Title: Causal Discovery via Common Entropy

Speakers: Murat Kocaoglu

Collection: Causal Inference & Quantum Foundations Workshop

Date: April 18, 2023 - 2:00 PM

URL: https://pirsa.org/23040113

Abstract: Distinguishing causation from correlation from observational data requires assumptions. We consider the setting where the unobserved confounder between two observed variables is simple in an information-theoretic sense, captured by its entropy. When the observed dependence is not due to causation, there exists a small-entropy variable that can make the observed variables conditionally independent. The smallest such entropy is known as common entropy in information theory. We extend this notion to Renyi common entropy by minimizing the Renyi entropy of the latent variable. We establish identifiability results with Renyi-0 common entropy, and a special case of (binary) Renyi-1 common entropy. To efficiently compute common entropy, we propose an iterative algorithm that can be used to discover the trade-off between the entropy of the latent variable and the conditional mutual information of the observed variables. We show that our algorithm can be used to distinguish causation from correlation in such simple two-variable systems. Additionally, we show that common entropy can be used to improve constraint-based methods such as the PC algorithm in the small-sample regime, where such methods are known to struggle. We propose modifying these constraint-based methods to assess if a separating set found by these algorithms is valid using common entropy. We finally evaluate our algorithms on synthetic and real data to establish their performance.

Causal Discovery via Common Entropy

Murat Kocaoglu
Purdue University
ECE

Sanjay Shakkottai
Alex Dimakis
Constantine Caramanis
Sriram Vishwanath

Causal Inference and
Quantum Foundations Workshop
Perimeter Institute
Waterloo, Canada

April 18, 2023

ation

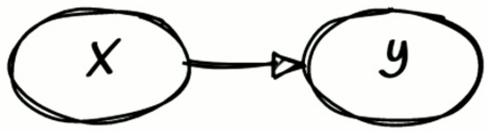
Modeling Probabilistic Causation

X : Percentage of population w/ access to clean water

Y : Child mortality

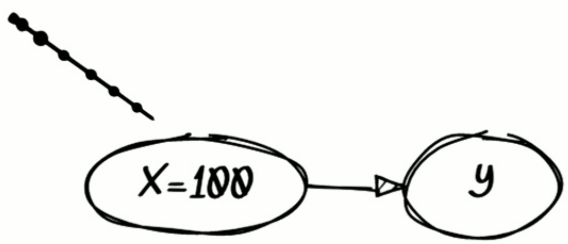
X	Y
22	165
97	15
85	33
100	3
51	154
....	....

http://data.un.org



$$Y = f(X, E)$$

Magic wand to intervene/do:



Y
3
1
5
....

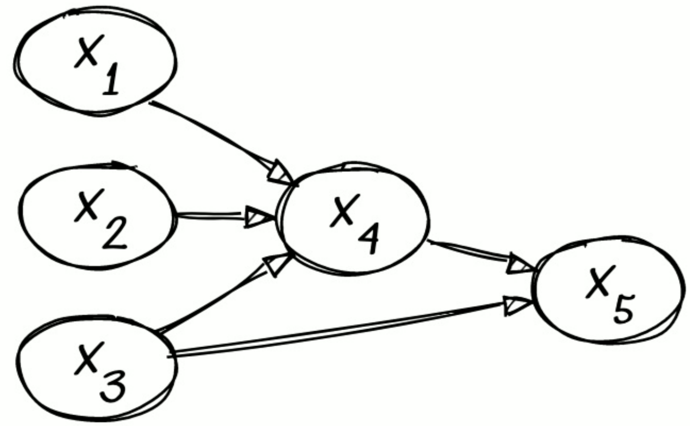
3

Modeling Probabilistic Causation

X is said to cause Y
if
intervening on X
changes
the
distribution of Y

ation

Causal Graphs



Vertices: Random variables
Edges : Causal relations

$$X_i = f_i(Pa_i, E_i)$$

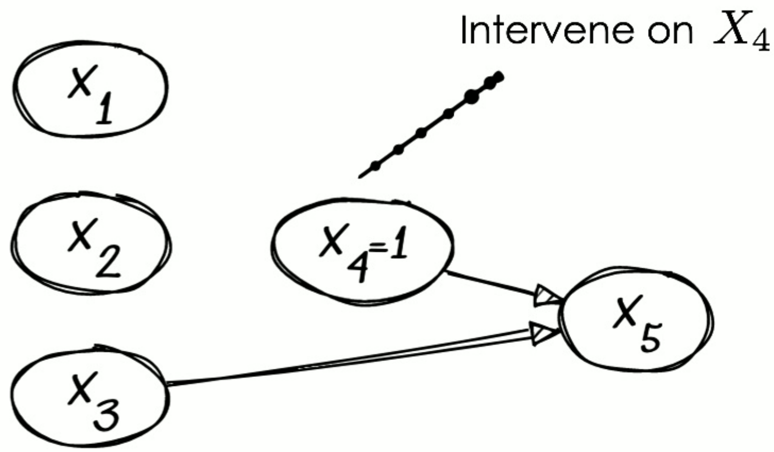
Pa_i : Set of parents of X_i in the causal graph

