Causal Inference with Panel Data

Lecture 4: Matching/Balancing and Hybrid Methods

Yiqing Xu (Stanford University)

Washington University in St. Louis

27 August 2021

This Lecture

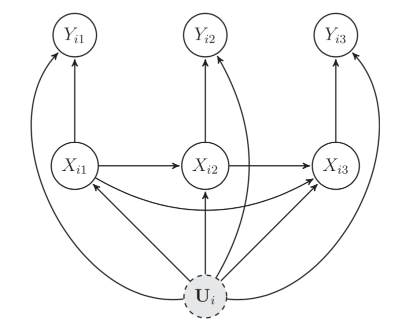
- Identification under sequential ignorability
    - Marginal structural models
    - Panel matching
    - Trajectory balancing
- Hybrid methods
    - Augmented synthetic control
    - Synthetic DiD

Identification under Sequential Ignorability

DAG for 2WFE

Recall that 2WFE require strong identification assumptions (Imai and Kim 2019)

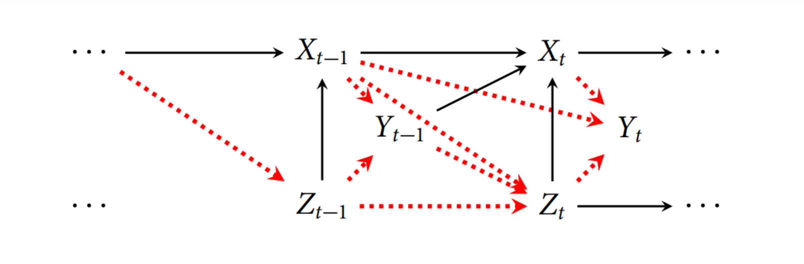
- No time-varying confounder
- No carryover effect
- No feedback



DAG under Sequential Ignorability

In reality, the more likely scenario (Blackwell and Glynn 2018):

- Contemporaneous effect: $X_t \to Y_t$
- Lagged effects:
  - $X_{t-1} \to Y_t$
  - $X_{t-1} \to Z_t \to Y_t$
  - $X_{t-1} \to Y_{t-1} \to Z_t \to Y_t$



Two Identification Regimes

- Strict exogeneity, which (roughly) corresponds to baseline randomized experiments

$$\{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp X_{is} \mid Z_{it}, \mathbf{U}_{it} \text{ (extractable)}$$

  - DiD, 2WFE, DiD_M, ...
  - Factor-augmented models (`fect`)
  * SCM (imho)

- Sequential ignorability, which corresponds to sequentially randomized experiments

$$\{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp X_{it} \mid Z_{i,1:t}, X_{i,1:(t-1)}, Y_{i,1:(t-1)}$$

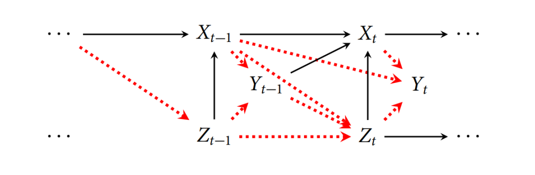
  - Marginal structural models (MSM)
  - Panel matching
  - Trajectory balancing (*)

MSM

Blackwell and Glynn (2018); Robins, Hernan & Brumback (2000)

- Motivation: conventional regression methods are biased

$$Y_{it} = \beta_0 + \alpha Y_{i,t-1} + \beta_1 X_{it} + \beta_2 X_{i,t-1} + Z_{it}'\delta + \varepsilon_{it}$$



β_2 is inconsistently estimated because Z_{it} is posttreatment

- Basic idea of MSM: model the "marginal" mean of potential outcomes as a function of treatment history

MSM

- Goal: estimating the average causal effect of a treatment history:

$$\tau(x_{1:t}, x'_{1:t}) = \mathbb{E}[Y_{it}(x_{1:t}) - Y_{it}(x'_{1:t})]$$

- Strategy: flexibly estimate $\mathbb{E}[Y_{it}(x_{1:t})] = g(x_{1:t}; \beta)$

- Challenge: the relationship between Y_{it} and $x_{1:t}$ is confounded by time-varying covariates and past outcomes

