COMS 6998-5: Synthetic Control
Liushiya Chen (lc3501)

Optimal Treatment Units Selection for Synthetic Control

Motivation

When conducting experiments, many treatments have effects that can vary across the population of units. Here are few motivating examples:

1. Drug to lower blood pressure
   The drug might have different effects on different patients. Without knowing the details of each patient, we assume the effects are random.
2. Tutoring program for students
   The program has varying effectiveness for students at improving test scores.
3. Advertisement to increase spending on online store
   There are a variety of effects that ads could have on different customers. Some spend more as a result and some might spend less.

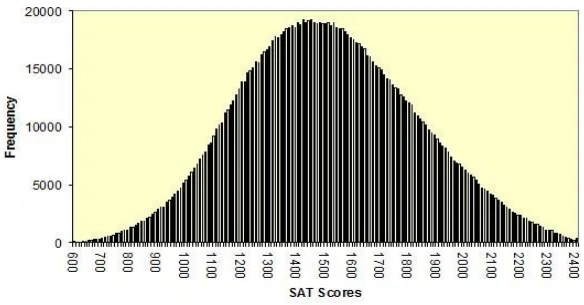
Normally Distributed Effects

For simplicity, treatments are assumed to have effects that are drawn from a known normal distribution. For instance, given k treatments given to k units, we can assume each treatment effect E_i is drawn I.I.D. from a normal distribution with known mean and variance.

$$E_{1..k} \sim Normal(\mu, \sigma^2)$$

This is a reasonable assumption because many human traits are normally distributed so their response to treatments is also likely to be normally distributed. For example, the distribution of SAT scores in 2010 (below) is distributed quite normally.

Figure 1 SAT Scores in 2010



COMS 6998-5: Synthetic Control
Liushiya Chen (lc3501)

Problem Formulation

Estimator of Average Treatment Effect

In evaluation experiments, the ultimate goal is to estimate the average treatment effect (ATE) of an effect over the population. I will denote the average treatment effect as μ in this case to be consistent with the above notations.

When using synthetic control to evaluate treatments, we are taking the difference between the outcomes of the treated units and the synthetic outcomes without the treatment (Equation 1). Only the former is observed and the latter is constructed using control units.

$$\begin{aligned}\widehat{\mu} &= \frac{1}{k}\sum_{i=1}^{k} Y_i^{treat} - Y_i^{synthetic} \\ &= \frac{1}{k}\sum_{i=1}^{k} E_i - error_i\end{aligned}$$

Equation 1: re-expression of the average treatment effect estimator

In the case of synthetic control, we can reformulate the estimator as the average difference between the randomly drawn treatment effect E_i and the prediction error of the synthetic control, effectively cancelling out the true counterfactuals.

Unbiased Assumption

The ATE estimator is assumed to be unbiased. The first component with E_i is simply the unbiased estimator of the mean. Then, we assume that the synthetic control's prediction error is also unbiased. Given a linear prediction model and assuming random variation in the underlying outcome Y, this is a reasonable assumption similar to the assumptions of Ordinary Least Squares.

If the synthetic control's predictions are biased in any way, it is worth investigating into the cause of the bias and correcting it before proceeding.

Estimator Variance

Given an unbiased estimator, we would always prefer an estimator with lower variance, since it provides more confidence in our estimates. This is especially important when the number of units is small.

Firstly, we assume that the treatment effect E_i is independent of the synthetic control prediction error. There is no reason why they will be dependent. This allows us to rewrite the variance as a linear combination of the variance of each of the two terms.

Furthermore, we assume that the variance of the prediction errors across treatment units are independent. This is a stronger assumption that might not always hold in practice because treatment units can deviate from predictions the same way. However, this assumption is no more unreasonable than the classic OLS assumptions.

$$\begin{aligned} Var[\widehat{\mu}] &= \frac{1}{k^2}\sum_{i=1}^{k} Var[E_i - error_i] \\ &= \frac{1}{k^2}\sum_{i=1}^{k} Var[E_i] - Var[error_i] \\ &= \frac{1}{k^2}\cdot k(\sigma^2 + \sigma^2_{error}) \\ &= \frac{1}{k}\cdot(\sigma^2 + \sigma^2_{error}) \end{aligned}$$

Equation 2: variance of the estimator

This gives us a final objective function to minimize. Since σ is a fixed quantity, we can vary k by choosing more or less treatment units, and vary the prediction error, by selecting appropriate units for treatment and control groups. The assumption is that not all units are the same and some make for better controls.

