Covariate Selection for the Synthetic Control Method

Byoung G. Park[1]; Yongki Yoon[1]
[1]University at Albany, SUNY

Abstract

This paper proposes a new procedure to select appropriate predictors for the synthetic control method (SCM) using the adaptive group LASSO. Traditional SCM assumes that a set of time-invariant predictors share the same synthetic control weights as the outcome of interest. We consider time-varying covariates, some of which have different synthetic control weights from the outcome of interest. Our method selects the valid predictors and estimates the synthetic control weights in a single step. We present a data-driven procedure to jointly optimize the selection of the penalty terms and the weights on each covariate. Our method is robust against potential biases from mis-specified predictors and enhances efficiency by fully exploiting the appropriate predictors. We prove that our method inherits the oracle property from LASSO, ensuring consistency in predictor selection and estimation. In Monte Carlo simulations, our method demonstrates strong finite-sample performance across various data generating processes.

Introduction

The Synthetic Control Method creates a counterfactual unit using a weighted average of controlled units. Often covariates are employed to estimate synthetic weight efficiently. SCM assumes there exists the synthetic weight, w^*, such that $Y_1 = \sum_{j=2}^{J} w_j^* Y_j$ and $X_1^k = \sum_{j=2}^{J} w_j^* X_j^k$ where $k \in \{1, \dots, q\}$ denotes the k-th covariate. While previous literature highlighted the importance of covariates, there was little guidance on how to select appropriate covariates. We propose a new, single-step procedure to select *valid* covariates and estimate the synthetic weight based on the above assumption using group LASSO.

Model & Methodology

Setup: Following potential outcome framework, we have to estimate $Y_{1t}(0)$ after the treatment, for $t > T_0$, where T_0 is the number of the pre-treatment period. SCM addresses this issue by $\widehat{Y_{1t}}(0) = \sum_{j=2}^{J} \widehat{w_j} Y_{jt}(0)$.

Key Innovation: We allow covariates to be time-varying to use full information:

$$X_{1t}^k = \sum_{j=2}^{J} (w_j^* + \delta_j^k) X_{jt}^k,$$

where δ^k represents the difference between outcome weight and k-th covariate's weight. If $\delta^k = (\delta_2^k, \dots, \delta_J^k)^\top = 0$, k-th covariate satisfies the assumption. If $\delta^k = (\delta_2^k, \dots, \delta_J^k)^\top \neq 0$, k-th covariate does NOT satisfy the assumption. We aim to choose only the *valid* covariates that the assumption holds.

We define new design matrices and weight vector such that:

$$Z_1 = \begin{pmatrix} \boldsymbol{\mathcal{Y}_1} \\ \mathcal{X}_1^1 \\ \vdots \\ \mathcal{X}_1^q \end{pmatrix}, Z_0 = \begin{pmatrix} \boldsymbol{\mathcal{Y}_0} & 0 & \cdots & 0 \\ \mathcal{X}_0^1 & \mathcal{X}_0^1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{X}_0^q & 0 & 0 & \mathcal{X}_0^q \end{pmatrix}, \omega = \begin{pmatrix} \boldsymbol{w^*} \\ \delta^1 \\ \vdots \\ \delta^q \end{pmatrix}$$

where $\mathcal{Y}_1$ and $\mathcal{Y}_0$ are outcome vector of treated unit and controlled units for pre-treatment period, $\mathcal{X}_1^1$ and $\mathcal{X}_0^1$ are covariate vector of treated unit and controlled units for pre-treatment period, respectively. Our parameter of interest is $\boldsymbol{w^*}$, the synthetic weight, and δ^1 to δ^q represent the deviation between two weights.

Estimation: We estimate weights by minimizing following optimization problem:

$$\widehat{\omega} = \text{argmin}_w \frac{1}{2} \parallel Z_1 - Z_0 w \parallel + n \sum_{k=1}^{q} \lambda \sqrt{p^k} \parallel w^k \parallel.$$

For the $k = 0$ group, which represents the outcome group, $\widehat{\omega} = \widehat{w}$ (primary interest). For covariate group, where $k = 1, \dots, q$, group LASSO automatically makes $\widehat{\omega^k} = \widehat{\delta^k} = \left(\widehat{\delta_2^k}, \dots, \widehat{\delta_J^k}\right) = 0$ for all $j = 2, \dots, J$ if δ^k is close to zero.

Theoretical Properties: Define $a_n = \min\{\lambda_k : 1 \leq k \leq q_1\}$ and $b_n = \max\{\lambda_k : k > q_1\}$ where $q_1 \leq Q$ is the number of important covariates.