- Solution: use IPW to balance them out

$$\Pr[X_{it} = 1 | Z_{it}, Y_{i,t-1}, X_{i,t-1}] = f(Z_{it}, Y_{i,t-1}, X_{i,t-1}; \alpha)$$

$$\hat{w}_{it} = \Pi_{s=1}^{t} \frac{\widehat{\Pr}[X_{is} | X_{i,s-1}; \hat{\gamma}]}{\widehat{\Pr}[X_{is} | Z_{is}, Y_{i,s-1}, X_{i,s-1}; \hat{\alpha}]}$$

$$\text{plim } \mathbb{E}_{\hat{w}}[Y_{it} | X_{i,1:t} = x_{1:t}] = \mathbb{E}[Y_{it}(x_{1:t})]$$

- Limitations: many modeling choices; unstable weights (consider balancing weights); no additional confounding "fixed effects"

Panel Matching (Imai, Kim & Wang 2021)

Assumptions

- Sequential ignorability (past info can affect today's treatment)
- No cross-sectional spillover
- Allow limited carryover effect

Estimand

- Average Treatment Effect of Policy Change for the Treated (ATT):

$$\mathbb{E}[Y_{i,t+F}(X_{it}=1, X_{i,t-1}=0, \{X_{i,t-l}\})_{l=2}^{L}) - \\ Y_{i,t+F}(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-l}\})_{l=2}^{L}) \mid X_{it}=1, X_{i,t-1}=0]$$

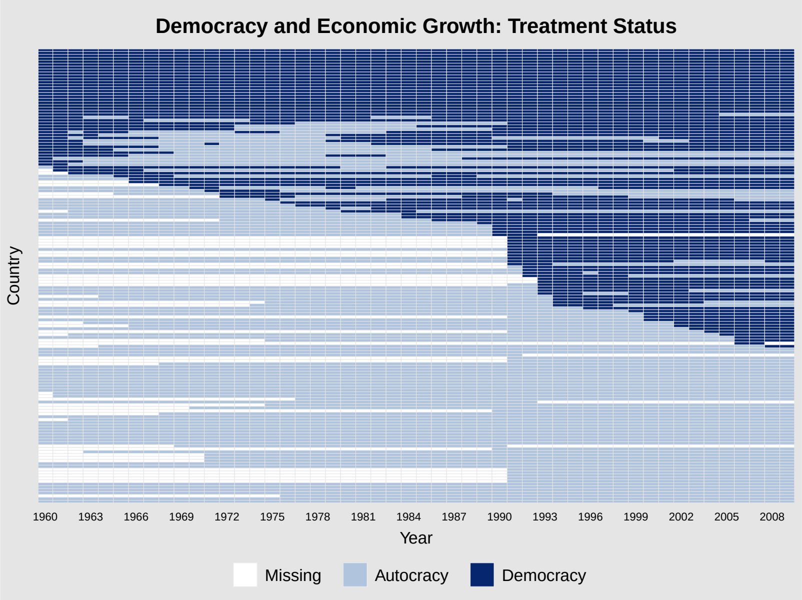
- Note that this is less ambitious than MSM as it focuses on "switches" only and forces the reminder of the treatment history to be the same or irrelevant

Panel Matching

Procedure

1. Create a matched set for each transition based on treatment history
2. Refine the matched set via any matching or weighting method
    - Mahalanobis distance matching
    - Propensity score weighting
3. Compute the ATT using the refined set
4. Calculate standard errors using block bootstrap or theoretical approximation

