].

Algorithm 1 Synthetic principal component Design

```
Require: Pre-treatment Observations $Y \in \mathbb{R}^{N\times T}$
  Set initial treatment assignment guess through $y^0 = \mathrm{sgn}(v)$, where $v$ is the smallest eigenvector of matrix $(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)$, where $\alpha, \lambda$ are two pre-defined hyper-parameter.
                                                    ▷ Spectral Initialization
  while Converged do
      Select one of the following two boxes to iterate
      For SPCD, update the design via               ▷ Generalized power methods
          $y^{t+1} = \mathrm{sgn}\left[\left((YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta I\right) y^t\right],$        (3)
      where $\beta$ is a pre-defined hyper-parameter.
      For NormSPCD, update the design via   ▷ Normalize the inverse covariance matrix
          $y^{t+1} = \mathrm{sgn}\Big[\left[(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta I\right] (y^t/d)\Big],$        (4)
      where $d = \sqrt{\mathrm{diag}((YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1})}$ and / denotes element-wise divide.
  end while
  Solve the following *convex* optimization problem
      $\{w_i\}_{i=1}^n = \arg\min_{\{w_i\}_{i=1}^n} \frac{1}{T}\sum_{t=1}^T \left(\sum_{i:y(i)=1} w_i Y_{it} - \sum_{i:y(i)=-1} w_i Y_{it}\right)^2 + \sigma \sum_{i=1}^N w_i^2$        (5)
      $\text{s.t.} \quad w_i \geq 0,\ \forall i \in [N], \sum_{i:y(i)=1} w_i = \sum_{i:y(i)=-1} w_i = 1.$
  Treat Unit $i$ if $y(i) = -\mathrm{sgn}\left(\sum_{i=1}^N y(i)\right)$ and run the experiment.
                ▷ To ensure the size of the treated group is smaller than the control group
  Estimate the treatment effect via
      $\hat{\tau} = \sum_{t=1}^S \left(\sum_{i:y(i)=-\mathrm{sgn}\left(\sum_{i=1}^N y(i)\right)} w_i Y_{i,T+t} - \sum_{i:y(i)=\mathrm{sgn}\left(\sum_{i=1}^N y(i)\right)} w_i Y_{i,T+t}\right).$
```

Theorem 2 (Equivalence between Synthetic Design and Phase Synchronization)**.** *If $x^* \in \mathbb{R}^n$ is the global solution of $\min_{\|x\|_1=1} \|Ax\|_2^2$ for some matrix $A \in \mathbb{R}^{D\times n}$ $(D > n)$ and the matrix $A^\top A \in \mathbb{R}^{n\times n}$ is invertible, then $y^* = \mathrm{sgn}(x^*)$ is the global solution of $\max_{y\in\{-1,+1\}^n} y^\top ((A^\top A)^{-1})^\top y$.*

The proof of Theorem 1 and Theorem 2 is omitted in the main text and is shown in Appendix B.

Remark 2. *Phase synchronization [11, 13] aims to recover n phases $z_i = e^{i\theta_i}$, $i \in [n]$ via solving the following optimization problem $\max_{|x_1|=\cdots=|x_n|=1} \quad x^\top Cx$ where C_{ij} is the noisy observation of $z_i\bar{z}_j = e^{i(\theta_i-\theta_j)}$. Our problem is symbolically equivalent. However, the data generating process is quite different from the Phase synchronization for we are considering the inverse of the Gram matrix. In Appendix, we will show that our design is actually the first ℓ_1-principal component [22, 23].*

3 Algorithm description

In this section, we propose a normalized version generalized power method [24, 14, 15, 16] with spectral initialization [17] to solve our problem.

3.1 Generalized Power Methods

Spectral relaxation [11] is the first simple and efficient approach to solve the phase synchronization problem. [11] relaxed the N constraints $|x_i| = 1, i \in [N]$ to $\|x\|_2^2 = n$. Then the solution becomes the leading eigenvector. [15, 16] showed that the eigenvector estimator is almost close to the global optima under certain data generating process. Following these works, we take our initial guess of the optimal experiment to be $\text{sgn}(v)$, where v is the smallest eigenvector of the matrix $(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)$ with $\alpha, \lambda > 0$ as two pre-defined hyper-parameters.

To further improve the experiment assignment, we utilize the generalized power method [24, 25], which considers the linearization of the objective function at the current point and moves towards a minimizer of this linear function over the non-convex set $\mathcal{C}$. The generalized power method can be also understood as projected gradient descent [26]. Indeed, the update

$$y^{t+1} = \text{sgn}[((\frac{1}{\beta}YY^\top + \frac{\sigma}{\beta}I + \frac{\alpha}{\beta}\mathbb{1}\mathbb{1}^\top)^{-1} + I)y^t]$$

can be understood as a projection step (sgn) after a gradient descent update with step size $\frac{1}{\beta}$. Thus the algorithm shares a sufficient ascent condition for each iteration. Our algorithm is called Synthetic principal component Design (SPCD) and it is summarized in Algorithm 1.

Normalized Variant In [14, 16, 15], the global optimality result is highly dependent on the assumption that the top eigenvector of the iteration matrix lies in $\{-1, 1\}^N$. However, in our setup, the top eigenvector of the iteration matrix is the smallest eigenvector of the covariance matrix which may not be a $\{-1, 1\}^N$ vector. This is also the case appearing in the phase retrieval [27, 28] and the degree corrected stochastic block model [29, 30]. Inspired by the SCORE [30, 31] method for degree corrected stochastic block models, we further introduce a normalization step to the generalization power method and call the new algorithm Normalized SPCD (cf. NormSPCD, see (4) for details). We use the diagonal component of the inverse covariance as an estimate of the true normalization component. In Appendix C, we show that NormSPCD can be interpreted as a Riemannian gradient descent with a specific metric. Empirical results show that it is better than the original GPW in Figure 1b. This normalization technique may be of independent interest in other applications.

3.2 Global Guarantee

In this subsection, we provide the global optimization guarantee for the (normalized) generalized power method. [13, 14, 15, 16] have shown that phase retrieval is globally solvable under certain generative models. We will show that generalized power method can globally converge under certain data generating processes, which are quite different from the ones assumed in the previous works. Following [7], we consider a realizable linear factor model (also referred to as "interactive fixed-effects model") [9, 18, 19, 20], which has already been commonly employed in the literature as a benchmark model to analyze the properties of synthetic control estimators [32, 33]. Recently, [34] justify the linear assumption from an independent causal mechanism viewpoint. The linear latent factor model is stated in the following assumption. .

Assumption 1 (Linear Latent Factor Model [9, 18])**.** *The outcomes are generated via the following linear factor model*

$$Y_{jt} = \delta_t + \frac{D_{jt}+1}{2}\tau + \theta_t^T \mu_j + e_{jt}, \qquad \mathbb{E}[e_{jt}|\delta_t, \mu_j, D_{jt}] = 0, \qquad \text{Var}[e_{jt}|\delta_t, \mu_j, D_{jt}] = \sigma.$$

Here δ_t is the time fixed effect; μ_j is the unobserved common factors; θ_t is a vector of unknown factor loading; e_{jt} is the unobserved i.i.d. idiosyncratic noise; τ is the treatment effect that we aim to estimate and D_{jt} is the $\{-1, 1\}$ variable according to the treatment assignment to unit j at time t. More specifically, in the pre-treatment period, $D_{jt} = -1$ for all $\forall j \in [N], t \in [T]$.

To obtain the global optimality result, we further make the following realizable assumption that there is only one realizable experiment (zero error experiment) in population.

Assumption 2 (Realizable Assumption)**.** *There exists a unique parameter $(w_i, D_i)_{i=1}^n (D_i \in \{-1, 1\})$ that satisfies the following conditions:*

- $w_i \geq 0$ *and* $\sum_{i=1}^n D_i w_i = 0$. $\|w\|_2^2 = N$ *and* $\epsilon \leq |w_i| \leq \frac{1}{\epsilon}$ *for all* $\forall i \in [N]$.
- *The weights will balance the covariates,* i.e. $\sum_{i=1}^n w_i D_i \mu_i = 0$.

Remark 3. *This realizable assumption is similar to [33], [34, (5)] and [7, Assumption 3]. The difference is that [7, Assumption 3] assumes the weight will cancel noisy observation of the untreated outcome $Y_{jt}(-1)$ which is not realistic when the pre-treatment period T is larger than the number of units N. Our assumption is closer to [33] and [34, (5)], but we further assume the uniqueness of the realizable experiment that makes the optimization problem easier (in terms of no need to distinguish different realizable experiments).*

Under Assumptions 1 and 2, we can show the following global optimality result and the proof is shown in the Appendix C.

Theorem 3. *(Informal) Suppose that Assumptions 1 and 2 hold and that the latent time factor $[\theta_i^\top \delta_i]^\top$ is sampled from a underlying distribution with mean $\tilde{\theta}$ and covariance $\tilde{\Sigma}$. Under regularity assumptions (see Appendix C.2.4), if σ is small enough and $T \geq \mathrm{poly}(N, \frac{1}{\epsilon})$, then*

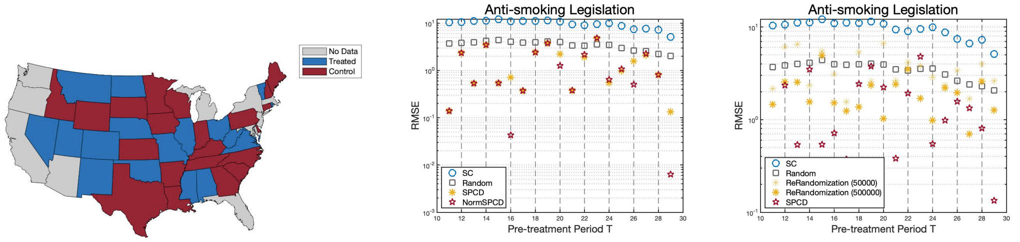
- *If $\epsilon > \frac{\sqrt{3}}{2} - 1$, then SPCD converges to the global optima.*
- *If $\epsilon > 0$, then NormSPCD converges to the global optima at a linear rate.*

4 Numerical Study

This section report the numerical tests of our algorithm. In subsection 4.1, we demonstrate our algorithm on two real world datasets. Both experiments have shown the effectiveness of our proposed algorithm in terms of the the root-mean-square error (RMSE), where the squared differences between the true values of the treatment effects and the respective estimates are computed for each treatment period and averaged. We also validate our algorithm on the latent factor model in Appendix D.