Optimization Problem

This is a classic subset selection problem where we have to select two subsets of the set of all units. This is NP-hard. We will need to evaluate the synthetic control predictions for every treatment group selection and this can be extremely time consuming.

Let U denote all units:

$$T^*, C^* = \arg\min \frac{1}{k^2} \cdot k(\sigma^2 + \sigma^2_{error})$$

Subject to constraints:

$$T, C \subset U, T \cap C = \emptyset, T \cup C = U$$

Equation 3: optimization problem formulation

Greedy Algorithm

An exact algorithm is likely infeasible, so heuristics are more practical. The most easily implemented and often sufficiently effective is the greedy algorithm. In our case, for each iteration, we want to pick the unit that minimizes the estimator's variance when we add that unit from the control group to the treatment group.

Pseudocode
For **C** in **control :**
1. Use **control \ {C}** to create synthetic **treatment + {C}**
2. Compute average estimation error using validation data

Set **C*** = **C** with the lowest average estimation error
Move **C*** from control to treatment
REPEAT UNTIL **control** IS EMPTY

Complexity

Assuming we run to completion, i.e. compute variance for 1, 2, …, |U|-1 treatment units, this algorithm requires $O(|U|^2)$ repetition of the synthetic control process. This polynomial time algorithm is much more scalable than exact algorithms.

Experimentation

Data

The data used is time series denoting the GDP of 30 major countries from 1995 - 2016 (22 years). Each data is normalized (1995 number is 1) and log differenced to produce a more stationary time series (Figure 4). This is necessary for accurate predictions using synthetic control.

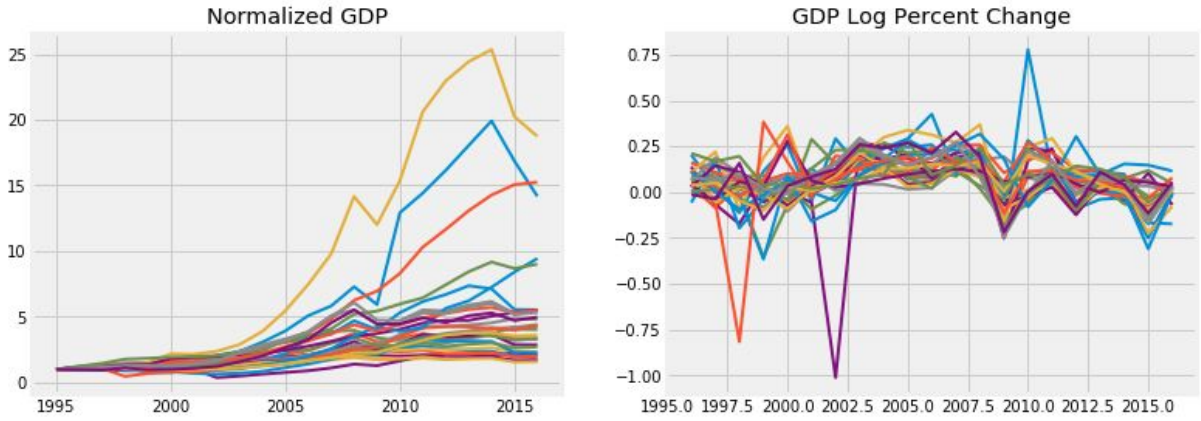


Figure 4: normalized time series (left) and final data after transformation (right)

Denoising

Standard Singular Value Decomposition (SVD) as introduced in class is used to denoise the data, removing some idiosyncratic variations. A total of 6 singular vectors were retained and this contributes to 98% of the cumulative energy (Figure 5).

