Causal Inference with Text Data

Elliott Ash

Bocconi, 2018

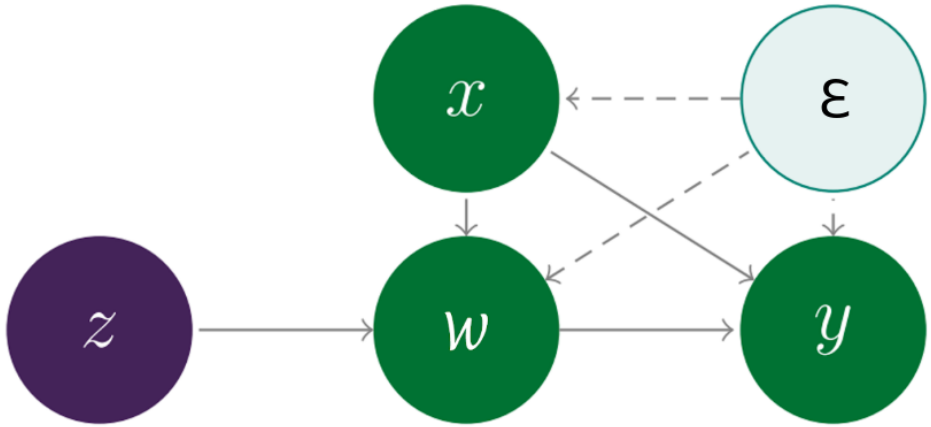
Research Design

- The goal of social-science research with text data is the same as other social-science research:
  - provide credible tests of socials-science hypotheses
  - estimate policy parameters to inform policymakers

Four Roles for Text in the Causal Pipeline

- Text as outcome: assess a text-based response
  - e.g. text-predicted ideology changes according to senate election schedule (Ash, Morelli, and Van Weeldden 2017)
- Text as treatment: assess the effect of a text
  - discover treatments (Fong and Grimmer 2016)
- Text as confounder: condition on text
  - matching with text (Roberts, Stewart and Nielsen 2017)
- Text as source of heterogeneity:
  - estimate conditional average treatment effects as function of text (Wager and Athey 2017).

The Empirical Problem



- y, outcome
- w, text treatments
- x, observable covariates
- ε, confounders
- z, instruments

Econometrics and Machine Learning

- We would like to learn

$$f(w;\theta) = \mathbb{E}\{y|w\}$$

  the conditional expectation function for y, where θ represents the true parameter vector.
- If we assume linearity and run OLS, the estimates for $\hat{\theta}$ are biased because of the confounder.
- Similarly, we could take a machine learning (ML) approach and learn a nonlinear approximation $\hat{f}(w, x; \theta)$ to predict y in held-out data.
  - if we collected more data on the text w_i for new individual i, this would give us a good prediction about the associated y_i.
  - But the ML estimates $\hat{\theta}$ are not causal: i.e., one could not use them to make a counterfactual prediction about the effect on y of exogenously altering the text features w.

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Lab/Field Experiments

- Lab/field experiments provide a gold standard for obtaining causal estimates.
  - ask your subjects to fill out open-ended survey responses before and after the experiment
  - subjects can talk to each other in a chatroom
  - subjects view randomly assigned text treatments (Fong and Grimmer 2016)

Differences in differences

- Estimate changes in an outcome due to changes in text.
  - a lot of strong assumptions for this, probably won't work.
- More realistic: measure changes in a text-based metric due to a treatment
  - Include fixed effects and trends for the individuals.
    - Try to illustrate effect with event-study graph
  - Ash, Morelli, and Van Weelden (2017): effect of senate elections on divisiveness
  - Ash (2016), effect of political control on tax code

Fixed Effects with Sparse Predictor Matrix

- Recall that standardizing data breaks sparsity structure.
  - fixed effects or other residualization steps will also do this.
- Some solutions:
  - Can residualize outcomes but not predictors.
  - Can use use first-differences rather than fixed-effects.
  - Can center on the mode after residualizing
- What do fixed-effects transformations accomplish with non-linear models?