4.1 Real World Data

To exam our algorithm on real data, we follow [35, 6] and utilize the US Bureau of Labor Statistics and the Anti-Smoking Legislation data to examine the validity of our algorithm. Besides synthetic control (SC), which randomly selects one unit to implement the treatment, we also implement an additional random baseline, which randomly select units as control/group with probability $1/2$. The final result is shown in Table 1. SPCD surpasses SC a large margin on both of the datasets.



(a) Treatment Selected when $T = 25$ (b) Comparison in terms of RMSE (c) Comparison with rerandomization

Figure 1: A typical design selected via synthetic principal component design (SPCD) and its performance.

The Abadie–Diamond–Hainmueller Smoking Data. [9] uses SC to study the effects of Proposition 99, a large-scale anti-smoking legislation program that California implemented in 1988. To simulate the bias of SC and SPCD on this application, following [19], we consider observations for 38 states (excluding California due to Proposition 99) from 1970 through 2000. We regard the first T year as pre-treatment periods to produce the design and use the last $31 - T$ years as post-treatment periods to test the performance of the treatment assignment. The final result is shown in Table 1 and Figure 1b. Our design surpasses the random design by a large margin on most of the selection of time T. We also compare our method with the rerandomization design [4, 36, 37] in Figure 1c, which shows that our algorithm is still better than 500000 times of rerandomization.

One typical design produced by our algorithm is shown in Figure 1. The experiment design for different pre-treatment length T is shown in Figure 6. The plots show that our selection of the control group is robust to different pre-treatment time period and has the ability to represent all different geographic, demographic, racial, and social structure of states in the United State.

US Bureau of Labor Statistics. We also apply our algorithm on the unemployment rate of 50 states in 40 months from the US Bureau of Labor Statistics (BLS). We run 50 simulations such that

Table 1: Root-mean-square errors of the average treatment effect estimates by both synthetic control (SC) and synthetic principal component design (SPCD) on real data. The random design is simulated 10 times and 95% confidence interval is demonstrated. The reported RMSE for BLS dataset are multiplied by 10^3 for readability.

US Bureau of Labor Statistics						
RMSE	$T = 5$			$T = 10$		
	SC	Random	SPCD	SC	Random	SPCD
	14.5	7.5	**0.9**	11.6	5.6	**0.6**
Anti-smoking legislation						
RMSE	$T = 15$			$T = 25$		
	SC	Random	SPCD	SC	Random	SPCD
	11.65	4.32±0.21	**1.14**	7.89	3.13±0.19	**0.98**

each simulation utilizes a 20-by-$T + S$ matrix sampled from the original 50-by-40 dataset. More specifically, we randomly select 20 units and use the first T time period to select the synthetic design and synthetic weight. The remaining S time periods are the consecutive months that follow. In our experiment, we fix $S = 5$ and run both experiment for $T = 5, 10$. The final result in terms of the RMSE is shown in Table 1.

References

[1] Donald B Rubin. For objective causal inference, design trumps analysis. *The annals of applied statistics*, 2(3):808–840, 2008.

[2] Maximilian Kasy. Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338, 2016.

[3] Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

[4] Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

[5] Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The gram-schmidt walk: a cure for the banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–597, 2018.

[6] Nick Doudchenko, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess, et al. Synthetic design: An optimization approach to experimental design with synthetic controls. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Alberto Abadie and Jinglong Zhao. Synthetic controls for experimental design. *arXiv preprint arXiv:2108.02196*, 2021.

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

[9] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

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

[11] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

[12] Shuzhong Zhang and Yongwei Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16(3):871–890, 2006.

[13] Afonso S Bandeira, Nicolas Boumal, and Amit Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 163(1):145–167, 2017.

[14] Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

[15] Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017.

[16] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

[17] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

[18] Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

[19] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

[20] Bruno Ferman. On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*, 116(536):1764–1772, 2021.

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

[22] Michael McCoy and Joel A Tropp. Two proposals for robust pca using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

[23] Peng Wang, Huikang Liu, and Anthony Man-Cho So. Linear convergence of a proximal alternating minimization method with extrapolation for ℓ_1-norm principal component analysis. *arXiv preprint arXiv:2107.07107*, 2021.

[24] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

[25] Ronny Luss and Marc Teboulle. Conditional gradient algorithmsfor rank-one matrix approximations with a sparsity constraint. *siam REVIEW*, 55(1):65–98, 2013.

[26] Steven T Smith. Optimization techniques on riemannian manifolds. *Fields institute communications*, 3(3):113–135, 1994.

[27] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[28] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Advances in Neural Information Processing Systems*, 28, 2015.

[29] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

[30] Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

[31] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Improvements on score, especially for weak signals. *Sankhya A*, 84(1):127–162, 2022.

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

[33] Kathleen T Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083, 2020.

[34] Claudia Shi, Dhanya Sridhar, Vishal Misra, and David M Blei. On the assumptions of synthetic control methods. *arXiv preprint arXiv:2112.05671*, 2021.

[35] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

[36] Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

[37] Xinran Li, Peng Ding, and Donald B Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

[38] Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

[39] Nojun Kwak. Principal component analysis based on l1-norm maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(9):1672–1680, 2008.

[40] Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016.

[41] Peng Wang, Zirui Zhou, and Anthony Man-Cho So. Non-convex exact community recovery in stochastic block model. *Mathematical Programming*, pages 1–37, 2021.

[42] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[43] Terence Tao. Topics in random matrix theory. *Graduate Studies in Mathematics*, 132, 2011.

[44] Alberto Abadie and Jaume Vives-i Bastida. Synthetic controls in action. *arXiv preprint arXiv:2203.06279*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Organization of the Appendix

We organize the appendix as following:

- In Appendix B, we demonstrated the equivalence between Synthetic Design, ℓ_1-PCA and Phase Synchronization. We also briefly introduce the literature of solving ℓ_1-PCA and Phase Synchronization in Appendix B.
- In Appendix C, we demonstrated the proof of global convergence of the generalized power method. The road-map of the proof is following. In Appendix C.2.1, we analyze the spectral initialization. We showed that it provide as accuracy estimate as the global optima to the ground truth signal. In Appendix C.2.2, we verify the global optimality of the stationary point of GPW via the Riemann Hessian. In Appendix C.2.3, we demonstrated the linear convergence rate of the GPW method. In Appendix C.2.4, we analyze the data generating process to match the assumption needed in the global optimality of the GPW method.

B Equivalent to Phase Synchronization

Theorem 4. *For large enough λ, the global solution W^* of (2) satisfies*

$$\operatorname{sgn}(W^*) = \operatorname{sgn}\left(\underset{W\in\mathbb{R}^n, \|W\|_1=1}{\arg\min} W^\top (YY^\top + \sigma I + \lambda \mathbb{1}\mathbb{1}^\top) W \right)$$

Proof. We denote $W_\lambda = \arg\min_{W\in\mathbb{R}^n,\|W\|_1=1} W^\top (YY^\top + \sigma I + \lambda \mathbb{1}\mathbb{1}^\top) W$ for $\lambda > 0$. From [38, Theorem 17.1], we know that $W_\lambda \to W^*$ as $\lambda \uparrow \infty$. Thus there exists λ^*, such that for all $\lambda > \lambda^*$, we have

$$\|W_\lambda - W^*\|_\infty \leq \min\{|W^*(i)| : W^*(i) \neq 0\}.$$

Thus for all $\lambda > \lambda^*$, we have $\operatorname{sgn}(W^*) = \operatorname{sgn}(W_\lambda)$. □

Remark 4. *In the above discussion, we consider both* $\operatorname{sgn}(0) = 1$ *and* $\operatorname{sgn}(0) = -1$ *are right. The reason is that we plug in both sign selection in to the convex programming (5) can both produce the true global optimum.*

Theorem 5 (Equivalence between Synthetic Design, ℓ_1-PCA and Phase Synchronization)**.** *If $x^* \in \mathbb{R}^n$ is the global solution of* $\min_{\|x\|_1=1} \|Ax\|_2^2$ *for some matrix $A \in \mathbb{R}^{D\times n}$ $(D > n)$ and matrix $A^\top A \in \mathbb{R}^{n\times n}$ is invertible, then $y^* = \operatorname{sgn}(x^*)$ is the global solution of* $\max_{y\in\{-1,+1\}^n} y^\top ((A^\top A)^{-1})^\top y$.

Proof. Firstly, the problem $\min_{\|x\|_1=1} \|Ax\|_2^2$ is equivalent to $\min_{\|x\|_1=1} \|(A^\top A)^{\frac{1}{2}} x\|_2^2$ and can be further transformed to $\min_{\|x\|_2=1} \|(A^\top A)^{-\frac{1}{2}} x\|_1$.

At the same time, for any matrix $T \in \mathbb{R}^{n\times n}$, we have

$$\max_{\|x\|_2=1} \|Tx\|_1 = \max_{\|x\|_2=1, y\in\{-1,+1\}} y^\top Tx = \max_{y\in\{-1,+1\}} \|T^\top y\|_2 = \max_{y\in\{-1,+1\}} y^\top TT^\top y. \quad (6)$$

and thus leads to $\arg\max_{y\in\{-1,+1\}} \|T^\top y\|_2 = \operatorname{sgn}(Tx^*)$ where $x^* = \arg\max_{\|x\|_2=1} \|Tx\|_1$. Combining the two facts, we can prove the theorem. □

Remark 5. *Although we formulated the mixed integer programming as a well-known compact matrix form, the two problems,* i.e. *ℓ_1-PCA and phase synchronization, are still known to be NP-hard [22, 12]. However phase synchronization can be globally solved under certain data generative models [13, 14, 15, 16]. As far as the author known, there is still no data generative models for ℓ_1-PCA been found can be globally solved. [23] show that for the Kurdyka-Lojasiewicz exponent of the ℓ_1-PCA problem at any of the limiting critical points is $\frac{1}{2}$. This allows one to establish the linear convergence to the local stationary point of certain algorithm. Although, Generalized Power Method is also proposed for ℓ_1-PCA [39], but only local convergence is guaranteed.*

C Optimization Theory

Out theory mostly follows [13, 14]. But we have slightly different optimization problem (optimization over $\mathbb{C}^n$ in [13, 14] but $\mathbb{R}^n$ in ours) and uses different data generating process (gram matrix in [13, 14] and inverse of gram matrix in our paper. All the entries of ground truth vector norm equals to 1 in [13, 14], *i.e.* $|z_i| = 1$. But this is not assumed in our paper.). For completeness, we complete all the proof details here in the appendix.