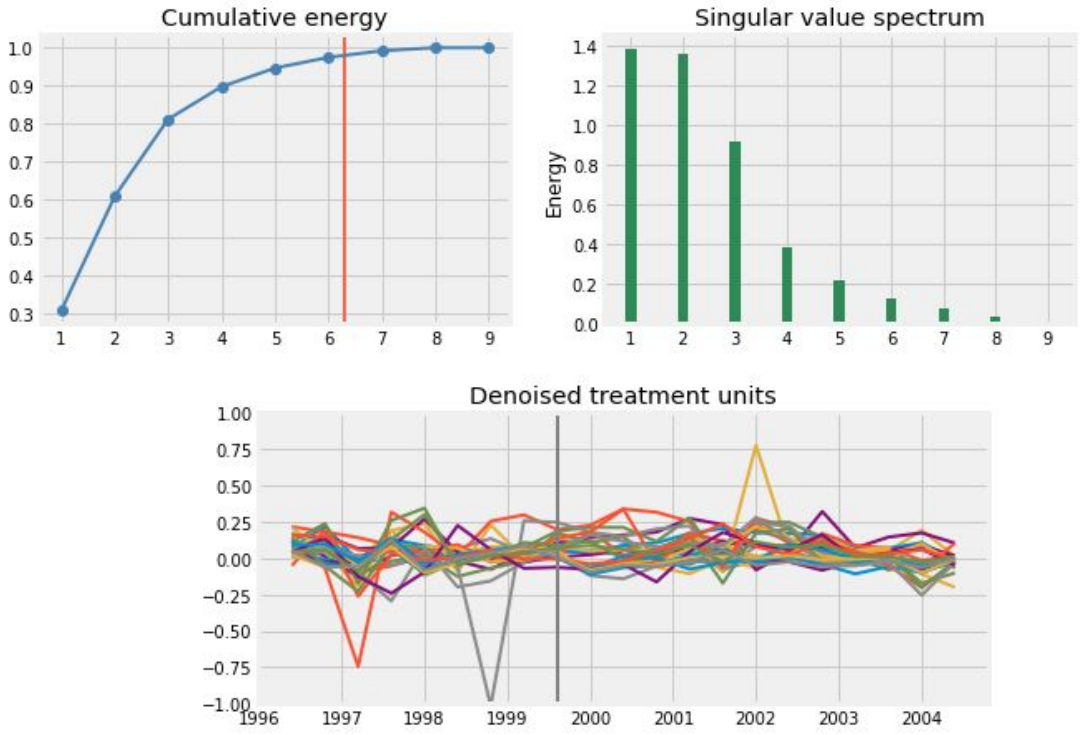


Figure 5: singular value spectrum (top) and denoised time series (bottom)

Experiments

Data splitting

The data is split into 3 sets in chronological order: *train*, *validation*, and *test* set. The train and validation are used to select the treatment units while the test set is the out-of-sample set that will be used to evaluate the performance of our algorithm.

The treatment is simulated and drawn from a known normal distribution.

Baseline algorithm

The random algorithm will be used as a baseline for comparison. Instead of picking the best unit at each iteration, the random algorithm selects a random unit to be assigned to the treatment set. The hypothesis is that the greedy algorithm should outperform the random algorithm on average except in extreme cases such as when we have a single control unit.

Results

To evaluate the algorithm, we will look at the estimator variance at each step. Each step k denotes the number of treatment units and the y-axis represents the variance. (Figure 6) The black line is the objective function, while the blue line represents the variance contribution from the treatment variation (σ) and the red line represents the variance contribution from synthetic control prediction error.

At each point along the x-axis, the algorithm is adding 1 unit to the treatment group.

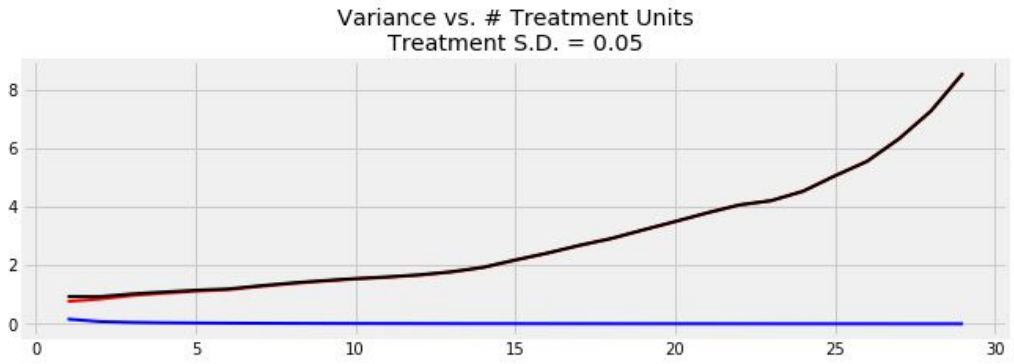


Figure 6: in-sample (validation set) variance across different k
treatment standard deviation (σ) = 0.05