- **Theorem 1 (Consistency)**: If $\sqrt{n} b_n \to_p 0$, then $\widehat{\omega} - \omega = O_p(n^{-\frac{1}{2}})$.
- **Theorem 2 (Selection Consistency)**: If $\sqrt{n} b_n \to_p 0$ and $\sqrt{n} a_n \to_p \infty$, then $P(\widehat{\omega}_a = 0) \to 1$.
- **Theorem 3 (Oracle Properties)**: If $\sqrt{n} b_n \to_p 0$ and $\sqrt{n} a_n \to_p \infty$, then $\sqrt{n}(\widehat{\omega}_b - \omega_b) \to_d N(0, \Sigma_b)$,

where ω_a and ω_b represent the weights for valid/invalid covariates, respectively.

Simulation

We generate i.i.d. sample $\{Y_{jt}, X_{jt}^k\}$ with $J = 21$ units for $T_0 = 50$ period. We set $q_1 = q_0 = 3$ thus we have 6 covariates. We generate the treated units' data as follows:

$$Y_{1t} = \sum_{j=2}^{J} w_j^* Y_{jt} + \varepsilon_{\{1t\}}^Y, \varepsilon_{1t}^Y \sim N(0,1)$$

$$X_{1t}^k = \sum_{j=2}^{J} (w_j^* + \delta_j^k) X_{jt}^k + \varepsilon_{\{1t\}}^k, \varepsilon_{1t}^k \sim N(0,1)$$

True weight w^* are set as $w_j^* = 1/(J-1)$ for all $j = 2, \dots, J$ and δ^k is set as $\delta^k = \left(0.5 - \frac{1}{J-1}, 0.5 - \frac{1}{J-1}, -\frac{1}{J-1}, \dots, -\frac{1}{J-1}\right)$. We compare our method with **a)** SCM with all covariates, **b)** pre-testing method, **c)** Post-LASSO, **d)** infeasible SCM, and **e)** SCM w/o any covariates. We evaluate the performance with **i)** no constraints, **ii)** non-negativity, **iii)** non-negativity with sum-to-one constraint based on RMSE and bias.

RMSE	No constraints	Non-negativity	Non-negativity & Sum-to-one
Infeasible	0.334	0.269	0.236
Park and Yoon	0.387	0.338	0.268
Pre-testing	0.736	0.304	0.281
SCM w/ all covariates	0.389	0.344	0.320
Post-LASSO	0.560	0.435	0.340
SCM w/o covariates	0.813	0.554	0.349

Table 1. Simulation results: RMSE.

bias	No constraints	Non-negativity	Non-negativity & Sum-to-one
Infeasible	0.025	0.056	0.018
SCM w/o covariates	0.060	0.206	0.022
Post-LASSO	0.171	0.186	0.078
Park and Yoon	0.240	0.220	0.110
Pre-testing	0.054	0.079	0.175
SCM w/ all covariates	0.287	0.281	0.270

Table 2. Simulation results: Bias.

Work in Progress

- We are currently working on simulation studies, especially on pre-testing. Likelihood Ratio test under inequality constraints does not work when the constraints are binding. Thus, we are studying Bootstrap method or sub-sampling to handle this issue.
- We are working on the role of V matrix. V matrix in SCM papers may play similar role to select covariates. Our research will primarily focus on two objectives: 1) precisely defining the role of the V matrix and 2) investigating methods for jointly optimizing V, ω, and the penalty parameter λ within our current algorithm,

Conclusions

We propose a new method to select valid covariates and estimate the synthetic weight in a single step. We show that our method inherits oracle properties. Moreover, our method performs well compared to alternative methods in finite samples.

Contact

Yongki Yoon
University at Albany, State University of New York
Email: yyoon@albany.edu
Website: yongkiyoon.github.io
Phone: 425-750-7012

References

1. Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, *59*(2), 391–425.
2. Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505.
3. Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method: COMPARATIVE POLITICS AND THE SYNTHETIC CONTROL METHOD. *American Journal of Political Science*, *59*(2), 495–510.
4. Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*(1), 113–132.
5. Botosaru, I., & Ferman, B. (2019). On the role of covariates in the synthetic control method. *The Econometrics Journal*, *22*(2), 117–130.
6. Ferman, B., Pinto, C., & Possebom, V. (2020). Cherry Picking with Synthetic Controls. *Journal of Policy Analysis and Management*, *39*(2), 510–532.
7. Imbens, G. and Rubin, D. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press, New York.
8. Wang, H., & Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, *52*(12), 5277–5286.
9. Vives-i-Bastida, J. (2022). *Predictor Selection for Synthetic Controls* (No. arXiv:2203.11576). arXiv.
10. Yuan, M., & Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67.
11. Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.