Counterfactual Data Generation using VAEs

Ayan Majumdar, Preethi Lahoti, Junaid Ali, Till Speicher,
Isabel Valera, Krishna Gummadi

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

max planck institut
informatik

UNIVERSITÄT
DES
SAARLANDES

Fairness and causation

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

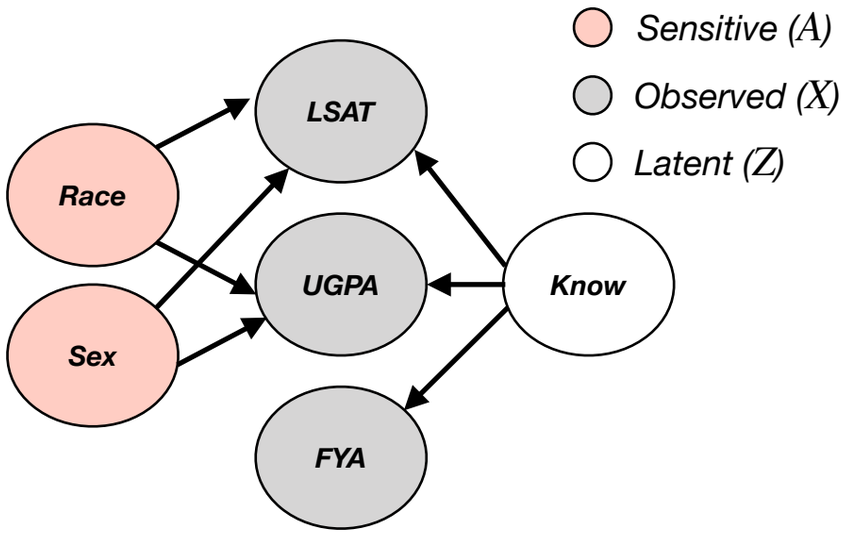
- Had Jacob been white instead, would he had been accepted?
  - *counterfactual*
- Did Jacob's race **cause** him to get negative outcome?
  - *counterfactual fairness* (*Kusner et al. 2017*)[1]

Such questions of fairness need counterfactual data!

How do we generate them?

[1]*Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Counterfactuals



Causal graph

+

LSAT $\sim \mathcal{N}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S), \sigma_L)$

UGPA $\sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G)$

FYA $\sim \mathcal{N}(w_F^K K, 1)$

Know $\sim \mathcal{N}(0,1)$

Structural equations

Generating counterfactual *(Pearl et al. 2009)*[2] —

1. Abduction: Given observed X, $A = a$ **estimate** Z
2. Action: **Intervene** on A by setting it to a'
3. Prediction: **Re-compute** X using Z under intervention $do(A = a')$

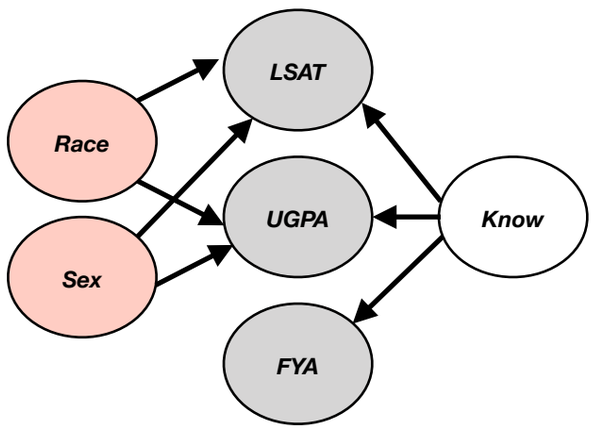
Need complete access to causal model!

Infeasible in real settings.

Can we generate counterfactual data for fairness in the absence of the whole causal model?