$\{E_i\}_i$: Jointly independent exogenous variables

5

ation

Causal Graphs



Vertices: Random variables
Edges : Causal relations

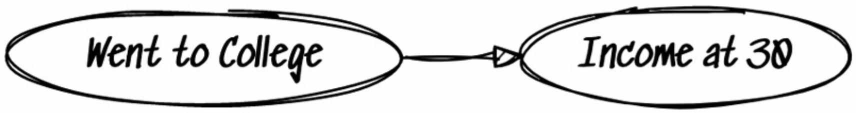
$$X_i = f_i(Pa_i, E_i)$$

Pa_i : Set of parents of X_i in the causal graph

$\{E_i\}_i$: Jointly independent exogenous variables

How to Infer Causation?

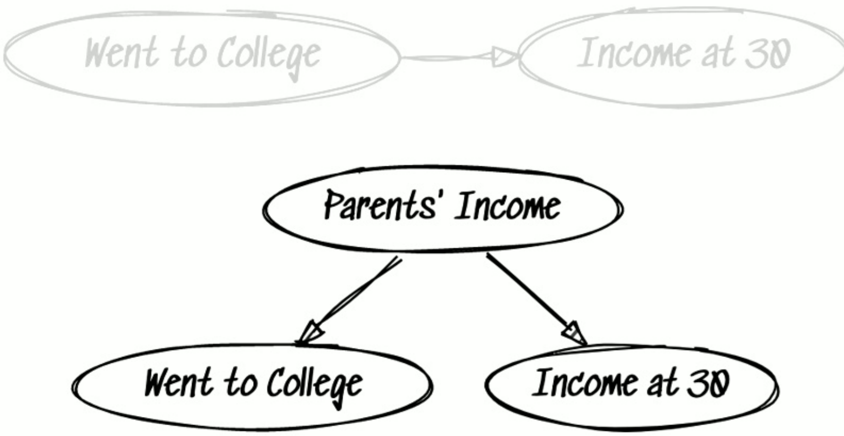
Does going to college have any causal effect on income at 30?



Went to College	Income at 30 > 50k
0	0
0	1
0	0
1	1
1	1
….	….

How to Infer Causation?

Does going to college have any causal effect on income at 30?

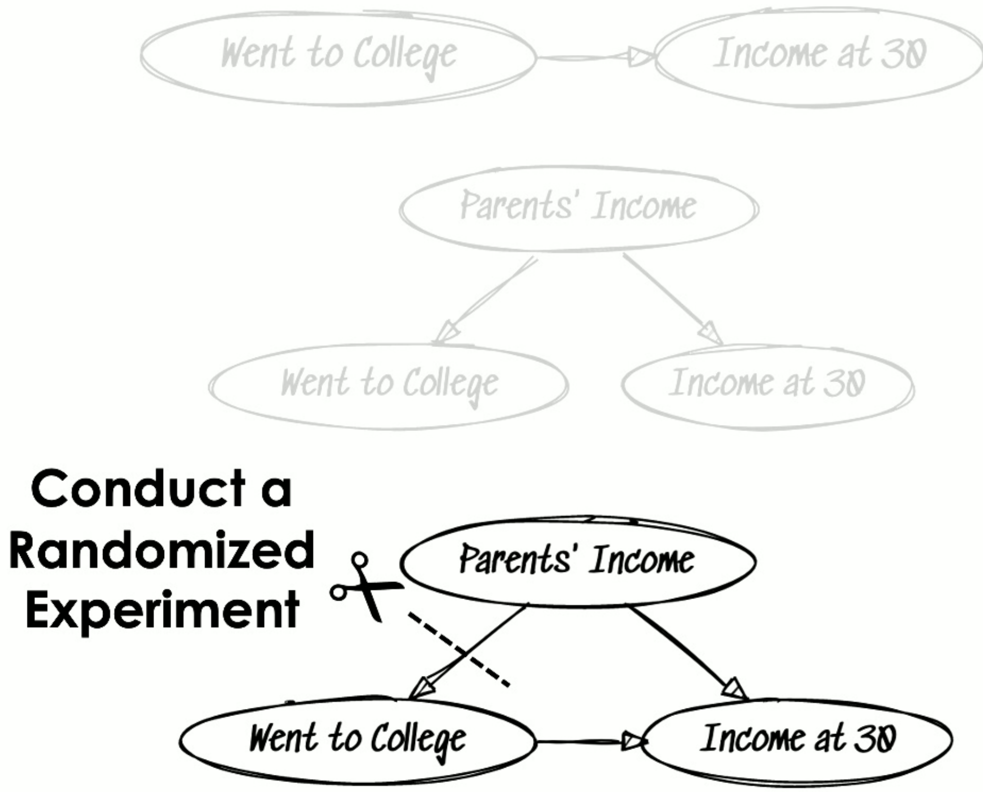


Went to College	Income at 30 > 50k	Parents' Income
0	0	0
0	1	1
0	0	0
1	1	1
1	1	1
....	....	....

ation

How to Infer Causation?

