On Counterfactual Inference with Unobserved Confounding

Abhin Shah
MIT

Raaz Dwivedi
Harvard & MIT

Devavrat Shah
MIT

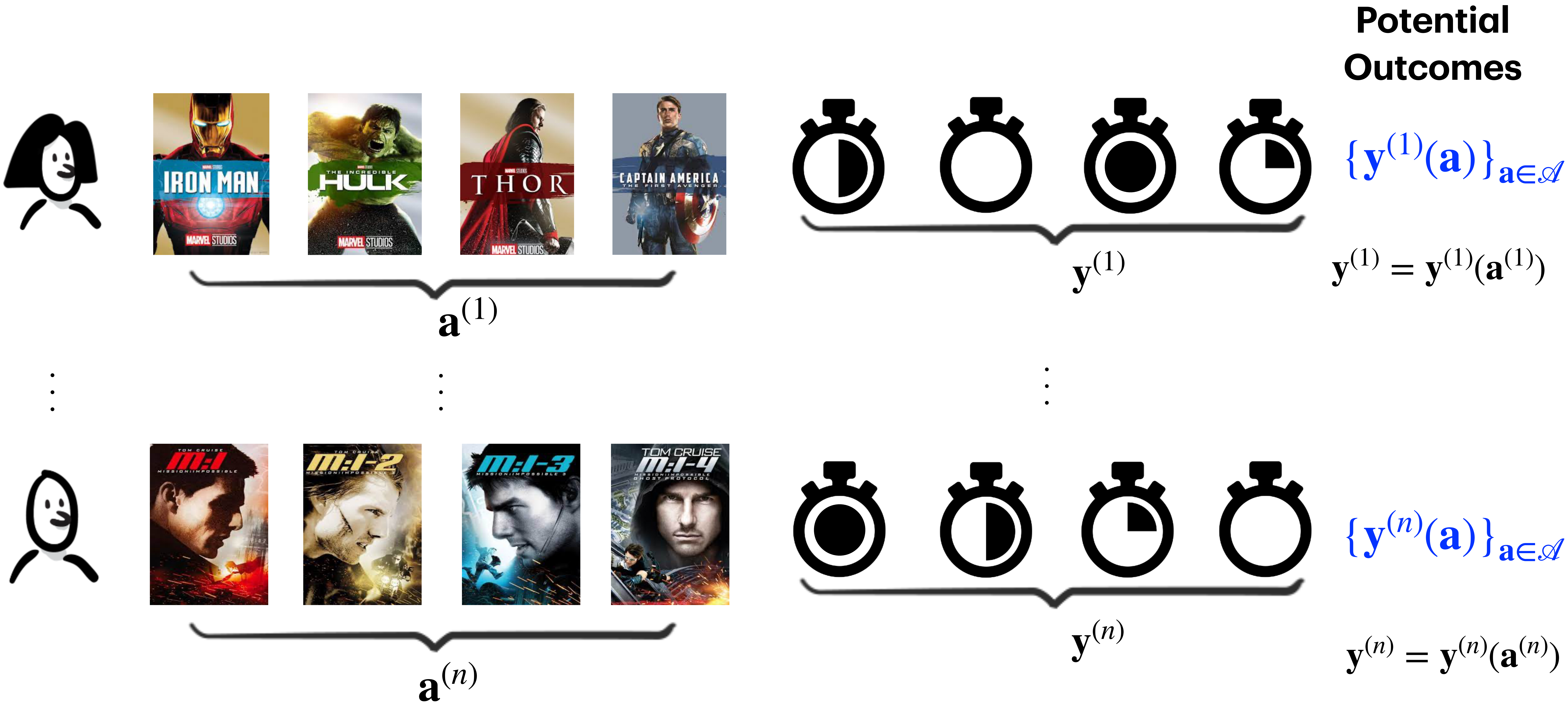
Greg Wornell
MIT

Observational Setting

a — action/intervention

y — outcome

Panel Data

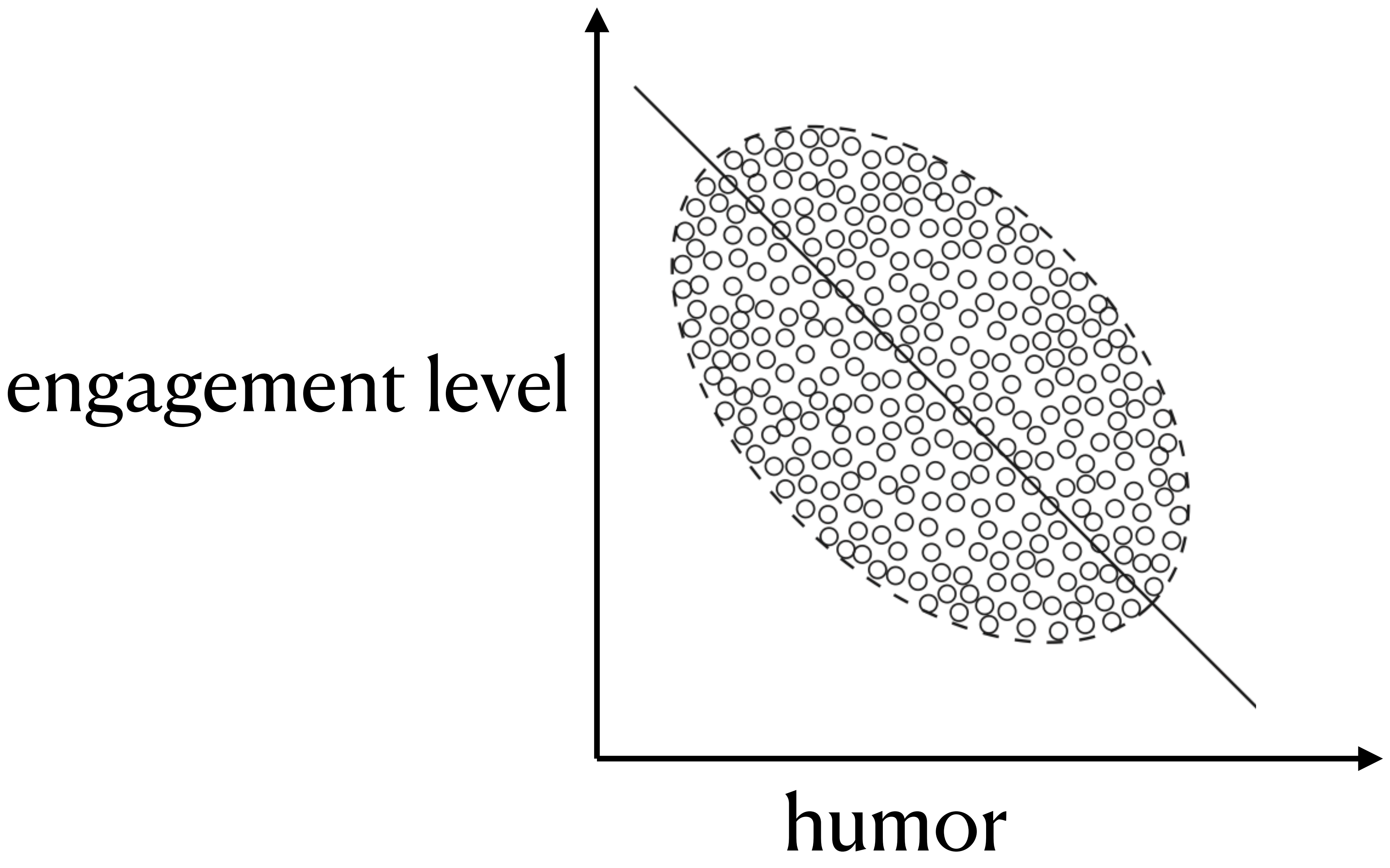

Potential Outcomes
IRON MAN
MARVEL STUDIOS
THE INCREDIBLE HULK
MARVEL STUDIOS
THOR
MARVEL STUDIOS
CAPTAIN AMERICA
THE FIRST AVENGER
$\{\mathbf{y}^{(1)}(\mathbf{a})\}_{\mathbf{a}\in\mathscr{A}}$
$\mathbf{a}^{(1)}$
$\mathbf{y}^{(1)}$
$\mathbf{y}^{(1)} = \mathbf{y}^{(1)}(\mathbf{a}^{(1)})$
TOM CRUISE
