Disaggregated Interventions to Reduce Inequality + Counterfactuals for the Future

DSGA-1017 Responsible Data Science

Spring 2023

Lucius Bynum, Joshua R. Loftus (LSE), Julia Stoyanovich (NYU)

+

Outline

1. Disaggregated interventions to reduce inequality
   - Problem definition
   - Causal inference and social categories
   - A causal framework for addressing pre-existing inequalities
2. Counterfactuals for the future
   - Motivating example
   - Forward-looking counterfactuals
   - Empirical exploration

Disaggregated Interventions to Reduce Inequality

Three types of algorithmic bias

[Friedman, Nissenbaum 1996], [Stoyanovich, Howe, Jagadish 2020]

- **Pre-existing bias**
  - Originates in society and exists independently of an algorithm
- **Technical bias**
  - Introduced or exacerbated by the technical properties of an algorithm
- **Emergent bias**
  - Arises in the context of an algorithm's use

Three types of algorithmic bias

[Friedman, Nissenbaum 1996], [Stoyanovich, Howe, Jagadish 2020]

- **Pre-existing bias**
  - Originates in society and exists independently of an algorithm
- **Technical bias**
  - Introduced or exacerbated by the technical properties of an algorithm
- **Emergent bias**
  - Arises in the context of an algorithm’s use

Problem definition

The “impact remediation problem”
Formalizing pre-existing bias

1. We observe an existing disparity. We consider it undesired.
2. We have the ability to perform an intervention.
3. We want to decrease the measured disparity.

Example:

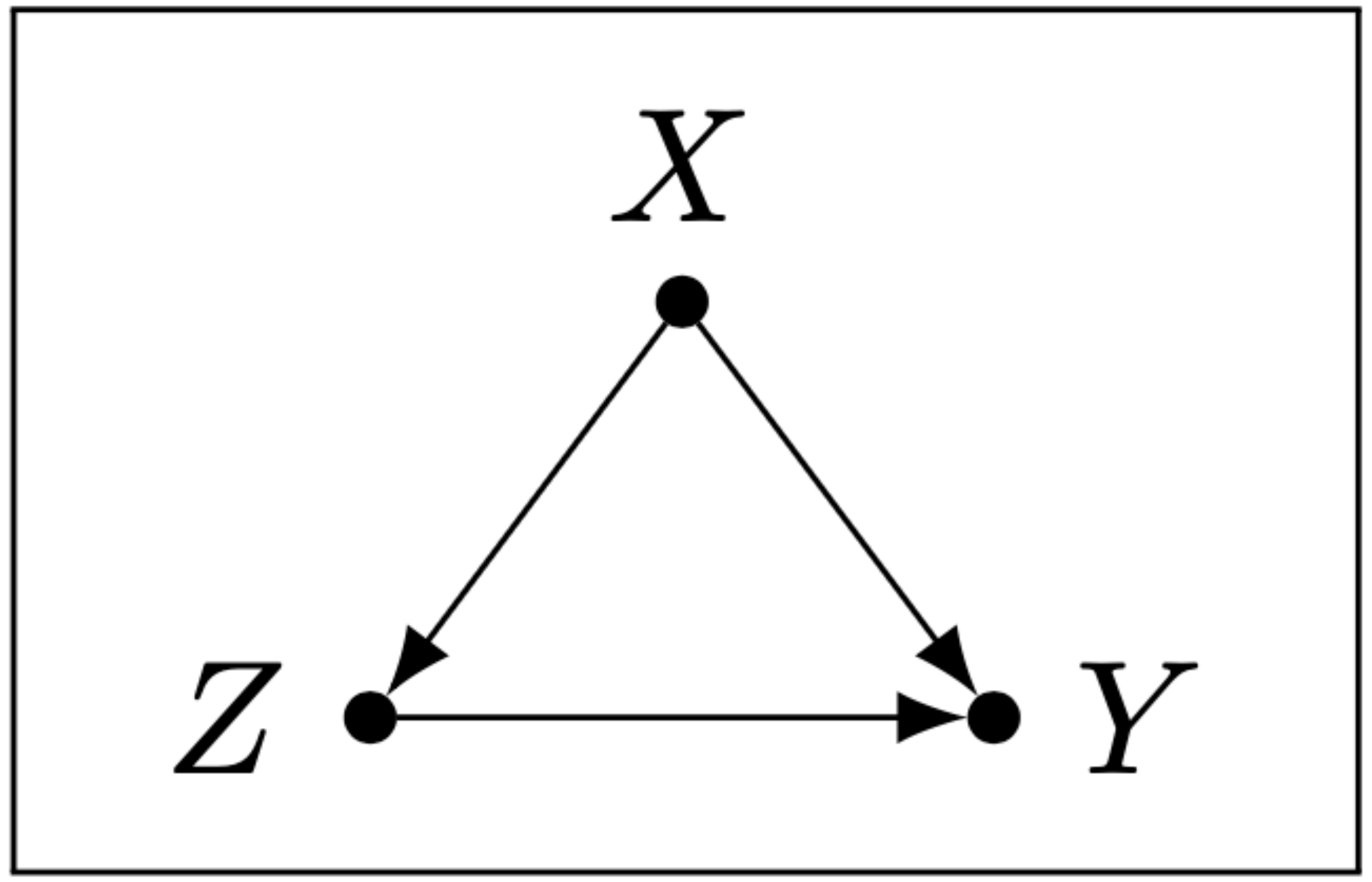
1. Gender imbalance in a job applicant pool
2. Hosting booths at different career fairs
3. Rebalance our applicant pool

Causal inference and social categories

Structural causal models (SCMs)

[Pearl 2009], [Peters et al. 2017]

- An SCM is a four-tuple (U, V, F, P_U)
    - U: a set of exogenous background variables
    - V: a set of endogenous observed variables
    - F: a set of functions (structural equations) for each $V_i \in V$
    - P_U: a distribution over the exogenous variables U
- Each SCM entails a directed acyclic graph (DAG)

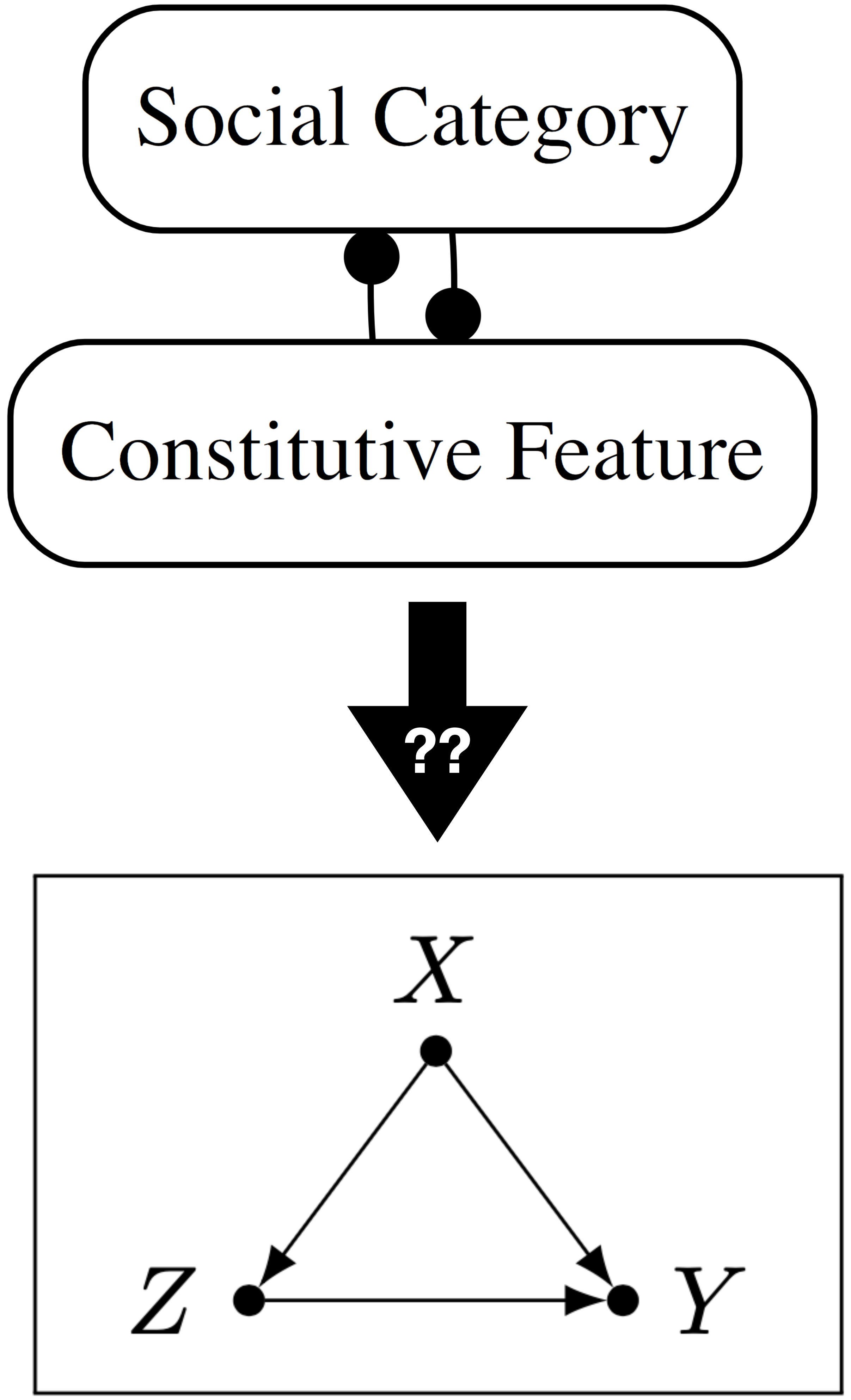


$$X = \epsilon_X$$
$$Z = -1 + X + \epsilon_Z$$
$$Y = 2 \cdot Z + X + \epsilon_Y$$
$$\epsilon_X, \epsilon_Y, \epsilon_Z \sim \mathcal{N}(0,1)$$

Social categories and constitutive features

[Benthall and Haynes 2019], [Hanna et al. 2020], [Sen and Wasow 2016], [Hu and Kohler-Hausmann 2020], [Jacobs and Wallach 2021], [Kasirzadeh and Smart 2021]

- Inequality involves social categories (e.g. race, gender)
- Two types of features:
  - **Regular feature:** causal, diachronic
    - Unfolding over time via cause-and-effect
    - “If A then B then C…”
  - **Constitutive feature:** a feature that defines a social category
    - Synchronous, definitional [Hu and Kohler-Hausmann 2020]
    - “If A then the definition of B has changed…”
- How to write down a DAG? —> instantaneous constitutive **cycle**
- Examples:
  - Intuition: water + number of hydrogen atoms
  - Racial categorization + socioeconomic history
  - Simple variables (e.g., race + net worth), complex constructs