Example: Democracy and Economic Growth



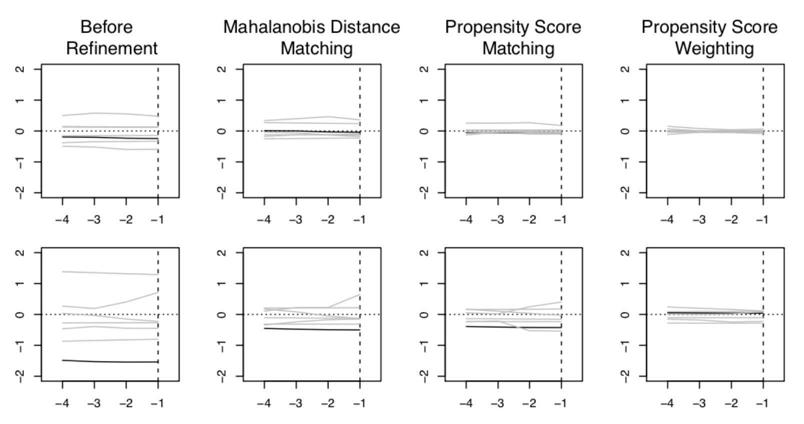
Example: Democracy and Economic Growth

- Match based on treatment history for the past L periods

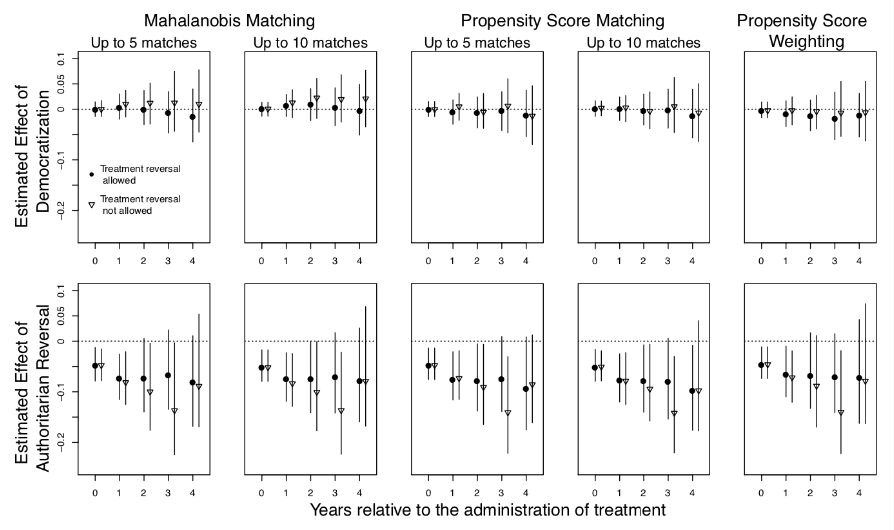
	Country	Year	Democracy	logGDP	Population	Trade
1	Argentina	1974	**1**	888.20	29.11	14.45
2	Argentina	1975	**1**	886.53	29.11	12.61
3	Argentina	1976	**0**	882.91	29.15	12.11
4	Argentina	1977	**0**	888.09	29.32	15.15
5	**Argentina**	**1978**	**0**	**881.99**	**29.57**	**15.54**
6	Argentina	1979	0	890.24	29.85	15.93
7	Argentina	1980	0	892.81	30.12	12.23
8	Argentina	1981	0	885.43	30.33	11.39
9	Argentina	1982	0	878.82	30.62	13.40
10	Thailand	1974	**1**	637.24	43.32	37.76
11	Thailand	1975	**1**	639.51	42.90	41.63
12	Thailand	1976	**0**	645.97	42.44	42.33
13	Thailand	1977	**0**	653.02	41.92	43.21
14	**Thailand**	**1978**	**1**	**660.57**	**41.39**	**42.66**
15	Thailand	1979	1	663.64	40.82	45.27
16	Thailand	1980	1	666.57	40.18	46.69
17	Thailand	1981	1	670.27	39.44	53.40
18	Thailand	1982	1	673.52	38.59	54.22

- Refine the matched set based on covariates and pre-treatment outcomes

The number of matched control units



Estimated treatment effects



Actually, I slightly misrepresented the method...

- The authors assume what they call sequential exogeneity instead of sequential ignorability (and use DiD to estimate the ATT)

$$\mathbb{E}[\epsilon_{it}|\{X_{i,1:t}\}, \mathbf{V_{i,t-1}}, \alpha_i, \gamma_t] = 0$$

- It implies parallel trends after conditioning

$$\mathbb{E}[Y_{it} - Y_{i,t-1}|X_{it} = 1, X_{i,t-1} = 0, \{X_{i,1:(t-2)}\}, \mathbf{V_{i,t-1}}] = \mathbb{E}[Y_{it} - Y_{i,t-1}|X_{it} = 0, X_{i,t-1} = 0, \{X_{i,1:(t-2)}\}, \mathbf{V_{i,t-1}}]$$

- Note that this assumption embeds functional-form requirements, e.g., the following outcome model works

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \sum_{l=1}^{4} \rho_l Y_{i,t-l} + \epsilon_{it}$$

  I don't know how specific or demanding they need to be

- Moreover, conditioning on past outcome in a DiD setting can lead to biaes if transitory shocks are an important part of $Y_{i,t-1}$ (Chabé-Ferret 2021)

Panel Matching: Pros and Cons

Advantages

- Require sequential ignorability/exogeneity instead of strict exogeneity
- Allow treatment reversal and limited carryover
- Weaker functional form assumptions
- Allow a variety of matching/reweighting methods

