Quant II

Missing Data and Interference

Ye Wang

4/29/2020

Outline

- Multiple testing and missing data
- Interference
  - Interference with known structure
  - General interference
  - Random network
- Contagion

Multiple testing

- PCA or ICW? (copyright to Cyrus)
- Suppose you have three variables: College math grade, math GRE, and verbal GRE.
- ICW will give 25% weights to each of the math scores and 50% to the verbal score.
- PCA will generate two new variables, one for mathematical capability and the other for linguistic capability.
- Which one is more proper depends on the context.

Missing data as a causal inference problem

- Notice the duality between causal inference and missing data.
- Machine learning could also be used to predict missing outcomes.
- Liu (2019): honest inference on missing data (causal forest).

When SUTVA fails

- One of the pioneering works in this field is Manski (1993).
- A linear-in-means model:

$$Y_i = \alpha + \beta \frac{\sum_{j \in P} Y_j}{n_i} + \gamma X_i + \delta \frac{\sum_{j \in P} X_j}{n_i} + \varepsilon_i.$$

When SUTVA fails

- One of the pioneering works in this field is Manski (1993).
- A linear-in-means model:

$$Y_i = \alpha + \beta \frac{\sum_{j \in P} Y_j}{n_i} + \gamma X_i + \delta \frac{\sum_{j \in P} X_j}{n_i} + \varepsilon_i.$$

- β: Endogenous effect; δ: Exogenous effect.
- In addition, ε_i may not be independent to each other due to homophily.

When SUTVA fails

- One of the pioneering works in this field is Manski (1993).
- A linear-in-means model:

$$Y_i = \alpha + \beta \frac{\sum_{j \in P} Y_j}{n_i} + \gamma X_i + \delta \frac{\sum_{j \in P} X_j}{n_i} + \varepsilon_i.$$

- β: Endogenous effect; δ: Exogenous effect.
- In addition, ε_i may not be independent to each other due to homophily.
- Many econometric works are based on this framework (e.g. Bramoullé et al., 2009; Graham, 2014).
- The model can be arbitrarily complicated: dynamics, spatial autoregression, network formation, etc.

When SUTVA fails

- The outcome-based approach provides a conceptual benchmark.
- When the treatment of others matters: interference.
- When the outcome of others matters: contagion.
- Homophily is often an important source of confounding.
- Challenge: how could we ensure the correctness of the outcome model?

When SUTVA fails

- The outcome-based approach provides a conceptual benchmark.
- When the treatment of others matters: interference.
- When the outcome of others matters: contagion.
- Homophily is often an important source of confounding.
- Challenge: how could we ensure the correctness of the outcome model?
- We can defend the model via structural approaches.
- Or we can switch to the design-based perspective.

Interference with known structure

- We assume that the treatment is randomly assigned.
- We possess the knowledge of the "social network" among individuals.
- i's treatment affects j's outcome if and only if there is a tie between them in the network.

Interference with known structure

- We assume that the treatment is randomly assigned.
- We possess the knowledge of the "social network" among individuals.
- i's treatment affects j's outcome if and only if there is a tie between them in the network.
- Partial interference: all units in the same stratum are connected, but no connection between strata.
- Aronow and Samii (2017): with the network we can construct "exposure mapping."
- The network is actually a moderator: $D_i \mapsto W_i$.

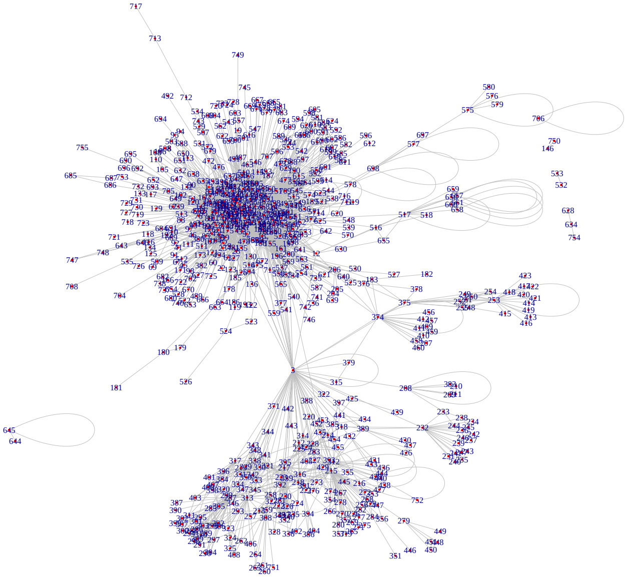
Interference with known structure

- The network allows us to draw the DAG.
- Then everything is identifiable.
- van de Laan (2014) and Ogburn et al. (2018): a general framework to deal with spillover effects in dynamic social networks.