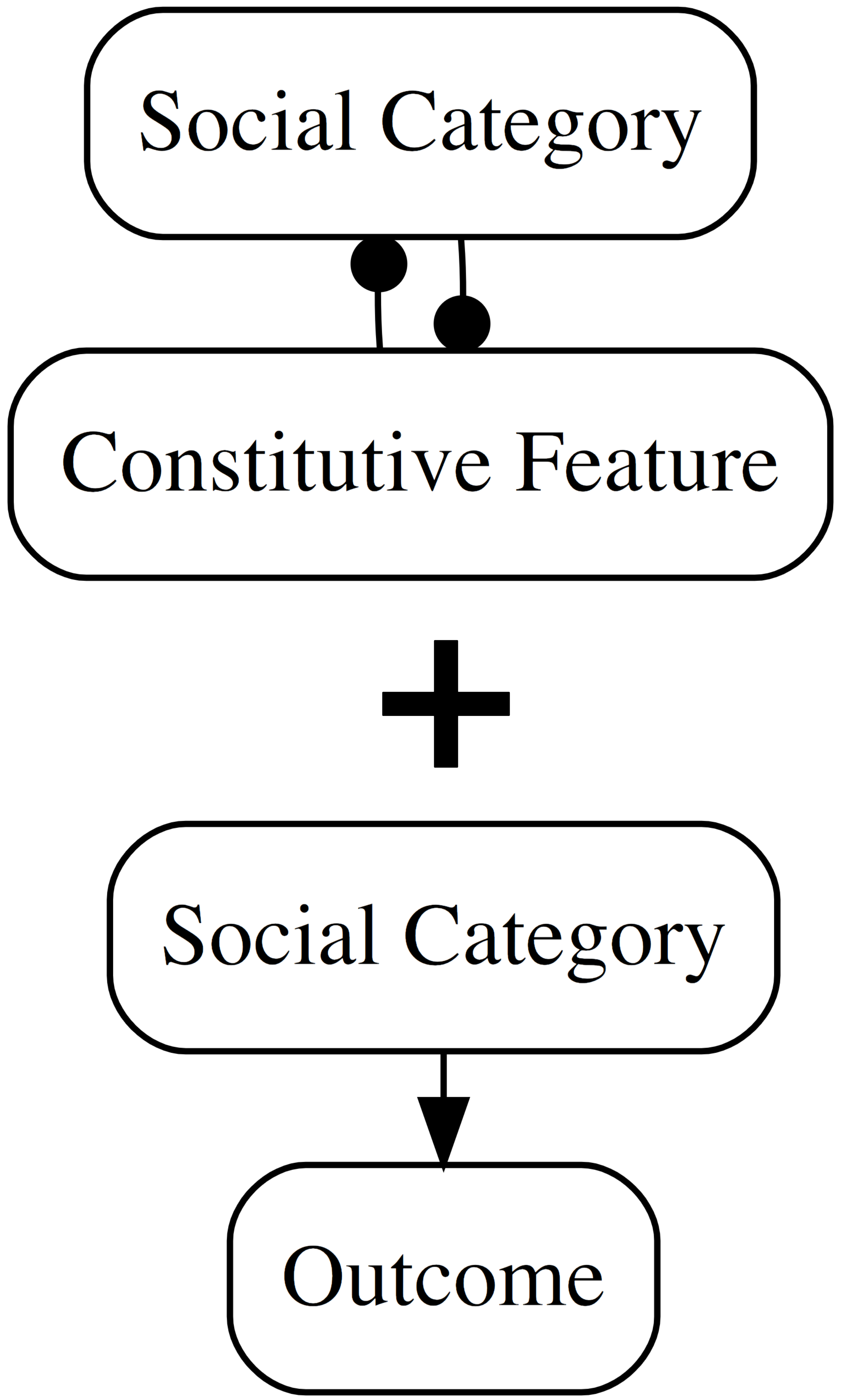
Interventions on social categories

[Benthall and Haynes 2019], [Hanna et al. 2020], [Sen and Wasow 2016], [Hu and Kohler-Hausmann 2020], [Jacobs and Wallach 2021], [Kasirzadeh and Smart 2021]

- Questions for interventions on race:
    - Are manipulations defined?
    - Post-treatment bias
    - Is the social category well-defined? Stable?
- **Simplified positions:**
    1. Defining counterfactuals via exposure to a racial cue

       vs.

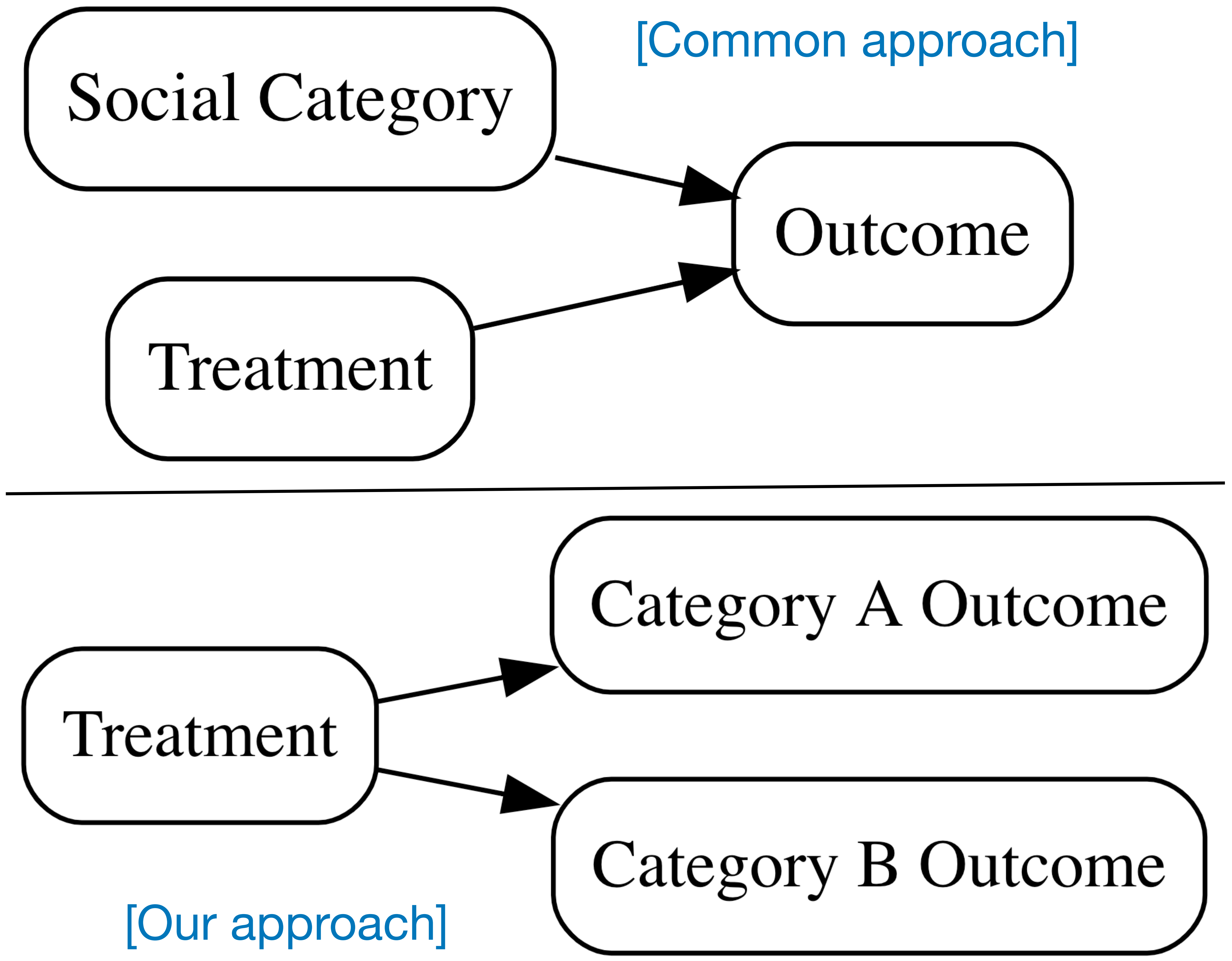
    2. Not using causal models with race at all
- **Unavoidable problem:** racial disparities still exist (and other social category disparities)



Social categories in impact remediation

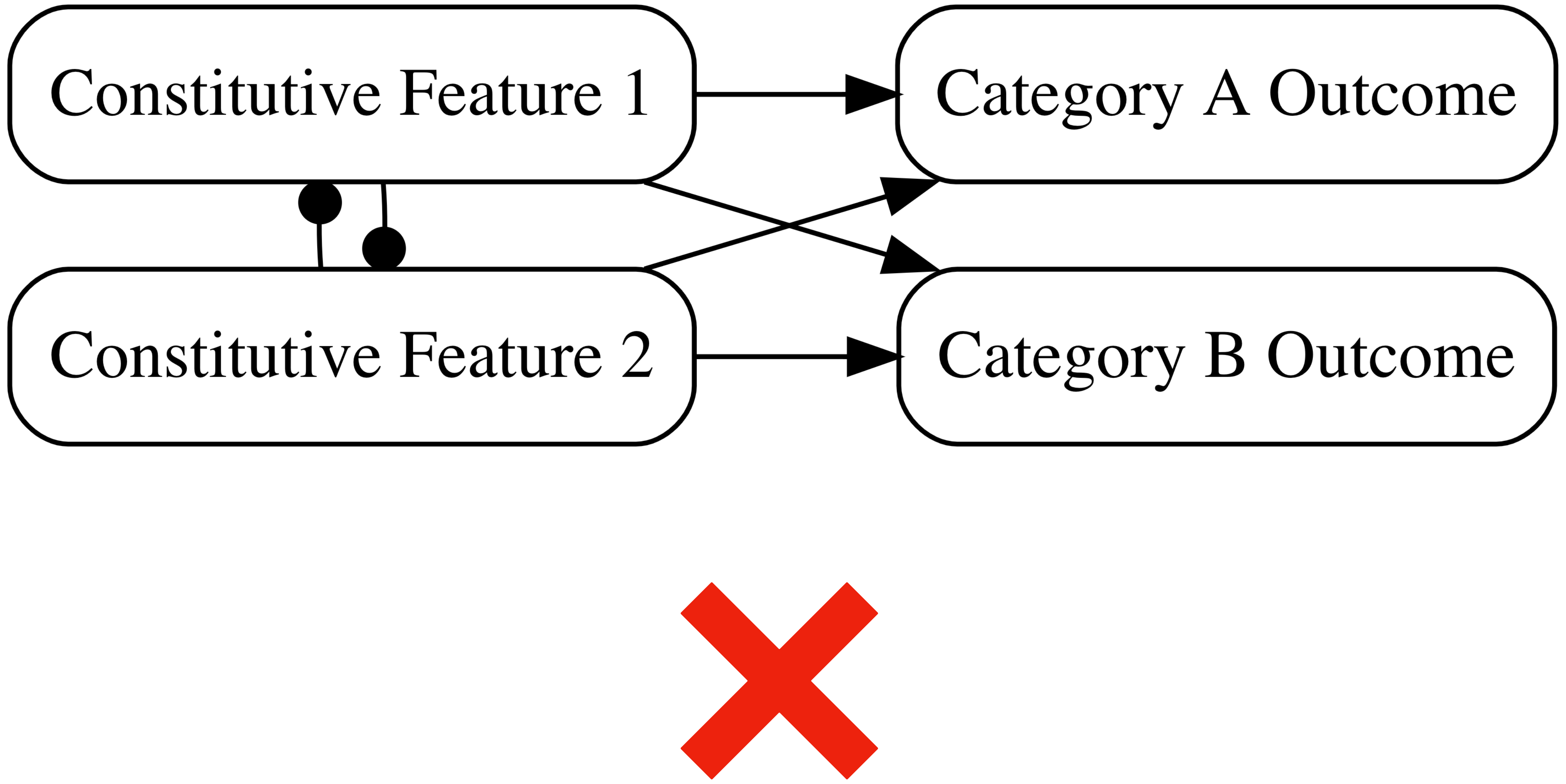
Disaggregation to the rescue!