Regression Discontinuity

- Local impact of a thresholded treatment on a text-based metric.
  - Electoral RD effect on type of speech used by Congressmen
- Local impact of a thresholded text treatment:
  - Electoral RD effect on economy of ballot referenda

Instrumental Variables

- Use exogenous variation from instruments:
  - Relevance ($F > 10$) and exclusion restriction (instrument only affects outcome through endogenous regressor channel)
- Text-based outcome:
  - Ash, Morelli, and Van Weelden (2017): effect on divisiveness of news coverage instrument from Snyder and Stromberg (2010)
- Instruments for text:
  - see Ash (2016) and Ash, Morelli, Vannoni (2018) for examples of IV using text data
  - in law, can use random assignment of judges, combined with differences across judges in writing style.

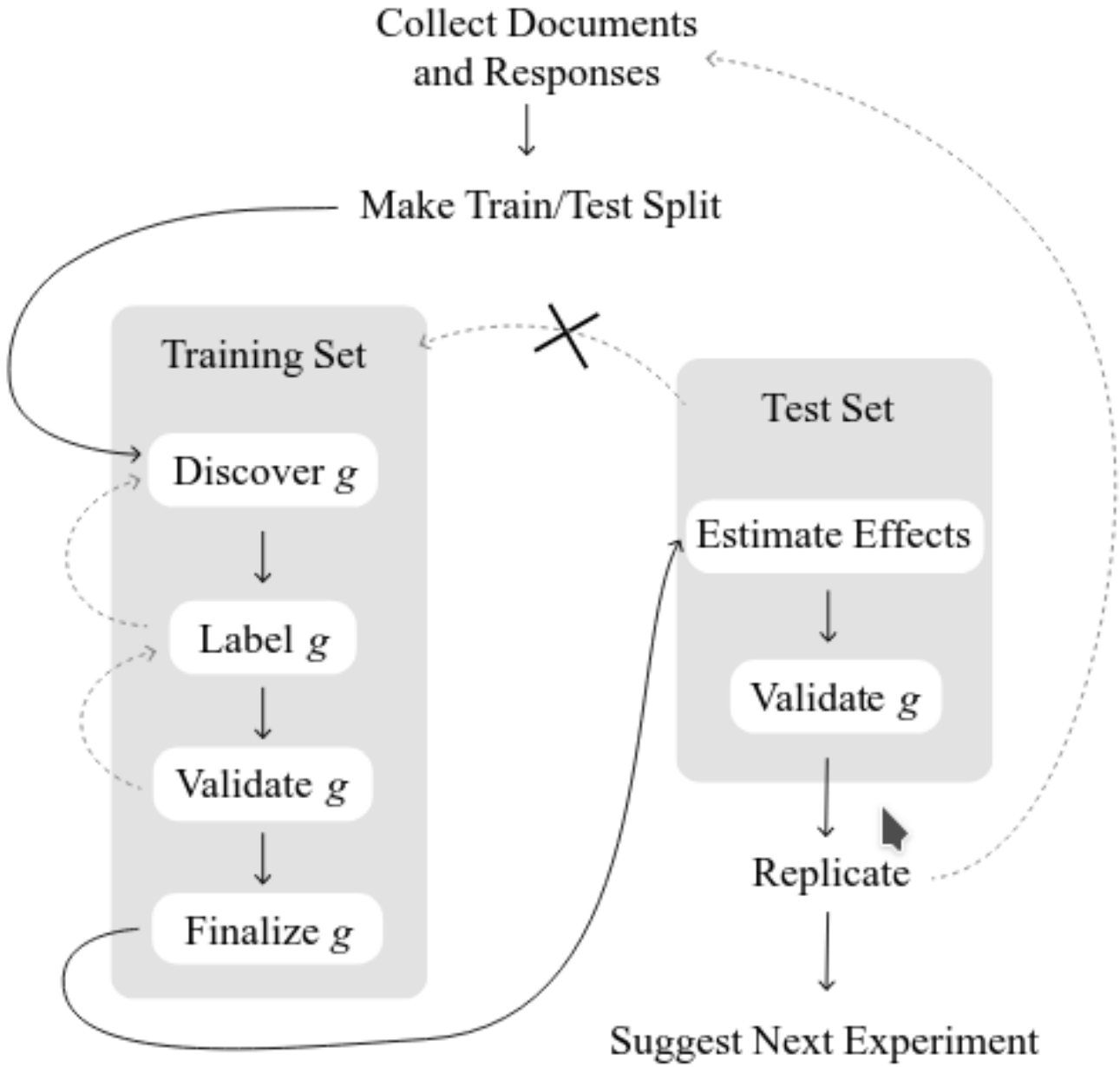
Matching / high-dimensional controls

- "matching" is the use of covariates to weight other observations as better controls.
- Can imagine the text documents associated with individuals as a set of covariates.
  - e.g., compare impacts of state-level policing-policy changes on crime rates, while matching on (controlling for) a high-dimensional representation of the rest of a state's laws. (Roberts, Stewart, and Nielsen 2016).

Heterogeneous Treatment Effects

- *Wager and Athey (2017): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*
  - provide a non-parametric causal forest for estimating heterogeneous treatment effects
  - see also Athey, Tibshirani, and Wager (2018)
  - e.g., could look at heterogeneous effects of minimum wage changes by text of operative employment regulations