Does going to college have any causal effect on income at 30?



Went to College	Income at 30 > 50k	Parents' Income
0	0	0
0	1	1
0	0	0
1	1	1
1	1	1
….	….	….

ation

How to Infer Causation?

Conduct intervention (RCT)

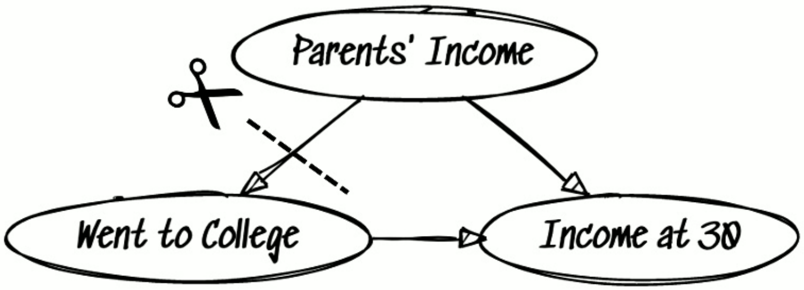
- Force half the people to go

Went to College	Income at 30 > 50k
1	1
1	1
1	0
....	....

- Force other half to NOT go

Went to College	Income at 30 > 50k
0	0
0	1
0	0
....	....

- Compare income of both populations



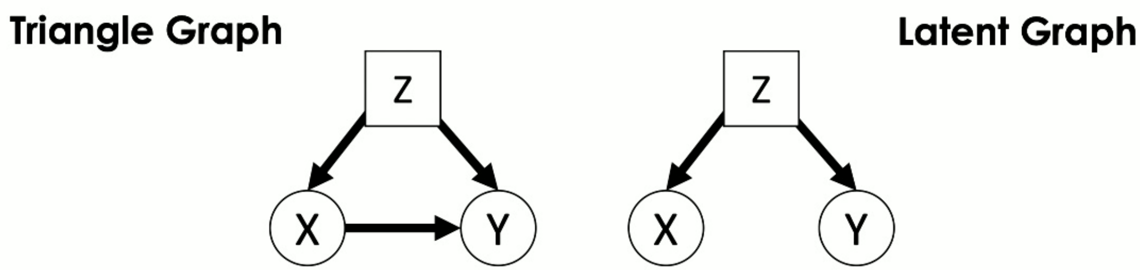
ation

Talk Outline

- Motivation
- Introduction to Probabilistic Causality
- Causal Discovery and Common Entropy

ation

Motivation:
Distinguish Causation from Correlation



- Z is an unobserved (latent) confounder.

- Can we distinguish them from observational data?

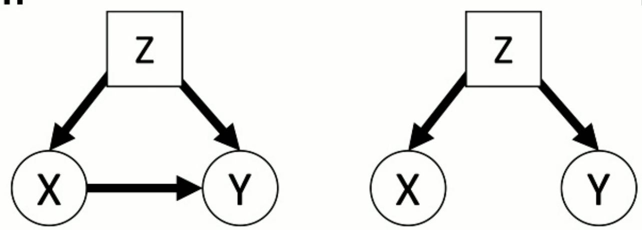
No.

ation

Motivation:
Distinguish Causation from Correlation

Triangle Graph

Latent Graph



- Z is an unobserved (latent) confounder.

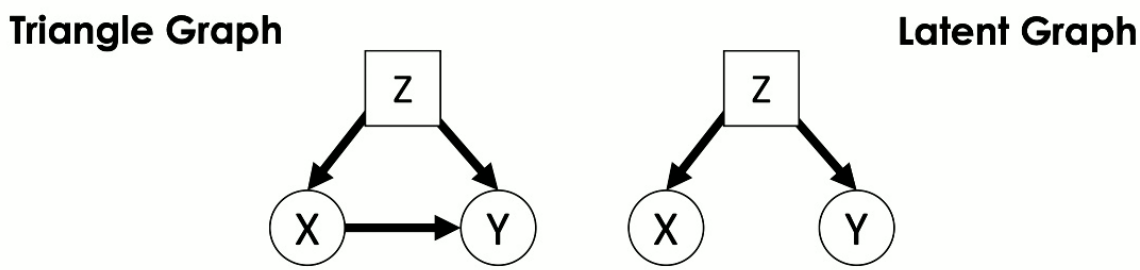
- Can we distinguish them from observational data?

No.

- What if the latent confounder is **simple**?

Maybe!

Motivation:
Distinguish Causation from Correlation



- Z is an unobserved (latent) confounder.

- Can we distinguish them from observational data?

No.

- What if the latent confounder is **simple**?

