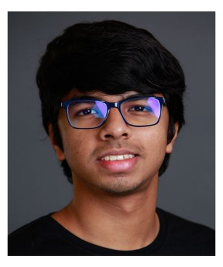
Finding Valid Adjustments under Non-ignorability with Minimal DAG Knowledge


Abhin Shah
MIT
abhin@mit.edu

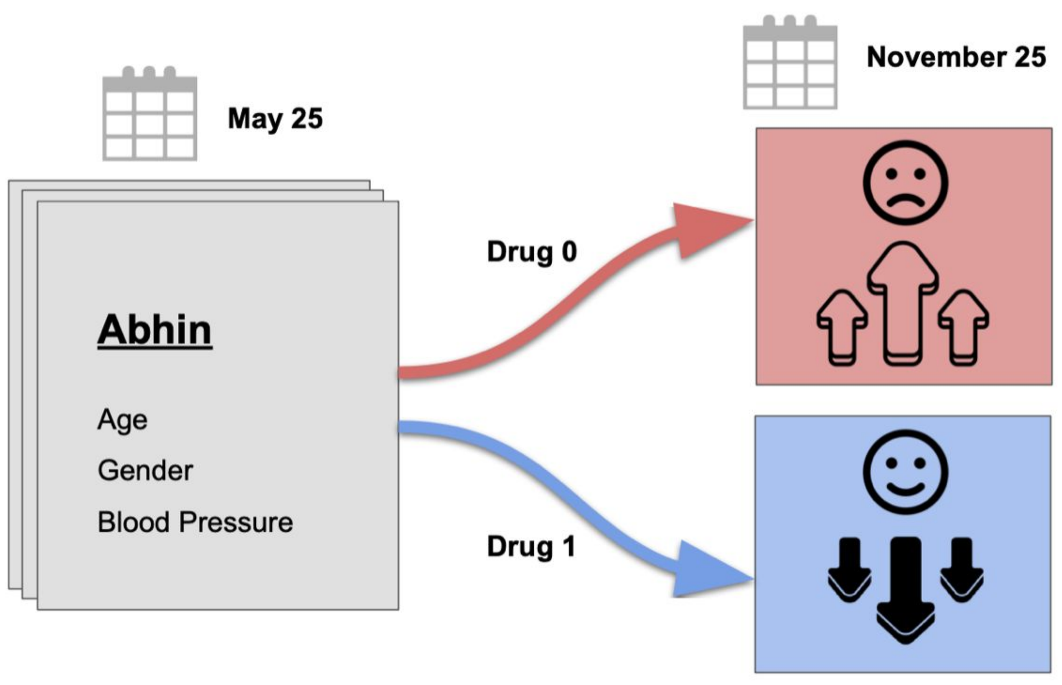

Karthikeyan Shanmugam
Google Research


Kartik Ahuja
Meta AI

arXiv Link

Causal Effect Estimation

Causal effect of a drug on cholesterol level from observational data



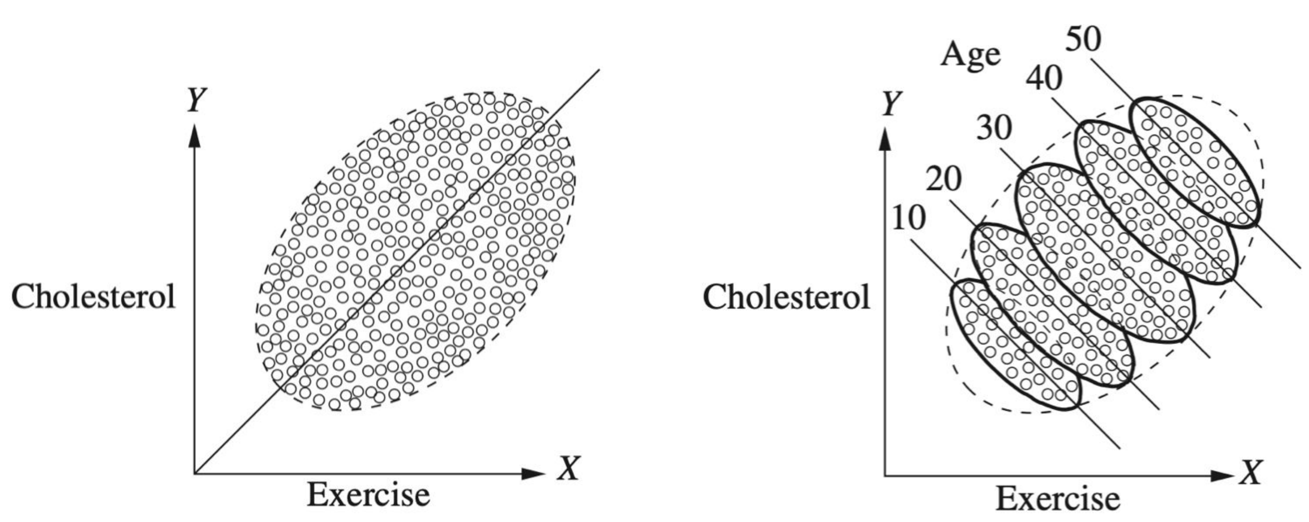
$\mathbb{P}(\text{cholesterol} \mid do(\text{drug}))$?