Limitations

- An arguably narrower focus
- Lots of data (w/ info on outcome dynamics) are dropped
- Normally, imbalances remain
- Many choices require user discretion

Balancing under Sequential Ignorability

- In Lecture 2, we briefly discussed a balancing algorithm for the SCM
- An algorithm can be used under different (treatment) designs
- Under sequential ignorability:

  $$\{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp X_{it} \mid Z_{i,1:t}, X_{1,1:(t-1)}, Y_{i,1:(t-1)}$$

  We want to balance on $\mathbf{V}_{it} = \{Z_{i,1:t}, X_{1,1:(t-1)}, Y_{i,1:(t-1)}\}$
- Challenge: we don't know the functional form of either $\Pr(X_{it} = 1|\mathbf{V_{it}})$ or $\mathbb{E}(Y_{it} = 1|\mathbf{V_{it}})$
- In other words, weights that achieve mean balancing can leave treated and control different on non-linear functions of $\mathbf{V}_{it}$

Trajectory Balancing

- Mean balancing: on original features (Robbins et al. 2017)

$$\sum_{i \in \mathcal{T}} q_i \mathbf{Y}_{i,pre} = \sum_{j \in \mathcal{C}} w_j \mathbf{Y}_{j,pre}$$

- Trajectory balancing: feature mapping $\mathbf{Y}_{i,pre} \mapsto \phi(\mathbf{Y}_{i,pre})$, then balance on the expanded features (Hazlett and Xu 2018):

$$\phi : \mathbb{R}^P \mapsto \mathbb{R}^{P'}$$
$$\sum_{i \in \mathcal{T}} q_i \phi(\mathbf{Y}_{i,pre}) = \sum_{j \in \mathcal{C}} w_j \phi(\mathbf{Y}_{j,pre})$$

- In practice: seek approximate balance, working from largest toward smallest principal components of $\mathbf{Y}_{pre}(\mathbf{Y}_{pre})'$ with a stopping rule of minimizing the upper bound of biases

Implementation

A good choice of $\phi()$ is one that:

- requires little or no user discretion
- includes all continuous functions (at the limit)
- perhaps, prioritizes low frequency, smoother functions
- allows covariates to play a role

Gaussian kernel then approximation via principal components

- form kernel matrix $\mathbf{K}_{i,j} = k([V_i],[V_j]) = exp(-||[V_i]-[V_j]||^2/h)$
- Replaces each unit's $[V_i]$ with a vector k_i encoding how similar observation i is to observation 1, 2, ...
- SVD this matrix to obtain components/ eigenvectors
- Choose weights to get mean balance on these, starting from largest
- We choose the number of principal components to include by minimizing the upper bound of bias in the ATT estimates

When Averages Fail and $\phi()$'s Thrive

Intuition: mean balancing is okay but may emphasize "wrong" features of the pre-treatment trend

- Trajectory balancing gets you similarity of whole trajectories rather than just equal means at each time point $\rightarrow$ balance on "higher-order" features such as variance, curvature, etc.
- Approximately, trajectory balancing gets multivariate distribution of $\mathbf{V}_i$ for the controls equal to that of the treated, whereas mean balancing only gets equal marginals
- This can matter when non-linear functions of $\mathbf{V}_i$ are confounders, especially when T_0 short

When Mean Balancing Fail: A Severe Example

- $N = 200$ countries with simulated GDP over years $T \in \{1, 2, ..., 24\}$
- Two "types" of countries:
  Volatile with no growth:

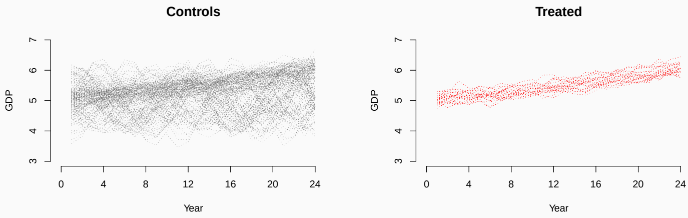
$$GDP_{it} = 5 + a_i sin(.2\pi t) + b_i cos(.2\pi t) + .1\epsilon_{it}$$
$$\epsilon_{it} \sim N(0,1), \quad a_i, b_i \sim U(-1,1)$$

  Or steady growing:

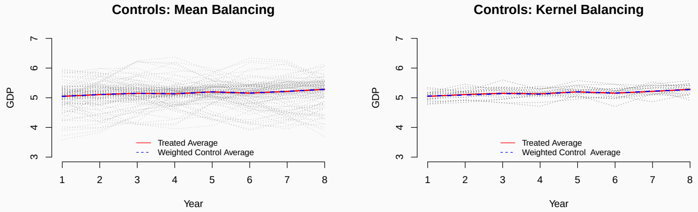
$$GDP_{it} = 4 + c_i 1.03^t + .1\epsilon_{it}$$
$$\epsilon_{it} \sim N(0,1), \quad c_i \sim U(0.9, 1.1)$$

- A randomly selected 25% of the stable type take the treatment.