M:I
M:I-2
M:I-3
TOM CRUISE
M:I-4
$\{\mathbf{y}^{(n)}(\mathbf{a})\}_{\mathbf{a}\in\mathscr{A}}$
$\mathbf{a}^{(n)}$
$\mathbf{y}^{(n)}$
$\mathbf{y}^{(n)} = \mathbf{y}^{(n)}(\mathbf{a}^{(n)})$

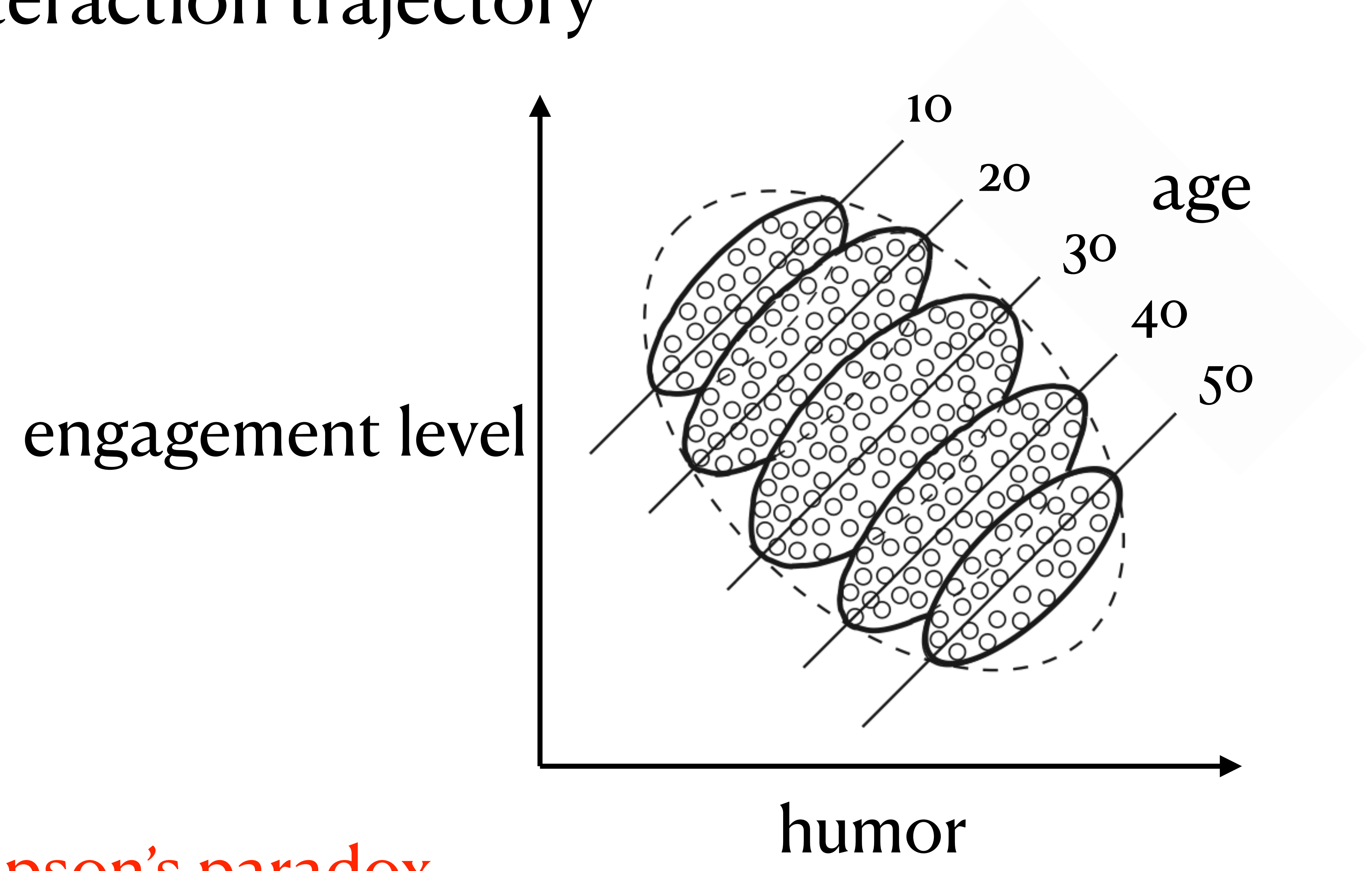
Goal : What-if?


TOY STORY
TOM HANKS TIM ALLEN
TOY STORY 2
THE TOYS ARE BACK! THANKSGIVING
TOY STORY 3
TOY STORY 4
$\tilde{\mathbf{a}}^{(1)}$
$\mathbf{y}^{(1)}(\tilde{\mathbf{a}}^{(1)})$?
ICE AGE
ICE AGE 2
THE MELTDOWN
ICE AGE 3
DAWN OF THE DINOSAURS
ICE AGE 4
CONTINENTAL DRIFT
$\tilde{\mathbf{a}}^{(n)}$
$\mathbf{y}^{(n)}(\tilde{\mathbf{a}}^{(\mathbf{n})})$?