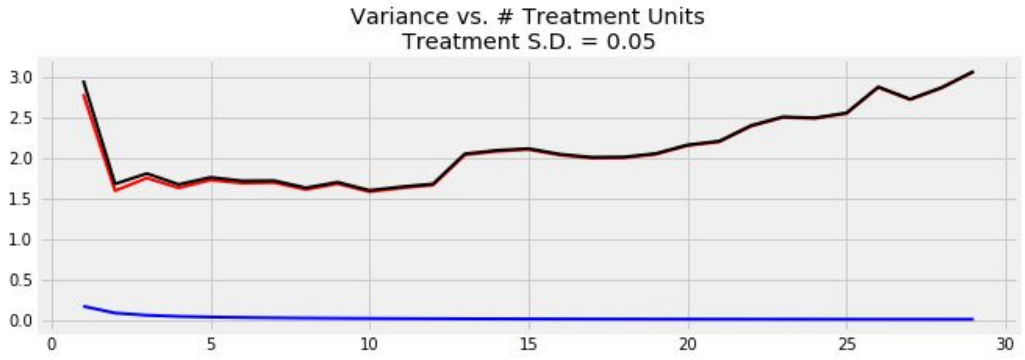


Figure 7: out-of-sample (test set) variance across different k
treatment standard deviation (σ) = 0.05

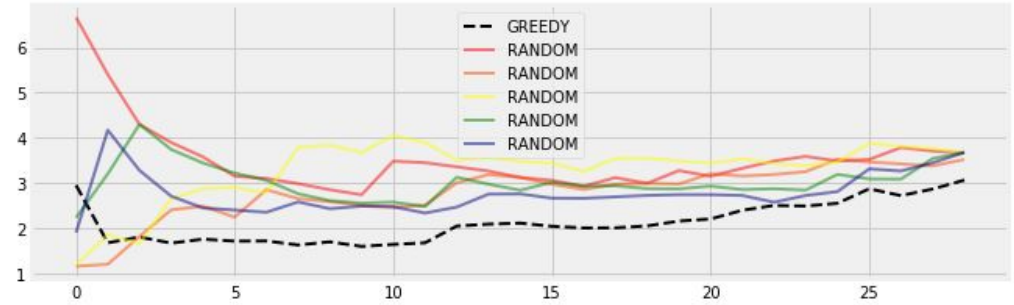


Figure 8: comparison of out-of-sample variance with 5 random algorithms

The out-of-sample variance (Figure 7) is noisier than in-sample as expected. In this case, the optimal selection of treatment units has size around 10 since that is where variance is optimized. Compared against 5 random algorithms, the greedy algorithm consistently performs better and is more stable.

Next, the treatment standard deviation is increased to 0.2 as in Figure 9. Naturally, we observe greater contribution from the treatment variance (decaying blue line).

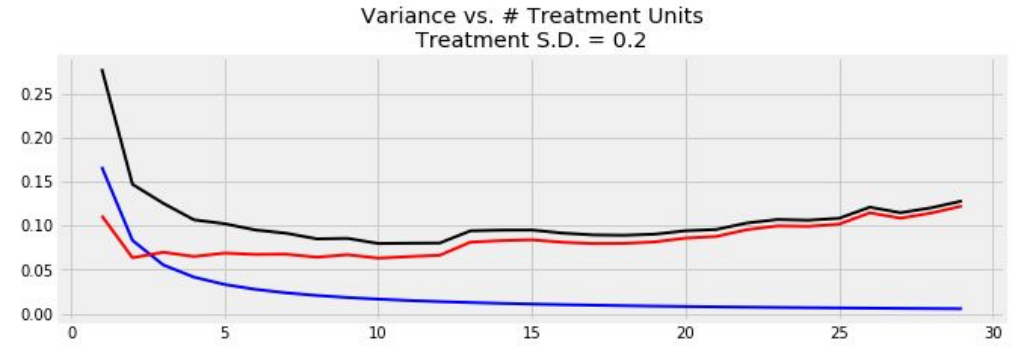


Figure 9: out-of-sample (test set) variance across different k
treatment standard deviation (σ) = 0.05

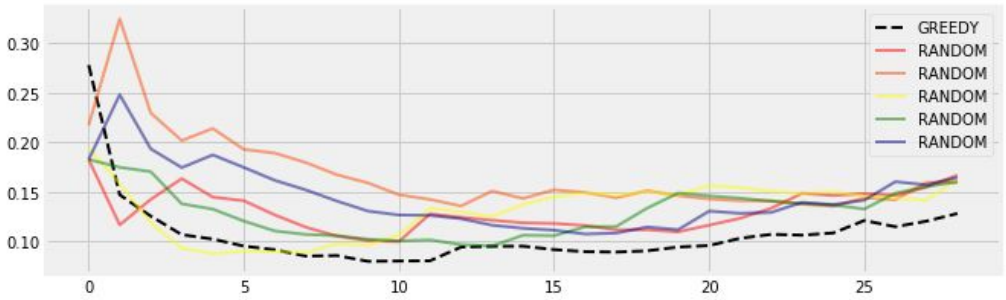


Figure 10: comparison of out-of-sample variance with 5 random algorithms

As we further increase the treatment standard deviation to 0.5, we see that the variance contribution from the treatment itself (blue) starts to dominate the variance from synthetic control prediction error (red) and make up a larger share of the total variance (Figure 11).