Egami, Fong, Grimmer, Roberts, and Stewart



Egami, Fong, Grimmer, Roberts, and Stewart

- Setup:
  - There are some latent treatments in the text, represented by W_i
  - Each individual has an outcome Y_i or a non-text treatment Z_i
- Text outcome, non-text treatment: $W_i = g(Z_i; \theta)$
- Text treatment, non-text oucome: $Y_i = f(W_i; \theta)$

- Learn functional form for $g(\cdot)$ in half the data, and then run causal inference in the other half.

Sample Split

- The insight/emphasis of Egami et al is that $g(\cdot)$ can take any form (you can use any featurization approach you like), and you get valid inference as long as its done in held-out data.

Double/Debiased ML

- Chernozhukov, Chetverikov, Duflo, Hansen, Demirer, and Newey: *Double/Debiased Machine Learning for Treatment and Structural Parameters:*

$$Y = \theta T + g(X) + \epsilon$$

  - low-dimensional treatment T, high-dimensional set of confounders X: $T = m(X) + \eta$.
- Because of confounders, forming a prediction $\hat{Y} = \hat{\theta} T + \hat{g}(X)$ will be biased.

Double ML method

1. Predict Y given X: $\hat{Y}(X)$, and T given X: $\hat{T}(X)$, using any ML method
2. Form residuals $\tilde{Y} = Y - \hat{Y}(X)$ and $\tilde{T} = T - \hat{T}(X)$
3. Regress $\tilde{Y}$ on $\tilde{T}$ to learn $\hat{\theta}$.

- Sample split:
  - Run 1 on sample A, then run 2 and 3 on sample B, to estimate $\hat{\theta}$
  - and vice versa, to learn a second estimate for $\hat{\theta}$.
  - average them to get a more efficient estimator.

How do voters evaluate candidates?