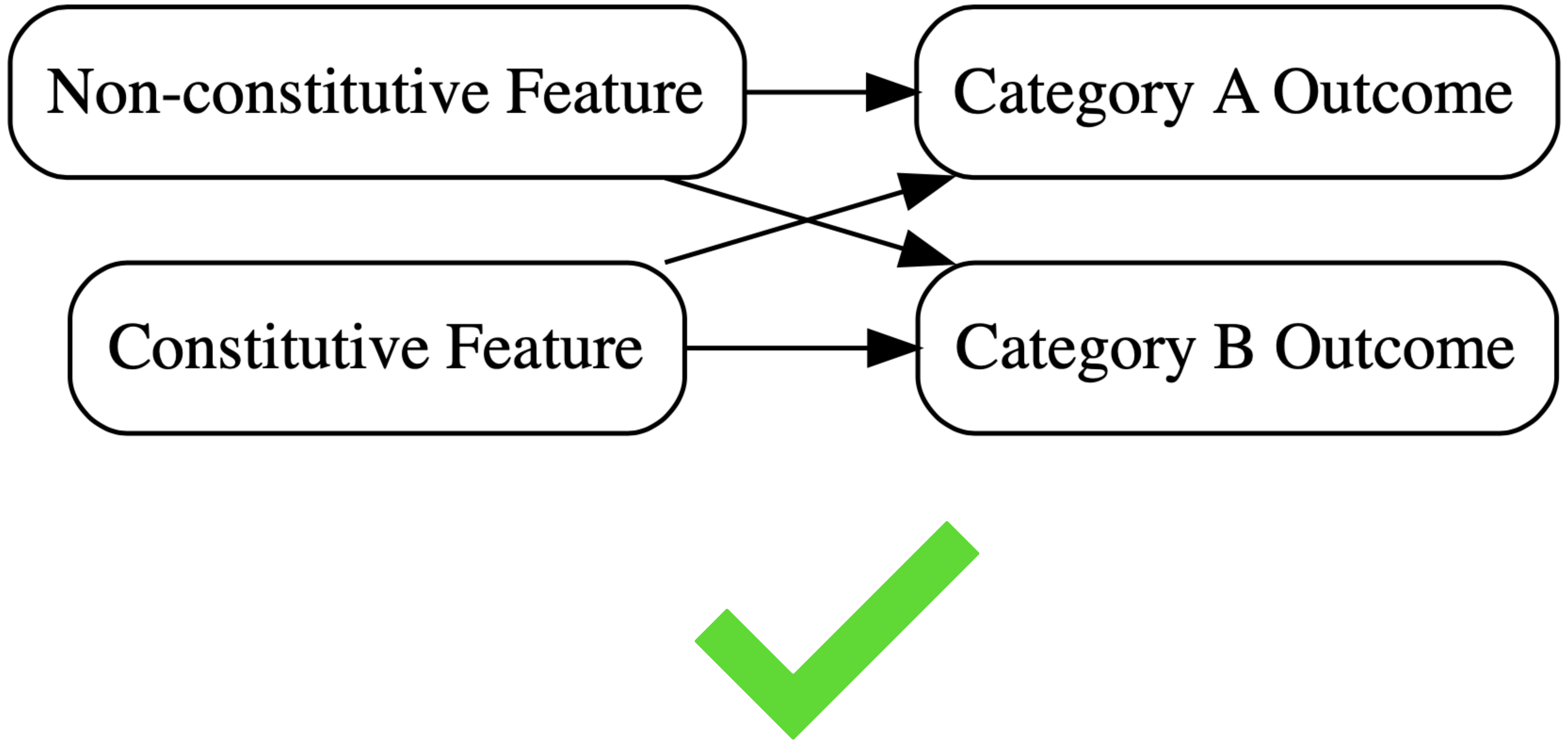
- **Approach:** *measure* a disparity across groups of people
  - Racial categories, genders, disabilities
- Our required assumption about social categories:
  - "A social category consists of a group of people" — no shared attributes necessary
- **Takeaway:** we don't need to resolve the philosophical debate to tackle pre-existing disparities



Social categories in impact remediation

Nuance: one constitutive feature at a time





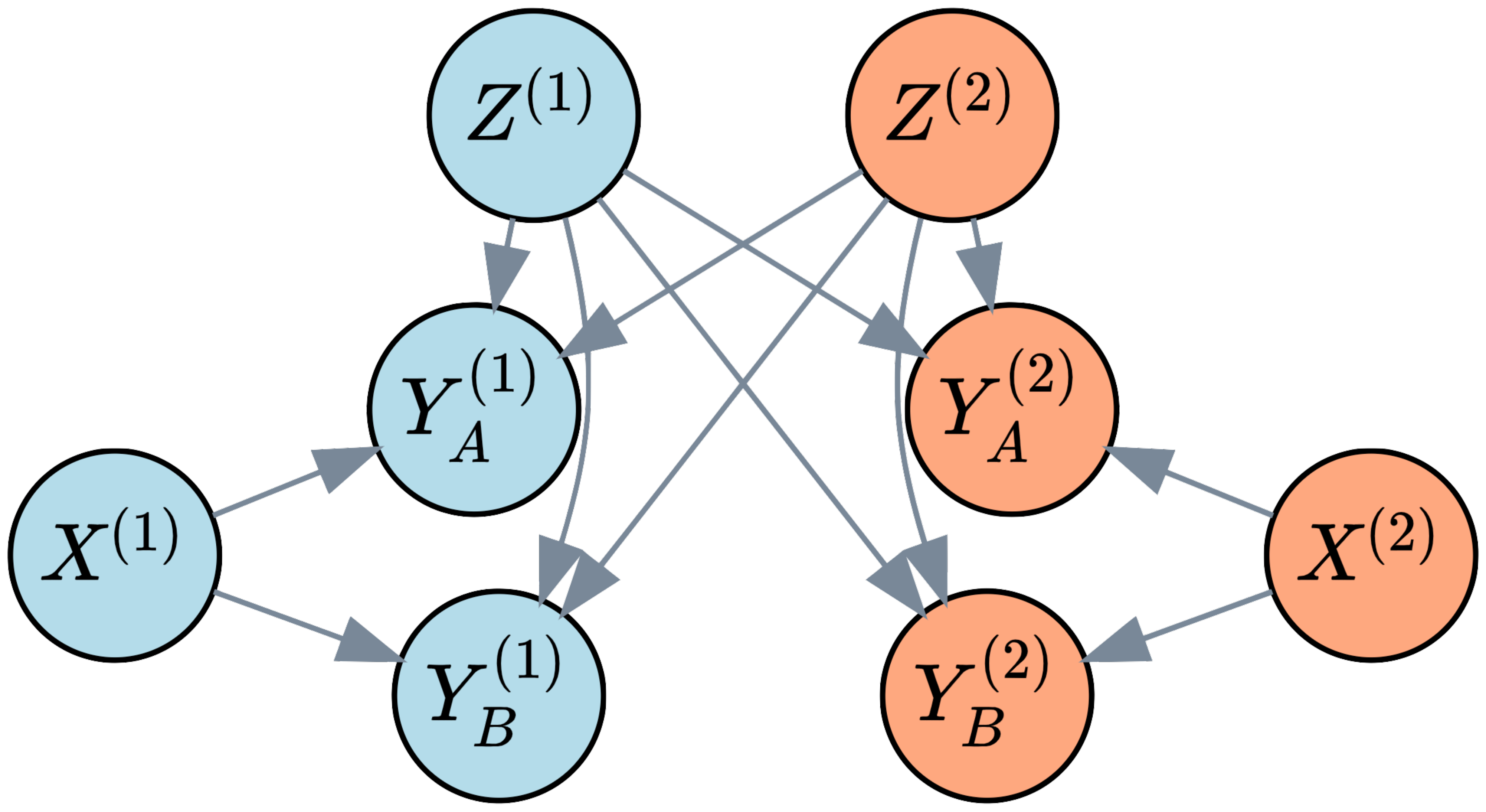
Framework formalization

Impact remediation — toy example

A multi-level, nested intervention structure

Framework	Example
n individuals	425 potential job applicants
m sub-populations on which we can intervene (“intervention sets”)	2 universities
r sub-populations across which we see disparity	Female (A) and Male (B) gender groups
outcome of interest Y, disaggregated across r groups	Y = fraction of students who applied for the job
real-world features X	X = number of career counselors
possible intervention Z	Z = whether or not we hosted a booth at the career fair

Causal graph relating X, Y, Z



extension of [Kusner et al. 2019] + disaggregation

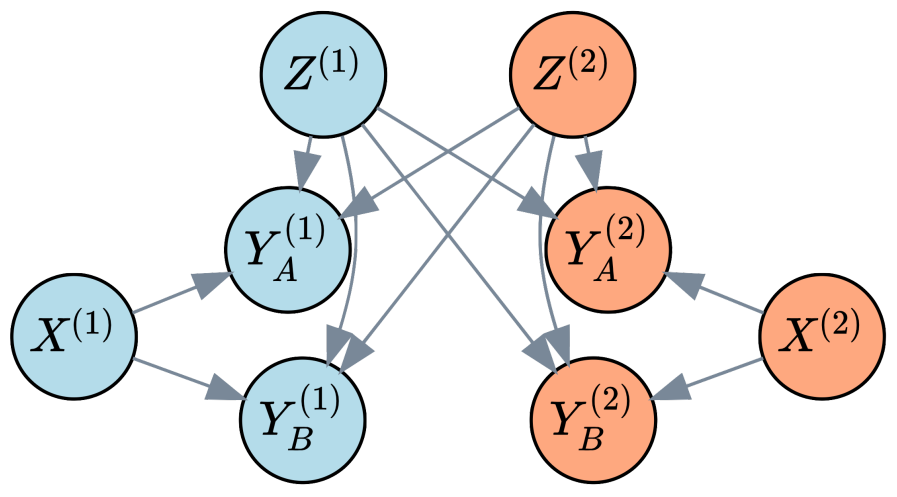
Finding optimal interventions to decrease disparity

Example: outreach in a job applicant pool

- Students in each gender group: $n_A^{(1)} = 100, \quad n_A^{(2)} = 75, \quad n_B^{(1)} = 150, \quad n_B^{(2)} = 100$
- Observed application rates: $(Y_A^{(1)}, Y_B^{(1)}) = (0.10, 0.20) \qquad (Y_A^{(2)}, Y_B^{(2)}) = (0.05, 0.10)$
- Measure of disparity:

$$\delta(z) = \left| \frac{1}{n_A} \sum_{i=1}^{2} n_A^{(i)} \mathbb{E}[Y_A^{(i)}(z)] - \frac{1}{n_B} \sum_{i=1}^{2} n_B^{(i)} \mathbb{E}[Y_B^{(i)}(z)] \right|$$

- Estimated application rates *after* intervention:



$$\mathbb{E}[Y_A^{(1)}([z^{(1)} = 1, z^{(2)} = 0])] = 0.20 \qquad \mathbb{E}[Y_A^{(1)}([z^{(1)} = 0, z^{(2)} = 1])] = 0.15$$
$$\mathbb{E}[Y_A^{(2)}([z^{(1)} = 1, z^{(2)} = 0])] = 0.10 \qquad \mathbb{E}[Y_A^{(2)}([z^{(1)} = 0, z^{(2)} = 1])] = 0.15$$
$$\mathbb{E}[Y_B^{(1)}([z^{(1)} = 1, z^{(2)} = 0])] = 0.30 \qquad \mathbb{E}[Y_B^{(1)}([z^{(1)} = 0, z^{(2)} = 1])] = 0.25$$
$$\mathbb{E}[Y_B^{(2)}([z^{(1)} = 1, z^{(2)} = 0])] = 0.15 \qquad \mathbb{E}[Y_B^{(2)}([z^{(1)} = 0, z^{(2)} = 1])] = 0.15$$

- Disparity after intervention:

$\delta \approx 0.08$ — no intervention

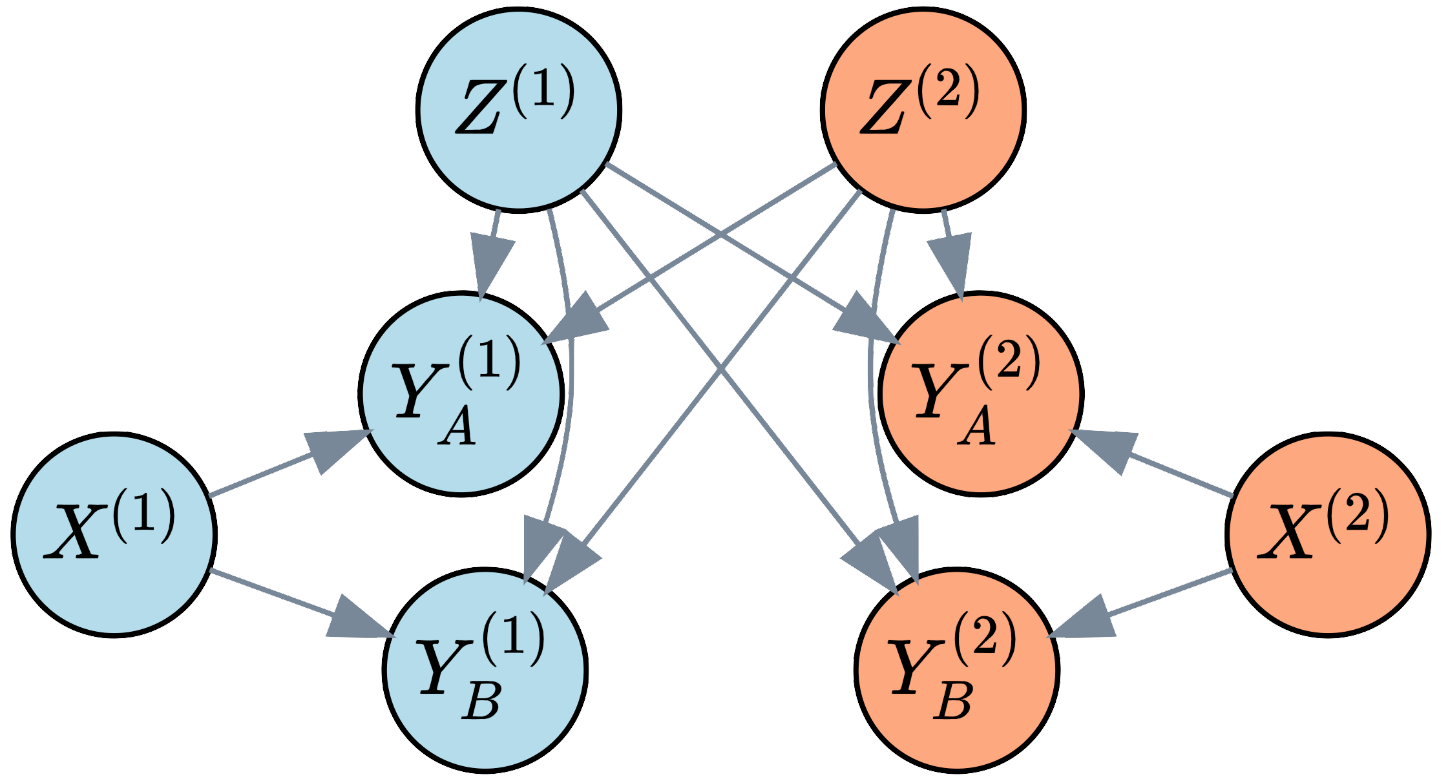
$\delta([z^{(1)} = 1, z^{(2)} = 0]) \approx 0.08$ — university one

$\boxed{\delta([z^{(1)} = 0, z^{(2)} = 1]) = 0.06}$ — university two

Impact remediation (IR) overview

Process overview:

1. Social categorization + data collection
2. Fit causal model to estimate intervention effects
3. Define our objective (how to mitigate disparity)
4. Find optimal interventions subject to constraints (budget, etc.)



$$\min_{z \in \{0,1\}^m} \left| \frac{1}{n_A} \sum_{i=1}^{2} n_A^{(i)} \mathbb{E}[Y_A^{(i)}(z)] - \frac{1}{n_B} \sum_{i=1}^{2} n_B^{(i)} \mathbb{E}[Y_B^{(i)}(z)] \right|$$

$$\text{s.t.} \quad \sum_{i=1}^{m} z^{(i)} \leq b$$

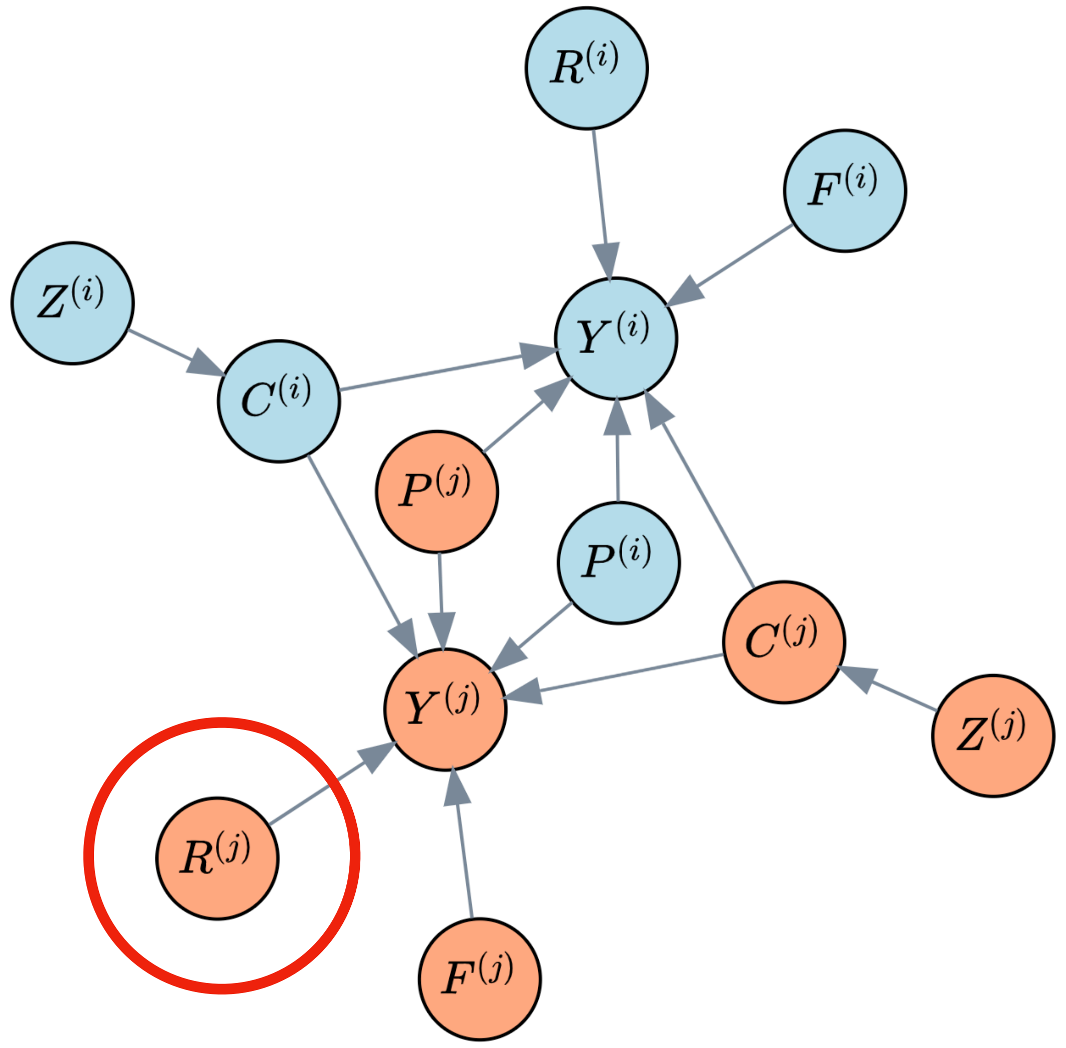
Case study

Stylized NYC schools example

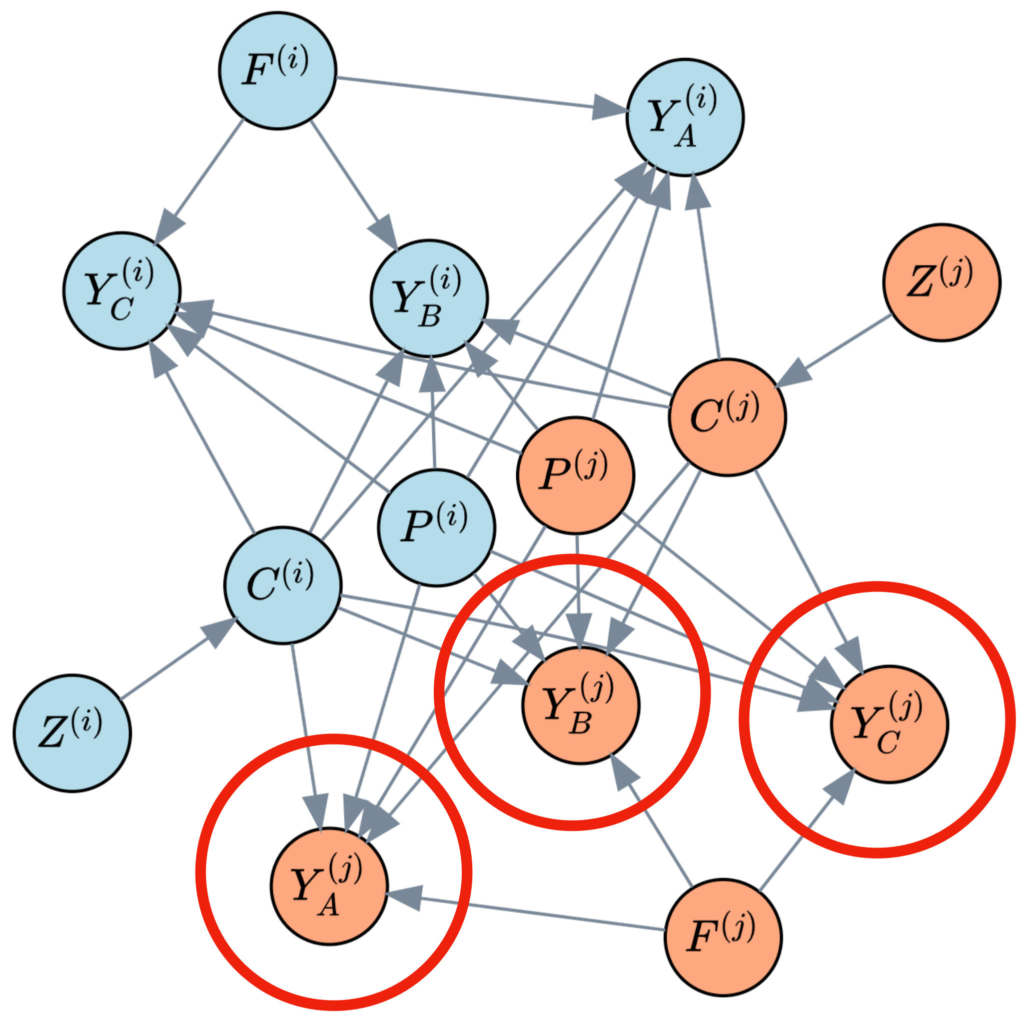
An IR case-study

- An example with realistic data:
  - **Setup:** The US DOE giving funding to NYC public schools to hire Calculus teachers
  - **Goal:** increase college attendance
  - **Subgroups:** racial and gender categories

Causal Fairness
[Kusner et al. 2019]



vs.

