Randomization-Based Inference for Synthetic Control Estimators

Dmitry Arkhangelsky[1] David Hirshberg[2]

[1]CEMFI

[2]Emory QTM

DataX Workshop on Synthetic Control Methods
Princeton

Starting point

- Synthetic Control (SC) algorithm is a major addition to the standard causal inference toolkit for observational data
- Initially developed for the applications with a single treated unit, related ideas are now used if the number of treated units is large
- Properties of SC-type estimators are often studied in a particular outcome model: "low-rank + noise"
- Inference relies on the existence of "good" weights that correct for confounding

This paper

- Consider a different regime: large n, relatively smaller T, many treated units (more DiD-type setup)
- Rely on the assignment model (hence randomization-based)
- Allow for selection on unobservables and past outcomes
- Show that when the unobservables are "learnable" from the past and the number of treated units is small the SC estimator is asymptotically normal, but biased

Related literature

- **Synthetic control**: Ben-Michael, Feller, and Rothstein (2018); Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2018), Bottmer, Imbens, Spiess, Warnick (2021)
- **Balancing**: Graham, Pinto, and Egel (2012), Zubizarreta (2015); Hirshberg and Wager (2017), Ben-Michael, Feller, Hirshberg, and Zubizarreta (2021)
- **Panel data models**: Nikkel (1981); Anderson and Hsiao (1981); Arellano and Bond (1991), Arellano and Carrasco (2003)

Data structure

- Observe outcomes Y_{it} for n units over T periods
- Some units are treated in the last period, D_i is the treatment indicator, $\pi := \mathbb{P}_n D_i$ is the share of treated units
- Notation: $Y_i^{T_0} := (Y_{i1}, \ldots, Y_{iT_0})$, where $T_0 = T - 1$ (pretreatment outcomes)

Causal model

- Observed outcomes are connected to underlying potential outcomes in the standard way:

$$Y_{it} = W_{it} Y_{it}(1) + (1 - W_{it}) Y_{it}(0),$$

  where $W_{it} := D_i\{t = T\}$, i.e., assume no interference

- Object of interest is the conditional treatment effect on the treated:

$$\tau := \mathbb{P}_n \frac{D_i}{\pi} \left(Y_{iT}(1) - Y_{iT}(0) \right)$$

Algorithm

- Linear estimator with weights $\hat{\gamma}_i$:

$$\hat{\tau} := \mathbb{P}_n Y_{iT} \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right)$$

- Weights solve the entropy version of synthetic control:

$$\hat{\gamma} := \arg\min_{\gamma} \left\{ \frac{\eta^2}{n} \mathbb{P}_n \gamma_i \log(\gamma_i) + \frac{1}{2} \left\| \mathbb{P}_n Y_i^{T_0} \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right) \right\|_2^2 \right\}$$

$$\text{subject to: } \mathbb{P}_n \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right) = 0$$

Discussion

- For simplicity focus on a version of SC rather than Synthetic DID
- Entropy regularization:
  1. Makes the problem smoother (compared to l_2 penalty) thus helping with non-negativity constraint
  2. Connects nicely with the assignment models we consider
- Can include additional info (e.g., covariates) in the constraint set

Balance

- Using the representation for the potential outcomes:

$$\hat{\tau} = \tau + \mathbb{P}_n Y_{iT}(0) \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right) = \tau + \text{error}$$

- Estimator is well-behaved when the weights "balance" the part of the future baseline outcomes related to the selection indicators D_i
- To understand the properties of the error need to make statistical assumptions that connect outcomes to selection
- Depart from the standard analysis and focus on the the assignment model

Selection

- Assume $\{D_i\}_{i=1}^n$ are independent Bernoulli, each with probability π_i
- Model for the selection probabilities:

$$\pi_i = \frac{\exp\left(Y_i^{T_0}\beta_0 + \alpha_i\right)}{1 + \exp\left(Y_i^{T_0}\beta_0 + \alpha_i\right)}$$

  where α_i is a unit-specific unobserved heterogeneity
- Since α_i is not observed, logit model is without loss of generality, we impose restrictions later
- $Y_{iT}(0)$ does not enter into the expression for π_i but can be correlated with it through the past outcomes and α_i