Interference with known structure

- The network allows us to draw the DAG.
- Then everything is identifiable.
- van de Laan (2014) and Ogburn et al. (2018): a general framework to deal with spillover effects in dynamic social networks.
- The more challenging part is inference.
- We cannot assume that everyone is connected to everyone else.
- Then there is only one observation, $N = 1$.
- We have to assume that the spillover is somehow "local."

Interference with known structure



Interference with known structure

```
##  dir_ind1  isol_dir      ind1
##  2437.561 -1128.116  2717.595
```

```
## dir_ind1 isol_dir     ind1
## 65653393  3837374 24041666
```

General interference

- But knowing the network is often a very strong assumption.
- Egami (2019): sensitivity analysis for the network's misspecification.
- Recent works proceed under the framework of "general interference."
- There are two questions we are interested in:
  - Is it still possible to estimate the ATE?
  - How can we estimate the interference effect?

General interference

- For the first question, the answer is yes under some conditions (Aronow, Hudgen and Savje, 2019; Chin, 2017).
- We require the average "perturbation" from changing the treatment status of one unit to be small.
- Notice that the ATE is no longer well-defined, so the limit is actually EATE:

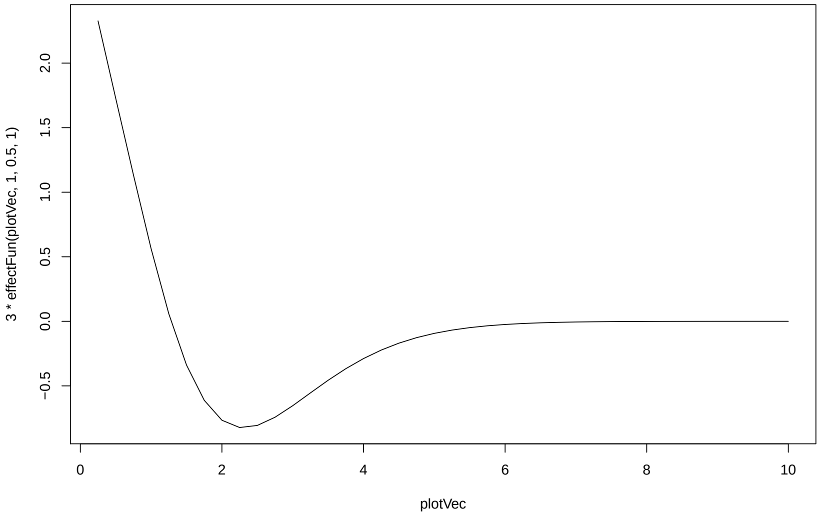
$$\tau^* = \frac{1}{N}\sum_{i=1}^{N}\tau_i = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1, \mathbf{Z}_{-i}) - Y_i(0, \mathbf{Z}_{-i})\right]$$

$$\tau = E_{\mathbf{Z}_{-i}}\tau^* = \frac{1}{N}\sum_{i=1}^{N}\left[E_{\mathbf{Z}_{-i}} Y_i(1, \mathbf{Z}_{-i}) - E_{\mathbf{Z}_{-i}} Y_i(0, \mathbf{Z}_{-i})\right]$$