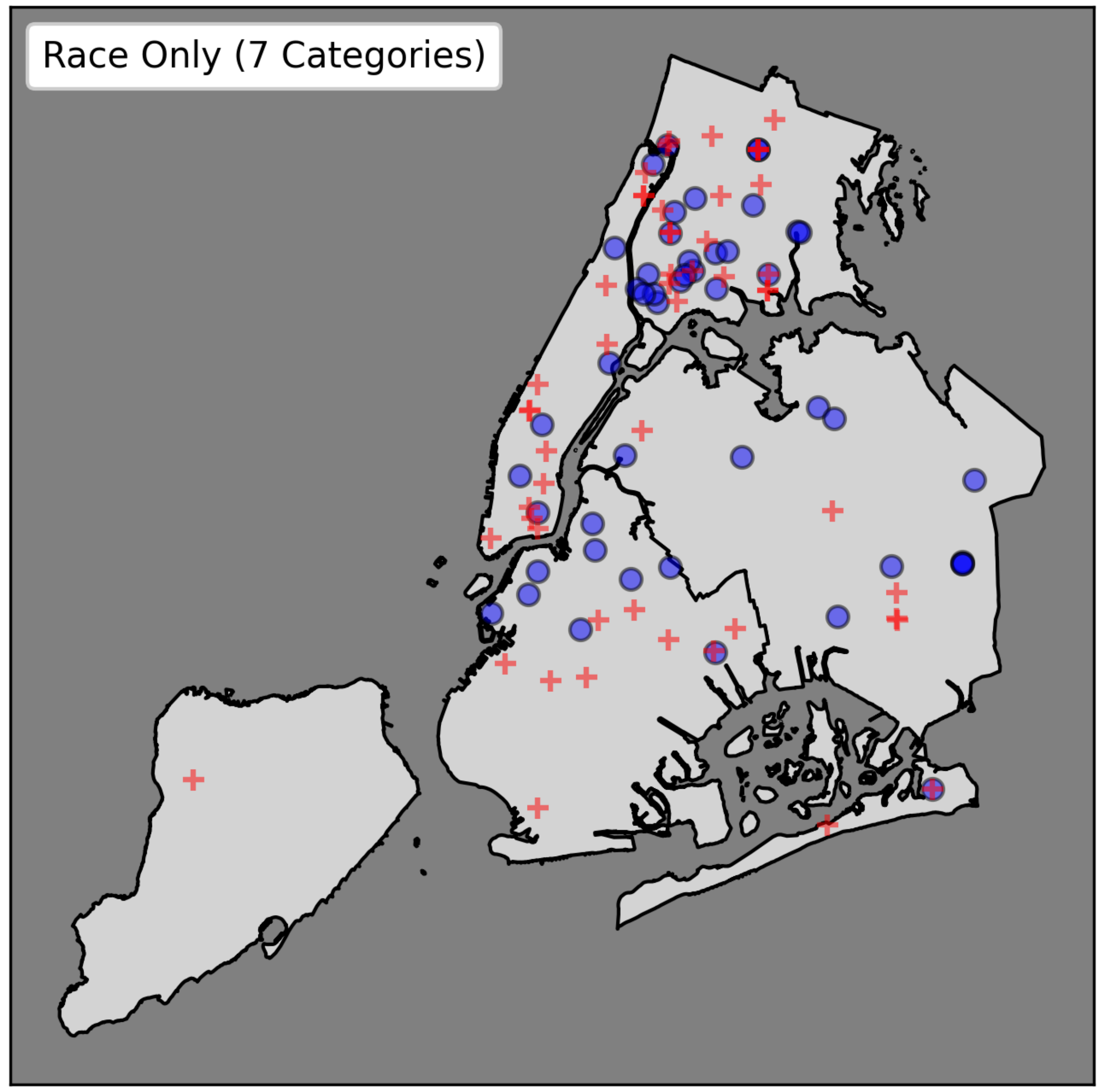


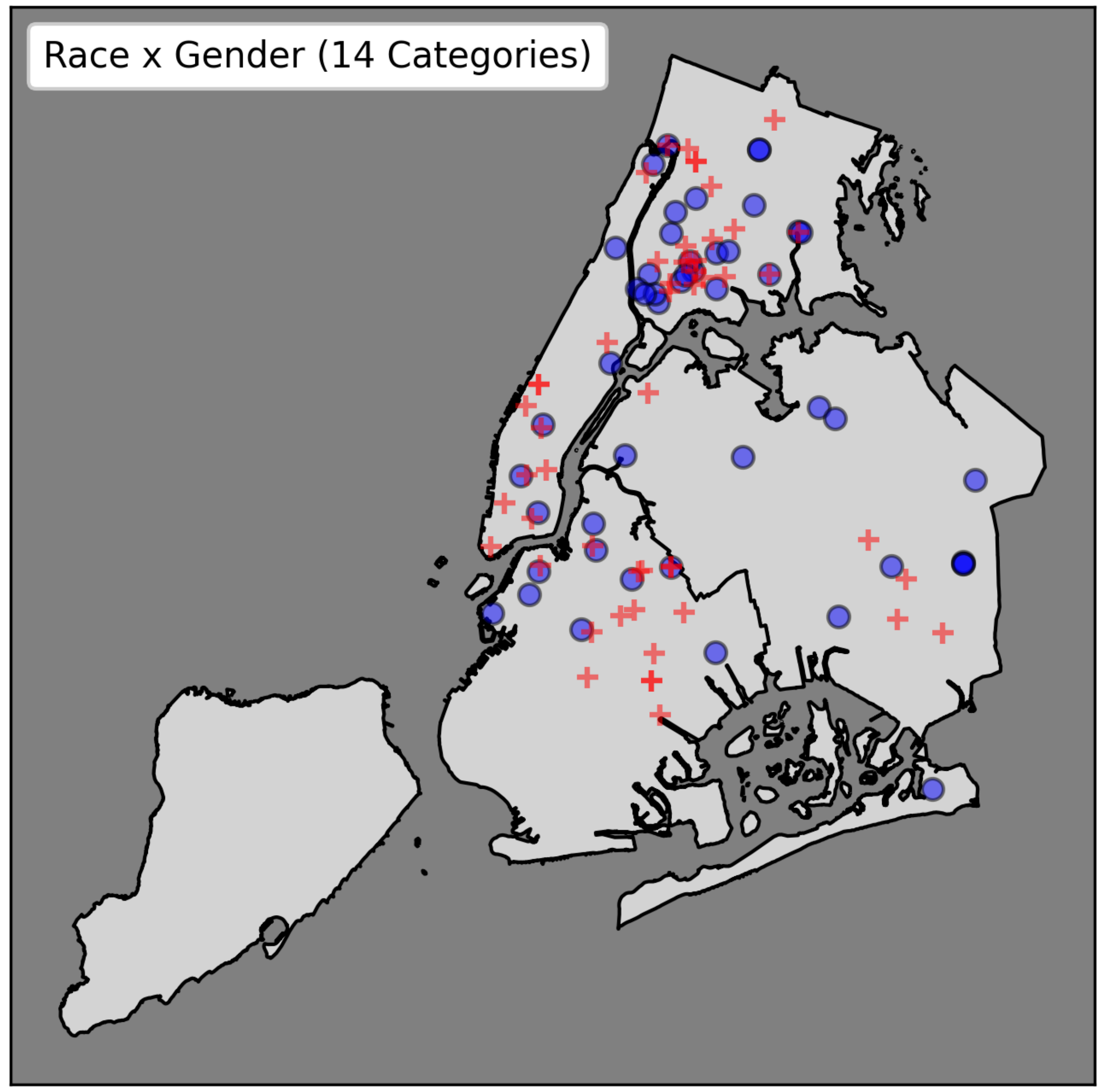
Focus on Inequality
(disaggregation)

Stylized NYC schools example

Takeaways from a case-study

- Social categorization (i.e., which partition) changes results
- Measuring subgroup outcomes better allows for focus on inequality
- With a focus on utility, inequality can increase
    - Even with strict fairness constraints
    - Even if group membership is known
    - Even if aggregate impact is larger





Approach	% Change in Impact Per-group							Aggregate % Impact	Disparity ($\delta(z)$)
	A	B	C	D	E	F	G		
No Intervention	±0.0	±0.0	±0.0	±0.0	±0.0	±0.0	±0.0	±0.0	1.429
IR	+1.76	+0.24	+0.42	+0.10	+0.16	+5.26	−1.35	+0.657	**1.386**
IR + ‘no harm’	**+1.78**	+0.54	+0.74	+0.11	+1.02	**+5.56**	±0.0	+0.848	1.394
DIP, $\tau = 0.567$	+1.20	+0.69	+1.46	**+0.63**	+1.61	+3.50	+0.43	+0.953	1.435
DIP, Unconstrained	+1.21	**+0.72**	**+1.48**	**+0.63**	**+1.64**	+3.51	**+0.47**	**+0.971**	1.435

Counterfactuals for the Future

Motivating toy example

Treatment choice in a fixed sample

- Individualized treatment choice focused on **entire distribution of outcomes**
- **Allocating tutoring to students**
  - We want to allocate tutoring to students at a school in order to improve test performance
    - One time-step of past data on every student of interest
    - No confounding
    - We have a model of the data generating process
  - *Unobserved* external factors about the students (e.g., family income, encouragement from parents) explain at least some of the variation in the outcome across units

Two approaches to the same problem

Which approach is better?

- **Approach 1:**
  - Use our model to identify which students to treat based on their covariate values only
- **Approach 2:**
  - Use our model to identify which students to treat based on their covariate values *and modeled exogenous variables (i.e., noise)*

Two approaches to the same problem

Which approach is better?

- **Approach 1:** Interventional
  - Use our model to identify which students to treat based on their covariate values only
- **Approach 2:** Counterfactual
  - Use our model to identify which students to treat based on their covariate values *and modeled exogenous variables (i.e., noise)*

Two approaches to the same problem

Which approach is better?

- **Approach 1:** Interventional
  - Use our model to identify which students to treat based on their covariate values only
- **Approach 2:** Counterfactual
  - Use our model to identify which students to treat based on their covariate values *and modeled exogenous variables (i.e., noise)*
- **Takeaway:** These two approaches can lead us to tutor a different set of students —> **two different policies**

Motivating toy example

Treatment choice in a fixed sample

- Individualized treatment choice focused on **entire distribution of outcomes**
- **Allocating tutoring to students**
  - We want to allocate tutoring to students at a school in order to improve test performance
    - One time-step of past data on every student of interest
    - No confounding
    - We have a model of the data generating process
  - *Unobserved* external factors about the students (e.g., family income, encouragement from parents) explain at least some of the variation in the outcome across units

Key question

What do we assume about family income and encouragement from parents year-to-year?

Interventional distributions

Notation

- Intuition:
  - "What will the distribution of test score Y be for a student described by SCM $\mathfrak{C}$ if they are enrolled in tutoring Z?"

the intervention do$(Z := 1)$

the SCM $\mathfrak{C}$ we start with

$$P_Y^{\mathfrak{C};do(Z:=1)}$$

the variable Y we are interested in

Image adapted from: [Peters et al. 2017]

Counterfactual distributions

Notation

- Intuition:
  - “What would the distribution of test score Y **have been** for a student described by SCM $\mathfrak{C}$ if they **had been** enrolled in tutoring Z?”

the intervention do$(Z := 1)$

the observed data $\mathbf{X} = \mathbf{x}$

the SCM $\mathfrak{C}$ we start with

$$P_Y^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(Z:=1)}$$

the variable Y we are interested in

Image recreated from: [Peters et al. 2017]

Counterfactuals

The retrospective view

- Counterfactuals are typically described as *retrospective*
  - We condition on *observed circumstances* before simulating an intervention
  - Use posterior $P_{U|\mathbf{X}=\mathbf{x}}$ instead of prior P_U to obtain $P_Y^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(\cdots)}$
- **Our work:** When can (or should) counterfactuals be forward-looking?

Why would we use past noise to make future decisions?