Challenges

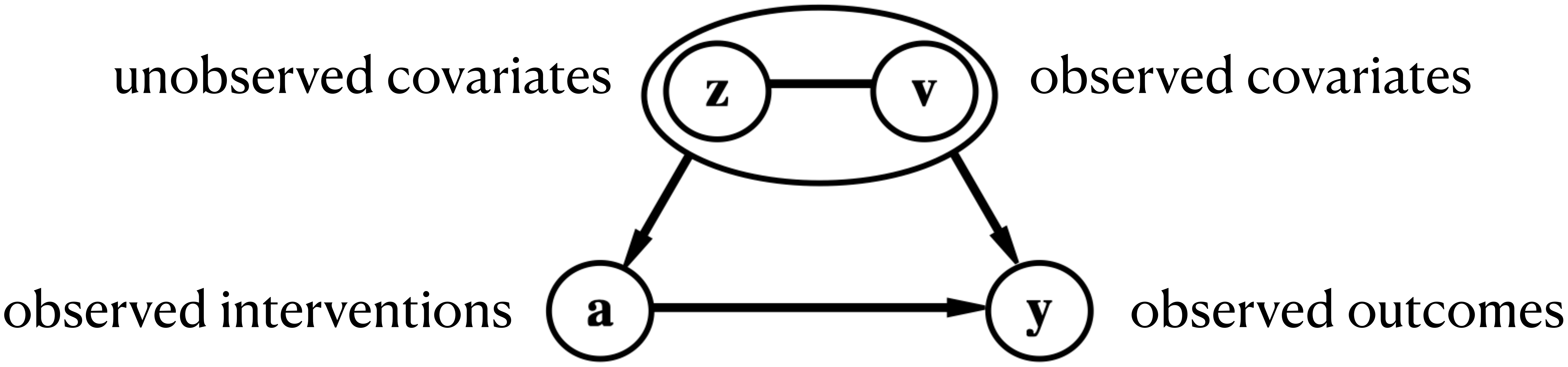
1. unobserved factors → spurious associations
2. users → heterogeneous
3. each user → a single interaction trajectory