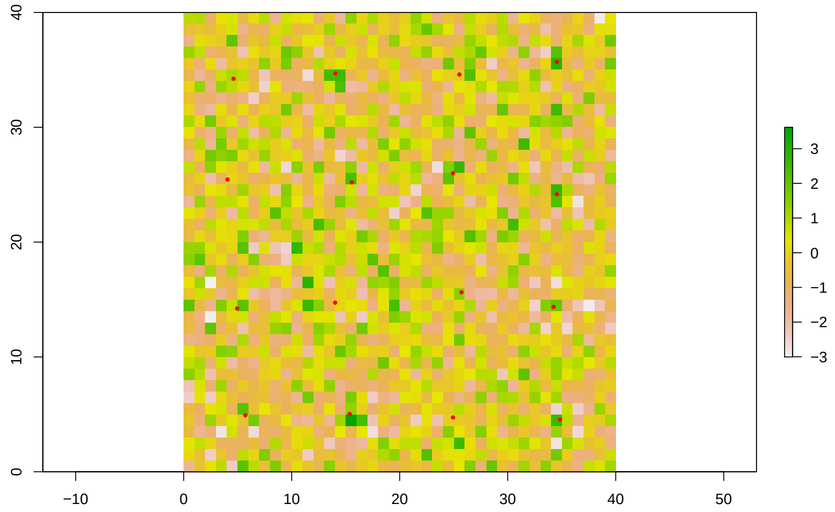
General interference

- For the second question, there is rapid progress in recent years.
- Aronow, Samii and Wang (2020): As the design is known, we can aggregate the individualistic effects in arbitrary ways.
- In spatial experiments, it is natural to do it by distance.
- Just draw circles at each distance d and apply the IPW estimator.
- Inference relies on Stein's method.

General interference

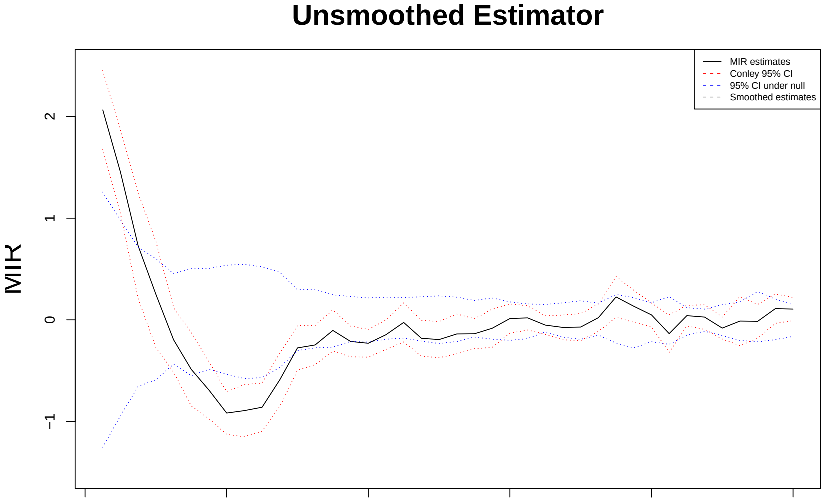


General interference



General interference

```
## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 1820 to perform exact estin
```



General interference

- Wang (2020): the same idea works in TSCS dataset under the assumption of sequential ignorability.
- Researchers have to trade off two sources of confounding: fixed effects vs. interference.
- Papadogeorgou et al. (2020): modeling treatment point process as a stochastic intervention strategy.
- They examine the effect of airstrike in Iraq on rebellions.

Within-unit interference

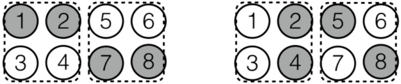
- Interference is also a big problem in TSCS data analysis.
- It is not surprising that the treatment in period $t - 1$ affects the outcome in period t.
- Then the assumption behind the two-way fixed effects model will be violated (no carryover).
- What can we do?

Within-unit interference

- Interference is also a big problem in TSCS data analysis.
- It is not surprising that the treatment in period $t-1$ affects the outcome in period t.
- Then the assumption behind the two-way fixed effects model will be violated (no carryover).
- What can we do?
- If we are fine with abandoning the unit fixed effects: just estimate a MSM.
- Or we have to add more variables into the regression and hope the model is correct.
- A special case is staggered adoption (Strezhnev, 2019; Liu, Wang and Xu, 2020).

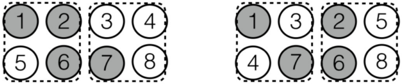
Random network

- All these approaches assume that there is a fixed network and the intervention is at random.
- Another possibility is that the assignment is pre-specified but the network is randomly generated.
- Li et al. (2019)



(a1) (a2)

(a) The first type of interference with a fixed network and random external interventions. (a1) and (a2) are two possible realizations of random external interventions (colors of the units).