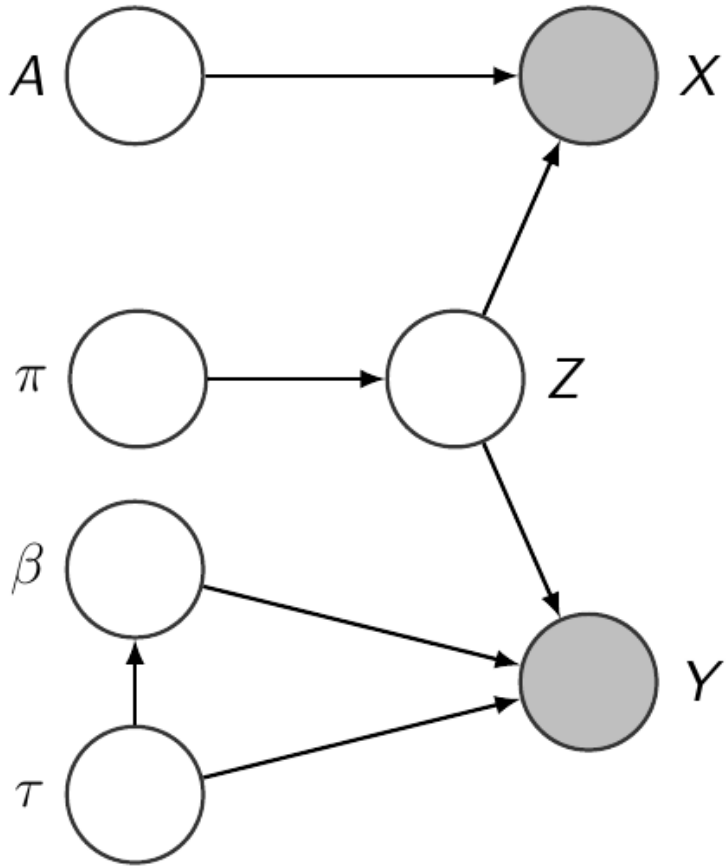
- What biographical facts affect voter evaluations?
- Could run a survey experiment:
  - `Document 1: He earned his Juris Doctor in 1997 from Yale Law School, where he operated free legal clinics for low-income residents of New Haven, Connecticut.`
  - `Document 2: He served in South Vietnam from 1970 to 1971 during the Vietnam War in the Army Rangers' 75th Ranger Regiment, attached to the 173rd Airborne Brigade. He participated in 24 helicopter assaults...`
- But hard to generalize what features drive differences.

Discovery of Treatments from Text Corpora

1. Randomly assign texts, X_i , to respondents
2. Obtain responses Y_i for each respondent
3. Randomly divide text/responses into training and test set
   - 3.1 Avoid technical issues with using entire sample
   - 3.2 Ensure we avoid "p-hacking" (false discovery)
4. In training set: Discover mapping from texts to treatments
5. In test set: infer treatments and measure their effects

Supervised Indian Buffet Process

The Supervised Indian Buffet Process (sIBP)



Text and response depend on latent treatments

- Treatment assignment

$$z_{i,k} \sim \text{Bernoulli}(\pi_k)$$
$$\pi_k \sim \prod_{m=1}^{k} \eta_m$$
$$\eta_m \sim \text{Beta}(\alpha, 1)$$

- Document Creation:

$$\boldsymbol{X}_i \sim \text{MVN}(\boldsymbol{Z}_i\boldsymbol{A}, \sigma_X^2 I_D)$$
$$\boldsymbol{A}_k \sim \text{MVN}(\boldsymbol{0}, \sigma_A^2 I_D)$$

- Response:

$$Y_i \sim \text{MVN}(Z_i\boldsymbol{\beta}, \tau^{-1})$$
$$\boldsymbol{\beta}|\tau \sim \text{MVN}(\boldsymbol{0}, \tau^{-1} I_K)$$
$$\tau \sim \text{Gamma}(a, b)$$