Simpson's paradox

Problem Setup

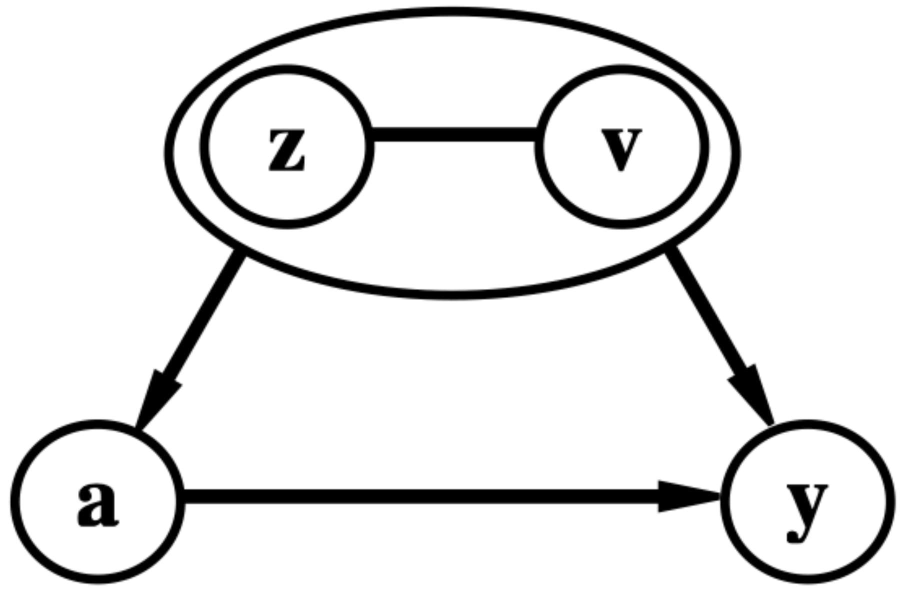


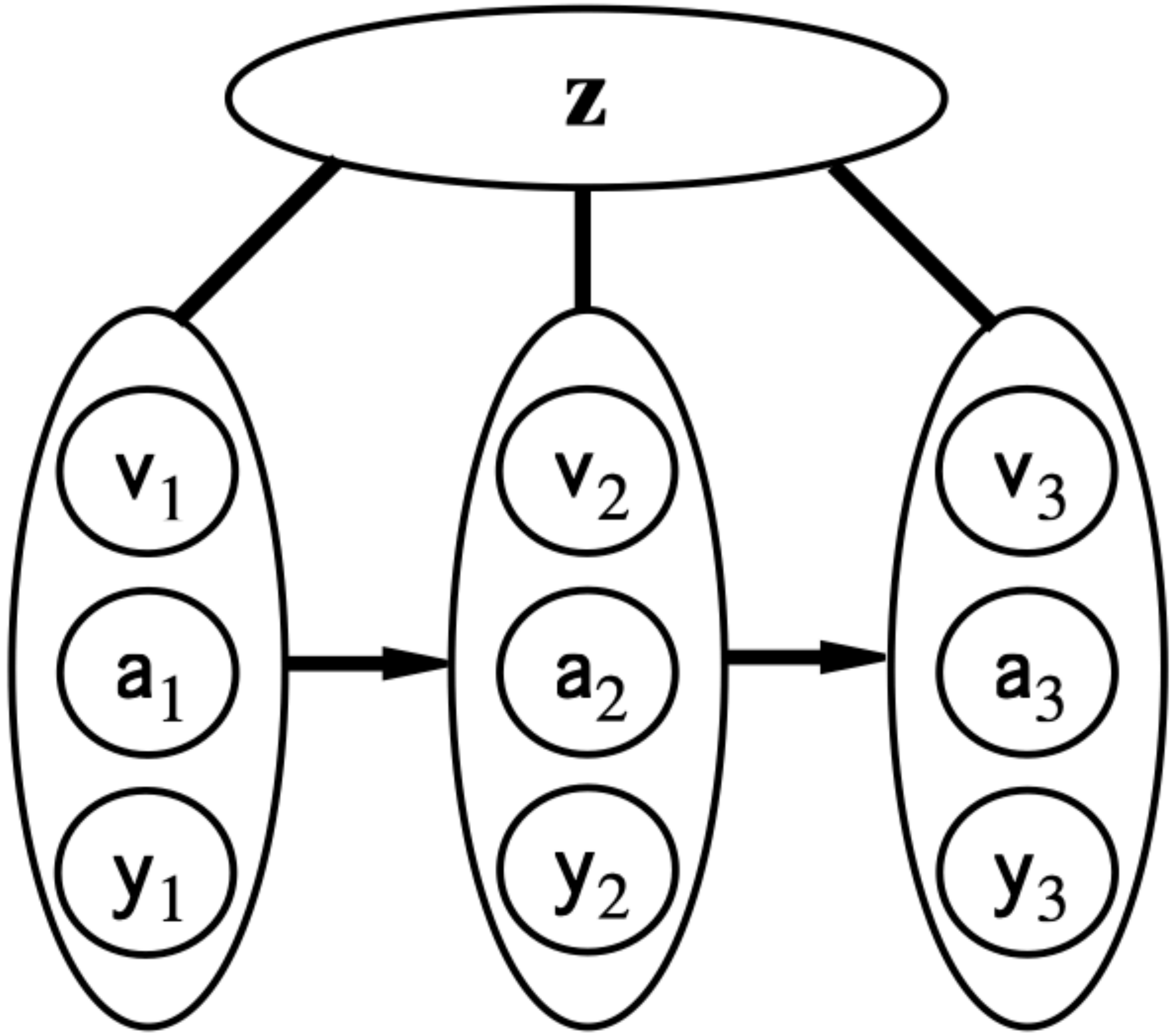
n heterogenous and independent users

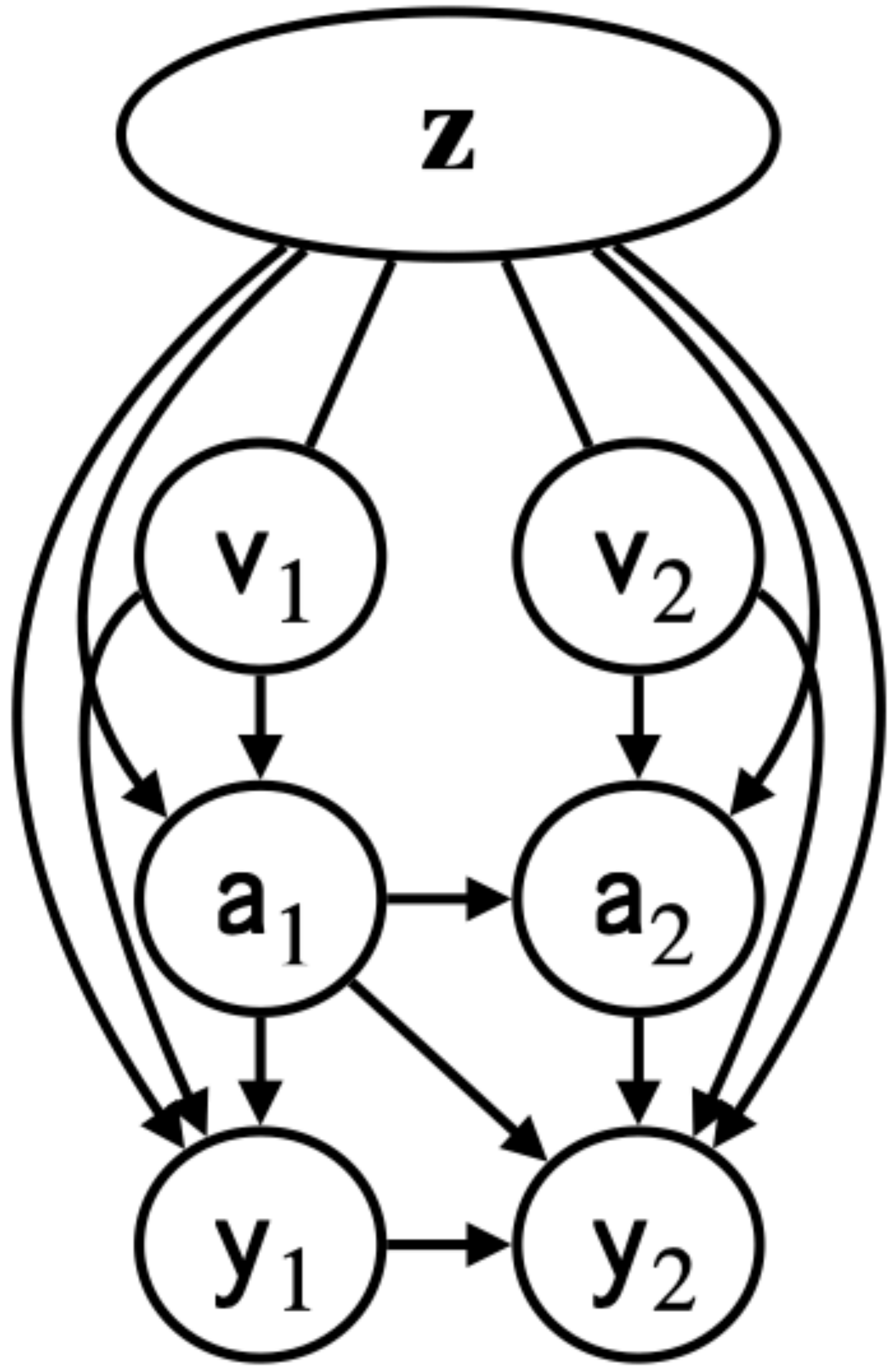
one $p-$dimensional observation per user - $\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{n}$