C.1 Preliminaries

In this section, we present some basics of Riemannian gradients. For $\{-1, 1\}^n$ is a degenerate manifold. In the proof, we will consider the global optimality of the synchronization problem over a larger space $\mathbb{T}^n = \{z \in \mathcal{C}^n : |z_1| = \cdots = |z_n| = 1\}$. Next we endow $\mathbb{T}^n$ with Euclidean metric $\left\langle y^1, y^2 \right\rangle = \sum_{i=1}^n \mathcal{R}\{y_i^1 {y_i^2}^H\}$ which is the equivalent to viewing $\mathcal{C}^n$ as $\mathbb{R}^{2n}$ and equip with the canonical inner product. Then $\mathbb{T}^n$ can be considered as a sub-manifold and the tangent space can be written as

$$\mathcal{T}_y \mathbb{T}^n = \{\dot{y} \in \mathcal{C}^n : \mathcal{R}(\dot{y}_i y_i^H) = 0, \forall i \in [n]\}.$$

The projector to the tangent space is $\text{Proj}_x : \mathbb{C}^n \to T_x\mathbb{T}^n : \dot{x} \to \dot{x} - \mathcal{R}\{\text{ddiag}(\dot{x}x^H)\}x$, where $\text{ddiag} : \mathbb{C}^n \to \mathbb{C}^n$ is a function set all off-diagonal entries of the input matrix to zero. Thus the Riemannian gradient of function $f(x) = x^H Cx$ is given as

$$\text{grad} f(x) = 2(\mathcal{R}\{\text{ddiag}(Cxx^H)\} - C)x$$

Following [13], we consider the Riemannian Hessian on the tangent space as the second-order necessary optimality condition. The Riemannian Hessian is defined as

$$\text{Hess} f(x)[\dot{x}] = \text{Proj}_x \text{Dgrad} g(x)[\dot{x}] = \text{Proj}_x 2S(x)\dot{x},$$

where $S(x) = \mathcal{R}\{\text{ddiag}(Cxx^H)\} - C$. If x is a (local) optimum, then $\langle \dot{x}, S(x)\dot{x} \rangle > 0$ for all $\dot{x} \in \mathcal{T}_x\mathbb{T}^N$.

For NormSPCD iteration 4, we consider the update as a Riemannian steepest-descent[40]. The Riemannian steepest-descent *search* direction to minimize objective function f as $\arg\min_{\xi_x \in \mathbb{R}^n} \langle \nabla f(x), \xi_x \rangle_{\mathcal{R}} + \frac{1}{2} \langle \xi_x, \xi_x \rangle_{\mathcal{R}}$. The Riemannian metric we consider for NormSPCD on $\mathcal{T}_x\mathbb{T}^N$ defined as

$$\langle y_1, y_2 \rangle_{\mathcal{R}} = \sum_{i=1}^N |z_i|^2 \mathcal{R}\left(y_1(i) y_2(i)^H\right) \quad, \forall y_1, y_2 \in \mathcal{T}_x\mathbb{T}^N.$$

Similarly, we can define the new Riemannian Hessian as $S_{\mathcal{R}}(x) = \mathcal{R}\{\text{ddiag}(\mathring{C}xx^H)\} - \mathring{C}$, where $\mathring{C} = \text{diag}(\frac{1}{|z|})C\text{diag}(\frac{1}{|z|})$. We'll show that $rS(x) \preccurlyeq S_{\mathcal{R}}(x) \preccurlyeq RS(x)$ for some constant $r, R > 0$

In our discussion, we consider our algorithm works in the Field of complexity numbers. However, from the closeness of the Field of real numbers, we know that the whole trajectory of our algorithm lies in the Field of real numbers. Global minimum in the complex domain is a harder problem and directly indicate the global optimality in $\{-1, +1\}^N$.

C.2 Global Optimality of (Normalzied) Generalized Power Methods

In this section, we first study a meta version of the optimization problem. Then we will show how our generative model can be fitted into this framework. We consider the following meta optimization problem

$$\min_{x \in \mathbb{T}^n} f(x) = x^H Cx \tag{7}$$

where $C = zz^H + \Delta$ is a Hermite perturbed rank-1 matrix. Different from [13, 14, 15, 16] which assumes $z \in \mathbb{T}^N$, instead, we have the following assumption on the ground truth vector $z \in \mathbb{R}^N$:

Assumption 3. *For some $\epsilon > 1$, we have*

$$\epsilon \le |z_i| \le \frac{1}{\epsilon}, \qquad \forall i \in [N]$$

This is a smooth optimization problem over a smooth Riemannian manifold $\mathcal{T}^n$. Then the Riemannian gradient grad$g(x) = 2(\mathcal{R}(\text{diag}(Cxx^H)) - C)x$. The first order necessary optimality condition is gard$g(x) = 0$. We will also make use the second order optimality via the Riemannian Hessian

$$\text{Hess}g(x)[\dot{x}] = 2\left\langle \dot{x}, S\dot{x} \right\rangle, \forall \dot{x} \in \mathcal{T}_y\mathbb{T}^n$$

where $S(x) = \mathcal{R}(\text{ddiag}(Cxx^H) - C)$ and ddiag : $\mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ zeros out all off-diagonal entries of a matrix. Although computing the global optimum of (7) is NP-hard [12], fortunately, global optimality of (7) can sometimes be certified through the Hermitian Hessian matrix $S(x) = \mathcal{R}(\text{ddiag}(Cxx^H) - C)$. This can be shown in the following lemma for sufficient optimality condition:

Lemma 1 (Optimality Gap)**.** *Let x^* be globally optimal for (7). For any $x \in \mathbb{T}^N$, the optimality gap at x can be bounded as*

$$0 \leq f(x^*) - f(x) \leq -N\lambda_{\min}(S(x)).$$

As a result, if $S(x) \succeq 0$, then x is the global optimality problem for (7).

Proof. See [13, Section 4.2] and [14, Lemma 2]. □

In the following lemma, we showed that similar property holds for the changed Riemannian metric.

Lemma 2 (Optimality Gap for Riemannian Formulation)**.** *Let x^* be globally optimal for (7). If $\epsilon < |z_i| < \frac{1}{\epsilon}$, then for any $x \in \mathbb{T}^N$, the optimality gap at x can be bounded as*

$$0 \leq f(x^*) - f(x) \leq -\frac{1}{\epsilon}N\lambda_{\min}(S_{\mathcal{R}}(x)).$$

As a result, if $S(x) \succeq 0$, then x is the global optimality problem for (7).

Proof. This is because

$$x^HCx - y^HCy = y^HS(x)y \geq \frac{N}{\epsilon}\lambda_{\min}(S_{\mathcal{R}}(x))$$

□

To solve this problem, we consider the following Generalized Power Method [14, 15] in Algorithm 2 and our normalized version in Algorithm 3.

Algorithm 2 Generalized Power Method

```
Set initialization through $x^0 = \text{sgn}(v)$, where $v$ is the leading eigenvector of matrix $C$. ▷ Spectral Initialization
Define $\tilde{C} = C + \alpha I_N$ where $\alpha = \|\Delta\|$
while Converged do
    $x^{t+1} = T(x^t) \triangleq \text{sgn}\left[\tilde{C}x^t\right]$        ▷ Generalized power methods
end while
```

Algorithm 3 Normalized Generalized Power Method

```
Set initialization through $x^0 = \text{sgn}(v)$, where $v$ is the leading eigenvector of matrix $C$. ▷ Spectral Initialization
Define $\tilde{C} = C + \alpha I_N$ where $\alpha = \|\Delta\|$
while Converged do
    $x^{t+1} = \mathring{T}(x^t) \triangleq \text{sgn}\left[\tilde{C}(x^t./\sqrt{\text{ddiag}(C)})\right]$,        ▷ Normalized Generalized power methods
    where ./ is the element-wise division.
end while
```

C.2.1 The Spectral Initialization

We make the initial guess in Algorithm 2 via spectral relaxation [11]. Denote v as the leading eigenvector of matrix C. From the following Lemma 3, we can know that the leading eigenvector v is close to the ground truth signal z.

Lemma 3. *Given vector $z \in \mathbb{R}^N$ satisfies $\|z\|_2 = 1$. For matrix $C = zz^\top + \Delta$, where $\Delta \in \mathbb{R}^{N\times N}$ is a symmetric perturbation matrix. Then for all $x \in \mathcal{C}^n$ and $\|x\|_2^2 = N$ satisfies $x^H Cx \geq z^H Cz$, we have*

$$\left\| \min_{\theta\in\{1,-1\}} \theta x - z \right\| \leq \frac{4\|\Delta\|}{\sqrt{N}}$$

where the $\|\Delta\|$ is matrix operator norm.

Proof. See [13, Lemma 4.1] and[14, Lemma 1]. □

Based on the top eigenvector v, we project v to the Riemann manifold $\mathbb{T}^n$ and make the initial guess. For C is a symmetric real matrix, the eigenvector $v \in \mathbb{R}^N$. Thus projection to $\mathbb{T}^n$ of vector v will simply become $\mathrm{sgn}(v)$. In the next lemma, we'll show the Spectral Estimator is almost as close to z as the global optima.