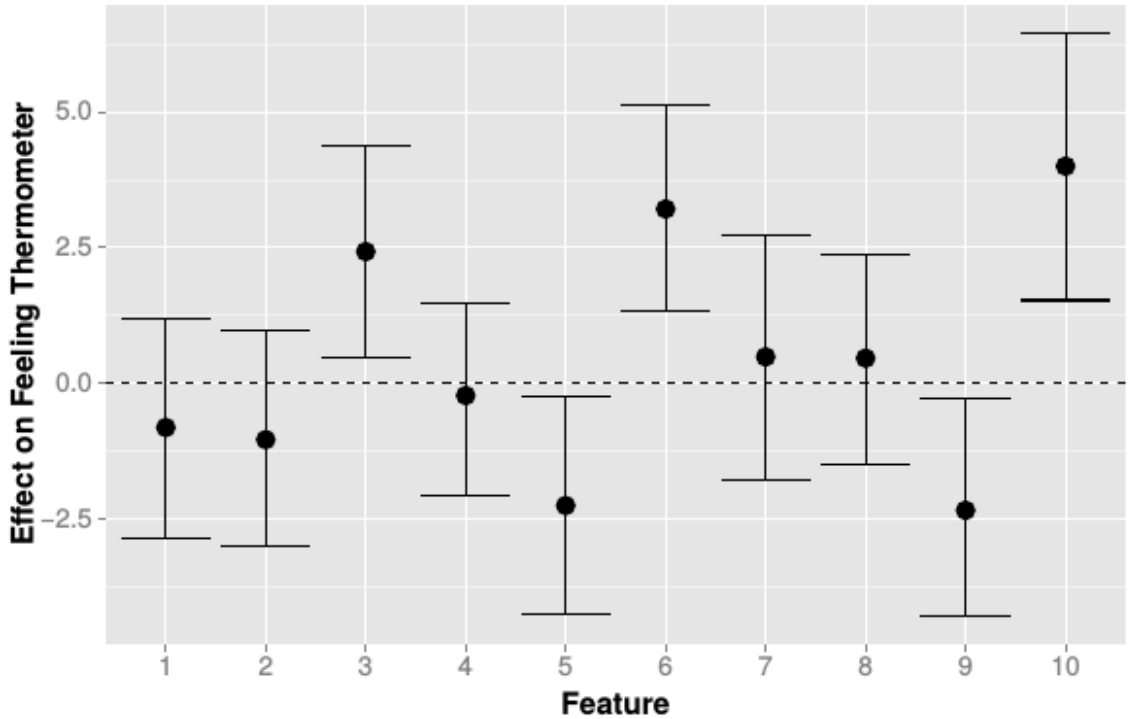
Candidate Biographies on Wikipedia

`Schumacher was born and raised in the Highlandtown neighborhood of East Baltimore, the eldest of the three daughters of Christine Eleanor (nee Kutz) and William Schumacher. Her parents were both of Polish descent; her immigrant great-grandparents had owned a bakery in Baltimore. During her high school years at the Institute of Notre Dame, she worked in her parents' grocery store...`

- Protocol: Each respondent sees up to 3 texts from the corpus of $>$ 2200 biographies
  - Observe text
  - Feeling thermometer rating: 0-100
- 1,886 participants, 5,303 responses
  - 2,651 training, 2,652 test

Results

Treatment	Keywords
3	director, university, received, president, phd, policy
5	elected, house, democratic, seat
6	united_states, military, combat, rank
9	law, school_law, law_school, juris_doctor, student
10	war, enlisted, united_states, assigned, army



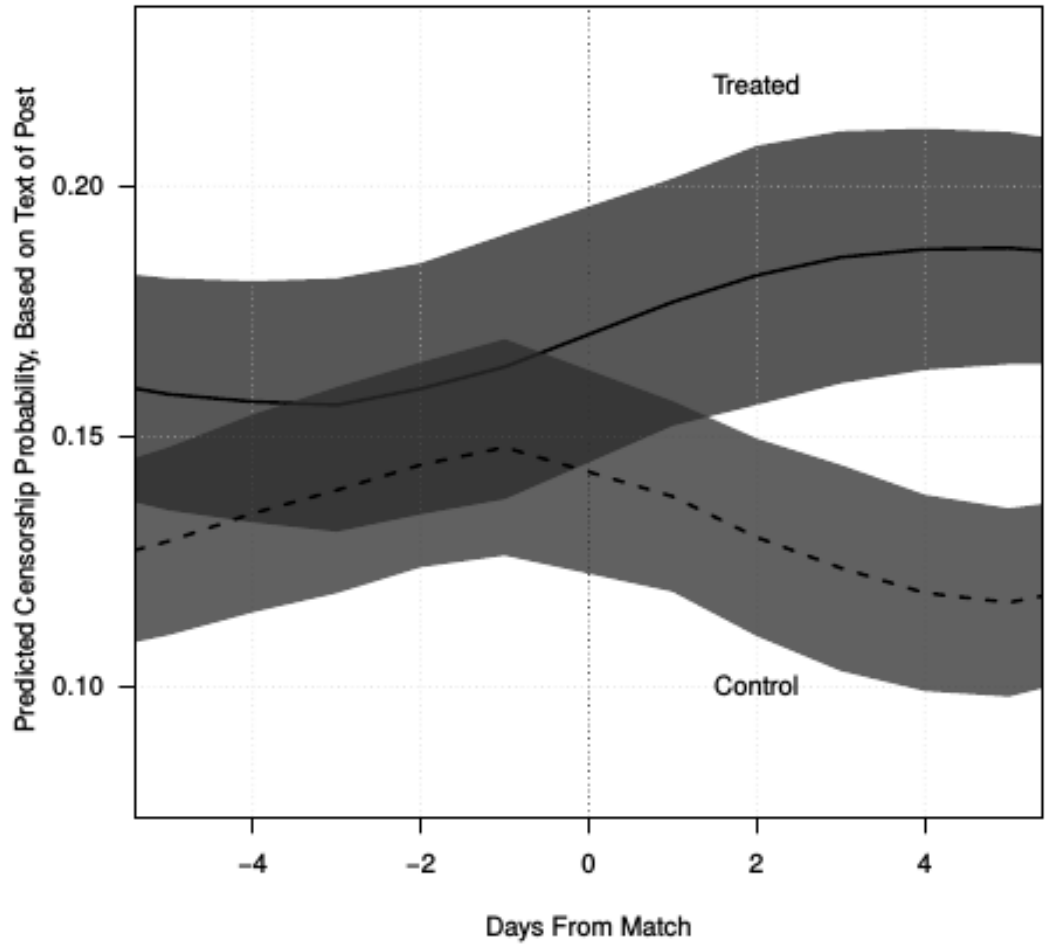
How do people react to online censorship?