Examples

Sequential setting

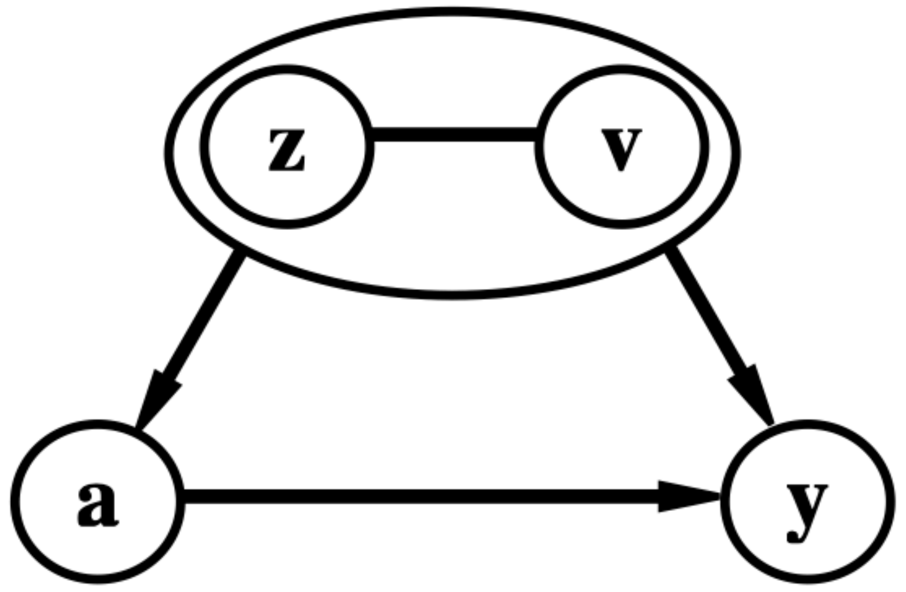


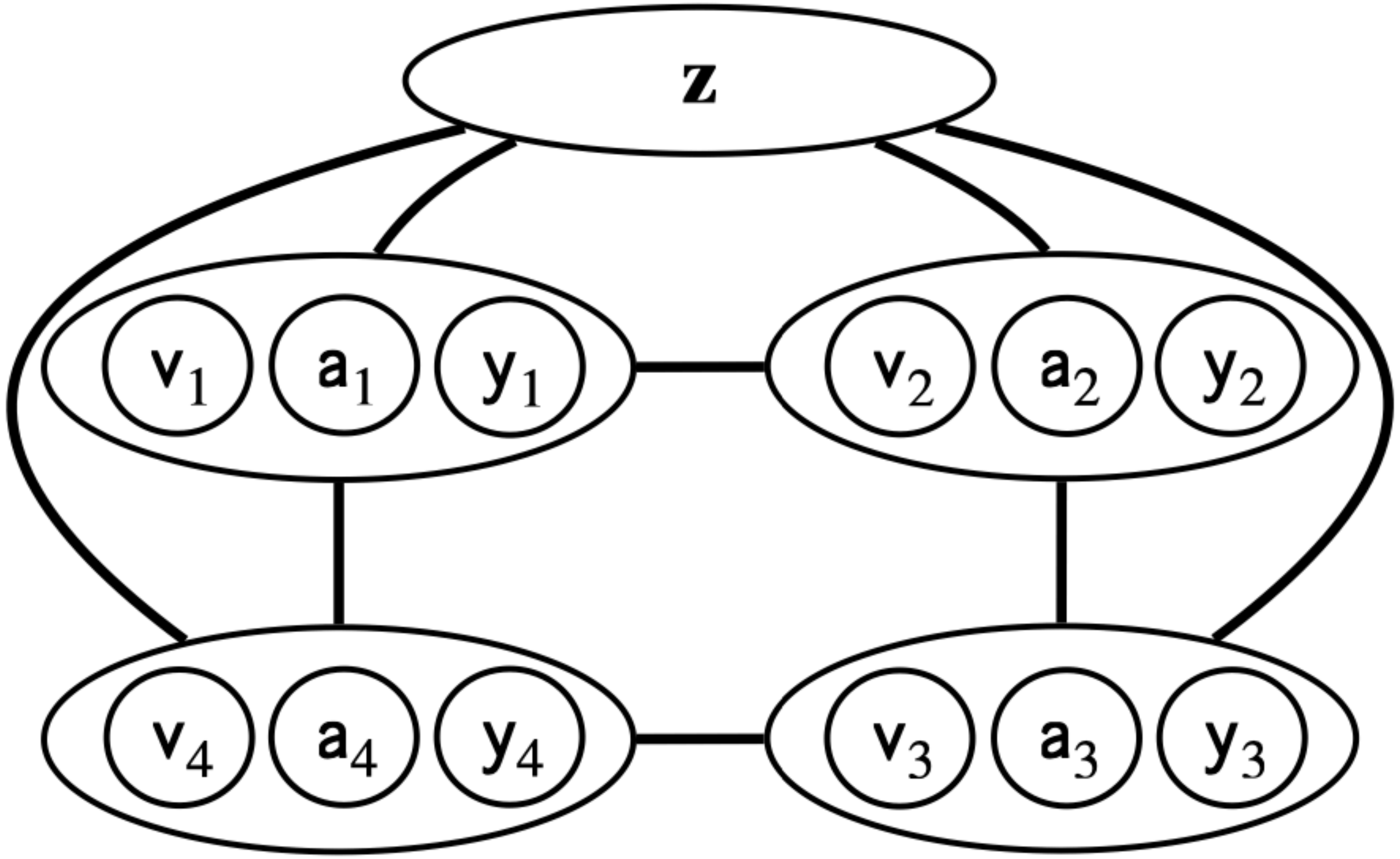


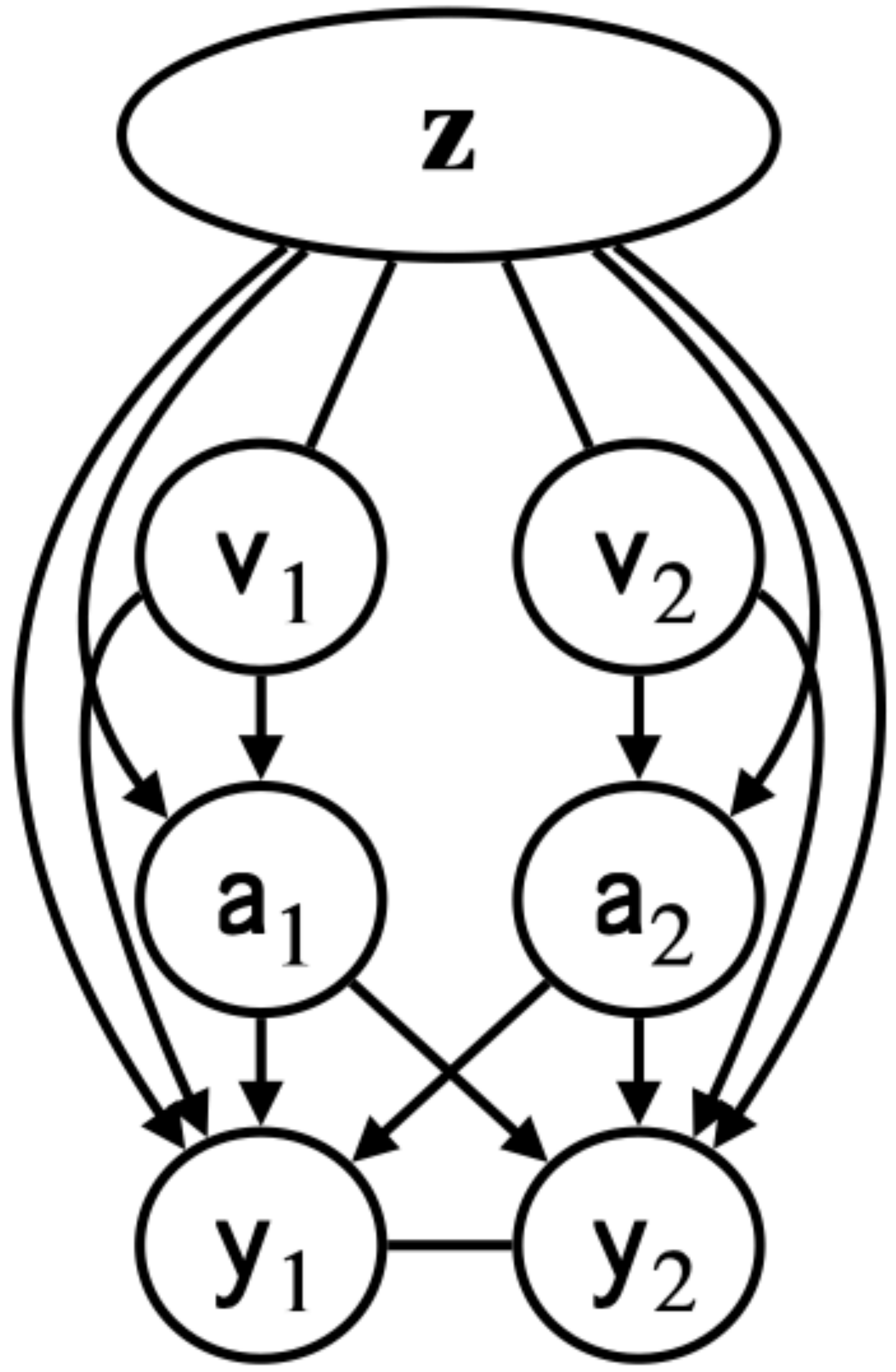


Examples

Network setting







Goal : Counterfactual Questions

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

$$\equiv$$

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

suffices to learn $p(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)})$ for all $i \in [n]$