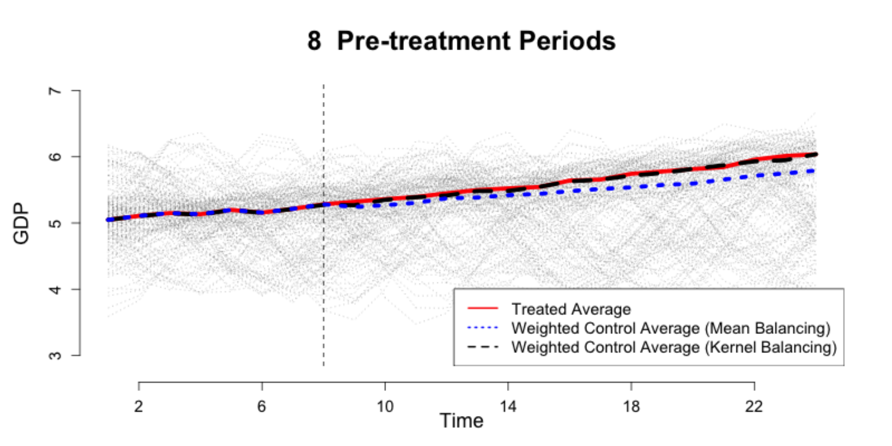
When Mean Balancing Fails: A Severe Example



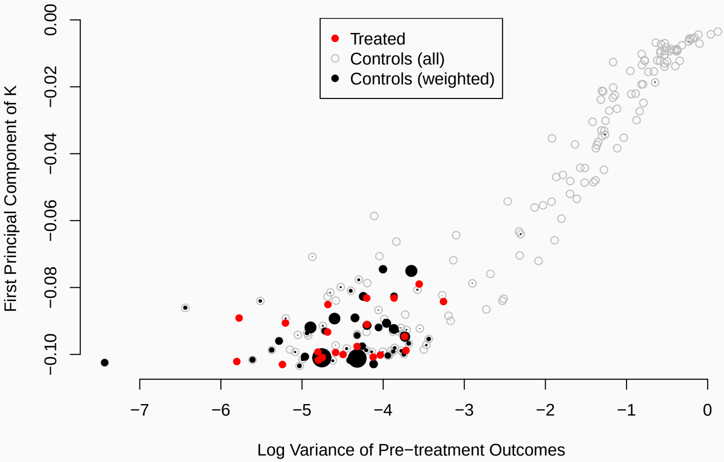
Heavily weighted control units (pre-treatment)



When Mean Balancing Fails: A Severe Example

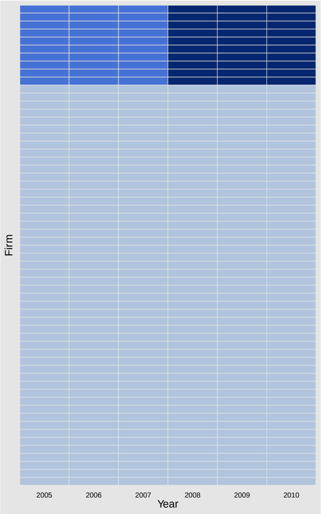


What Information is Encoded in the Kernel Matrix?