Lemma 4 (The Spectral Estimator is almost as accuracy as the Global Optima)**.**

$$\left\| \min_{\theta\in\{1,-1\}} \theta\mathrm{sgn}(v) - \theta\mathrm{sgn}(z) \right\| \leq \frac{8\|\Delta\|}{\epsilon\sqrt{N}}$$

Lemma 4 is the direct corollary of Lemma 3 and the following technical lemma, which is also important in the convergence rate proof in Section C.2.3.

Lemma 5. *For $w \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$ satisfies $\|z\|_2^2 = N$ and $\epsilon \leq |z_i| \leq \frac{1}{\epsilon}, \forall i \in [N]$ (or $1+\epsilon$), then we have*

$$\|\mathrm{sgn}(w) - \mathrm{sgn}(z)\|_2 \leq \frac{2}{\epsilon}\|w - z\|_2.$$

Proof. [16, Lemma 13] and [41, Lemma 3] □

C.2.2 The Generalized Power Method

Although, the Spectral Estimator produce good estimates. We still cannot obtain the global optimum of (7). Following [14, 15], we proceed the Generalized Power Method (GPM) to further improve the estimate. [14] showed that the Generalized Power Method will converge to the global optima of problem (7) and [15] showed that the proceeded estimate is always better than the initial spectral estimate. The procedure of the Generalized Power Method is shown in Algorithm 2. We also consider the Normalized GPM (Algorithm 3) in this section.

For the simplicity of description, we define an equivalence relationship $\sim$ over $\mathbb{T}^n$ as

$$x \sim y \quad \Longleftrightarrow \quad x = ye^{i\theta} \quad \text{for some } \theta \in \mathbb{R}.$$

The quotient space $\mathbb{T}^n/\sim$ is defined as all the corresponding equivalence class $\{xe^{i\theta} : \theta \in \mathbb{R}\}$ for some $x \in \mathbb{C}$. The error measure we are interested in

$$d_q(z, x) = \min_{\theta\in\mathbb{R}} \|xe^{i\theta} - z\|_q = \sqrt{2(n - |z^H x|)}, \quad q \in [1, \infty].$$

Lemma 6. *For all $x, y \in \{-1, 1\}^N$ and $q \in [1, \infty]$, then we have*

$$e^{i \arg\min_{\theta\in\mathbb{R}} \|xe^{i\theta} - z\|_q} \in \{-1, 1\}.$$

Proof. We use proof by contradiction to prove this statement. If $\theta^* = \arg\min_{\theta\in\mathbb{R}} e^{i\|xe^{i\theta}-z\|_q} \notin \mathbb{R}$, then we will have

$$\|x\mathrm{sgn}(\mathcal{R}(e^{i\theta^*})) - z\|_q \leq \|x(\mathcal{R}(e^{i\theta^*})) - z\|_q < e^{i\|xe^{i\theta}-z\|_q}.$$

This is contradicted with $\theta^* = \arg\min_{\theta\in\mathbb{R}} e^{i\|xe^{i\theta}-z\|_q}$. Thus $e^{i \arg\min_{\theta\in\mathbb{R}} \|xe^{i\theta}-z\|_q} \in \{-1, 1\}$.

□

Notice that (normalized) GPM iterates on the quotient space $\mathcal{T}^N/\sim$, *i.e.* if $x \sim y$, then $\mathrm{sgn}(\tilde{C}x) \sim \mathrm{sgn}(\tilde{C}y)$. Thus without further notice, all the equality in the following discussion is equality in the quotient, *i.e.* $x = y$ means $x \sim y$.

Lemma 7 (Monotonic Cost Improvement for GPM)**.** *The iterates $\{x^k\}_{k\in\mathbb{N}}$ produced by Algorithm 2 satisfies $f(x^{k+1}) > f(x^k)$ unless converged. Thus the iterates x^k do not cycle.*

Proof. See [14, Lemma 8] □

Although Algorithm 3 does not guarantee the Monotonic Cost Improvement on the original target function f. We can still prove that the produced iterates from Algorithm 3 do not cycle for it's monotonically improve another energy function.

Lemma 8 (Converging of Normalized GPM)**.** *The iterates $\{x^k\}_{k\in\mathbb{N}}$ produced by Algorithm 3 do not cycle.*

Proof. Consider the potential function $\mathring{f}(x) = \frac{1}{2}x^H\left(\text{diag}(\frac{1}{|z|})C\text{diag}(\frac{1}{|z|})\right)x$. Then the normalized power iteration can be considered as the frank-wolf algorithm for the potential function in the sense that

$$x^{t+1} = \mathring{T}(x^t) = \arg\max_{y\in\mathbb{T}^N}\left\langle y, \left(\text{diag}(\frac{1}{|z|})C\text{diag}(\frac{1}{|z|})\right)x^t\right\rangle.$$

Similar with [14, Lemma 8], we knows that the iterates $\{x^k\}_{k\in\mathbb{N}}$ monotonically improve the potential function $\mathring{f}$ and thus the iterates do not cycle. □

Lemma 9. *If x is a fixed point of Generalized Power Methods (Algorithm 2), at least one of the following holds*

$$|z^Hx| \geq \epsilon N - 4(\|\Delta\| + \alpha) \qquad \text{or} \qquad |z^Hx| \leq \frac{4(\|\Delta\| + \alpha)}{\epsilon}.$$

Furthermore, if $\|\Delta\| \leq \frac{\epsilon^2 N}{13}$ and $\alpha < \|\Delta\|$, all the accumulation points x of Algorithm 2 satisfies $|z^Hx| \geq \epsilon n - 8\|\Delta\|$.

Proof. The fixed point of the generalized power method satisfies $(\tilde{C}x)_i\bar{x}_i = |(\tilde{C}x)_i|$. Thus we have

$$x^H\tilde{C}x = \|\tilde{C}x\|_1$$

On one hand, the quadratic term $x^H\tilde{C}x$ can be upper bounded as

$$x^H\tilde{C}x = |z^Hx|^2 + x^H\Delta x + \alpha n \leq |z^Hx|^2 + (\|\Delta\| + \alpha)n.$$

On the other hand $\|\tilde{C}x\|_1$ can be lower bounded via

$$\|\tilde{C}x\|_1 = \sum_{i=1}^N \left|(z^Hx)z_i + (\Delta x)_i + \alpha x_i\right| \geq N\epsilon|z^Hx| - \|\Delta x\|_1 - \alpha N.$$

At the same time, $\|\Delta x\|_1 \leq \sqrt{N}\|\Delta x\|_2 \leq N\|\Delta\|$. Combine this with the two previous inequality, we get

$$|z^Hx|(\epsilon N - |z^Hx|) \leq 2N(\|\Delta\| + \alpha).$$

The above inequality enforces that one of $|z^Hx| \geq \epsilon N - 4(\|\Delta\| + \alpha)$ and $|z^Hx| \leq \frac{4(\|\Delta\|+\alpha)}{\epsilon}$. holds. We call all the stationary point satisfies $|z^Hx| \geq \epsilon N - 4(\|\Delta\| + \alpha)$ "good" stationary point and the stationary point satisfies $|z^Hx| \leq \frac{4(\|\Delta\|+\alpha)}{\epsilon}$ "bad" stationary point. In the following discussion, we use Lemma 4 to show that the spectral initialization sgn(v) outperforms all the "bad" fixed points. Due to Lemma 7, Generalized Power Method consistently improve the cost function and thus only converge to "good" stationary points. From Lemma 4, we have

$$\begin{aligned}\text{sgn}(v)^HC\text{sgn}(v) &= |\text{sgn}(v)^Hz|^2 + \text{sgn}(v)^H\Delta\text{sgn}(v)\\ &\geq \left(\epsilon N - \frac{32\|\Delta\|^2}{\epsilon^3 N}\right)^2 - N\|\Delta\| \geq \epsilon^2N^2 - \frac{64\|\Delta\|^2}{\epsilon^3} - N\|\Delta\|\end{aligned} \tag{8}$$

The first inequality is because of Lemma 4, which proved the estimation $\mathrm{sgn}(v)^H\mathrm{sgn}(z) \geq N - \frac{32\|\Delta\|^2}{\epsilon^2 N}$ and leads to the following results $\mathrm{sgn}(v)^H z \geq \epsilon N - \frac{32\|\Delta\|^2}{\epsilon^3 N}$. At the same time, all the bad fixed points x satisfies

$$x^H C x = |x^H z|^2 + x^H \Delta x \leq \frac{64\|\Delta\|^2}{\epsilon^2} + N\|\Delta\| \tag{9}$$

Combine (8), (9) with the assumption $\|\Delta\| \leq \frac{\epsilon^{3/2} N}{13}$ and $\alpha < \|\Delta\|$, we knows that the spectral initialization surpasses all the bad local points. □

For Normalized Generalized Power Method, we can prove a similar version.

Lemma 10. *If x is a fixed point of Normalized Generalized Power Methods (Algorithm 3), at least one of the following holds*

$$|z^H x| \geq N - \frac{4}{\epsilon}(\|\Delta\| + \alpha) \qquad \text{or} \qquad |z^H x| \leq \frac{4}{\epsilon}(\|\Delta\| + \alpha).$$

Furthermore, if $\|\Delta\| \leq \frac{n\epsilon}{13}$ and $\alpha < \|\Delta\|$, all the accumulation points x of Algorithm 2 satisfies $|z^H x| \geq n - 8\|\Delta\|$.

Proof. If x is a fixed point of Algorithm 3, then x satisfies $\|\mathring{C}x\|_1 = x^H \mathring{C} x$ (because $|(\mathring{C}x)_i| = \left\langle x_i, (\mathring{C}x)_i \right\rangle$ holds for all i) where $\mathring{C} = \mathrm{sgn}(z)\mathrm{sgn}(z)^H + \mathrm{diag}(\frac{1}{|z|})\Delta\mathrm{diag}(\frac{1}{|z|}) + \alpha\mathrm{diag}(\frac{1}{|z|^2})$. For simplicity, we denote $\mathring{\Delta} = \mathrm{diag}(\frac{1}{|z|})\Delta\mathrm{diag}(\frac{1}{|z|})$ in the following proof.