Heterogeneity: each user may have different *unobserved* z

Can we learn n different distributions

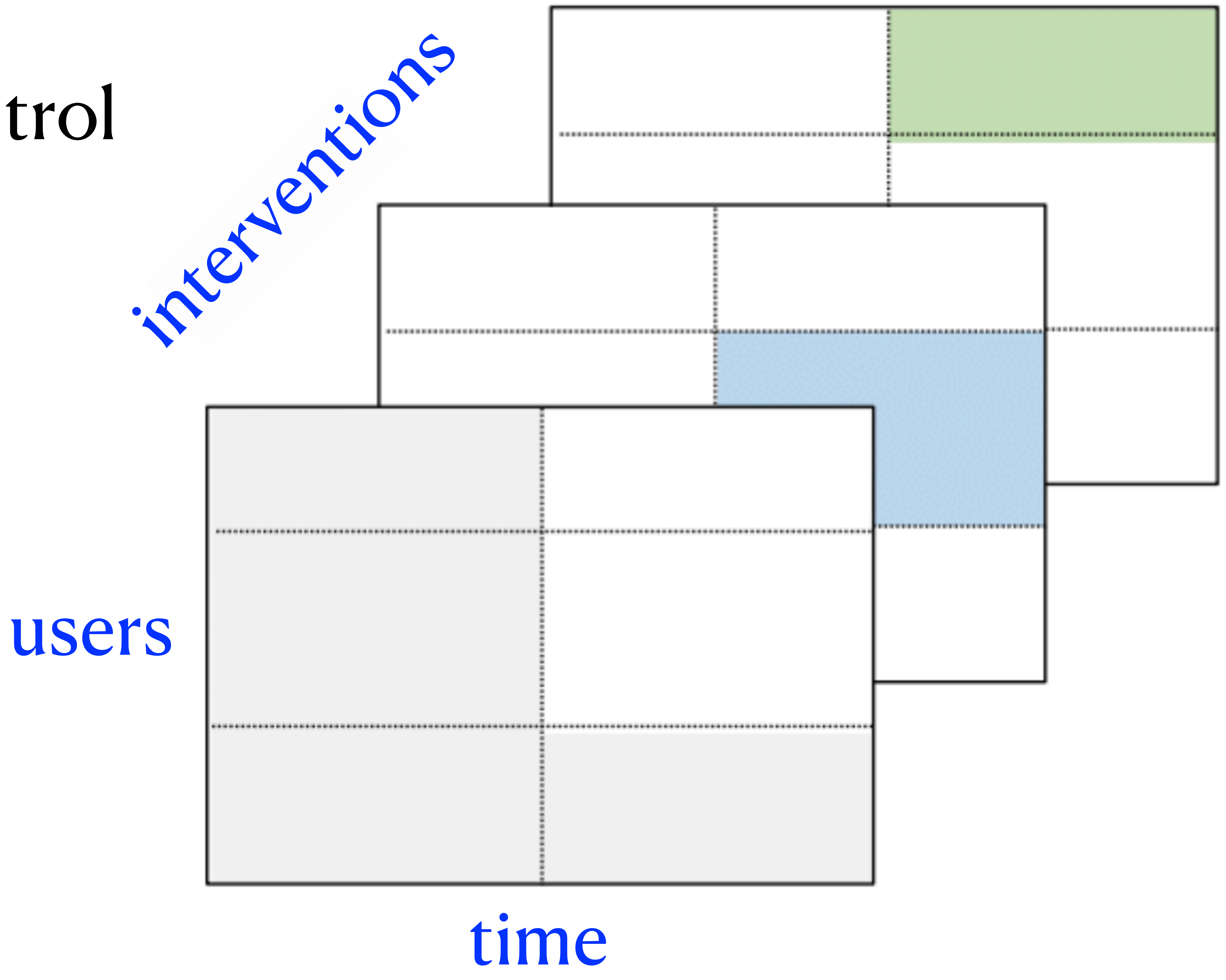
$$p(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, \boldsymbol{z}^{(i)}, \boldsymbol{v}^{(i)}) \quad i \in [n]$$

with *one* sample per distribution?

Related Work

Linear panel data

- Examples: difference in differences, synthetic control + variants, synthetic interventions + variants
- Only finitely many interventions
- Special structure on intervention assignment



Our Approach

Model the joint distribution of $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$ as a particular exponential family

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

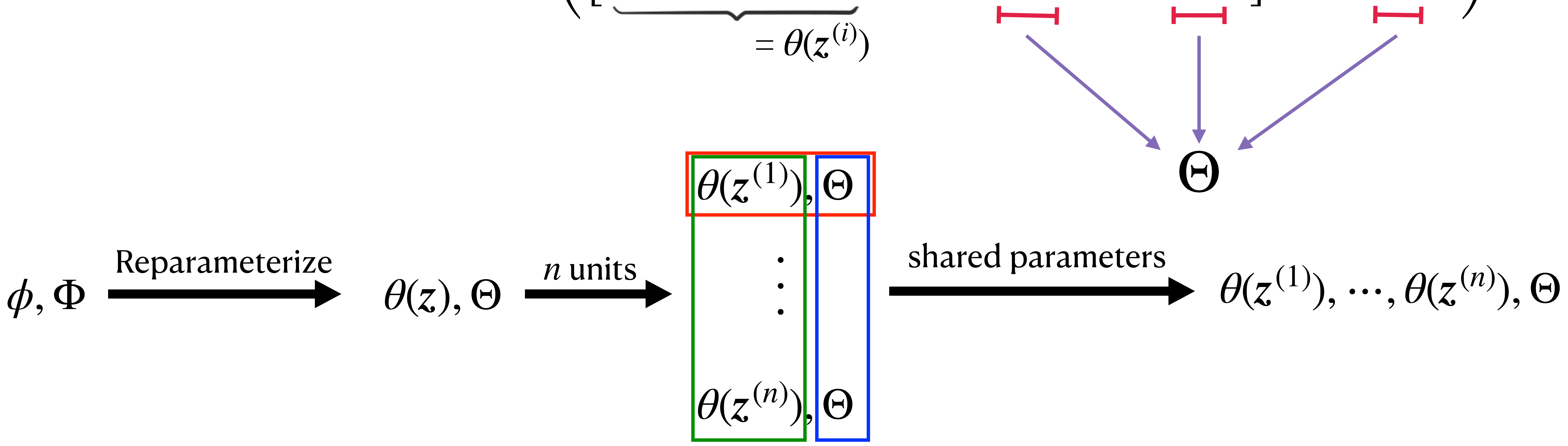
$$p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\,\phi_y^\top + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top \Phi_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top \Phi_{y,y}\boldsymbol{y}\right)$$

different for different users

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

Our Approach

$$p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\Big[\underbrace{\phi_y^\top + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y}}_{= \theta(\boldsymbol{z}^{(i)})} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top\Phi_{a,y}\Big]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y}\right)$$