(b1) (b2)

(b) The second type of interference with fixed attributes of all units and a random network. (b1) and (b2) are two possible realizations of random networks (dashed circles).

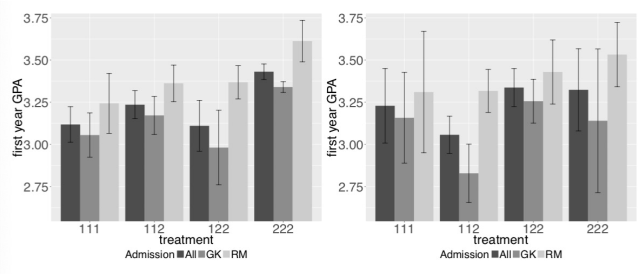
Random network

- A study using random dorm assignment in Peking University.
- Does living with IMO medalists boost your GPA?
- Notice that this is interference not contagion.
- A design-based approach to understand peer effect in dorms.

Random network

- A study using random dorm assignment in Peking University.
- Does living with IMO medalists boost your GPA?
- Notice that this is interference not contagion.
- A design-based approach to understand peer effect in dorms.
- Two assumption on "local interference:"
  - Partial interference
  - Only attributes matter, not identities
- They propose a Horvitz-Thompson estimator, and prove that it is equivalent to a regression with interaction.

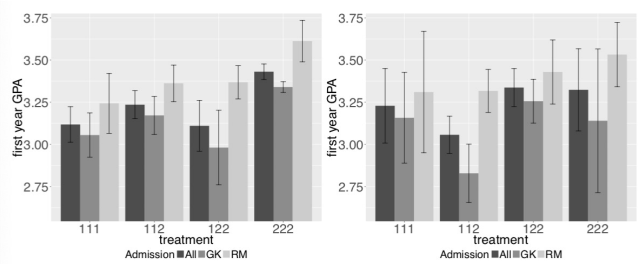
Random network



(a) Informatics (b) Physics

- They also discuss the optimal assignment.

Random network



(a) Informatics (b) Physics

- They also discuss the optimal assignment.
- Let champions play with champions...

Network formation

- A relevant question is whether interference affects the network's structure.
- There is a large literature on network formation in social network analysis.
- A familiar approach is to fit a two-way fixed effects model:

$$Y_{ij} = \mu + \alpha_i + \zeta_j + \beta X_{ij} + \varepsilon_{ij}$$

- You can further add the time dimension into the model.
- No design-based approaches yet.

Contagion

- The peril of peer effects (Angrist, 2011).
- Can you regression your GPA on the average GPA of your dorm?

Contagion

- The peril of peer effects (Angrist, 2011).
- Can you regression your GPA on the average GPA of your dorm?
- Can you regression your GPA on the average GPA of your roommates?

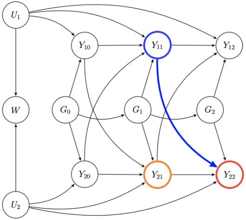
Contagion

- The peril of peer effects (Angrist, 2011).
- Can you regression your GPA on the average GPA of your dorm?
- Can you regression your GPA on the average GPA of your roommates?
- Fowler and Christakis: peer effect of obesity.

Contagion

- The peril of peer effects (Angrist, 2011).
- Can you regression your GPA on the average GPA of your dorm?
- Can you regression your GPA on the average GPA of your roommates?
- Fowler and Christakis: peer effect of obesity.
- Egami (2019): a causal framework using static DAG.
- He also has a placebo test on the identification assumption and a debiased estimator.

Contagion



(a) Example of Placebo Test

	C	C^P	Placebo Test
No Bias	Y_{21}, U_2, G_2	$Y_{20}, Y_{10}, U_2, G_2, G_1$	Accept
Contextual Confounding	Y_{21}, U_2	Y_{20}, Y_{10}, U_2	Reject
Homophily Bias	Y_{21}, G_2, G_1	$Y_{20}, Y_{10}, G_2, G_1, G_0$	Reject
Both	Y_{21}, Y_{20}	Y_{20}, Y_{10}	Reject

(b) Control and Placebo Sets