Discussion

- Selection model nests several edge cases:
  1. Experiments: $\beta = 0, \alpha_i = \alpha$
  2. Selection on the past outcomes: $\alpha_i = \alpha$
  3. Selection on unobservables: $\beta_0 = 0$
- Low-rank + noise model assumes selection on unobservables, but not past outcomes
- In applications it is natural to allow idiosyncratic shocks to past outcomes to affect selection into treatment (sequential rather than strict exogeneity)

IPW weights

- Define "IPW" weights:

$$\gamma_i^{IPW} = \frac{1}{\pi} \times \frac{\pi_i}{1-\pi_i} \propto \exp(\beta_0^\top Y_i^{T_0} + \alpha_i)$$

- With IPW weights instead of $\hat{\gamma}_i$, the error is well-behaved regardless of the outcome model:

$$\text{error} = \frac{1}{\pi}\mathbb{P}_n(D_i - \pi_i)\frac{Y_{iT}(0)}{1-\pi_i} + o_p\left(\frac{1}{\sqrt{n}}\right)$$

- Are our empirical weights $\hat{\gamma}_i$ close to these weights in large samples?

Dual problem

- Consider the dual problem:

$$(\hat{\lambda}, \hat{\beta}) := \arg\min_{\lambda,\beta} \left\{ \mathbb{P}_n (1 - D_i) \exp\left(Y_i^{T_0} \beta + \lambda - 1 \right) - \mathbb{P}_n \frac{D_i}{\pi} \left(Y_i^{T_0} \beta + \lambda \right) + \frac{\eta^2}{n} \|\beta\|_2^2 \right\}$$

- Correspondence between primal and dual solutions:

$$\hat{\gamma}_i = (1 - D_i) \exp\left(Y_i^{T_0} \hat{\beta} + \hat{\lambda} - 1 \right)$$

- For any (β, λ) have the first order conditions:

$$\mathbb{P}_n \left((1 - D_i) \hat{\gamma}_i - \frac{D_i}{\pi} \right) \left(Y_i^{T_0} \beta + \lambda \right) = -\frac{\eta^2}{n} \hat{\beta}^\top \beta$$

Discussion

- Empirical weights approximately balance linear functions of the past
- Depend only on $Y_i^{T_0}$ and thus, in general, do not converge to IPW weights
- Key issue – selection on unobservables
- But, unobservables might be "learnable" from the past

Simple factor model

- Suppose the baseline outcomes follow a factor model:

$$Y_{it}(0) = \alpha_i \psi_t + \epsilon_{it},$$

  where α_i is has variance 1, and i.i.d ϵ_{it} have variance σ^2

- Consider the variance of the α_i after projecting it linearly on the past outcomes:

$$\mathbb{E}_0[(\alpha_i - \alpha_0 - Y_i^{T_0}\alpha_1)^2] = \frac{\sigma^2}{\sigma^2 + \|\Psi_{T_0}\|_2^2},$$

  where $\Psi_{T_0} = (\psi_1, \ldots, \psi_{T_0})$

- If α_i is visible in the past – $\|\Psi_{T_0}\|_2$ is large relative to the noise – then the variance is small

Localization

- Redefine parameters of the selection model using (π_i-weighted) projection of α_i on the past into account:

$$\tilde{\alpha}_i = \alpha_i - \alpha_0 - Y_i^{T_0}\alpha_1$$
$$\tilde{\beta}_0 = \beta_0 + \alpha_1$$

- Can express probabilities differently now:

$$\pi_i = \frac{\exp(Y_i^{T_0}\tilde{\beta}_0 + \alpha_0 + \tilde{\alpha}_i)}{1 + \exp(Y_i^{T_0}\tilde{\beta}_0 + \alpha_0 + \tilde{\alpha}_i)}$$

- If variance of $\tilde{\alpha}_i$ is small we can localize:

$$\pi_i \approx \pi_i^0 + \pi_i^0(1 - \pi_i^0)\tilde{\alpha}_i$$

where $\pi_i^0 = \frac{\exp(Y_i^{T_0}\tilde{\beta}_0 + \alpha_0)}{1 + \exp(Y_i^{T_0}\tilde{\beta}_0 + \alpha_0)}$; let $\gamma_i^0 = \frac{1}{\pi} \times \frac{\pi_i^0}{1 - \pi_i^0}$ – IPW weights in the model without unobservables