- Lots of governments try to control online information
- But, censoring the whole internet is hard (# of bloggers $\gg$ # of censors)
- Limited external enforcement $\rightsquigarrow$ self-policing

Text Matching

- Roberts et al (2016) construct a corpus of chinese blog posts, some of which are censored.
  - 593 bloggers, 150,000 posts, 6 months
- They use a variation of propensity score matching to identify almost identical blog posts, some of which were censored, and some of which were not.
- Outcome:
  - Using text of subsequent posts, measure how likely they are to be censored (how censorable)
  - Can see whether censorship has a deterrence or backlash effect.

Censorship has a backlash effect



- Bloggers who are censored respond with more censorable content.

Hartford et al (2017): Deep IV

- *Deep IV: A Flexible Approach for Counterfactual Prediction*
  - use ML algorithms to extend 2SLS to high-dimensional settings
- Causal effect of interest:

$$f(w;\theta) = \mathbb{E}\{y|w\}$$

- Predictors are a function of some instruments:

$$w \sim g(w|z)$$

First stage

- Deep IV allows arbitrarily high-dimensional w and z.
- In first stage, approximate $g(w|\gamma(z))$, the distribution of w:
  - assume that $g(\cdot)$ is a mixture density network (a mixture of gaussian distributions) where the parameter vector $\gamma(\cdot)$ includes the weights, means, and variances (Bishop 2006).
  - $\gamma(z)$ is any function of the instruments – can use an MLP, for example.
  - $g(\cdot)$ has to be a parametrized distribution because Deep IV requires that the distribution be integrated in the second stage.

Second Stage

- In second stage, want to predict $\hat{y}(w;\theta)$, where $\hat{y}(w;\theta)$ should be a flexibly specified DNN to allow for non-linearities and interactions.
- Hartford et al (2017) show that causal estimates for θ are obtained by minimized the conditional loss function

$$\mathcal{L}(\theta) = \sum_i [y_i - \int \hat{y}(w;\theta) d\hat{g}(w|\gamma(z_i))]^2$$

  - this is true y minus predicted $\hat{y}$, but $\hat{y}$ is conditioned on the instrument-predicted treatment distribution $\hat{g}$.

Second Stage Loss Approximation

- The integral in $\mathcal{L}(\theta)$ is approximated by

$$\int \hat{y}(w;\theta) d\hat{g}(w|\gamma(z_i)) \approx \frac{1}{m}\sum_{j}^{m} \hat{y}(\tilde{w}(z_i);\theta)$$

  where you make m draws from the estimated treatment distribution given z_i (the instruments for observation i).
  - Like 2SLS, a prediction for the endogenous regressor with the instruments is used during second-stage estimation.