Observational Data

Age	Gender	Blood Pressure	Drug	Cholesterol (0)	Cholesterol (1)
22	Male	145/95	0	↑↑↑	?
26	Female	135/80	0	↓↓↓	?
58	Female	130/70	1	?	↑↑↑
50	Male	145/80	1	?	↓↓↓
24	Female	150/85	1	?	↑↑↑

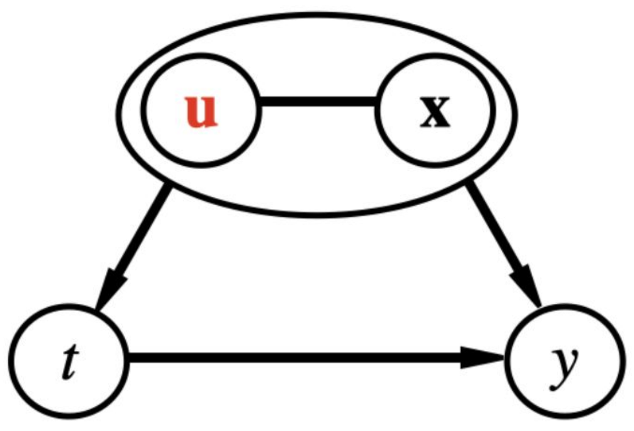
Challenge — Unobserved Confounding

Simpson's paradox: Which subsets of the observed features should be used?



Problem Formulation

- $\mathbf{u}$: unobserved exogenous variables
- $\mathbf{x}$: observed features
- t : observed binary treatment variable
- y : observed outcome
- $\mathcal{G}$: DAG over the set of vertices $\{\mathbf{u}, \mathbf{x}, t, y\}$



Valid adjustments

$\mathbf{z}$ is a valid adjustment set if $\mathbb{P}(y \mid do(t = t)) = \mathbb{E}_{\mathbf{z}}[\mathbb{P}(y \mid \mathbf{z} = z, t = t))]$

Pearlian Framework

DAG knowledge

Given the complete knowledge of the DAG, graphical criteria could be used to check whether $\mathbf{z}$ is valid for adjustment

Potential Outcomes

Ignorability

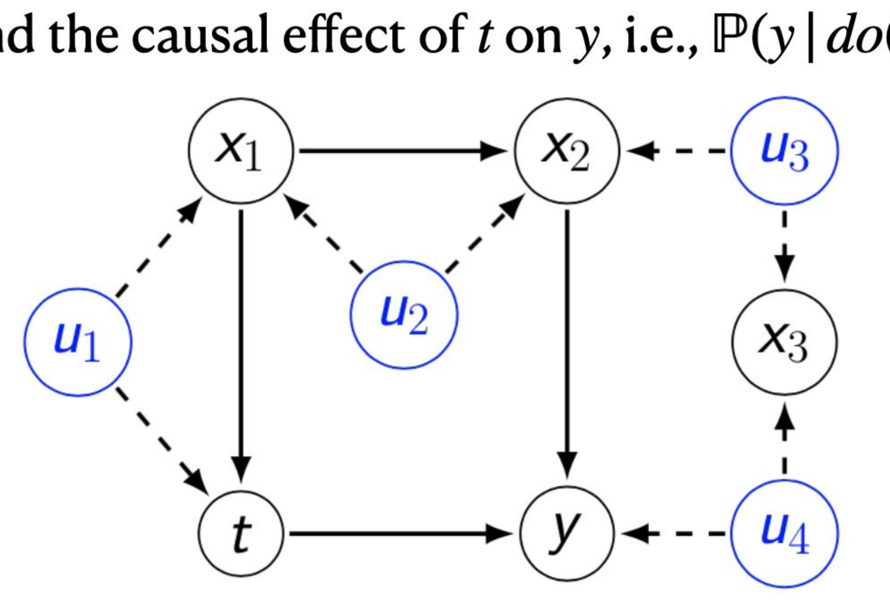
$\mathbf{x}$ satisfies ignorability

⇓

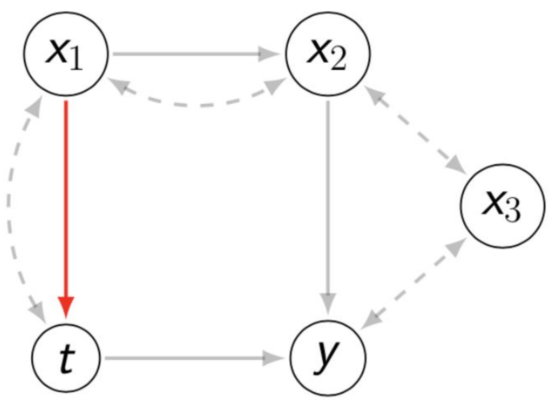
$\mathbf{x}$ is a valid adjustment.