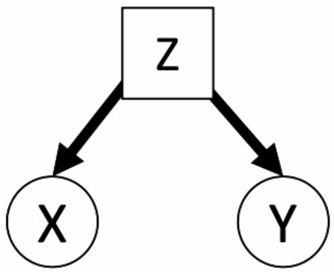
simple = low Rényi entropy

Maybe!

ation

Simple Confounder
Case1: Low support size

- Two variables $X \in \mathcal{X}, Y \in \mathcal{Y}$.



- $Z \in \mathcal{Z}$ unobserved (latent)
  $X \perp\!\!\!\perp Y \,|\, Z$

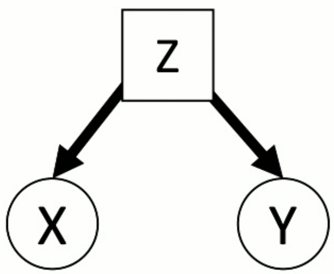
IID Datapoints	Farmer X	Farmer Y
1992		
1993		
1994		
1995		
1996		
1997		
1998		
1999		
2000		
	....	....

17

ation

Simple Confounder
Case1: Low support size

- Two variables $X \in \mathcal{X}, Y \in \mathcal{Y}$.

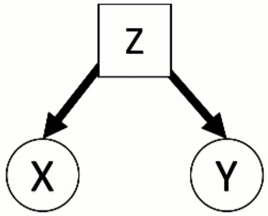


- $Z \in \mathcal{Z}$ unobserved (latent)
  $X \perp\!\!\!\perp Y \,|\, Z$

IID Datapoints	Farmer X	Farmer Y
1992	🍋	🍒
1993	🍋	🍋
1994	🍅	🍒
1995	🍒	🍒
1996	🍋	🍋
1997	🍅	🍒
1998	🍋	🍋
1999	🍅	🍒
2000	🍅	🍅
	....	....

- **Q:** How does this graph manifest itself in the observed distribution $p(x, y)$ when Z has small support size?

Footprints of Latent Graph



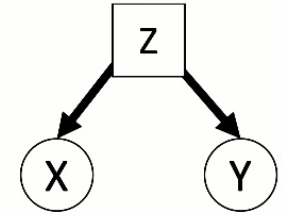
- Suppose Z has k states

$$p(x,y) = \sum_z p(x,y|z)p(z)$$
$$= \sum_z p(x|z)p(y|z)p(z)$$

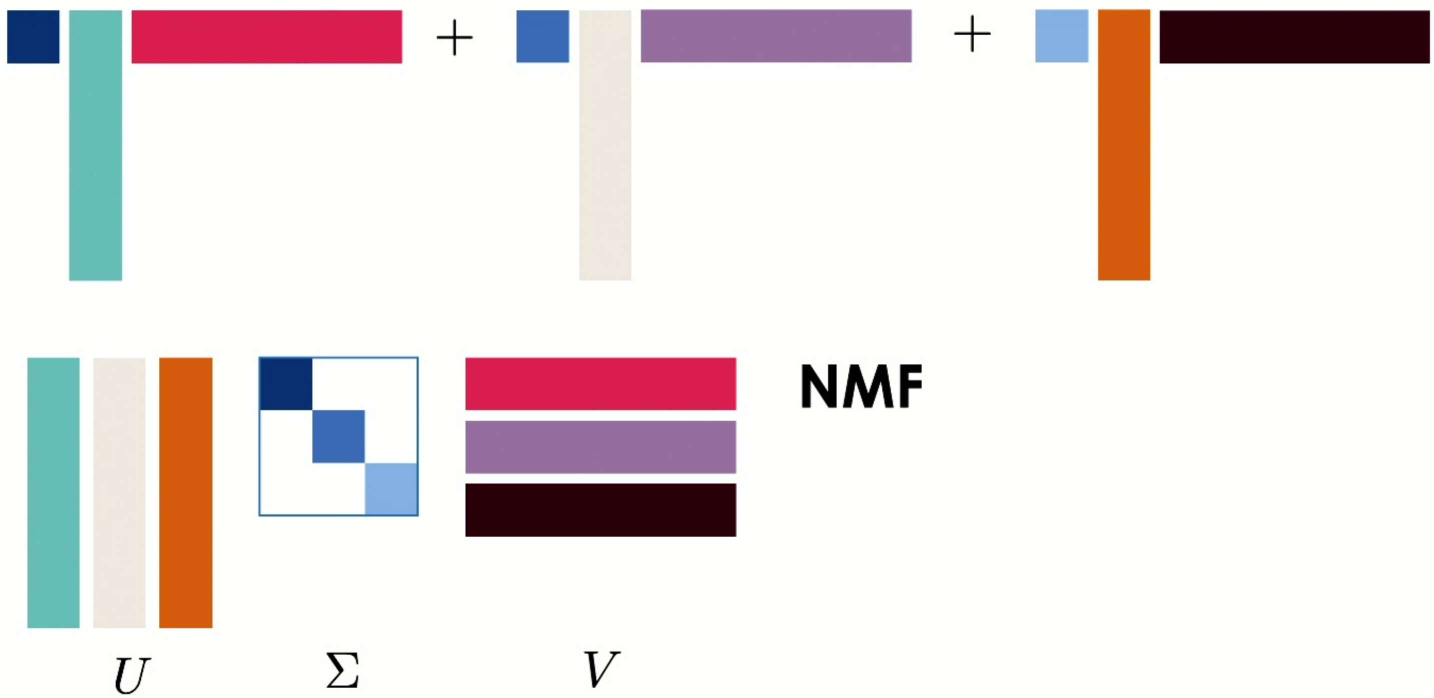
$$[p(x,y)]_{x,y} = \underbrace{\boxed{p(X,Y)}} = p(z=1)\underbrace{\boxed{p(X,Y|Z=1)}}_{\textbf{Rank 1}} + p(z=2)\underbrace{\boxed{p(X,Y|Z=2)}}_{\textbf{Rank 1}} + \dots$$