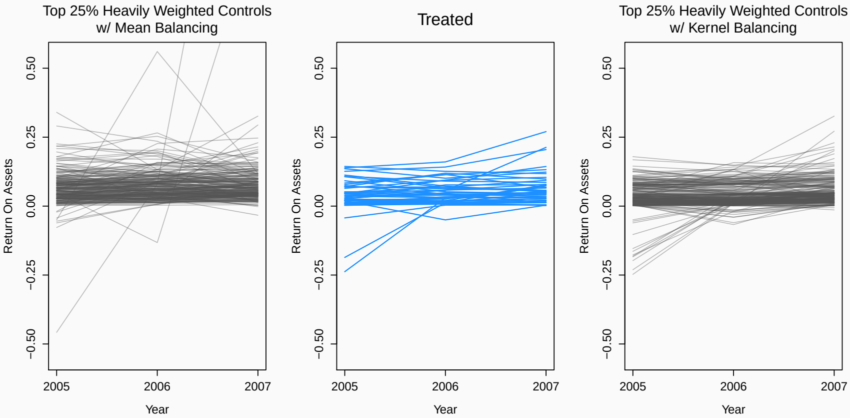


Truex (2014): Return to office in China's Parliament

- **Treatment**: CEO taking a seat in the National People's Congress (NPC)
  Outcome: Return on assets (ROA)
- 48 treated firms, 984 controls
  Pre-treatment: 2005-2007
  Post-treatment: 2008-2010
- Two covariates: state ownership, revenue in 2007
- Balancing on: `roa2005`, `roa2006`, `roa2007`, `so_portion`, `rev2007` (and higher order terms through a kernel transformation)



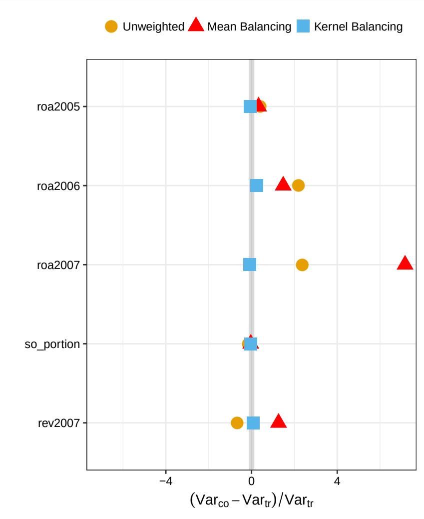
Balance on Pre-treatment Outcome Trajectories



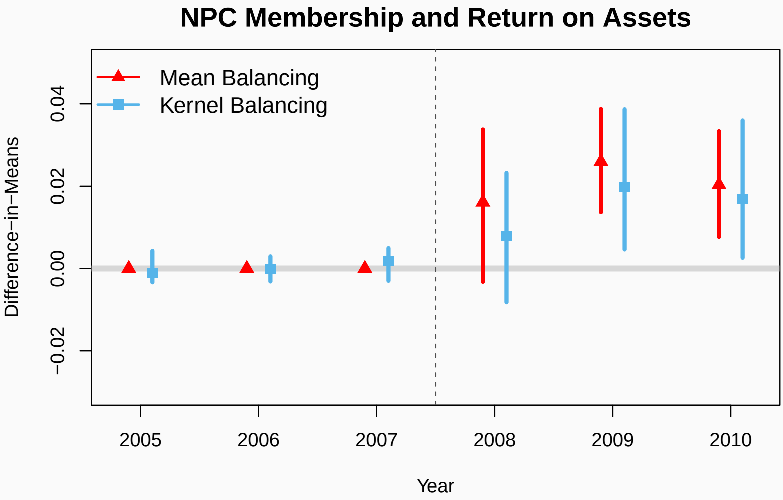
Balance Check

Mean

Variance



Truex (2014): Main Results



Summary

- Removing time-invariant confounders is costly, e.g., *no feedback*
- Sequential ignorability may be more desirable than strict exogeneity in many applied settings
- MSMs are nice but often require strong functional-form assumptions
- Panel non-parametric and semi-parametric methods are appealing but have limited applicability or are data hungry
- Things quickly get more complex when the number of different treatment histories grows
- Inference is hard with a small number of treated units

Hybrid Methods

Hybrid Methods

- So far, we've surveyed two group of methods: (1) those constructing balancing weights; (2) those modeling the conditional outcomes
- Combining the two approaches will likely produce doubly robust estimators
- Some methods we discussed, including semi-parametric DiD, panel matching, trajectory balancing, are already doing a simple version of it (balancing plus regression)
- We review two new methods that formally adopt this idea
  - Augmented synthetic control (Ben-Michael et al 2018): modeling first
  - Synthetic DiD (Arkhangelsky et al. 2019): weighting first

Augmented Synthetic Control (Ben-Michael et al 2017)