Assumptions about what we haven't observed

- Common for data with multiple time steps
  - "There are unobserved variables that play an important role in our model"
  - Large literature: time-series cross sectional data, mixed effects models, latent variable models, etc.
  - Can use repeated observations for estimation
- **Our setting: data with one time-step**
  - Noise decomposition is no longer an estimation problem
    - No repeated observations
  - Accounting for unobserved variables is instead *based on assumptions*

Forward-looking counterfactuals (FLCs)

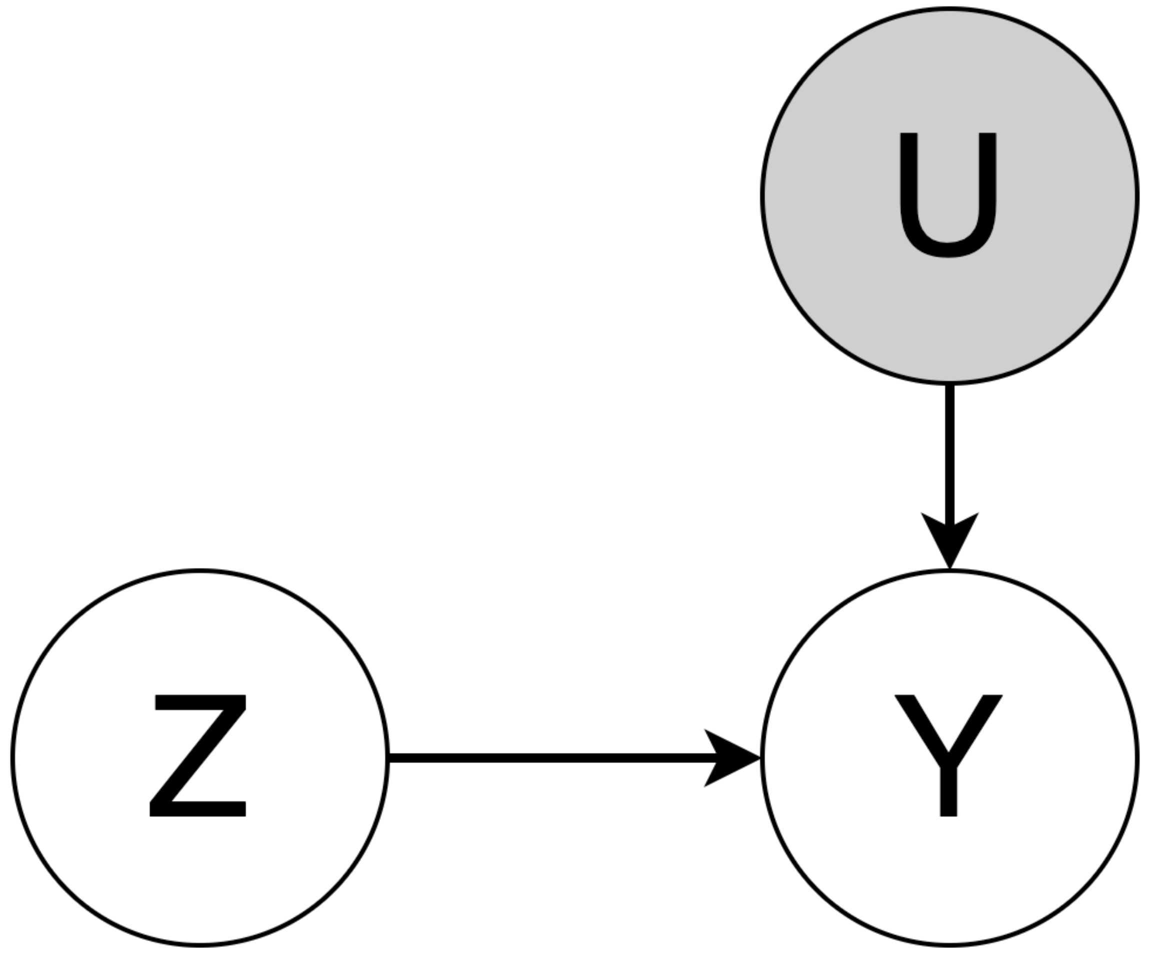
An alternate view

- The ‘retrospective’ view is connected to assumptions about the **structure** and **stability** of exogenous variables (noise)
  - Structure:
    - How does a unit look exogenously compared to other units?
  - Stability:
    - How does a unit look exogenously compared to itself over time?
- **Spoiler:** FLCs useful when units’ exogenous factors are (1) sufficiently stable over time OR (2) sufficiently dissimilar to other units

Exploring FLCs empirically

An illustrative parameterization

- Outcome Y, treatment Z, exogenous factors U, observed data $\{Z_0^{(i)}, Y_0^{(i)}\}_{i=1}^n$
- Intervention on unit i will increase Z by amount δ
- **Goal:** recover distribution P_{Y_1} after intervention on those for whom $Y_0 < 0$



$$(t=0): \begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases} \qquad (t=1): \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{(i)} \end{cases}$$

Treatment choice

Exploring FLCs empirically

Model for exogenous noise terms

$$(t=0):\begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases}$$

$$(t=1):\begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{(i)} \end{cases}$$

$$\mu_U^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2)$$

$$U_0^{(i)}, U_1^{(i)} \overset{iid}{\sim} \mathcal{N}(\mu_U^{(i)}, \sigma_U^2)$$

Exploring FLCs empirically

Model for exogenous noise terms

$$(t=0): \begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases}$$

$$(t=1): \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{(i)} \end{cases}$$

We can now explore structure (σ_μ) and stability (σ_U)

$$\boxed{\begin{array}{c} \mu_U^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2) \\ U_0^{(i)}, U_1^{(i)} \overset{iid}{\sim} \mathcal{N}(\mu_U^{(i)}, \sigma_U^2) \end{array}}$$

Parameterizing exogenous structure and stability

Connecting assumptions to parameters

Assumption	Model	Interpretation
(A1) Exogenous factors are constant over time.	$\sigma_U = 0$	Among the relevant variables we haven't measured, each unit looks exactly the same next year as it does this year.
(A2) Exogenous factors vary over time.	$\sigma_U > 0$	Among the relevant variables we haven't measured, each unit looks somewhat the same next year as it does this year. Similarities grow weaker with larger σ_U values.
(A3) Exogenous factors exhibit unstructured variation.	$\sigma_\mu = 0$	Among the relevant variables we haven't measured, each unit looks the same as any other unit, apart from random variability with time.
(A4) Exogenous factors exhibit structured (unit-specific) variation.	$\sigma_\mu > 0$	Among the relevant variables we haven't measured, there are units that look unlike other units, in addition to random variability with time. Units look less like each other with larger σ_μ.

$$\mu_U^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2)$$

$$U_0^{(i)}, U_1^{(i)} \overset{iid}{\sim} \mathcal{N}(\mu_U^{(i)}, \sigma_U^2)$$

Counterfactual vs. interventional distributions

What happens with a 'correct' model?

$$\mu_U^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2)$$

$$\text{Truth } (t=0): \begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ U_0^{(i)} \sim \mathcal{N}(\mu_U^{(i)}, \sigma_U^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases}$$

$$\text{Truth } (t=1): \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ U_1^{(i)} \sim \mathcal{N}(\mu_U^{(i)}, \sigma_U^2) \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{(i)} \end{cases}$$

$$\text{Model } (t=0) = \begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ U_0^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2 + \sigma_U^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases}$$

$$\text{Interventional } (t=1) = \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ U_1^{\prime(i)} \sim \mathcal{N}(0, \sigma_\mu^2 + \sigma_U^2) \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{\prime(i)} \end{cases}$$

$$\text{Counterfactual } (t=1) = \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ \tilde{U}_1^{(i)} = U_0^{(i)} \\ Y_1^{(i)} = Z_1^{(i)} + \tilde{U}_1^{(i)} \end{cases}$$

Counterfactual vs. interventional distributions

What happens with a 'correct' model?

$$\mu_U^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2)$$

$$\text{Truth } (t=0): \begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ U_0^{(i)} \sim \mathcal{N}(\mu_U^{(i)}, \sigma_U^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases}$$

$$\text{Truth } (t=1): \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ \boxed{U_1^{(i)} \sim \mathcal{N}(\mu_U^{(i)}, \sigma_U^2)} \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{(i)} \end{cases}$$

p_{Y_1}

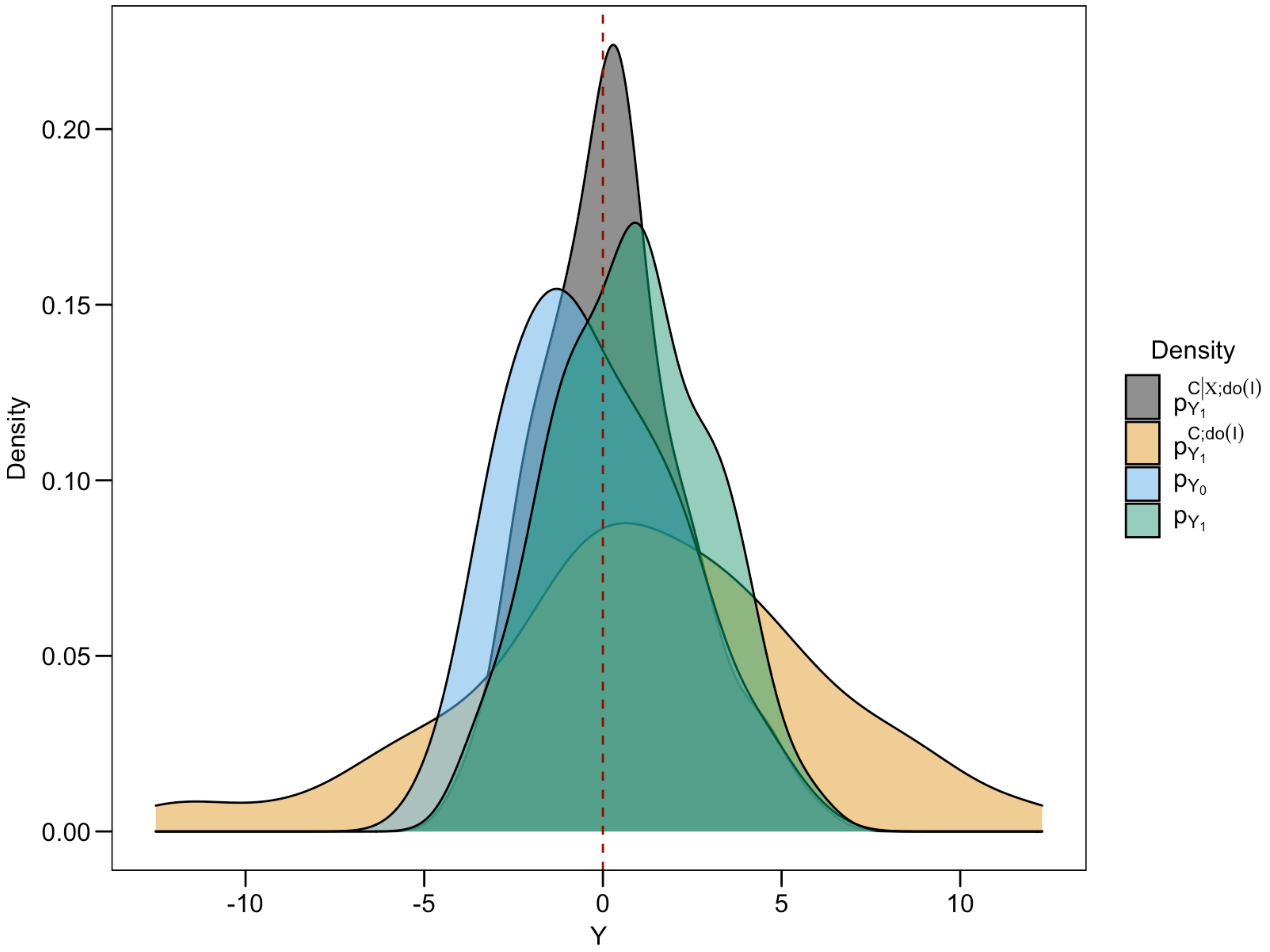
$$\text{Model } (t=0) = \begin{cases} Z_0^{(i)} \sim \mathcal{N}(\mu_Z, \sigma_Z^2) \\ U_0^{(i)} \sim \mathcal{N}(0, \sigma_\mu^2 + \sigma_U^2) \\ Y_0^{(i)} = Z_0^{(i)} + U_0^{(i)} \end{cases}$$

$$\text{Interventional } (t=1) = \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ \boxed{U_1^{\prime(i)} \sim \mathcal{N}(0, \sigma_\mu^2 + \sigma_U^2)} \\ Y_1^{(i)} = Z_1^{(i)} + U_1^{\prime(i)} \end{cases}$$