1. If $z^{(1)} = \cdots = z^{(n)} \rightarrow$ a single exponential family with n samples **[B '15, KM '17, SSW '21, VML '22]**
2. If $n = 1 \rightarrow$ a single exponential family with one sample (assume Θ is known) **[KDDGD '21, MHBM' 21]**

Inference Tasks

1. Parameters
   A. User-level — $\theta^\star(z^{(i)})$ for all $i \in [n]$ → counterfactual distribution
   B. Population-level — $\Theta^\star$
2. Expected potential outcomes — $\mathbf{E}\left[y^{(i)}(\tilde{a}^{(i)}) \mid \mathbf{z} = z^{(i)}, \mathbf{v} = v^{(i)}\right]$ → counterfactual mean

Complexity of Parameter Space

Λ_θ — the set containing the true model parameters $\{\theta^\star(z^{(i)})\}_{i\in[n]}$

	Linear combination of k-known vectors	s-sparse linear combination of k-known vectors
$M(\epsilon) = logN(\Lambda_\theta, \epsilon)$	$O\left(k \cdot log\left(1 + \frac{1}{\epsilon}\right)\right)$	$O\left(s \log k \cdot log\left(1 + \frac{1}{\epsilon}\right)\right)$
$M_n(\epsilon) = nM(n\epsilon)$	$O\left(\frac{k}{\epsilon}\right)$	$O\left(\frac{s \log k}{\epsilon}\right)$

Parameter Estimation

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

pool all n samples

Loss function

$\hat{\theta}(\boldsymbol{z}^{(1)}), \cdots, \hat{\theta}(\boldsymbol{z}^{(n)}), \hat{\Theta}$

estimates

$\|\Theta^\star - \hat{\Theta}\|_{2,\infty} \leq \epsilon$ when $n \geq O\left(\frac{p^2\left(p + M_n(\epsilon^2)\right)}{\epsilon^4}\right)$

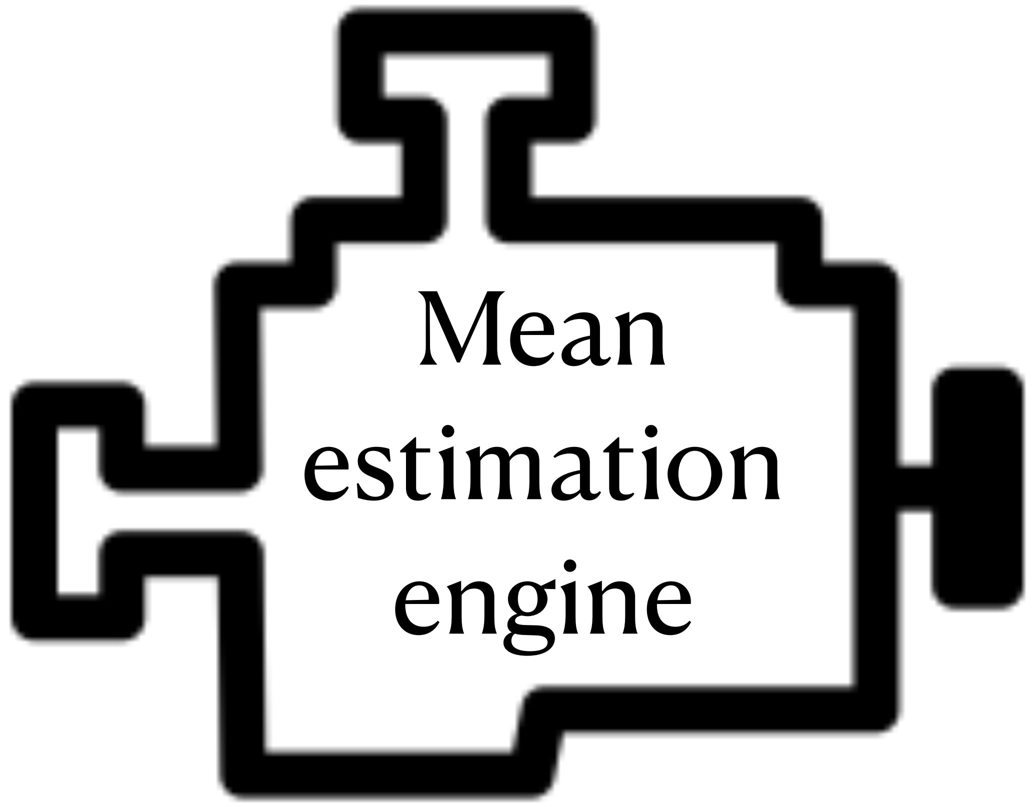
For all i, $\text{MSE}\left(\theta^\star(\boldsymbol{z}^{(i)}), \hat{\theta}(\boldsymbol{z}^{(i)})\right) \leq \max\left\{\epsilon^2, \frac{M(c)}{p}\right\}$ when $n \geq O\left(\frac{p^2\left(pM(c) + M_n(\epsilon^2)\right)}{\epsilon^4}\right)$

Outcome Estimation

Expected potential outcomes — $\mu^{(i)} \triangleq \mathbf{E}\left[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \mid \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}\right]$

$(\boldsymbol{v}^{(\boldsymbol{i})}, \tilde{\boldsymbol{a}}^{(i)})$

$\hat{\theta}(\boldsymbol{z}^{(i)}), \hat{\Theta}$



$\hat{\mu}^{(i)}$

For all i, $MSE\left(\mu^{(i)}, \hat{\mu}^{(i)}\right) \leq \epsilon^2 + \frac{M(c)}{p}$ when $n \geq O\left(\frac{p^2\left(pM(c) + M_n(\epsilon^2)\right)}{\epsilon^4}\right)$

$\mathbf{x} \triangleq (\mathbf{v}, \mathbf{a}, \mathbf{y})$

An Application

Denoise user-wise data

No systematically unobserved covariates $\mathbf{z}$

Noisy observed data = true data + measurement error

$\bar{\mathbf{x}}$ $\qquad$ $\mathbf{x}$ $\qquad$ $\Delta\mathbf{x}$

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

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

Goal: Estimate the true data

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

Remainder Of The Talk

The Loss Function, And Why It Works

Log likelihood as the loss function = Max. Likelihood Est.

not the right answer (computational challenge)