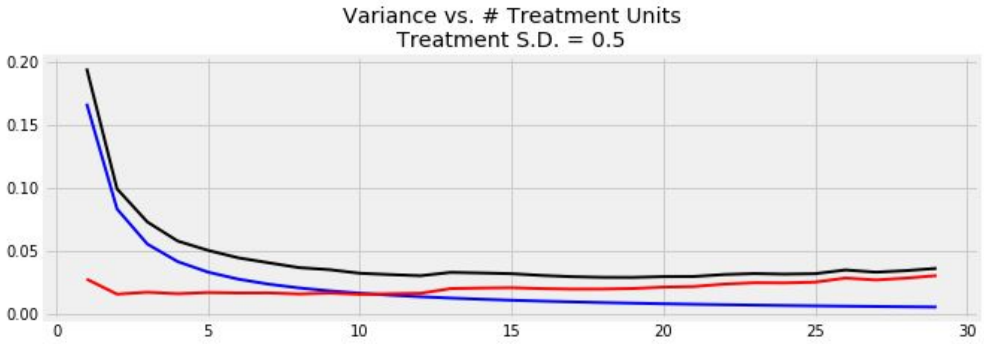


Figure 11: out-of-sample (test set) variance across different k
treatment standard deviation (σ) = 0.05

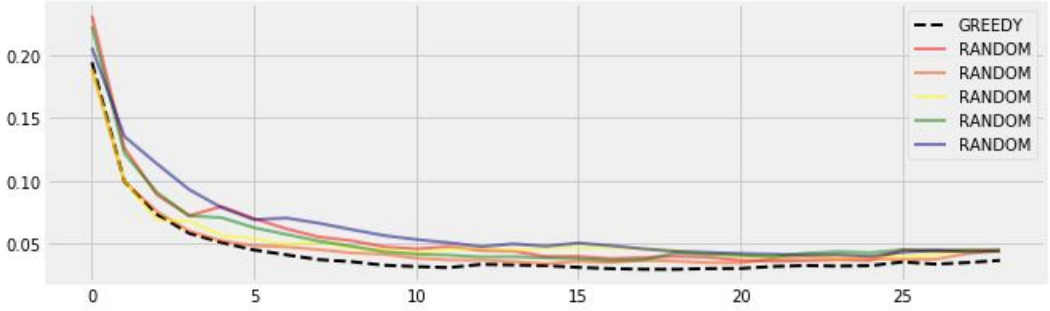


Figure 12: comparison of out-of-sample variance with 5 random algorithms

Compared with the 5 random algorithms, the difference is now a lot closer. This agrees with the theoretical results because as the treatment effect variance becomes dominant, it matters less which units we select as the treatment and which we retain as control. It is more important to simply have a large number of treatment units so we can sample more from the normal distribution of sample effects.

Conclusion

The greedy algorithm is better than the baseline random algorithm. The results are also consistent with theoretical results as we vary the variance of the treatment effect.

Future Work

Assumptions

The definition of estimator variance contains a lot of shaky assumptions, e.g. independence between synthetic control prediction error. In the GDP data for example, GDP of neighboring countries might be highly dependent. It might be worth investigating how these assumptions are violated and how to revise our objective function to account for correlated prediction errors.

Submodularity

From the algorithmic approach, the greedy algorithm does not guarantee optimal results and there is little theoretical bounds on this algorithm. However, there is clearly an element of submodularity in this problem which could be exploited to show a tight upper bound on the results (Lin and Blimes, 2010).

References

Amjad, Muhammad, et al. "mrsc: Multi-dimensional robust synthetic control." Proceedings of the ACM on Measurement and Analysis of Computing Systems 3.2 (2019): 1-27.

Klößner, Stefan, et al. "Comparative politics and the synthetic control method revisited: A note on Abadie et al.(2015)." Swiss journal of economics and statistics 154.1 (2018): 11.

Angrist, Joshua D., and Guido W. Imbens. Identification and estimation of local average treatment effects. No. t0118. National Bureau of Economic Research, 1995.

COMS 6998-5: Synthetic Control
Liushiya Chen (lc3501)

Lin, Hui, and Jeff Bilmes. "Multi-document summarization via budgeted maximization of submodular functions." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010.