How much of the DAG do we need to know?

To find the causal effect of t on y, i.e., $\mathbb{P}(y \mid do(t = t))$



Can we significantly reduce the structural knowledge required about the DAG and yet find valid adjustment sets?

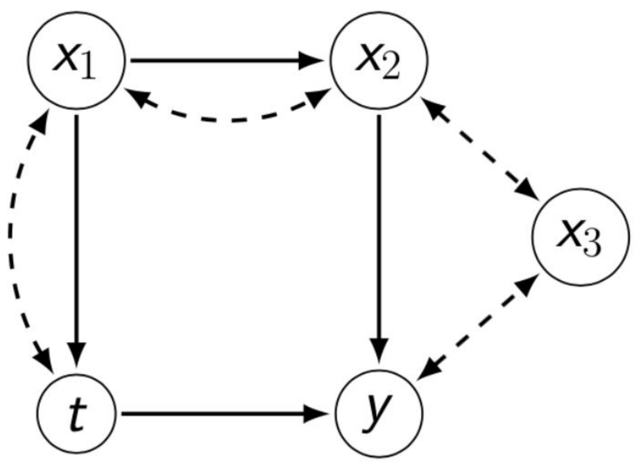


The knowledge of one causal parent of the treatment is sufficient to find a class of valid adjustment sets!

Assumptions

Semi-Markovian model

1. The treatment t has the outcome y as its only child.
2. The outcome y has no child.



Back-door Criterion

A popular sufficient graphical criterion for finding valid adjustments

Under our assumptions, a set $\mathbf{z}$ satisfies the back-door criterion in $\mathcal{G}$ if

1. $\mathbf{z}$ blocks every path between t and y in $\mathcal{G}$ that contains an arrow into t.

Sets satisfying back-door: $\{x_1, x_2\}$ and $\{x_2\}$

Conditional Independence $\Longleftrightarrow$ Back-door

- x_t : an observed feature that is a direct causal parent of t.
- Consider any subset of the remaining observed features i.e., $\mathbf{z} \subseteq \mathbf{x} \setminus \{x_t\}$.
- $\mathbf{z}$ satisfies the back-door criterion if and only if $x_t \perp y \mid \mathbf{z}, t$.

Algorithms

- Subset Search:

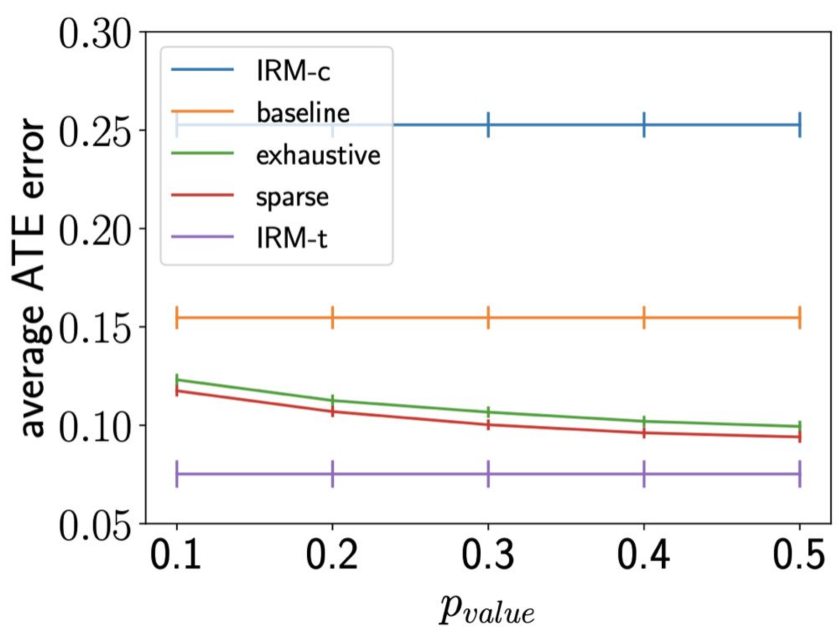
  ➡ Use a subset based search procedure that exploits conditional independence (CI) testing to check our invariance criterion.

- IRM-based:

  ➡ Use a sub-sampling trick to leverage Invariant Risk Minimization (IRM) as a scalable approximation for CI testing.