$p_{Y_1}^{\mathfrak{C};do(\cdots)}$

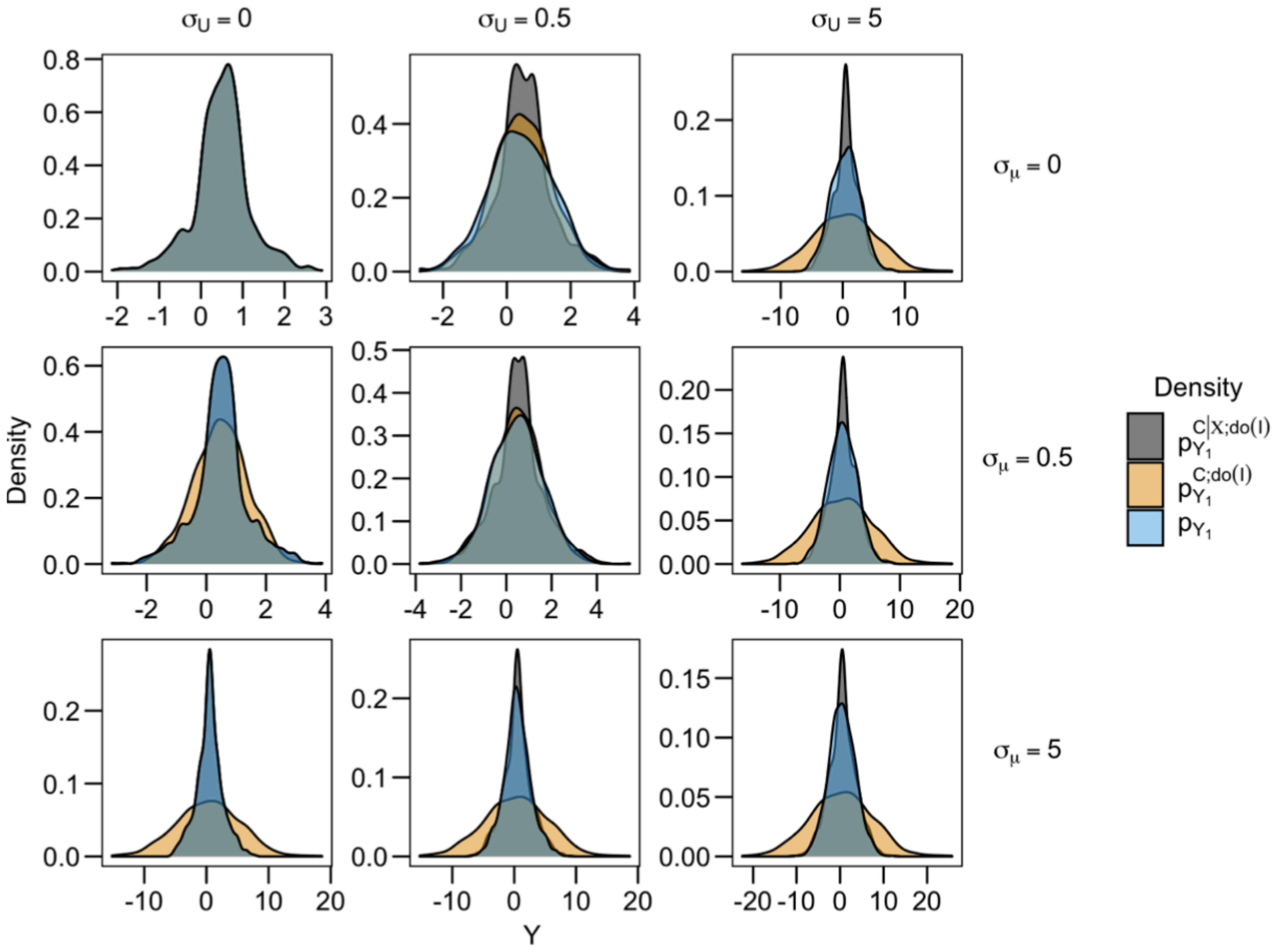
$$\text{Counterfactual } (t=1) = \begin{cases} Z_1^{(i)} = Z_0^{(i)} + \delta \cdot w(i) \\ \boxed{\tilde{U}_1^{(i)} = U_0^{(i)}} \\ Y_1^{(i)} = Z_1^{(i)} + \tilde{U}_1^{(i)} \end{cases}$$

$p_{Y_1}^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(\cdots)}$

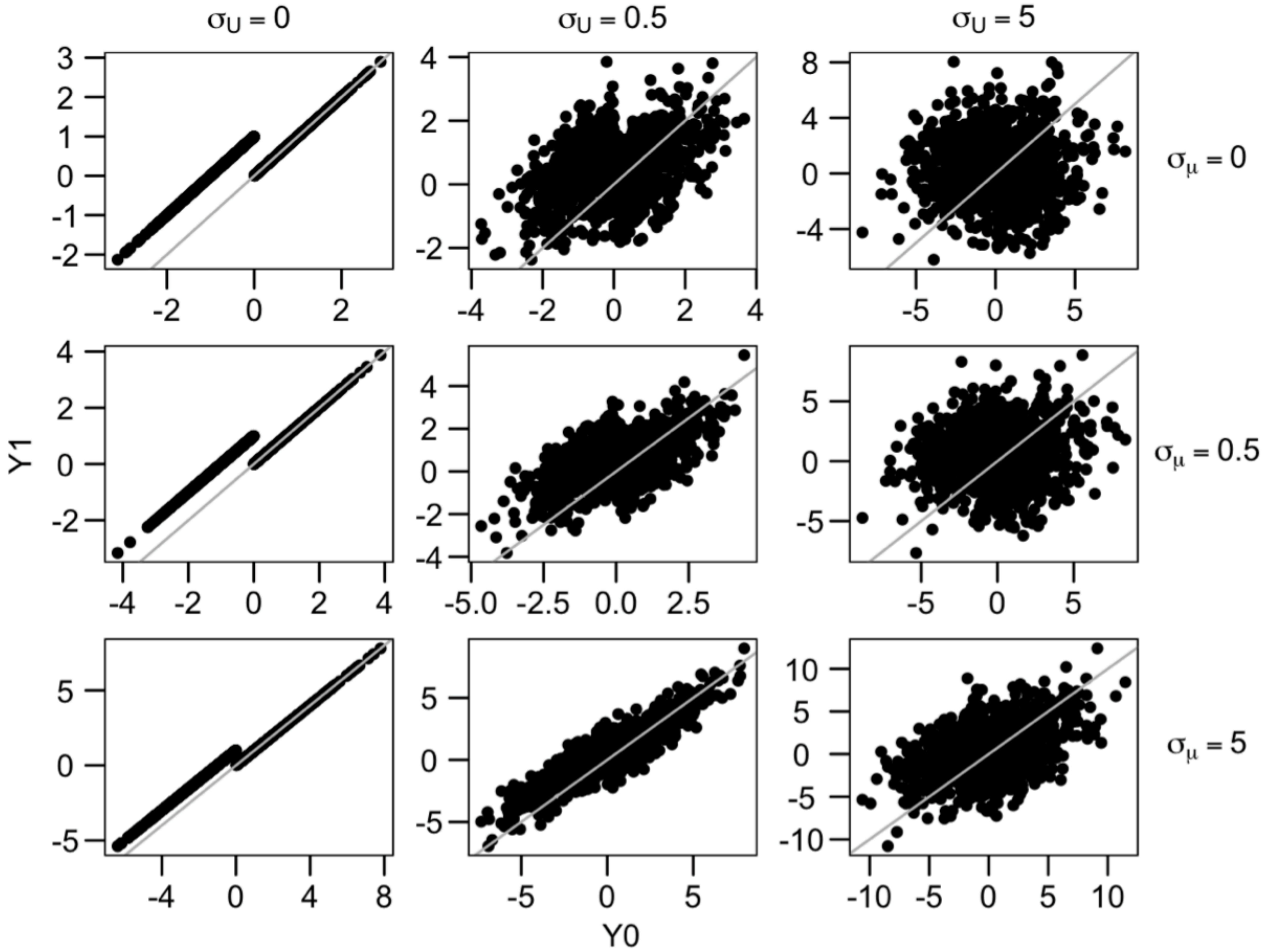

Density
0.20
0.15
0.10
0.05
0.00
-10
-5
0
5
10
Y
Density
$p_{Y_1}^{C|X;do(I)}$
$p_{Y_1}^{C;do(I)}$
p_{Y_0}
p_{Y_1}

Takeaways

Unit-specific structure OR stability over time —> FLCs



(a)

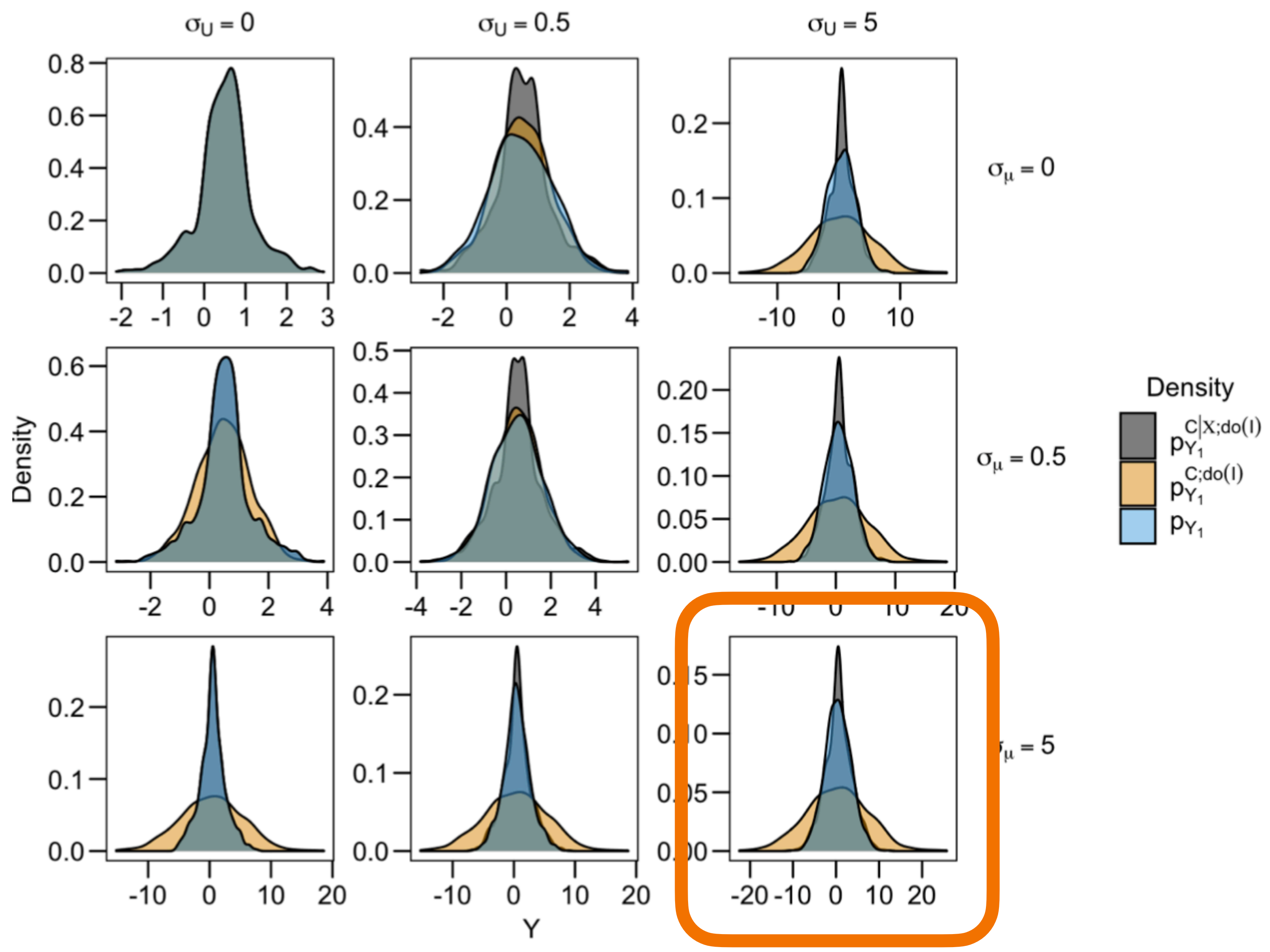


(b)

Why do we care?