Convex combination of ***k*** rank 1 matrices.

Model Decomposition and NMF

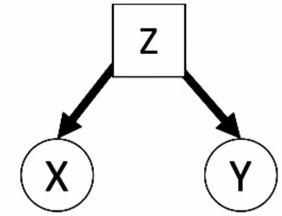


$$[p(x,y)]_{x,y} =$$

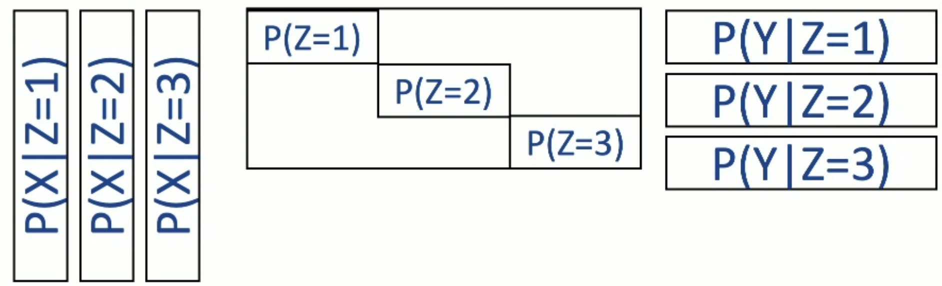


$$U \in \mathbb{R}_+^{n \times k}, V \in \mathbb{R}_+^{k \times n} \qquad \Sigma = diag(d), d \in \mathbb{R}_+^k$$

Model Decomposition and NMF



$$[p(x,y)]_{x,y} = M = U\,V = \tilde{U}\Sigma\tilde{V} =$$



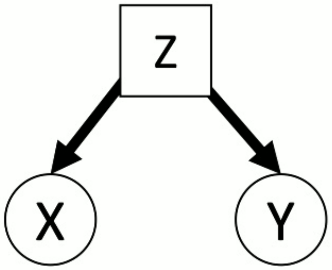
NMF

row/column normalization

$\Rightarrow$ Nonnegative rank gives minimum support size for the confounder!

21

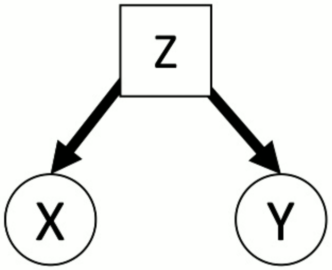
Story with Support Size



Latent Graph

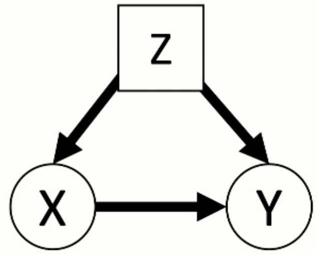
$[p(x,y)]_{x,y}$ *has non-negative rank* $\leq |\mathcal{Z}|$

Story with Support Size



Latent Graph

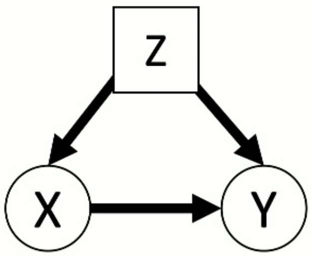
$[p(x,y)]_{x,y}$ *has non-negative rank* $\leq |\mathcal{Z}|$



Triangle Graph

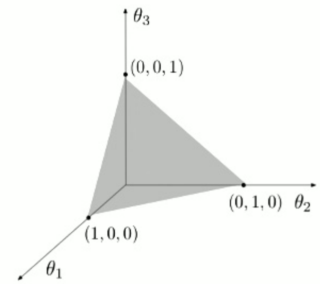
What can we say about rank?

A typical distribution from Triangle Graph



Triangle Graph

$$p(x, y, z) = p(z)p(x|z)p(y|x, z)$$



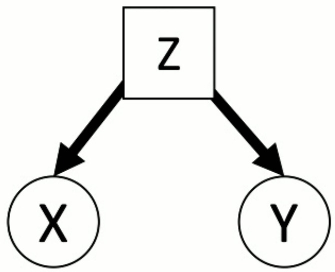
2D simplex

Theorem: Suppose each conditional in **triangle graph** is uniformly randomly chosen from the simplex.