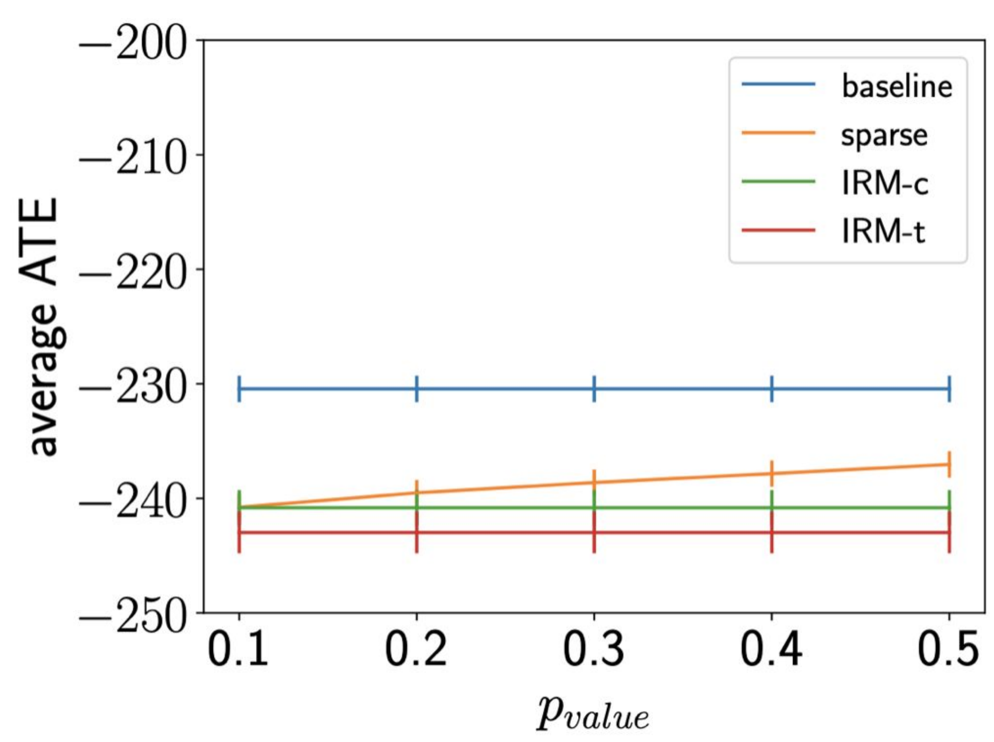
IHDP

A RCT studying cognitive test score of low-birth-weight, premature infants.

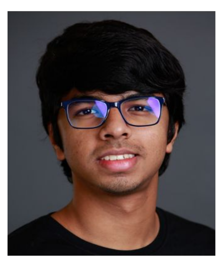


Cattaneo

Studies the effect of maternal smoking on babies' birth weight.