On one hand

$$x^H \mathring{C} x = |\mathrm{sgn}(z)^H x|^2 + x^H \mathring{\Delta} x + \alpha \sum_{i=1}^{N} \frac{1}{|z_i|} \leq |\mathrm{sgn}(z)^H x|^2 + \frac{N}{\epsilon}(\|\Delta\| + \alpha)$$

On the other hand

$$\|\mathring{C}x\|_1 = \sum_{i=1}^{N} |(\mathrm{sgn}(z)^H x)\mathrm{sgn}(z_i) + (\mathring{\Delta}x)_i + \alpha x./|z|| \geq N|\mathrm{sgn}(z)^H x| - \|\mathring{\Delta}x\|_1 - \frac{\alpha N}{\epsilon}$$

At the same time, $\|\mathring{\Delta}x\|_1 \leq \sqrt{N}\|\mathring{\Delta}x\|_2 \leq N\|\mathring{\Delta}\|$. Combine this with the two previous inequality, we get

$$|z^H x|(N - |z^H x|) \leq N\|\mathring{\Delta}\| + \frac{N}{\epsilon}(\|\Delta\| + 2\alpha) \leq \frac{2N}{\epsilon}(\|\Delta\| + \alpha).$$

The above inequality enforces that one of $|z^H x| \geq N - \frac{4}{\epsilon}(\|\Delta\| + \alpha)$ and $|z^H x| \leq \frac{4}{\epsilon}(\|\Delta\| + \alpha)$ holds. We call all the stationary point satisfies $|z^H x| \geq N - \frac{4}{\epsilon}(\|\Delta\| + \alpha)$ good stationary point and the stationary point satisfies $|z^H x| \leq \frac{4}{\epsilon}(\|\Delta\| + \alpha)$ bad stationary point. In the following discussion, we use Lemma 4 to show that the spectral initialization $\mathrm{sgn}(v)$ outperforms all the bad fixed points in terms of the potential function $\mathring{f}(x) = \frac{1}{2}x^H \left(\mathrm{diag}(\frac{1}{|z|})C\mathrm{diag}(\frac{1}{|z|})\right) x$. From Lemma 4, we have

$$\begin{aligned} \mathring{f}(\mathrm{sgn}(v)) &= |\mathrm{sgn}(v)^H z|^2 + \mathrm{sgn}(v)^H \mathring{\Delta} \mathrm{sgn}(v) \\ &\geq \left(N - \frac{32\|\Delta\|^2}{\epsilon^2 N}\right)^2 - \frac{N}{\epsilon}\|\Delta\| \geq N^2 - \frac{64\|\Delta\|^2}{\epsilon^2} - \frac{N}{\epsilon}\|\Delta\| \end{aligned} \tag{10}$$

The second equality is because $\frac{64\|\Delta\|^2}{\epsilon^2 N} \leq \|\mathrm{sgn}(v) - z\|^2 \leq 2(N - |z^H \mathrm{sgn}(v)|)$. At the same time, all the bad fixed points x satisfies

$$\mathring{f}(x) = |x^H \mathrm{sgn}(z)|^2 + x^H \mathring{\Delta} x \leq \frac{64\|\Delta\|^2}{\epsilon^2} + \frac{N\|\Delta\|}{\epsilon} \tag{11}$$

Combine (10), (11) with the assumption $\|\Delta\| \leq \frac{\epsilon N}{13}$ and $\alpha < \|\Delta\|$, we knows that the spectral initialization surpasses all the bad local points. □

Lemma 1 and Lemma 2 guaranteed the global optimality of the second order stationary point of problem (7). Thus in the next theorem, we verify the Hessian $S(x) = \mathrm{ddiag}(Cxx^H) - C$ the Riemann Hessian $S_{\mathcal{R}}(x) = \mathrm{ddiag}(\mathring{C}xx^H) - \mathring{C}$ is P.S.D over $\mathcal{T}_x\mathbb{T}^n$ at the final stationary point. Then we can conclude the global optimality of the converging point of the Generalized Power Method.

Theorem 6. *Given vector $z \in \mathbb{R}^N$. For matrix $C = zz^\top + \Delta$, where $\Delta \in \mathbb{R}^{N\times N}$ is a symmetric perturbation matrix. If $1 - \frac{\sqrt{3}}{2} < \epsilon \leq \min_{i\in[N]} |z_i|$, $\|\Delta\| \leq \frac{\epsilon'}{28}N$ and $\|\Delta\|_\infty \leq \frac{\epsilon'}{28}N$, where $\epsilon' = (\epsilon^2 + 2\epsilon - 2)$. When $\alpha \leq \|\Delta\|$, then the GPM converge to the unique global optimum in the quotient space $\mathbb{R}^n / \sim$.*

Proof. From Lemma 7, the Generalized Power Method must converge to a stationary point x. The fixed point of the generalized power methods satisfies $Sx = 0$, which leads to $(\tilde{C}x)_i x_i = |(\tilde{C}x)_i|$ and Lemma 9 guarantees convergence to the good stationary points satisfies $|z^H x| \geq \epsilon N - 4(\|\Delta\| + \alpha)$.

From Lemma 1, we also know that, to prove global optimality of x, it suffices to show that $u^H S u > 0$ holds for all $u \in \mathbb{C}^n$ such that $u \neq 0$ and $u^H x = 0$. This is because

$$\begin{aligned}
u^H S u &= \sum_{i=1}^{N} |u_i|^2 |(C_i x)_i| - u^T C u \\
&= \sum_{i=1}^{n} |u_i|^2 \left| |z^H x| z_i + (\Delta x)_i \right| - |u^H z|^2 - u^H \Delta u \\
&\geq \sum_{i=1}^{n} |u_i|^2 \left(\epsilon |z^H x| - |(\Delta x)_i| \right) - |u^H (z - x)|^2 - u^H \Delta u \\
&\geq \|u\|^2 \left(\epsilon |z^H x| - \|\Delta x\|_\infty - \|z - x\|_2^2 - \|\Delta\| \right) \\
&\geq \|u\|^2 \left((2+\epsilon)(\epsilon N - 4(\|\Delta\| + \alpha)) - 2N - \|\Delta x\|_\infty - \|\Delta\| \right) \\
&\geq \|u\|^2 ((\epsilon^2 + 2\epsilon - 2)N - (9 + 4\epsilon)\|\Delta\| - \|\Delta\|_\infty)
\end{aligned} \tag{12}$$

Based on the assumption $\|\Delta\| \leq \frac{\epsilon'}{28}N$ and $\|\Delta z\|_\infty \leq \frac{\epsilon'}{28}N$, we know that $u^H S u > 0$. □

Theorem 7. *Given vector $z \in \mathbb{R}^N$ and there exists a constant $\epsilon > 0$ such that $\epsilon \leq \min_{i\in[N]} |z_i|$. For matrix $C = zz^\top + \Delta$, where $\Delta \in \mathbb{R}^{N\times N}$ is a symmetric perturbation matrix. If $\|\Delta\| \leq \frac{\epsilon}{28}N$ and $\|\Delta\|_\infty \leq \frac{\epsilon}{28}N$, when $\alpha \leq \|\Delta\|$, then the normalized GPM converge to the unique global optimum in the quotient space $\mathbb{R}^n / \sim$.*

Proof. From Lemma 8, the Normalized Generalized Power Methods must converge to a stationary point x. The first order condition of the stationary point is $\|\mathring{C}x\|_1 = x^H \mathring{C} x$ and Lemma 10 guarantees convergence to the good stationary points satisfies $|z^H x| \geq N - \frac{4}{\epsilon}(\|\Delta\| + \alpha)$.

Observe that $S_{\mathcal{R}}(x)x = 0$, according to Lemma 2, the only thing we need to prove global optimality of the converged stationary point x is to verify the $u^H S_{\mathcal{R}}(x) u > 0$ for all $u^H x = 0$.

$$\begin{aligned}
u^H S u &= \sum_{i=1}^{N} |u_i|^2 |(\mathring{C}_i x)_i| - u^T \mathring{C} u \\
&= \sum_{i=1}^{n} |u_i|^2 \left| |\mathrm{sgn}(z)^H x| \mathrm{sgn}(z_i) + (\mathring{\Delta} x)_i \right| - |u^H \mathrm{sgn}(z)|^2 - u^H \mathring{\Delta} u \\
&\geq \sum_{i=1}^{n} |u_i|^2 \left(|\mathrm{sgn}(z)^H x| - |(\mathring{\Delta} x)_i| \right) - |u^H (\mathrm{sgn}(z) - x)|^2 - u^H \Delta u - \alpha \|u\|_2^2 \\
&\geq \|u\|^2 \left(|\mathrm{sgn}(z)^H x| - \|\mathring{\Delta} x\|_\infty - \|\mathrm{sgn}(z) - x\|_2^2 - \|\mathring{\Delta}\| \right) \\
&\geq \|u\|^2 \left(N - \frac{12}{\epsilon}(\|\Delta\| + \alpha) - \frac{1}{\epsilon}\|\Delta\|_\infty - \frac{1}{\epsilon}\|\Delta\| \right)
\end{aligned} \tag{13}$$

Based on the assumption $\|\Delta\| \leq \frac{\epsilon}{28}N$ and $\|\Delta z\|_\infty \leq \frac{\epsilon}{28}N$, we know that $u^H S u > 0$. □

C.2.3 Linear Rate Convergence

In this section, following [15], we provide the proof of linear rate convergence of the normalized Generalized Power Method on our problem. With out loss of generality, we assume $1 = \arg\min_{\theta \in \{1,-1\}} \|\theta y^k - z\|_2$ for all $k \in \mathbb{N}$, where $\{y^k\}_{k\in\mathbb{N}}$ is the iterates generated by the normalized generalized power method.