Kocaoglu et al., NeurIPS'20

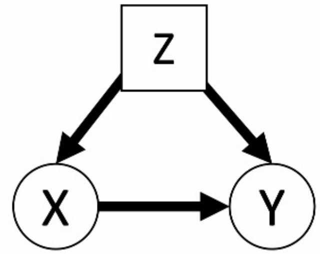
$\Rightarrow [p(x, y)]_{x,y}$ has non-negative rank $\min\{|\mathcal{X}|, |\mathcal{Y}|\}$ with prob. 1.

Story with Support Size



Latent Graph

$[p(x,y)]_{x,y}$ *has non-negative rank* $\leq |\mathcal{Z}|$

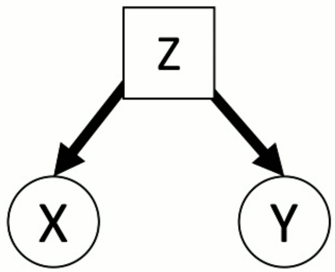


Triangle Graph

$[p(x,y)]_{x,y}$ *has non-negative rank* $= \min\{|\mathcal{X}|, |\mathcal{Y}|\}$

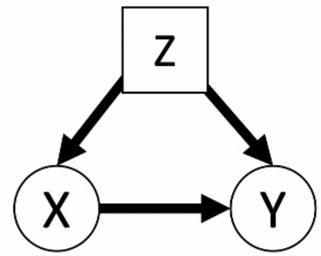
ation

Identifiability Result



Latent Graph

$[p(x,y)]_{x,y}$ *has non-negative rank* $\leq |Z|$



Triangle Graph

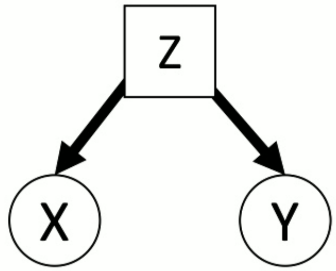
$[p(x,y)]_{x,y}$ *has non-negative rank* $= \min\{|\mathcal{X}|, |\mathcal{Y}|\}$

Corollary: If the support size of latent confounder is $< \min\{|\mathcal{X}|, |\mathcal{Y}|\}$ we can distinguish Latent Graph from Triangle Graph.

- Uses NMF rank. NP-Hard to calculate.
- **Next:** Assume low-entropy confounder.

Simple Confounder

Case2: Low entropy



$$H(Z) \leq \theta$$

- **Q:** How does such a causal graph manifest itself in the observed distribution $p(x, y)$ when Z has small entropy?

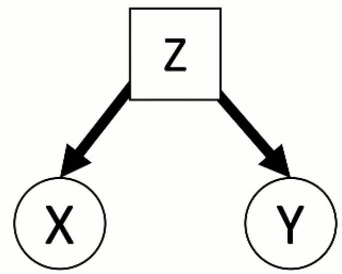
- **A:** This is related to **common entropy**.

27

ation

Common Entropy

- Given $p(x, y)$, find Z with **minimum entropy** such that $X \perp\!\!\!\perp Y \,|\, Z$.
- Common entropy is $G(X,Y) := H(Z)$. [related to Wyner info]
- Equivalent to fitting latent graph with smallest-entropy latent.



$\Rightarrow$ Entropy of true confounder upper-bounds common entropy!

ation

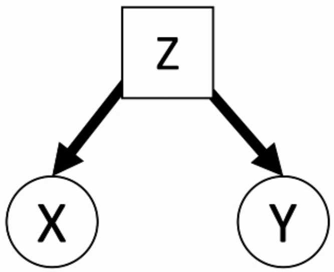
Common Entropy

- Given $p(x, y)$, find Z with **minimum entropy** such that $X \perp\!\!\!\perp Y \mid Z$.
- Common entropy is $G(X,Y) := H(Z)$. [related to Wyner info]
- Closed-form solution for binary variables by Kumar et al. 2014.
- Very hard problem in general.
  (But no formal hardness results)

G. R. Kumar, C. T. Li, A. El Gamal, "Exact common information," ISIT'14.

ation

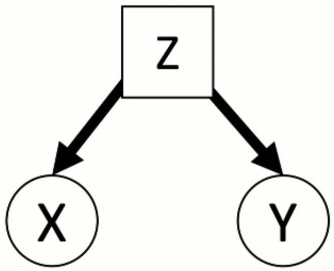
Story with Common Entropy



Latent Graph

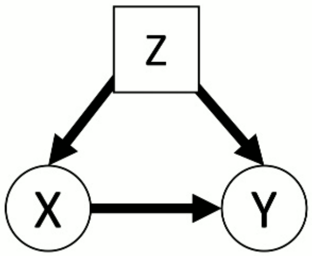
X, Y *has common entropy* $G(X,Y) \leq H(Z)$

Story with Common Entropy



Latent Graph

X, Y *has common entropy* $G(X, Y) \leq H(Z)$



Triangle Graph

What can we say about common entropy?