- Assuming unit 1 being treated ($D_1 = 1; D_{-1} = 0$), pretreatment covariates X_i
- **Basic Idea**
  1. Run an outcome model (e.g. Ridge, FEct, IFEct, MC, etc.) and obtain model fit $\hat{m}(X_i)$
  2. Balance on the residual averages, obtaining weights $\hat{\gamma}_i$ for the controls
  3. Treated average is constructed using:

$$\hat{Y}_1^{aug}(0) = \underbrace{\sum_{i \in \mathcal{C}} \hat{\gamma}_i Y_i}_{SCM} + \underbrace{\hat{m}(X_1) - \sum_{i \in \mathcal{C}} \hat{\gamma}_i \hat{m}(X_i)}_{debias}$$
$$= \hat{m}(X_1) + \sum_{i \in \mathcal{C}} \hat{\gamma}_i \left(Y_i - \hat{m}(X_i) \right)$$

- The balancing weights take care of the remaining biases from the outcome model; the estimator is thus doubly robust
- Inference via jackknife

Innovations

- Simplifying SCM (Robbins et al 2017)
    - computationally efficient
    - connection to IPW reweighting
- Combine outcome models with balancing weights
    - flexible and doubly robust
    - better balance, lower bias than either the outcome model or SCM alone
    - minimizing model dependency
- Example: Ridge-augmented SCM
    - better balance, lower bias than either ridge or SCM alone
    - can be represented as a weighting estimator (which allows negative weights)
    - connection to IPW reweighting

Simplifying SCM

- The original SCM

$$\min_{\gamma} \ (X_1 - X_0'\gamma)'\mathbf{V}(X_1 - X_0'\gamma)$$
$$s.t. \ \sum_{i \in \mathcal{C}} \gamma_i = 1; \quad \gamma_i \geq 0$$

- Entropy-penalized SCM (recall Robbins et al (2017) in Lecture 2)

$$\min_{\gamma} \ -\sum_{i \in \mathcal{C}} \gamma_i log \gamma_i$$
$$s.t. \ X_1 = X_0'\gamma; \quad \sum_{i \in \mathcal{C}} \gamma_i = 1$$

- Penalized SCM with exact balance is IPW

$$\hat{\gamma}_i = \frac{logit^{-1}(\hat{\alpha}+\hat{\beta}'X_i)}{1-logit^{-1}(\hat{\alpha}+\hat{\beta}'X_i)}$$

in which $\hat{\alpha}$, $\hat{\beta}$ are coefficients from a logit regression of D on X

Ridge-Augmented SCM

- The general form

$$\hat{Y}_1^{aug}(0) = \underbrace{\sum_{i \in \mathcal{C}} \hat{\gamma}_i Y_i}_{SCM} + \underbrace{\hat{m}(X_1) - \sum_{i \in \mathcal{C}} \hat{\gamma}_i \hat{m}(X_i)}_{debias}$$

- Ridge-augmented SCM

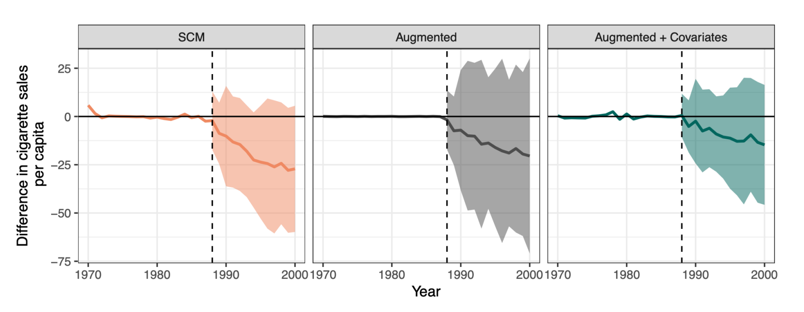
$$\hat{Y}_1^{aug}(0) = \underbrace{\sum_{i \in \mathcal{C}} \hat{\gamma}_i Y_i}_{SCM} + \underbrace{(X_1 - \sum_{i \in \mathcal{C}} \gamma_i X_i)\hat{\eta}}_{ridge\ \ debias}$$

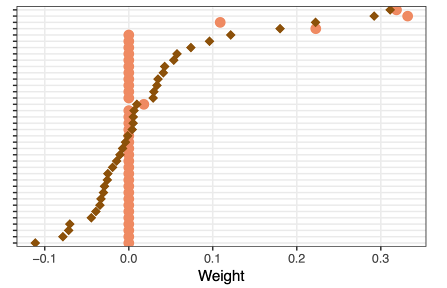
- Ridge-augmented SCM weights:

$$\hat{\gamma}_i^{aug} = \hat{\gamma}_i + \underbrace{(X_1 - X_0'\hat{\gamma})'(X_0'X_0 + \lambda I_{T_0})^{-1}}_{bias\ \ adjustment} X_i$$