Theorem 8 (Estimation Bound)**.** *Suppose that $\|\Delta\| \leq \frac{N\epsilon}{16}$ and $\alpha < \frac{N\epsilon}{6}$. Then the iterates $\{y^k\}_{k\in\mathbb{N}}$ generated by the normalized generalized power method satisfies*

$$\|y^{k+1} - \text{sgn}(z)\| \leq \mu^{k+1}\|y^0, \text{sgn}(z)\| + \frac{\nu}{1-\mu}\frac{8\|\Delta\|}{\epsilon\sqrt{N}}$$

for all $k \in \mathbb{N}$, where

$$\mu = \frac{16(\alpha + \|\Delta\|)}{(7N\epsilon - 8\alpha)} < 1, \nu = \frac{2N}{7N - 8\frac{\alpha}{\epsilon}}.$$

Proof. This Theorem is a direct adaptation of [15, Theorem 3.1]. □

Based on the previous estimation bound. We can build the local error bounds to guarantees global convergence. The local error bounds provide an estimation of the distance between any points in sgn(z)'s neighborhood and the global optima of the original optimization problem. To do this, we first define two mappings $\Sigma : \mathbb{T}^n \to \mathbb{H}^n$ and $\rho : \mathbb{T}^n \to \mathbb{R}+$ as

$$\Sigma(z) = \text{diag}(|\mathring{C}z|) - \mathring{C}, \quad \rho(z) = \|\Sigma(z)z\|_2.$$

Then we can have the following results

Lemma 11. *We denote z^* the global optimum of problem (7) and $\{y^{(k)}\}_{k\in\mathbb{N}}$ the iterates generated by the Normalized Generalized Power Method. If $\alpha \leq \|\Delta\| \leq \frac{\epsilon}{216}N$ and $\|\Delta\|_\infty \leq \frac{\epsilon}{12}N$, then we have*

- *(Local Error Bound) $\|y - z^*\| \leq \frac{N}{4}\rho(y)$ holds for all*
- *$\rho(y^k) \leq a\|y^{k+1} - y^k\|_2$ holds for some constant a.*

Proof. **Proof of Local Error Bound** To prove the local error bound, we make the following decomposition $\|\Sigma(y)y\| \geq \|\Sigma(z^*)y\| - \|(\Sigma(y) - \Sigma(z^*))y\|$. We first build the lower bound of $\|(\Sigma(y) - \Sigma(z^*))y\|$ following [15, Proposition 4.2]

$$\begin{aligned}
\|(\Sigma(y) - \Sigma(z^*))y\| &= \||\mathring{C}y| - |\mathring{C}z^*|\|_2 \leq \|\mathring{C}(y - \hat{z})\| \\
&\leq \sqrt{N}|z^H(y - z^*)| + \frac{\alpha + \|\Delta\|}{\epsilon}\|y - z^*\| \\
&\leq \sqrt{N}|(z^H - z^*)^H(y - z^*)| + \sqrt{N}|(z^*)^H(y - z^*)| + \frac{\alpha + \|\Delta\|}{\epsilon}\|y - z^*\| \\
&\leq \frac{\alpha + 5\|\Delta\|}{\epsilon}\|y - z^*\| + \frac{1}{2}\|y - z^*\|^2
\end{aligned} \tag{14}$$

Similar to Theorem 7, we can then lower bound $\|\Sigma(z^*)y\| = \|\Sigma(z^*)\hat{y}\|$ where $\hat{y} = (I - \frac{1}{n}z^*(z^*)^H)(y - z^*)$ is the projection of $y - z^*$ onto the orthogonal complement of span($\hat{z}$). At the same time

$$\|\hat{u}\| \geq \|y - z^*\| - \left\|\frac{1}{n}z^*(z^*)^H(y - z^*)\right\| = \|y - z^*\| - \frac{\|y - z^*\|^2}{2\sqrt{N}}$$

where the last equality is because $\|y - z^*\|^2 = 2(N - |y^H z^*|)$. Similar to Theorem 7, we have

$$\begin{aligned}
\|\hat{y}\|\|\Sigma(z^*)\hat{y}\| &\geq \hat{y}^H\Sigma(z^*)\hat{y} = \hat{y}^H(\text{diag}(|\mathring{C}z|) - \mathring{C})\hat{y} \\
&= \sum_{i=1}^{n} |\hat{y}_i|^2 \left||\text{sgn}(z)^H x|\text{sgn}(z_i) + (\mathring{\Delta}x)_i\right| - |\hat{y}^H \text{sgn}(z)|^2 - \hat{y}^H\mathring{\Delta}\hat{y} \\
&\geq \|\hat{y}\|^2\left(N - \frac{12}{\epsilon}(\|\Delta\| + \alpha) - \frac{1}{\epsilon}\|\Delta\|_\infty - \frac{1}{\epsilon}\|\Delta\|\right).
\end{aligned} \tag{15}$$

Thus

$$\|\Sigma(z^*)\hat{y}\| \geq \left(\|y-z^*\| - \frac{\|y-z^*\|^2}{2\sqrt{N}}\right)\left(N - \frac{12}{\epsilon}(\|\Delta\|+\alpha) - \frac{1}{\epsilon}\|\Delta\|_\infty - \frac{1}{\epsilon}\|\Delta\|\right).$$

At the same time $\|y-z^*\|^2 \leq \|y-z^*\|(\|y-z\|+\|z-z^*\|) \leq \left(\frac{\sqrt{N}}{2}+\frac{4\|\Delta\|}{\sqrt{N}}\right)\|y-z^*\|$. Combining all the results we get and finally we have

$$\rho(z) \geq \left[\frac{N}{2} - \frac{18(\|\Delta\|+\alpha)}{\epsilon} - \frac{\|\Delta\|}{\epsilon}\right]\|y-z^*\|. \tag{16}$$

Based on the assumptions $\alpha \leq \|\Delta\| \leq \frac{\epsilon}{216}N$ and $\|\Delta\|_\infty \leq \frac{\epsilon}{12}N$, we know that $\rho(z) \geq \frac{N}{4}d(z,\hat{z})$
Proof of $\rho(y^k) \leq a\|y^{k+1}-y^k\|_2$ By definition of y^{k+1}, $\rho(y^k) = \|\text{diag}(|\mathring{C}y^k|)(y^{k+1}-y^k)\| \leq \|\text{diag}(|\mathring{C}y^k|)\|_\infty\|y^{k+1}-y^k\|$. At the same time, we have

$$\begin{aligned}\|\text{diag}(|\mathring{C}y^k|)\|_\infty &\leq \|zz^Hy^k\|_\infty + \frac{\alpha+\|\Delta\|_\infty}{\epsilon}\\ &= |z^Hy^k| + \frac{\alpha+\|\Delta\|_\infty}{\epsilon} \leq 2N.\end{aligned} \tag{17}$$

This leads to the estimate $\rho(y^k) \leq 2N\|y^{k+1}-y^k\|_2$

□

Theorem 9. *We make the same assumption as Lemma 11. We further assumes $\mathring{C} \succeq a_0 I$ for some constant $a' > 0$, then the normalized generalized power method linearly converge to the global optimum z^*.*

Remark 6. *In [15], the data generating process doesn't ensures the matrix C is P.S.D. Thus [15] should apply a lower bound on α to ensure $\mathring{C}$ is P.S.D. In our case, the matrix is the covariance matrix of a noisy dataset. Thus it is nature have P.S.D. $\mathring{C}$.*

Proof. For $\mathring{C} \succeq a_0 I$, it's obvious to have sufficient ascent $\mathring{f}(y^{k+1}) - \mathring{f}(y^k) \geq a_0\|y^{k+1}-y^k\|_2^2$ holds for every iteration ([14, Lemma 8],[15, Proposition 4.3(a)]). Thus $f(y^{k+1}) - f(y^k) \geq \epsilon a_0\|y^{k+1}-y^k\|_2^2$. Before we present the final linear convergence proof, we first prove that $f(z^*) - f(y^k) \leq a_1\|y^k - z^*\|^2$. This is because

$$\begin{aligned}f(z^*) - f(y^k) &\leq \frac{1}{\epsilon}(\mathring{f}(z^*) - \mathring{f}(y^k))\\ &= (y^k)^H\left(\text{diag}(|\mathring{C}z^*|) - \mathring{C}\right)y^k\\ &= (y^k - z^*)^H\left(\text{diag}(|\mathring{C}z^*|) - \mathring{C}\right)(y^k - z^*)\\ &\leq (\|\mathring{C}\| + \|\mathring{C}\|_\infty)\|y^k, z^*\|^2.\end{aligned} \tag{18}$$

Now we are equipped with all the inequality needed to provide a global convergence proof. According to [15, Proof of Theorem 4.1], we knows that the normalized generalized power methods convergence to the global optimum linearly. □

C.2.4 Generative Models

In this section, we'll discuss how the random data sampled form the linear fixed effect model (also referred to an "interactive fixed-effect model") satisfies the discordant assumptions we used to prove the global optimization results. We assume the outcomes are generated via the following linear factor model

$$Y_{jt} = \delta_t + D_{jt}\tau + \theta_t^T\mu_j + e_{jt}, \qquad \mathbb{E}[e_{jt}|\delta_t, \mu_j, D_{jt}] = 0, \text{Var}[e_{jt}|\delta_t, \mu_j, D_{jt}] = \sigma$$

where δ_t is the time fixed effect, μ_j is the unobserved common factors and θ_t is a vector of unknown factor loadings. ϵ_{jt} is the unobserved idiosyncratic noise. τ is the treatment effect we aim to estimate and D_{jt} is the 0-1 variable according to the treatment assignment to unit j at time t. In specific, in the pre-treatment period, $D_{jt} = 0$ for all $\forall j \in [N], t \in [T]$. Thus the outcome matrix $Y \in \mathbb{R}^{N\times T}$ can be written in the following compact matrix form

$$Y = \underbrace{\begin{bmatrix} \mu_1^\top & 1 \\ \mu_2^\top & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \mu_N^\top & 1 \end{bmatrix}}_{\mu} \underbrace{\begin{bmatrix} \theta_1 & \cdots & \theta_t \\ \delta_1 & \cdots & \delta_t \end{bmatrix}}_{\theta} + W$$

where W is a matrix whose entries are i.i.d. standard normal random variables denote the measurement noise. We consider the time factor is sampled from a underlying distribution $\begin{pmatrix} \theta_i \\ \delta_i \end{pmatrix} \sim p(\tilde{\theta}, \tilde{\Sigma})$ and $\Sigma_\theta \triangleq \tilde{\theta}\tilde{\theta}^\top + \tilde{\Sigma} = \mathbb{E} \begin{pmatrix} \theta_i \\ \delta_i \end{pmatrix} \begin{pmatrix} \theta_i \\ \delta_i \end{pmatrix}^\top$. Then we knows that $\mathbb{E}YY^\top = \mu\Sigma_\theta\mu^\top + \sigma I_N$. We first assume that Σ_θ is a non-degenerate covariance matrix.