On Counterfactual Inference with Unobserved Confounding


Abhin Shah
abhin@mit.edu
MIT

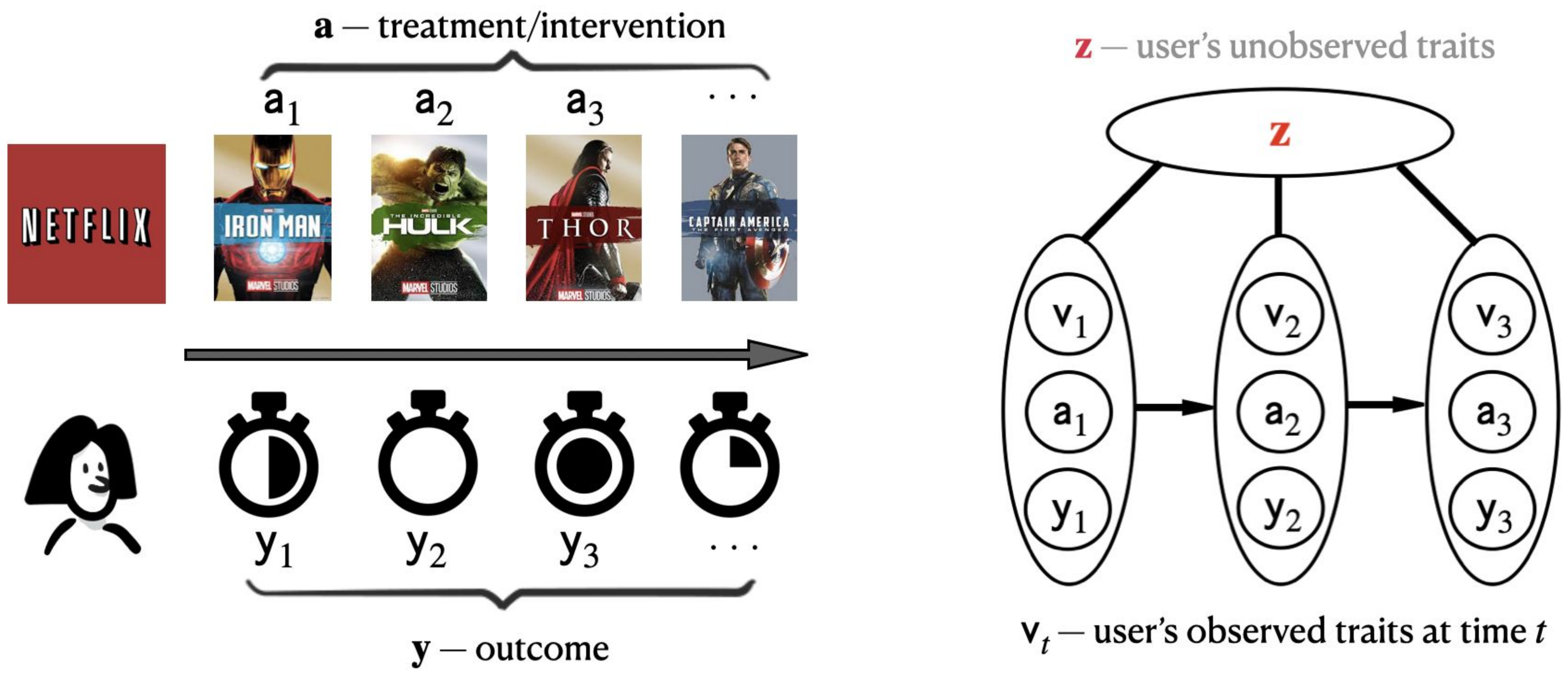

Raaz Dwivedi
raaz@mit.edu
MIT, Harvard University


Devavrat Shah
devavrat@mit.edu
MIT

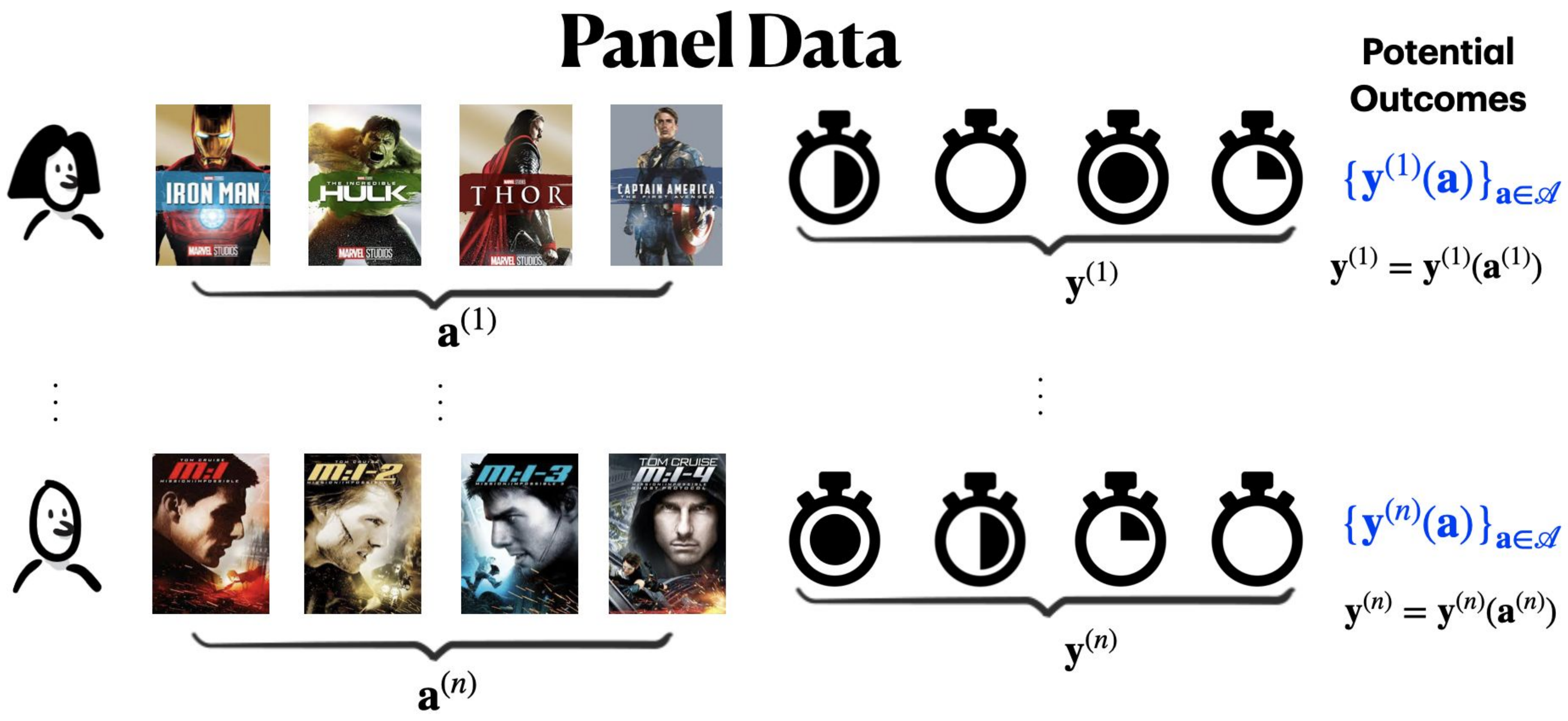

Greg Wornell
gww@mit.edu
MIT

arXiv Link

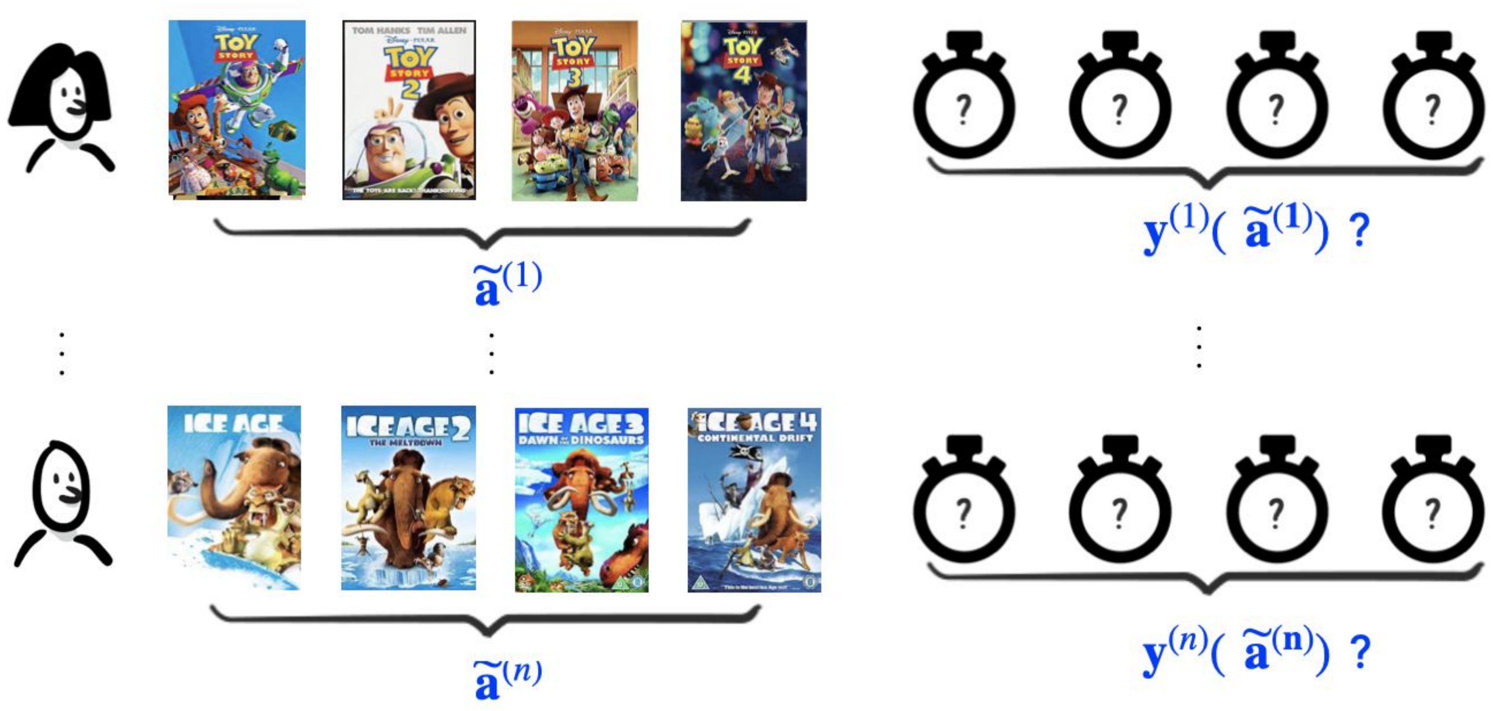
Observational Setting



Panel Data

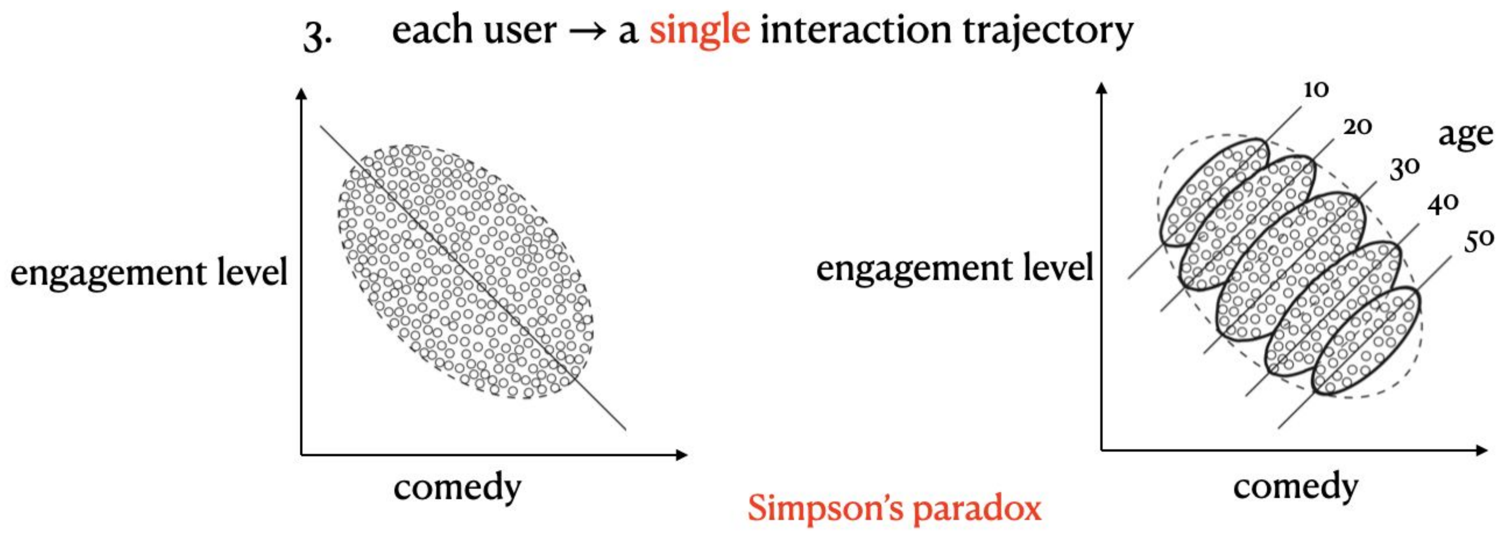


Goal

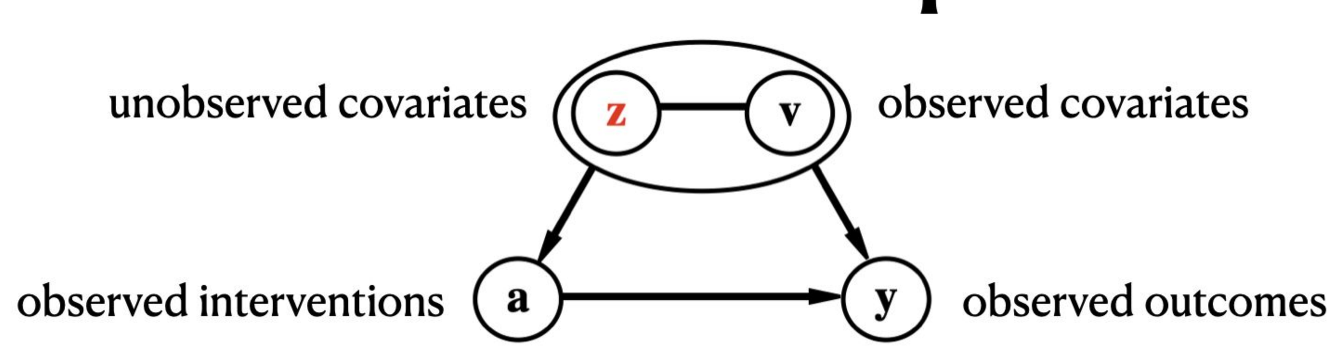


Challenges

1. unobserved factors → spurious associations
2. users → heterogeneous
3. each user → a single interaction trajectory



Problem Setup