Error decomposition

- Split $Y_{iT}(0)$ into Y_i^{old} and Y_i^{new} – linear projection on the past (π_i-weighted) and residuals
- Decompose:

$$\hat{\tau} - \tau = \mathbb{P}_n Y_i^{old} \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right) + \mathbb{P}_n Y_i^{new} \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right)$$

- Expect the first term to be small because the weights balance linear functions of the past
- The second term will depend on the relationship between $\tilde{\alpha}_i$ and Y_i^{new}

Balancing the past

- Y_i^{old} is a linear function of the past:

$$Y_i^{old} = Y_i^{T_0}\beta_{old} + \lambda_{old}$$

- Using the first order conditions for the dual program can bound the past error:

$$\left|\mathbb{P}_n Y_i^{old}\left((1-D_i)\hat{\gamma}_i - \frac{D_i}{\hat{\pi}}\right)\right| = \frac{2\eta^2}{n}|\hat{\beta}^\top\beta_{old}| \leq \frac{2\eta^2}{n}\|\hat{\beta}\|_2\|\beta_{old}\|_2 \leq \frac{2\eta^2}{n}\left(\|\tilde{\beta}_0\|_2 + \|\hat{\beta}-\tilde{\beta}_0\|_2\right)\|\beta_{old}\|_2$$

- As soon as $\|\beta_{old}\|_2$ is bounded the second term is $O_p\left(\frac{\eta^2}{n}\right)$

Balancing the future

- Under additional technical conditions can substitute $\hat{\gamma}_i$ with γ_i^0:

$$\mathbb{P}_n Y_i^{new} \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right) \approx \mathbb{P}_n Y_i^{new} \left(\frac{D_i}{\pi} - (1 - D_i)\gamma_i^0 \right)$$

- The future is well-behaved:

$$\mathbb{P}_n Y_i^{new} \left(\frac{D_i}{\pi} - (1 - D_i)\hat{\gamma}_i \right) \approx \frac{1}{\pi} \mathbb{P}_n (D_i - \pi_i) \frac{Y_i^{new}}{1 - \pi_i^0} + \frac{1}{\pi} \mathbb{P}_n \pi_i^0 Y_i^{new} \tilde{\alpha}_i$$

Discussion

- Collecting the terms we get the linear expansion:

$$\hat{\tau} - \tau \approx \frac{1}{\pi}\mathbb{P}_n(D_i - \pi_i)\frac{Y_i^{new}}{1-\pi_i^0} + \frac{1}{\pi}\mathbb{P}_n\pi_i^0 Y_i^{new}\tilde{\alpha}_i$$

- If π is fixed and $T_0 = o(\sqrt{n})$ then the first term is centered, converges at $\sqrt{n}$ rate, and is asymptotically normal
- The second term generates bias as soon as π_i^0-weighted correlation between α_i and Y_i^{new} is non-zero
- In the simplest factor model this bias is $\frac{1}{T_0}$ – the second term dominates

Bringing terms together

- To make bias small need large T_0 but with large T_0 learning good weights is hard – the past is too complicated
- Solutions:
  1. Impose structure on the coefficients (sparsity) and regularize appropriately
  2. Alternatively – assume that π is small
- In this regime:

$$\hat{\tau} - \tau \approx \mathcal{N}\left(b, \frac{V}{\#\text{ treated units}}\right),$$

where $b = \mathbb{E}[Y_i^{new}\tilde{\alpha}_i | D_i = 1]$ and $V = \mathbb{E}[(Y_i^{new})^2 | D_i = 1]$

Conclusion

- In DiD regime (constant π, T_0) SC estimator is in general inconsistent when there is selection on both observables and unobservables
- If T_0 is large, and share of treated units is small – the SC estimator is asymptotically normal but biased
- Projection on the past outcomes does not eliminate all unobserved heterogeneity – leads to bias
- Bias is larger when the signal from the past outcomes is less relevant for the current outcomes
- Next steps: exact relationship between T_0 and π and analysis for SDID