- Augmentation improves balance: $\|X_1 - X_0'\hat{\gamma}^{aug}\|_2 \leq \|X_1 - X_0'\hat{\gamma}\|_2$

Revisiting California Prop 99





Synthetic DiD (Arkhangelsky et al 2019)

- Assuming one treated unit (unit N) and one post-treatment period (period T); weights add up to 1
- **Procedure**
  1. Estimate "synthetic control weight" for each control unit: $\hat{\omega}^{sc} = \arg\min_{\omega} \sum_t^{T-1} \left(\sum_{i=1}^{N-1} \omega_i Y_{it} - Y_{Nt} \right)$
  2. Estimate "synthetic control weight" for each time period: $\hat{\lambda}^{sc} = \arg\min_{\lambda} \sum_i^{N-1} \left(\sum_{t=1}^{T-1} \lambda_t Y_{it} - Y_{iT} \right)$
  3. Estimate a weighted DiD by minimizing:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - X_{it}\gamma - D_{it}\tau)^2 \hat{\omega}_i \hat{\lambda}_t$$

- Either the SC weights or the outcome model is correct, the causal effect will be identified (doubly robust)
- Inference via jackknife

Conclusions

Concluding Remarks

- The identification assumptions required by DiD are not necessarily weak: functional form, no feedback, no spillover or general equilibrium effects
- 2WFE models are often problematic: on top of DiD assumptions, homogeneity (failure leads to negative weighting); limited carryover
- Counterfactual estimators, including the SCM, can be helpful but are not assumptions free
- Methods under sequential ignorability are (relatively speaking) underdeveloped and underutilized
- Doubly robust methods have appealing statistical properties, but so far have relatively few user cases (e.g. staggered adoption)

Practical Recommendations

- Plotting raw data, especially the distribution of treatment status, helps us see obvious problems
- Think harder on how the treatment is assigned; ask yourself: "what's the hypothetical experiment?"
- If you think feedback is weak, start from estimators under parallel trends (e.g., DiD, DiD_M, FEct, augsynth) and check "pre-trend"
- If you think feedback is strong, consider methods under sequential ignorability (e.g., MSM, PanelMatch, tjbal)
- Testing, testing, testing... Whichever method you use, conduct placebo tests to check if your identification assumptions are reasonable
- And of course, don't screw up uncertainty estimates; cluster-bootstrap and jackknife (esp. when N_{tr} is small) are relatively safe choices

Future Work and Uncovered Topics

- Rethinking of panel models from a design-based perspective (just getting started)
- Spatial-temporal data — see, e.g., Wang (2021); Sanford (2021)
- Policy diffusion — see, e.g. Egami (2021)
- Continuous treatment — see, e.g. Callaway et al. (2021)
- New development in Bayesian models Carlson (2018); Feller et al. (2021)
- New development w.r.t. MSMs
- The intersection of machine learning & causal inference

Packages

- panelView: panel data visualization
- gsynth: IFEct/MC approach with non-reversible treatments
- fect: IFEct/MC methods with diagnostic tests
- tjbal: trajectory balancing

–

- lfe (Simen Gaure): fast panel linear fixed effects estimation
- PanelMatch (Kim et al): panel matching
- Synth (Abaide et al): SCM
- augsynth (Ben-Michael et al): augmented SCM

Thank you!

yiqingxu@stanford.edu

https://yiqingxu.org

github.com/xuyiqing

References

- Blackwell, Matthew, and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data Under Selection on Observables." The American Political Science Review 112 (2): 1067–82.
- Imai, Kosuke, In Song Kim, and Erik Wang. 2021. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." American Journal of Political Science, no. forthcoming.
- Chabé-Ferret, Sylvain. 2021. "Should We Combine Difference In Differences with Conditioning on Pre-Treatment Outcomes?"
- Hazlett, Chad and Yiqing Xu (2018). "Trajectory Balancing: A Kernel Method for Causal Inference with Time-Series Cross-Sectional Data." Working Paper, UCLA.
- Ben-Michael Eli, Avi Feller, Jesse Rothstein (2018). "The Augmented Synthetic Control Method." Working Paper, UC Berkeley.
- Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens and Stefan Wager (2019). "Synthetic Difference In Differences." Working Paper, UC Berkeley.