Very difficult to answer. We can show a bound for binary case:

Theorem: For binary X, Y, all but a vanishing fraction* of models from Triangle Graph has $G(X, Y) > H(Z)$.

Kocaoglu et al., NeurIPS'20

* $\rightarrow 1$ as $H(Z) \rightarrow 0$

31

ation

A Conjecture

Conjecture: Let $p(X,Y,Z) = p(Z)p(X|Z)p(Y|X,Z)$ s.t. each conditional is uniformly randomly chosen from the simplex.

$\Rightarrow$ For any $q(X,Y,Z)$ that satisfies $p(X,Y) = q(X,Y)$ and $X \perp\!\!\!\perp Y \,|Z$

$$H(Z) > \alpha \min\{H(X), H(Y)\}$$

for some constant α.

Open Problem!

In words, most X, Y from triangle graph have large common entropy.

- **Next:** Propose an algorithm to estimate common entropy.

32

ation

A Relaxation of the Problem

$$\begin{aligned} \underset{q(x,y,z)}{\text{minimize}} \quad & H(Z) \\ \text{subject to} \quad & \sum_{x,y,z} q(x,y,z) = 1, \\ & q(x,y) = p(x,y),\ \forall x,y. \quad I(X;Y|Z) \leq \delta \end{aligned}$$

ation

Loss Function

$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

ation

Loss Function

$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

- Given $p(X,Y)$, **construct** $q(Z|X,Y)$

Takes care of the constraint

$q(x,y) = p(x,y), \ \forall x,y.$

Joint: $q(X,Y,Z) = p(X,Y)q(Z|X,Y)$

ation

Loss Function

$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

Takes care of the constraint

- Given $p(X,Y)$, **construct** $q(Z|X,Y)$ → $q(x,y) = p(x,y), \ \forall x, y.$

Joint: $q(X,Y,Z) = p(X,Y)q(Z|X,Y)$

- **Variables**:

$$q(z|x,y) \qquad \forall x \in [n], y \in [n], z \in [k]$$

kn^2 variables

- **Constraints**: Non-negative, slices sum to 1.

ation

A Practical Way to Estimate Common Entropy

- Regularize with the constraint:

$$\min \ \mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

- Still need to search over $q(z|x,y)$.

- Still non-convex in $q(z|x,y)$.

37

ation

Characterizing Stationary Points of Loss

- Find Lagrangian of $\mathcal{L} = I(X;Y|Z) + \beta H(Z)$

  take partial derivative and set to zero.

$$q(z|x,y) = \left(\frac{1}{2}\right)^{\delta_{x,y}-\beta} \frac{q(z|x)q(z|y)}{q(z)^{1-\beta}}$$

- Turn into an iterative update algorithm called *LatentSearch*.

  [Similar in spirit to Blahut-Arimoto, EM, Information bottleneck]

ation

Latent Variable Discovery Algorithm
LatentSearch

- Randomly initialize $q_0(z|x,y)$, Set $i=0$
- Repeat until convergence:
  - Set
    $$q_{i+1}(z|x,y) \leftarrow \frac{q_i(z|x)q_i(z|y)}{q_i(z)^{1-\beta}}$$
  - Update marginals
    $$q_{i+1}(z|x) \leftarrow q_{i+1}(z|x,y)$$
    $$q_{i+1}(z|y) \leftarrow q_{i+1}(z|x,y)$$
    $$q_{i+1}(z) \leftarrow q_{i+1}(z|x,y)$$
  - Set $i \leftarrow i+1$

- **Theorem 1:** Stationary points of LatentSearch are stationary points of the loss.
- **Theorem 2:** LatentSearch converges to local minimum or saddle point for $\beta = 1$.