What about relevance/inference?

- Both stages of Deep IV can be validated by out-of-sample prediction in held-out data
  - in the first stage, this guards against weak-instruments bias in the same way that first-stage F-statistics thresholds do for 2SLS

Blessings of Multiple Causes

- Wang and Blei (2018) provide a powerful insight:
  - causal inference with multiple causes (treatments) requires weaker assumptions than classical (single-treatment) causal inference.
- In particular, unbiased causal inference is possible if confounders are shared across multiple treatments.
- Wang and Blei (2018) provide an ML method to construct a "deconfounder" from the treatment data and allow inference.

How does the deconfounder work?

- Assume multiple treatments $A_1, ..., A_m$
  - Assume there is a latent factor Z that, when taken out from the A_j, renders them conditionally independent.
  - If we can learn Z, this will decounfound the treatments.

Argument for Deconfounder Z

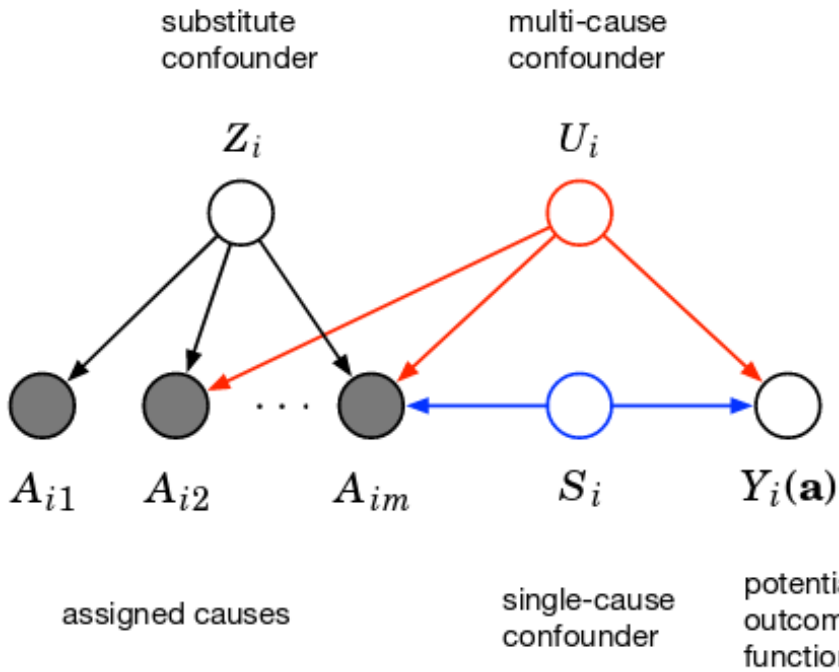


Figure 1: A graphical model argument for the deconfounder. The punchline is that if Z_i renders the A_{ij}'s conditionally independent then there cannot be a multi-cause confounder. The proof is by contradiction. Assume conditional independence holds, $p(a_{i1},...,a_{im}\,|\,z_i) = \prod_j p(a_{ij}|\,z_i)$; if there exists a multi-cause confounder U_i (red) then, by d-separation, conditional independence cannot hold (Pearl, 1988). Note we cannot rule out the single-cause confounder S_i (blue).

Constructing and validating the deconfounder

- Learning the deconfounder is the same as learning any factor model:
  - can use PCA, LDA, or DNN
- To check whether your deconfounder is working, check whether your factor model is capturing distribution of treatment assignment:
  - fit the factor model on training data; it should be able to predict treatment assignment in the test data.

Best Actors: Causal Evidence

- Top revenue actors, non-causal estimates:
  - Tom Cruise, Tom Hanks, Will Smith, Arnold Schwartzenegger, Robert De Niro, Brad Pitt.
- Top revenue actors, causal estimates:
  - Owen Wilson, Nick Cage, Cate Blanchett, Antonio Banderes.
- Most under-valued actors:
  - Stanley Tucci, Willem Dafoe, Susan Sarandon, Ben Affleck, Christopher Walken.