Assumption 4. *Σ_θ is positive semi-definite.*

Assumption 2 means that the matrix $\Sigma \triangleq \mu\Sigma_\theta\mu^\top$ is rank $n - 1$. We assume $v = (w_i D_i)_{i=1}^n$ to be vector in the null space of Σ, where $(w_i, D_i)_{i=1}^n$ is the only realizable experiment profile in Assumption 2. To verify that the data generating processing satisfies the assumptions we made for global convergence. We further made the following assumptions to the regularity of the problem.

Assumption 5 (Regularity of the Problem). *We further assume the following regularity properties of the covariance matrix and random sample Y_t*

- $\|\Sigma^\dagger\| \leq C_1$ *holds for some constant* C_1.
- $\|\Sigma^\dagger\|_\infty \leq O(N^{c_1})$ *holds for some constant* $c_1 \geq 0$.
- $\|Y_t Y_t^\top\| \leq C_2$ *holds almost surely holds for some constant* C_2.
- $\|Y_t Y_t^\top\|_\infty \leq O(N^{c_2})$ *holds almost surely for some constant* $c_2 \geq 0$.

Bound $\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\|$ In the following paragraph, we bound the error between the iteration matrix with the rank one ground truth in ℓ_2 operator norm. To do this, we make the following decomposition

$$\begin{aligned} \|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\| &\leq \sigma N\|(YY^\top + \sigma I)^{-1} - (\Sigma + \sigma I)^{-1}\| + \|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\| \\ &\leq \sigma N\|(\Sigma + \sigma I)^{-1}\|\|YY^\top - \Sigma\| + \|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\| \end{aligned} \tag{19}$$

We first bound $\|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\|$. To bound this term, we use the geometric series expansion $\frac{\lambda}{\lambda + X} = \sum_{j=0}^\infty (-1)^j \left(\frac{\lambda}{X}\right)^{j+1}$. If $\sigma\|\Sigma^\dagger\| < 1$, then

$$\|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\| \leq N \sum_{j=0}^\infty \sigma^{j+1}\|\Sigma^\dagger\|^{j+1} = \frac{N\sigma\|\Sigma^\dagger\|}{1 - \sigma\|\Sigma^\dagger\|} \tag{20}$$

To bound $\|\sigma N(\Sigma + \sigma I)^{-1}\|\|YY^\top - \Sigma\|$, we first use the matrix Bernstein inequality [42, 43] to bound $\|YY^\top - \Sigma\|$. We know

$$\|YY^\top - \Sigma\| \leq \sqrt{\frac{C_2^2 \log(\delta)}{T}} + \frac{2C_2 \log(\delta)}{T}$$

with high probability $1 - \delta$. At the same time, we have

$$\|\sigma N(\Sigma + \sigma I)^{-1}\| \leq \|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\| + \|uu^\top\| \leq \frac{N}{(1 - \sigma\|\Sigma^\dagger\|_\infty)}.$$

Finally, we achieve

$$\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\| \leq \frac{N\sigma\|\Sigma^\dagger\|}{1 - \sigma\|\Sigma^\dagger\|} + \frac{N}{(1 - \sigma\|\Sigma^\dagger\|_\infty)}\sqrt{\frac{C_2^2 \log(\delta)}{T}}$$

holds with high probability $1 - \delta$.

Bound $\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\|_\infty$ In the following paragraph, we bound the error between the iteration matrix with the rank one ground truth in ℓ_∞ operator norm.

$$\begin{aligned}\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\|_\infty &\le \sigma N\|(YY^\top + \sigma I)^{-1} - (\Sigma + \sigma I)^{-1}\|_\infty + \|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\|_\infty \\ &\le \sigma N\|(\Sigma + \sigma I)^{-1}\|\|YY^\top - \Sigma\|_\infty + \|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\|_\infty \end{aligned} \tag{21}$$

We first bound $\|\sigma N(\Sigma+\sigma I)^{-1} - uu^\top\|_\infty$. To bound this term, we use the geometric series expansion $\frac{\lambda}{\lambda+X} = \sum_{j=0}^{\infty}(-1)^j \left(\frac{\lambda}{X}\right)^{j+1}$. If $\sigma\|\Sigma^\dagger\|_\infty < 1$, then

$$\|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\|_\infty \le N\sum_{j=0}^{\infty}\sigma^{j+1}\|\Sigma^\dagger\|_\infty^{j+1} = \frac{N\sigma\|\Sigma^\dagger\|_\infty}{1-\sigma\|\Sigma^\dagger\|_\infty} \tag{22}$$

To bound $\|\sigma N(\Sigma + \sigma I)^{-1}\|\|YY^\top - \Sigma\|_\infty$, we first use the matrix Bernstein inequality [42, 43] to bound $\|YY^\top - \Sigma\|_\infty$. We know

$$\|YY^\top - \Sigma\|_\infty \le \sqrt{\frac{N^{2c_2}\log(\delta)}{T}} + \frac{2N^{c_2}\log(\delta)}{T}$$

with high probability $1-\delta$. At the same time, we have

$$\|\sigma N(\Sigma + \sigma I)^{-1}\| \le \|\sigma N(\Sigma + \sigma I)^{-1} - uu^\top\|_\infty + \|uu^\top\|_\infty \le \frac{N}{\epsilon^2(1-\sigma\|\Sigma^\dagger\|_\infty)}.$$

We plug in all the bounds and finally get

$$\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\|_\infty \le \frac{N\sigma\|\Sigma^\dagger\|_\infty}{1-\sigma\|\Sigma^\dagger\|_\infty} + \frac{N}{\epsilon^2(1-\sigma\|\Sigma^\dagger\|_\infty)}\sqrt{\frac{N^{2c_2}\log(\delta)}{T}}$$

with high probability.

From the discussion in Appendix C.2.3, if we can bound both $\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\|$ and $\|\sigma N(YY^\top + \sigma I)^{-1} - uu^\top\|_\infty$ as $O(\epsilon N)$, then we can have global convergence results. It's easy to check that if we select $\sigma \le \Omega(\epsilon N^{-c_1}), T \ge \Omega(\epsilon^6 N^{2c_2})$, then the assumptions for global convergence holds.

Corollary 1. *If* $c_1 = 0$, i.e. *there exists some constant* C_1 *such that* $\|\Sigma^\dagger\|_\infty \le C_1$*, then the noise level* $\sigma \le \Omega(\epsilon)$ *and* $T > \Omega(\epsilon^6 N^{2c_2})$ *ensures global convergence of NormSPCD algorithm.*

D Supplementary Experiments

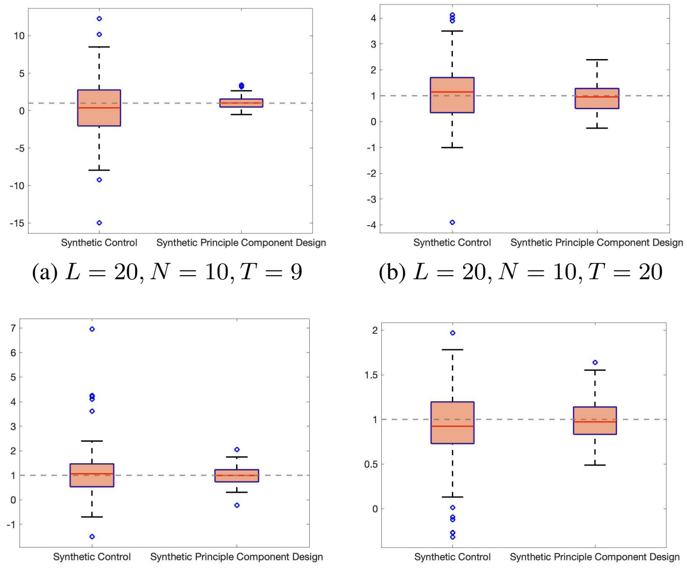
In this section, we'll introduce the experiment details and more experiments omitted from the main text due to the page limit.

We first introduce a simplified implementation of (Norm)SPCD, which although not guaranteed optimum but efficient, simple and effective in practice. In the simplified implementation, we don't solve the convex program (5) exactly, but using $w = \frac{2(YY^\top+\alpha I+\lambda\mathbb{1}\mathbb{1}^\top)^{-1}y^*}{\|(YY^\top+\alpha I+\lambda\mathbb{1}\mathbb{1}^\top)^{-1}y^*\|_1}$ to approximate instead. From 6, we know that once the optimal design profile y^* is obtained, then w is the optimal design weight. Notice that we don't exactly globally solve the problem (2) in the simplified implementation, although we obtained the right experiment profile y^* (Theorem 1). The weight we obtained here is the solve of the penalized approximation, but empirically it works good. The whole process is described in Algorithm 4. In all the experiment in this paper, we use this simplified implementation.