An alternative

a *proper* loss function, computationally efficient

ingredients of analysis

population-level parameter est through concentration due to many samples

user-level parameter est through *single* sample concentration due to

Log Sobolev + Dobrushin's criteria

Maximum Likelihood Estimation

Consider an exponential family $p_\theta(\boldsymbol{x}) = \dfrac{\exp(\theta^\top f(\boldsymbol{x}))}{Z(\theta)}$

Given independent $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim p_{\theta^\star}(\boldsymbol{x})$, learn $\theta^\star$

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

$$\hat{\theta}_{MLE} \triangleq \text{argmax}_\theta \ \text{Log-Likehood}(\theta)$$

$$\text{Log-likelihood}(\theta) = \frac{1}{n}\sum_{i\in[n]} \theta^\top f(\boldsymbol{x}_i) - \log Z(\theta)$$

A. Consistency ? ✓

B. Asymptotic normality ? ✓

C. Asymptotic efficiency ? ✓

D. Computational tractability ? ✗ How to compute $Z(\theta)$?

An Alternative

Consider an exponential family $p_\theta(\boldsymbol{x}) = \dfrac{\exp(\theta^\top f(\boldsymbol{x}))}{Z(\theta)}$

Given independent $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim p_{\theta^\star}(\boldsymbol{x})$, learn $\theta^\star$

$$\hat{\theta} \triangleq \text{argmin}_\theta \; \ell(\theta)$$

likelihood

$$\ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\big(-\theta^\top f(\boldsymbol{x}_i)\big) \qquad \prod_{i \in [n]} \frac{1}{Z(\theta)} \exp\big(\theta^\top f(\boldsymbol{x}_i)\big)$$

Convex constraint on θ ⟶ convex optimization problem

avoids $Z(\theta)$

but is it any good?

Properties

[SSW'21]

$$\hat{\theta} \in \arg\min_{\theta} \ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\big(-\theta^\top f(\boldsymbol{x}_i)\big)$$

- ✓ A. Strictly proper loss function
- ✓ B. Consistency ?
- ✗ C. Asymptotic normality ?
- ✓ D. Asymptotic efficiency ?
- ✓ E. Computational tractability ?
- ✓ F. Finite sample guarantees ?

$\hat{\theta}$ is MLE w.r.t. $p_{\theta^\star - \theta}$ (not p_θ)

No need to compute $Z(\theta)$!

Finite Sample Guarantees

[SSW'21]

$$\hat{\theta} \in \arg\min_{\theta} \ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\big(- \theta^\top f(\boldsymbol{x}_i)\big)$$

$$\|\theta^\star - \widehat{\theta}\|_2 \sim n^{-1/4}$$

Proof ingredients —

A. Concentration of gradient

B. Anti-concentration of Hessian (Restricted strong convexity)

} Hoeffding's inequality

Back To Our Setting

Condition on z

Recall the joint distribution of $\mathbf{w} = (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$

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

Letting $\mathbf{x} \triangleq (\mathbf{v}, \mathbf{a}, \mathbf{y})$, the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ can be written as

$$p_{\theta(z),\Theta}(\boldsymbol{x} \mid z) \propto \exp\left(\left[\theta(z)\right]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{a}, z = z^{(i)}, v = v^{(i)}) \propto \exp\left(\left[\phi_y + 2z^{(i)\top}\Phi_{z,y} + 2v^{(i)\top}\Phi_{v,y} + 2a^\top \Phi_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top \Phi_{y,y}\boldsymbol{y}\right)$$

$$p(x_t \mid \boldsymbol{x}_{-t}, z) \propto \exp\left(\left[\theta_t(z) + 2\Theta_t^\top \boldsymbol{x}\right] x_t\right)$$

Structure On Parameters

$$p_{\theta(z),\Theta}(\boldsymbol{x} \mid z) \propto \exp\left(\left[\theta(z)\right]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x}\right)$$

(A) Every element of $\theta^\star(z^{(1)}), \cdots, \theta^\star(z^{(n)})$, and $\Theta^\star$ are bounded

(B) Every row of $\Theta^\star$ is sparse

$\Lambda_\theta \triangleq \{\theta : \theta$ is consistent with (A) + low complexity$\}$ → required to provide meaningful guarantees

$\Lambda_\Theta \triangleq \{\Theta : \Theta$ is consistent with (A) and (B)$\}$

$$p(x_t \,|\, \boldsymbol{x}_{-t}, z) \propto \exp\Big(\big[\theta_t(z) + 2\Theta_t^\top \boldsymbol{x}\big] x_t\Big)$$

Learning Population-level Parameter

$$\mathscr{L}(\underline{\Theta}) = \sum_{t\in[p]} \frac{1}{n} \sum_{i\in[n]} \exp\Big(- \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ where } \underline{\Theta} \triangleq \big[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\big]$$

$\Lambda_\Theta \to$ (A) bounded elements, (B) sparse rows

Λ_Θ places independent constraints on the rows of Θ

p independent optimization problems

$$\mathscr{L}_t(\underline{\Theta}_t) = \frac{1}{n} \sum_{i\in[n]} \exp\Big(- \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ for all } t \in [p] \longrightarrow \widehat{\Theta}_t$$

$$\|\Theta_t^\star - \widehat{\Theta}_t\|_{2,\infty} \sim n^{-1/4}$$

A. Concentration of gradient
B. Anti-concentration of Hessian

} Hoeffding's inequality

$$p(x_t \mid \boldsymbol{x}_{-t}, z) \propto \exp\Big(\big[\theta_t(z) + 2\Theta_t^\top \boldsymbol{x}\big]x_t\Big)$$

Learning Unit-level Parameter

$$\mathscr{L}(\theta^{(1)}, \cdots, \theta^{(n)}) = \sum_{t\in[p]} \frac{1}{n} \sum_{i\in[n]} \exp\Big(- \big[\theta_t^{(i)} + 2\,\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}\big]x_t^{(i)}\Big)$$

$\Lambda_\theta \to$ (A) bounded elements, (B) low complexity

$\theta^{(1)}, \cdots, \theta^{(n)} \in \Lambda_\theta^n$ places independent constraints on units, i.e., $\theta^{(i)} \in \Lambda_\theta$ for all $i \in [n]$

n independent optimization problems

$$\mathscr{L}^{(i)}(\theta^{(i)}) = \sum_{t\in[p]} \exp\Big(- \big[\theta_t^{(i)} + 2\,\widehat{\Theta}_t^\top \boldsymbol{x}^{(i)}\big]x_t^{(i)}\Big) \text{ for all } i \in [n] \longrightarrow \widehat{\theta}^{(i)}$$

$$\|\theta^\star(z^{(i)}) - \widehat{\theta}^{(i)}\|_2 \sim \max\{n^{-1/4}, M\}$$

A. Concentration of gradient
B. Anti-concentration of Hessian

} Logarithmic Sobolev inequality