n heterogenous and independent users with one observation each - $\{\underbrace{\mathbf{v}^{(i)}, \mathbf{a}^{(i)}, \mathbf{y}^{(i)}}_{p-\text{dimensional}}\}_{i=1}^{n}$

Goal : Counterfactual Questions

For user $i \in [n]$, what would have happened if alternative treatments were assigned?

≡

Estimate $y^{(i)}(\tilde{a}^{(i)})$ for $\tilde{a}^{(i)} \in \mathcal{A}$?

Suffices to learn $f(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, z^{(i)}, \mathbf{v}^{(i)})$ for all $i \in [n]$, but each user may have *different* z

Can we learn n different distributions with *one* sample per distribution ?

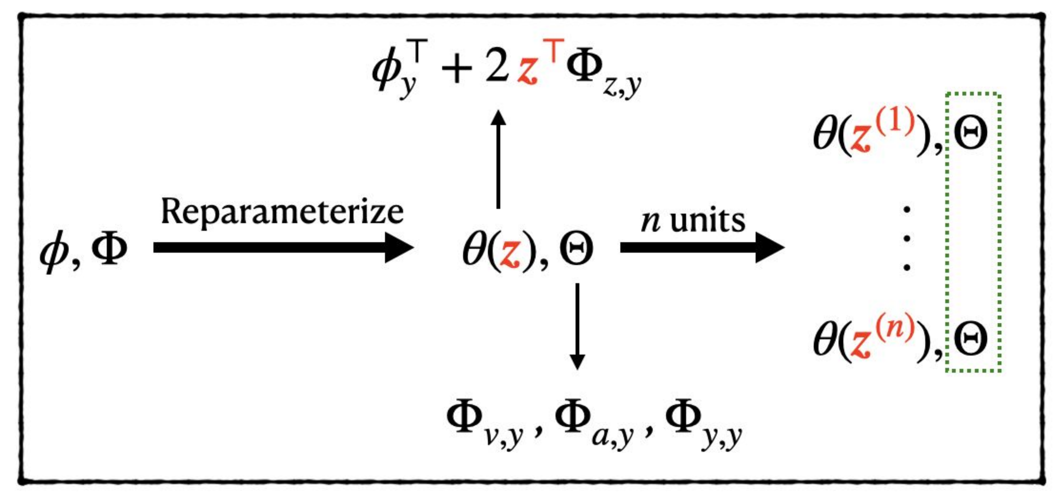
Our Approach

We posit a joint exponential family distribution for $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$

$$f(\boldsymbol{w}) \propto \exp\left(\boldsymbol{\phi}^\top \boldsymbol{w} + \boldsymbol{w}^\top \Phi \boldsymbol{w}\right)$$

$$f(\boldsymbol{y} \mid \boldsymbol{a}, z = z^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\ \underbrace{\phi_y^\top + 2 z^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y}}_{\text{different for different users}} + 2\boldsymbol{a}^\top\Phi_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y}\right)$$

n heterogeneous conditional distributions ➡ same exp. family but with diff. parameters