D.1 More simulated Examples

In this subsection, we run more simulated examples. On all examples, SPCD surpasses the original SC a large margin. All the data in this section is sampled from linear factor model (interactive fixed-effect model) [33, 19, 18]. The outcome Y comes from

$$Y_{it} = v_t^\top\gamma_i + \tau W_{it} + \epsilon_{it}, \forall i \in [N], \forall t \in [T+S].$$



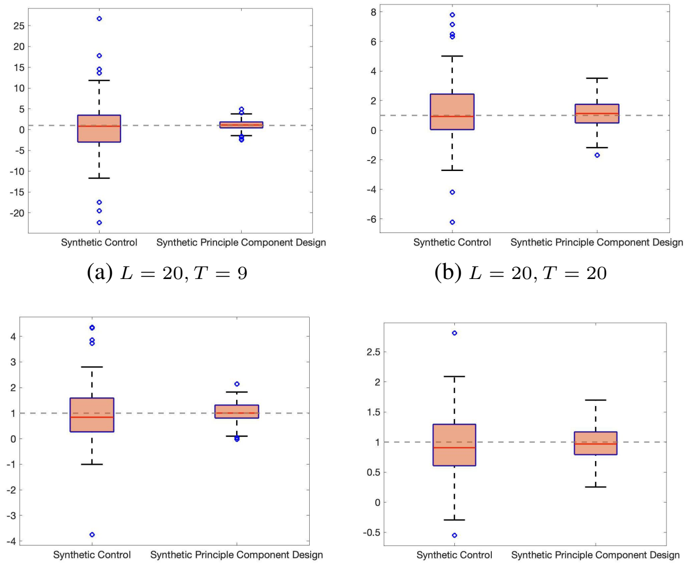
(a) $L = 20, N = 10, T = 9$

(b) $L = 20, N = 10, T = 20$

(c) $L = 8, N = 10, T = 9$

(d) $L = 8, N = 10, T = 20$

Figure 2: Treatment estimated via Synthetic Control and Synthetic Principle Design for data generated from pure random latent vector. We run the experiment over 100 runs of different seeds for different selections of L, T on data generated from purely random latent vector.In all cases, Synthetic Principle Design provides more robust estimate of the true treatment effect 1.



(a) $L = 20, T = 9$

(b) $L = 20, T = 20$

(c) $L = 8, T = 9$

(d) $L = 8, T = 20$

Figure 3: Treatment estimated via synthetic control (SC) and synthetic principal component design (SPCD) over 100 runs of different seeds for different selections of L, T.In all cases, Synthetic Principle Design provides more robust estimate of the true treatment effect 1.

where γ_i is a vector of latent unit factor of dimension L generated as a standard Gaussian and v_t is a vector of latent time factor. We tested different ways to select the following ways to generate the latent time factor. We fix the test time period $S = 10$ and number of units $N = 10$ and simulated different pairs of L, T selection.

Pure Random Latent Vector In this experiment, we follow [7] and run our algorithm on a synthetic dataset where all the time latent factors is sampled from random Gaussian. We sample the latent unit factor $v \in \mathbb{R}^{N \times T}$ and latent time factor $\gamma \in \mathbb{R}^{N \times T}$ both as random standard Gaussian matrices. We

Algorithm 4 Empirical Implementation of SPCD

```
Require: Pre-treatment Observations $Y \in \mathbb{R}^{T\times N}$
  Set initial treatment assignment guess through $y^0 = \mathrm{sgn}(v)$, where $v$ is the smallest eigenvector of
  matrix $(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)$, where $\alpha, \lambda$ are two pre-defined hyper-parameter.
                                                                    ▷ Spectral Initialization
  while Converged do
      Select one of the following two boxes to iterate

      For SPCD, update the design via                          ▷ Generalized power methods
          $y^{t+1} = \mathrm{sgn}\left[\left((YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta I\right) y^t\right],$      (23)
      where $\beta$ is a pre-defined hyper-parameter.

      For NormSPCD, update the design via   ▷ Normalize the inverse covariance matrix
          $y^{t+1} = \mathrm{sgn}\left[\left[(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta I\right] (y^t/d)\right],$      (24)
      where $d = \sqrt{\mathrm{diag}((YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1})}$ and $/$ denotes element-wise divide.
  end while
  Once obtained the optimal design $y^*$, one can select the design weight $w$ via
          $w = \frac{2(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} y^*}{\|(YY^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} y^*\|_1}$      (25)
                                          ▷ The optimality condition ensures $\mathrm{sgn}(w) = y$.
  Treat Unit $i$ if $y(i) = -\mathrm{sgn}\left(\sum_{i=1}^N y(i)\right)$ and run the experiment.
  Estimate the treatment effect via
          $\hat{\tau} = \sum_{t=1}^{S}\left(\sum_{i=1}^{N} w(i) Y_{i,T+t}\right)$
```

fix the test time period $S = 10$ and number of units $N = 10$ and simulated different pairs of L, T selection. The final results is shown in Figure 2.

Time Varying Factor In this experiment, we generate the time factor vector v_t as $t - \frac{T+S}{2} + \epsilon_t$, where $t - \frac{T+S}{2}$ is time trend term and e_{it} are i.i.d. standard Gaussian noise. The final results are shown in Figure 3.

AR(1) Process In this experiment, we follow [44] and run our algorithm on a synthetic dataset where the time latent factors is sampled from an AR(1) process. In particular the time factor $\gamma = [\gamma_1, \gamma_2, \cdots, \gamma_T]' \in \mathbb{R}^{N\times T}$ is sampled via

- $\gamma_1 \sim \mathcal{N}(0, I_N)$,
- $\gamma_{t+1} = A\gamma_1 + b + \sigma\epsilon, \epsilon \sim \mathcal{N}(0, I_N)$.

In our experiment, we take $A = 0.7I_N$, $b = \mathbb{1}$ and $\sigma = 1$. The final result is shown in Figure 5. The selected control and treatment group is shown in Figure 4. Synthetic principal component Design select the treated units whose features are representative of the whole aggregate market of interest [7] and the estimated treatment effect improves a lot respect to the original synthetic control estimator.

Details on Real World Data The Abadie–Diamond–Hainmueller Smoking data is first organized by [9]. In this paper, we use the organized version by [35] at `https://github.com/`

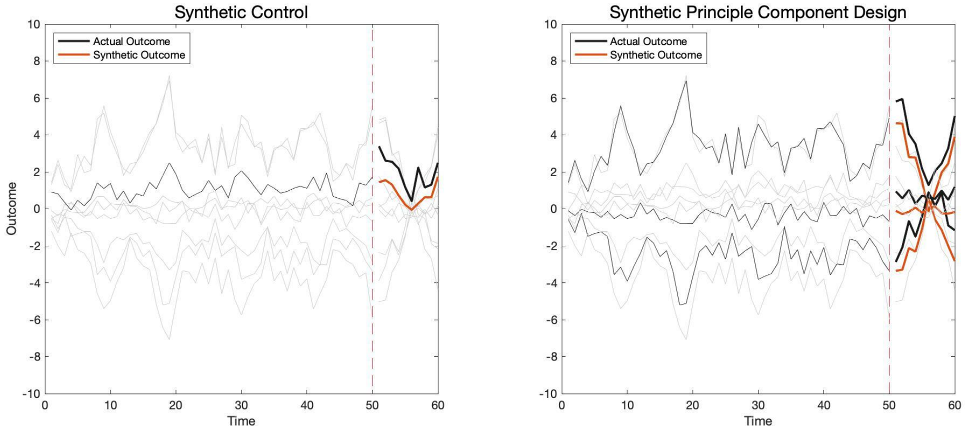
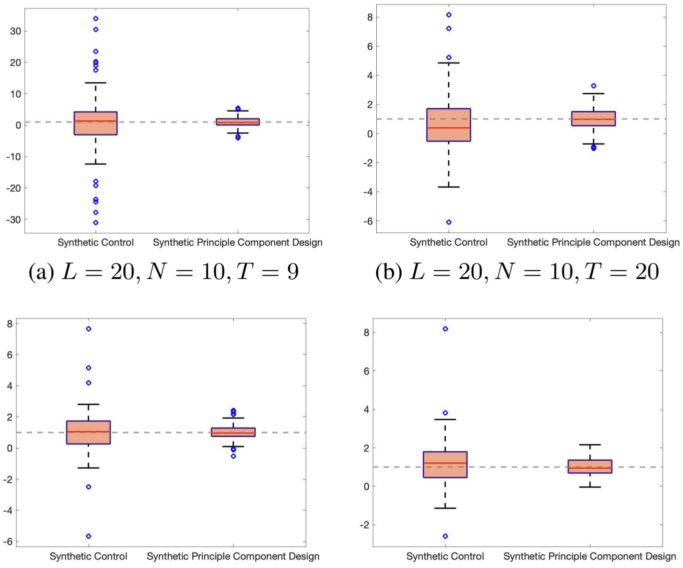


Figure 4: The experiment designed by SPCD for the Autoregressive model.



(a) $L = 20, N = 10, T = 9$

(b) $L = 20, N = 10, T = 20$

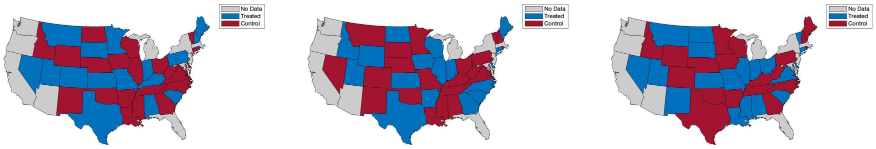
(c) $L = 8, N = 10, T = 9$

(d) $L = 8, N = 10, T = 20$

Figure 5: Treatment estimated via Synthetic Control and Synthetic Principle Design for data generated from an AR(1) process. We run the experiment over 100 runs of different seeds for different selections of L, T on data generated from purely random latent vector.In all cases, Synthetic Principle Design provides more robust estimate of the true treatment effect 1.

`synth-inference/synthdid/blob/master/data/california_prop99.csv` which drop the data of minimum wage laws, gun laws to abortion laws in the original data and only considers the smoking outcome data.

The BLS Statistics data is available from the BLS website. In this paper, we use the organized version by [35] at `https://github.com/synth-inference/synthdid/blob/master/experiments/bdm/data/urate_cps.csv`. We thank [35]'s authors carefully organize the data and open source it on github.



(a) $T = 15$ (b) $T = 20$ (c) $T = 30$

Figure 6: Selection of control and treatment group in the Abadie–Diamond–Hainmueller California Smoking Data when different pre-treatment period length T is available. The experiment design when $T = 25$ is shown in Figure 1a.