Back to motivation

- We often can't measure every relevant variable
- We might not be able to collect lots of data over time
- Our assumptions can lead to **different policies** and **incorrect conclusions**



What if we want to decrease variance?

$\mathbb{V}[P_{Y_0}]$	$\mathbb{V}[P_{Y_1}]$	$\mathbb{V}[P_{Y_1}^{\mathfrak{C}\vert\mathcal{X};\mathrm{do}(I)}]$	$\mathbb{V}[P_{Y_1}^{\mathfrak{C};\mathrm{do}(I)}]$
12.2	9.36	9.61	51.1

References 1

[1] Rediet Abebe, Jon Kleinberg, and S. Matthew Weinberg. 2020. Subsidy Allocations in the Presence of Income Shocks. Proceedings of the AAAI Conference on Artificial Intelligence 34, 05 (Apr. 2020), 7032–7039. https://doi.org/10.1609/aaai.v34i05.6188

[2] Peter M Aronow, Cyrus Samii, et al. 2017. Estimating average causal effects under general interference, with application to a social network experiment. The Annals of Applied Statistics 11, 4 (2017), 1912–1947.

[3] Susan Athey and Guido W Imbens. 2019. Machine learning methods that economists should know about. Annual Review of Economics 11 (2019), 685– 725.

[4] Sebastian Benthall and Bruce D. Haynes. 2019. Racial Categories in Machine Learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 289–298. https://doi.org/10.1145/3287560.3287575

[5] Silvia Chiappa and Thomas P. S. Gillam. 2018. Path-Specific Counterfactual Fairness. arXiv:1802.08139 [stat.ML]

[6] Cristobal de Brey, Lauren Musu, Joel McFarland, Sidney Wilkinson-Flicker, Melissa Diliberti, Anlan Zhang, Claire Branstetter, and Xiaolei Wang. 2019. Status and Trends in the Education of Racial and Ethnic Groups 2018. NCES 2019-038.

[7] Nick Doudchenko, Minzhengxiong Zhang, Evgeni Drynkin, Edoardo Airoldi, Vahab Mirrokni, and Jean Pouget-Abadie. 2020. Causal Inference with Bipartite Designs. arXiv:2010.02108 [stat.ME]

[8] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An Intersectional Definition of Fairness. In 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, Piscataway, New Jersey, 1918–1921. https: //doi.org/10.1109/ICDE48307.2020.00203

[9] Batya Friedman and H. Nissenbaum. 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14 (1996), 330–347.

[10] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-theLoop Analysis of Fairness in Risk Assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 90–99. https://doi. org/10.1145/3287560.3287563

[15] Lily Hu and Yiling Chen. 2020. Fair Classification and Social Welfare. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 535–545. https://doi.org/10.1145/3351095.3372857

[16] Lily Hu and Issa Kohler-Hausmann. 2020. What's Sex Got to Do with Machine Learning?. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 513. https://doi.org/10.1145/3351095.3375674

[17] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. Association for Computing Machinery, New York, NY, USA, 375–385. https: //doi.org/10.1145/3442188.3445901

[18] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 240–248. https://doi.org/10.1145/3287560.3287578

[19] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. arXiv:2006.06831 [cs.LG]

[20] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 228–236. https://doi.org/10.1145/3442188.3445886

[21] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 576–586. https: //doi.org/10.1145/3442188.3445919

[22] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., San Diego, California. https://proceedings. neurips.cc/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf

References 1 cont.

[23] Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. 2019. Making Decisions that Reduce Discriminatory Impacts. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97), Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, San Diego, California, 3591–3600. http://proceedings.mlr.press/v97/kusner19a.html

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., San Diego, California. https://proceedings. neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[25] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80), Jennifer Dy and Andreas Krause (Eds.). PMLR, San Diego, California, 3150–3158. https://proceedings.mlr.press/v80/liu18c.html

[26] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., San Diego, California. https://proceedings.neurips.cc/paper/2018/file/ 09d37c08f7b129e96277388757530c72-Paper.pdf

[27] Vishwali Mhasawade and Rumi Chunara. 2021. Causal Multi-Level Fairness. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Virtual Event, USA) (AIES '21). Association for Computing Machinery, New York, NY, USA, 784–794. https://doi.org/10.1145/3461702.3462587

[28] Sendhil Mullainathan. 2018. Algorithmic Fairness and the Social Welfare Function. In Proceedings of the 2018 ACM Conference on Economics and Computation (Ithaca, NY, USA) (EC '18). Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/3219166.3219236

[29] Razieh Nabi, Daniel Malinsky, and I. Shpitser. 2019. Learning Optimal Fair Policies. Proceedings of machine learning research 97 (2019), 4674–4682.

[30] Elizabeth L. Ogburn and T. VanderWeele. 2014. Causal diagrams for interference. Quality Engineering 60 (2014), 381–384.

[31] J. Pearl. 2000. Causality: Models, Reasoning and Inference.

[32] J. Pearl, Madelyn Glymour, and N. Jewell. 2016. Causal Inference in Statistics: A Primer.

[33] M. Sen and Omar Wasow. 2016. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. Annual Review of Political Science 19 (2016), 499–522.

[34] M. Sobel. 2006. What Do Randomized Studies of Housing Mobility Demonstrate? J. Amer. Statist. Assoc. 101 (2006), 1398 – 1407.

[35] Julia Stoyanovich, Bill Howe, and HV Jagadish. 2020. Responsible data management. Proceedings of the VLDB Endowment 13, 12 (2020), 3474–3488. Publisher: VLDB Endowment.

[11] LLC Gurobi Optimization. 2021. Gurobi Optimizer Reference Manual. http://www.gurobi.com.

[12] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 501–512. https: //doi.org/10.1145/3351095.3372826

[13] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. 2019. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. arXiv:1806.04959 [cs.AI]

[14] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/3287560.3287584

[36] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, California, 6035–6042. https://doi.org/10. 24963/ijcai.2019/836

[37] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2021. Causal Intersectionality and Fair Ranking. In 2nd Symposium on Foundations of Responsible Computing (FORC 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 192), Katrina Ligett and Swati Gupta (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 7:1–7:20. https://doi.org/10.4230/LIPIcs.FORC. 2021.7

[38] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in Decision-Making — The Causal Explanation Formula. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949

[39] Corwin M. Zigler and Georgia Papadogeorgou. 2021. Bipartite Causal Inference with Interference. Statist. Sci. 36, 1 (2021), 109 – 123. https://doi.org/10.1214/19STS749

References 2

Athey, S.; and Imbens, G. W. 2017. The State of Applied Econometrics: Causality and Policy Evaluation. Journal of Economic Perspectives, 31(2): 3–32.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. Fairness and Machine Learning. http://www.fairmlbook.org. Ac- cessed: 2022-12-07.

Beygelzimer, A.; Kakadet, S.; Langford, J.; Arya, S.; Mount, D.; and Li, S. 2022. FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1.3.1.

Buesing, L.; Weber, T.; Zwols, Y.; Racanie`re, S.; Guez, A.; Lespiau, J.-B.; and Heess, N. M. O. 2019. Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. ArXiv, abs/1811.06272.

Bynum, L. E.; Khan, F. A.; Konopatska, O.; Loftus, J. R.; and Stoyanovich, J. 2022. An Interactive Introduction to Causal Inference. https://lbynum.github.io/interactive-causal-inference. Accessed: 2022-12-07.

Bynum, L. E.; Loftus, J. R.; and Stoyanovich, J. 2021. Dis- aggregated Interventions to Reduce Inequality. Equity and Access in Algorithms, Mechanisms, and Optimization.

Friedman, B.; and Nissenbaum, H. 1996. Bias in Computer Systems. ACM Transactions on Information Systems, 14: 330–347.

Green, B.; and Chen, Y. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk As- sessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19, 90–99. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Heidari, H.; Ferrari, C.; Gummadi, K. P.; and Krause, A. 2019a. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. arXiv:1806.04959.

Heidari, H.; Loi, M.; Gummadi, K. P.; and Krause, A. 2019b. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In Proceedings of the Conference on Fairness, Accountabil- ity, and Transparency, FAT* ’19, 181–190. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Hu, L.; and Chen, Y. 2020. Fair Classification and Social Welfare. In Proceedings of the 2020 Conference on Fair- ness, Accountability, and Transparency, FAT* ’20, 535–545. New York, NY, USA: Association for Computing Machin- ery. ISBN 9781450369367.

Kannan, S.; Roth, A.; and Ziani, J. 2019. Downstream Ef- fects of Affirmative Action. In Proceedings of the Confer- ence on Fairness, Accountability, and Transparency, FAT* ’19, 240–248. New York, NY, USA: Association for Com- puting Machinery. ISBN 9781450361255.

Kasy, M. 2016. Partial Identification, Distributional Pref- erences, and the Welfare Ranking of Policies. Review of Economics and Statistics, 98: 111–131.

Kasy, M.; and Abebe, R. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

References 2 cont.

Khan, F. A.; Manis, E.; and Stoyanovich, J. 2021. Fairness as Equality of Opportunity: Normative Guidance from Political Philosophy. CoRR, abs/2106.08259.

Kitagawa, T.; and Tetenov, A. 2018. Who should be Treated? Empirical Welfare Maximization Methods for Treatment Choice. Econometrica, 86 2: 591–616.

Kitagawa, T.; and Tetenov, A. 2019. Equality-Minded Treat- ment Choice. Journal of Business & Economic Statistics, 39: 561 – 574.

Kusner, M.; Russell, C.; Loftus, J.; and Silva, R. 2019. Making Decisions that Reduce Discriminatory Impacts. In Chaudhuri, K.; and Salakhutdinov, R., eds., Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, 3591–3600. San Diego, California: PMLR.

Lei, L.; and Cande`s, E. J. 2021. Conformal Inference of Counterfactuals and Individual Treatment Effects. Journal of the Royal Statistical Society: Series B (Statistical Method- ology).

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed Impact of Fair Machine Learning. In Dy, J.; and Krause, A., eds., Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceed- ings of Machine Learning Research, 3150–3158. PMLR.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict Re- sponsibly: Improving Fairness and Accuracy by Learning to Defer. In Bengio, S.; Wallach, H.; Larochelle, H.; Grau- man, K.; Cesa-Bianchi, N.; and Garnett, R., eds., Advances in Neural Information Processing Systems, volume 31. Cur- ran Associates, Inc.

Manski, C. F. 2003. Statistical Treatment Rules for Hetero- geneous Populations. Econometrica, 72: 1221–1246.

Mullainathan, S. 2018. Algorithmic Fairness and the Social Welfare Function. In Proceedings of the 2018 ACM Confer- ence on Economics and Computation, EC ’18, 1. New York, NY, USA: Association for Computing Machinery. ISBN 9781450358293.

Nabi, R.; Malinsky, D.; and Shpitser, I. 2019. Learning Op- timal Fair Policies. Proceedings of machine learning re- search, 97: 4674–4682.

Oberst, M.; and Sontag, D. A. 2019. Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models. ArXiv, abs/1905.05824.

Pearl, J. 2009. Causality. Cambridge University Press.

Peters, J.; Janzing, D.; and Scho¨lkopf, B. 2017. Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. Causation, Prediction, and Search. MIT Press.

Thank you! Questions?