Inference Tasks

1. Parameters: User-level — $\theta^\star(z^{(i)})$ for all $i \in [n]$ → counterfactual distribution

Population-level — $\Theta^\star$

2. Potential Outcomes: $\mu^{(i)} \triangleq \mathbf{E}\left[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \mid \mathbf{z} = z^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}\right]$ → counterfactual mean

Parameter Estimation

$\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{n}$ → pool all n samples

→ $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(n)}, \widehat{\Theta}$ estimates

$$\min_{\theta^{(1)},\cdots,\theta^{(n)},\Theta} \sum_{t\in[p]} \frac{1}{n} \sum_{i\in[n]} \exp\left(-\left[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\right] x_t^{(i)}\right)$$

Assum 1: $\Theta^\star$ has sparse rows

Assum 2: $\theta^\star(z^{(i)}) \in$ set $\mathscr{B}$

$$\|\Theta^\star - \widehat{\Theta}\|_{2,\infty} \le \epsilon \quad \text{when } n \ge O\left(\frac{p^2(p + M_n(\epsilon^2))}{\epsilon^4}\right)$$

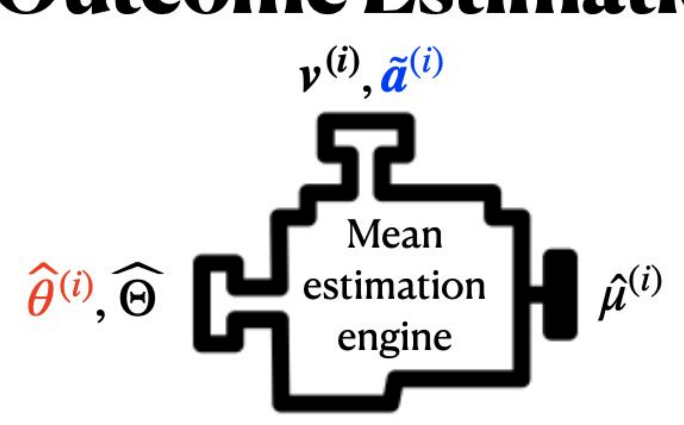
$$\text{For all } i, \text{MSE}\left(\theta^\star(z^{(i)}), \hat{\theta}^{(i)}\right) \le \max\left\{\epsilon^2, \frac{M(c)}{p}\right\} \quad \text{when } n \ge O\left(\frac{p^2(pM(c) + M_n(\epsilon^2))}{\epsilon^4}\right)$$

metric entropy of $\mathscr{B}$ — $M_n(\epsilon) = nM(n\epsilon)$

★ When $\mathscr{B}$ = $s-$sparse linear combinations of k known vectors,

$M(c) = O\left(s\log(k)\right)$ and $M_n(\epsilon) = O\left(\frac{s\log k}{\epsilon}\right)$

Outcome Estimation



$$\hat{f}(\boldsymbol{y} \mid \boldsymbol{a} = \tilde{\boldsymbol{a}}^{(i)}, z = z^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\ \hat{\theta}(z^{(i)}) + 2\boldsymbol{v}^{(i)\top}\widehat{\Phi}_{v,y} + 2\tilde{\boldsymbol{a}}^{(i)\top}\widehat{\Phi}_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top\widehat{\Phi}_{y,y}\boldsymbol{y}\right)$$

For all i and any $\tilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}$,

$$MSE\left(\mu^{(i)}, \hat{\mu}^{(i)}\right) \le \epsilon^2 + \frac{M(c)}{p} \quad \text{when } n \ge O\left(\frac{p^2(pM(c) + M_n(\epsilon^2))}{\epsilon^4}\right)$$

Application: Denoise User-wise Data

No systematically unobserved covariates

Noisy observed data = true data + measurement error
$\bar{\mathbf{x}}$ $\quad$ $\mathbf{x}$ $\quad$ $\Delta\mathbf{x}$

Assum 1: Only half users have error: $\Delta\mathbf{x}^{(i)} = \mathbf{0}$ for $i \in \{n/2, \cdots, n\}$

Assum 2: Data has sparse error: $\|\Delta\mathbf{x}^{(i)}\|_0 \le s$ for $i \in \{1, \cdots, n/2\}$

Goal: Estimate the true data

$$\text{For all } i, \ \|\Delta\mathbf{x}^{(i)}, \widehat{\Delta\mathbf{x}^{(i)}}\|^2 \le \max\left\{\frac{\epsilon^2}{s}, \frac{s}{p}\right\} + \epsilon^2 \quad \text{when } n \ge O\left(\frac{s^2 p}{\epsilon^4}\right)$$