[2]*Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

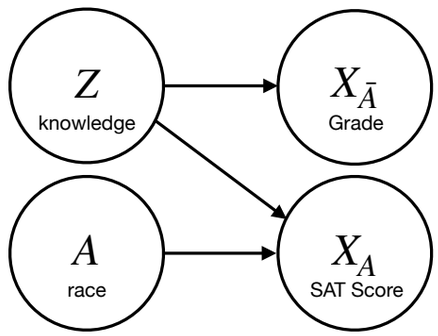
Approach



Causal graph

Assumptions

Simplify



Assumed causal structure

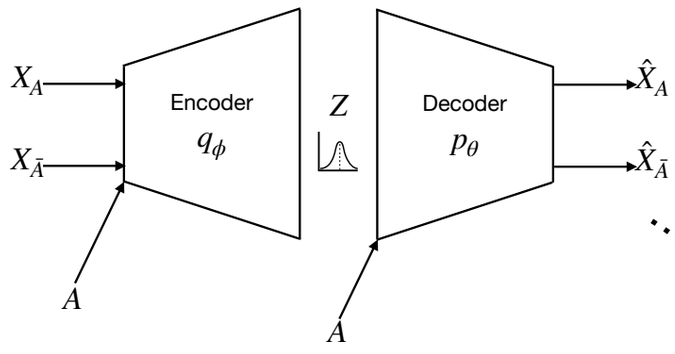
Training

LSAT $\sim \ldots$ **UGPA** $\sim \ldots$

FYA $\sim \ldots$ **Know** $\sim \ldots$

Structural equations

Learn from data



CVAE architecture

Generating counterfactual

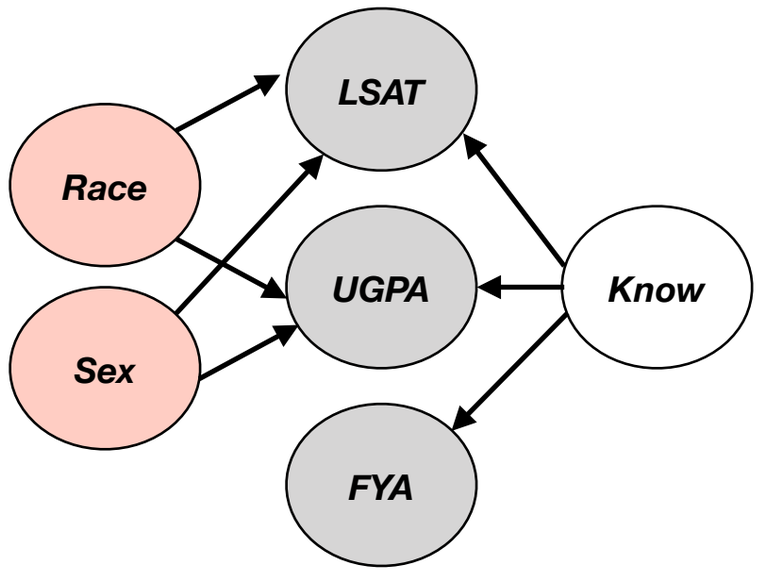
1. Abduction: …
2. Action: …
3. Prediction: …

Compute

1. Infer Z from $X, A = a$
2. Set $A = a'$ at decoder
3. Generate X from Z and $A = a'$

Can CVAE generate counterfactuals?

- **Train** CVAE on synthetic **generated data.**
- Condition on A (**race, sex**).
- **Metric:** Mean absolute error (MAE), cosine similarity b/w CVAE & causal counterfactuals.



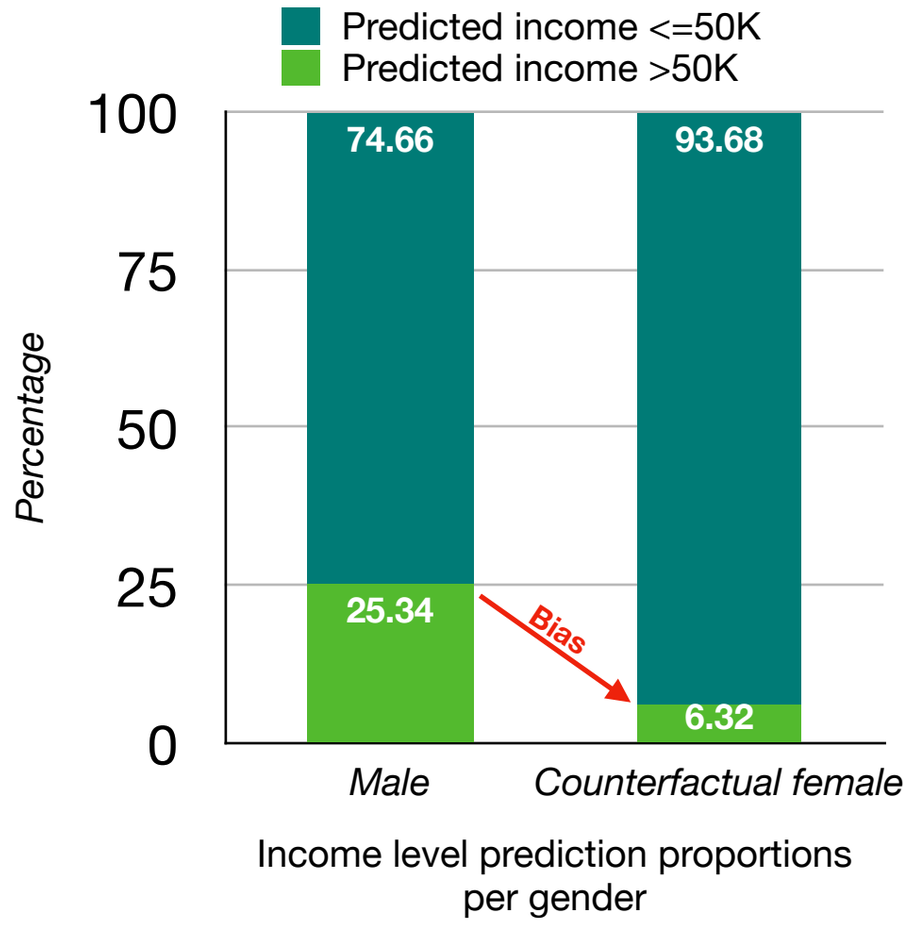
LSAT MAE	UGPA MAE	FYA MAE	Cosine Sim.
0.013	0.028	0.005	0.9997

Counterfactual generation quality (Race: White to Black).

CVAE can generate faithful counterfactuals!

Auditing counterfactual fairness

- Dataset: UCI Adult income
- **Trained** classification model
  - Predict income level (<=50K; >50K)
  - Sensitive feature (Gender: Male-Female)
- Audit counterfactual fairness:
  - ***Male*** *individual was predicted to have high income.*
  - *If individual was* ***female*** *instead, would the* ***prediction change****?*



Income level prediction proportions per gender

High-income males → **low-income** counterfactual females!

Model biased negatively towards females!