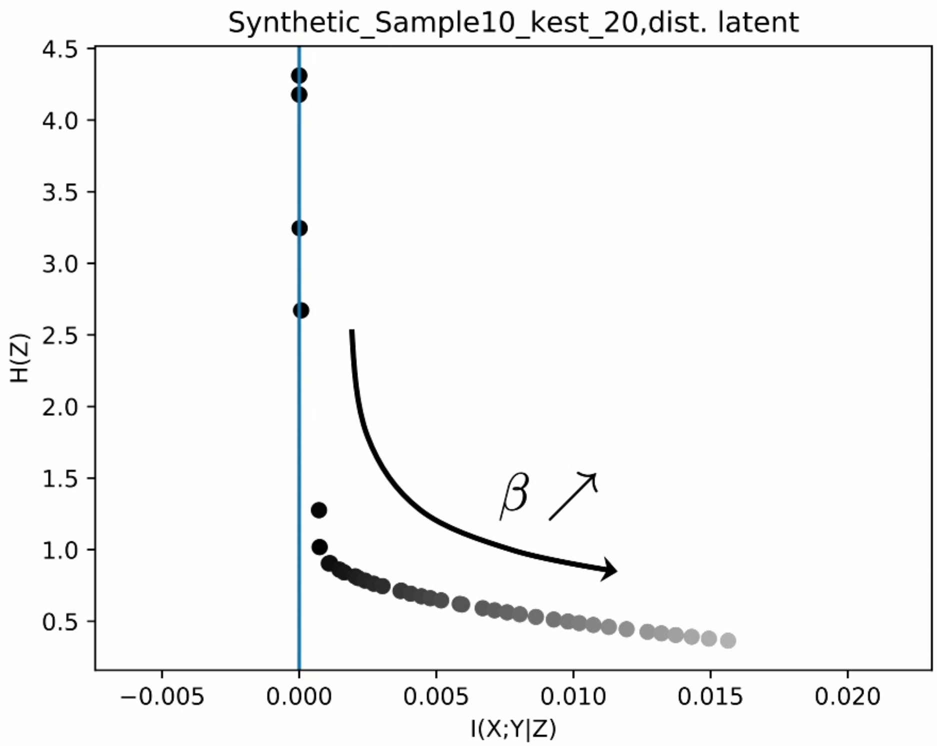
Latent Variable Discovery Algorithm
LatentSearch

- Recovers a point in the $H(Z)$ vs. $I(X;Y|Z)$ plane for each β.

Discover a fundamental tradeoff between Complexity of the Latent vs. Dependence explained away

$H(Z)$ $\qquad\qquad$ $I(X;Y|Z)$

Samples Output of Algorithm



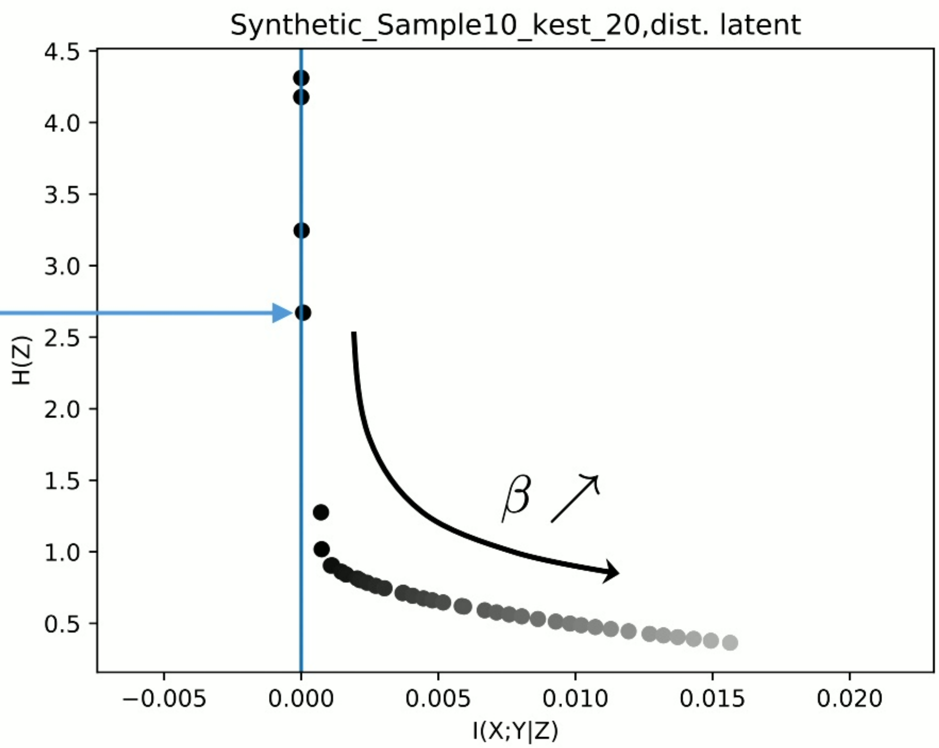
$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

42

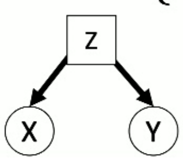
Samples Output of Algorithm

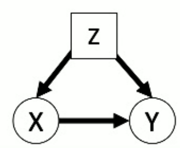
$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

Common entropy estimate

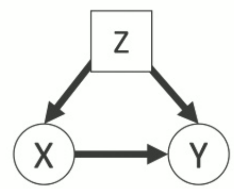


Causal Inference Algorithm
InferGraph

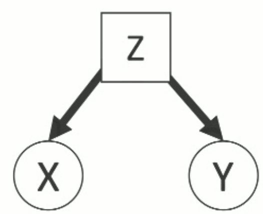
- Input: $p(X,Y)$
  Mutual information threshold: I_t
  Set of β: $\mathcal{B}$
- Set $\mathcal{S} = \emptyset$
- For β in $\mathcal{B}$:
  $q(X,Y,Z) \leftarrow$ LatentSearch($p(X,Y)$, β)
  $H(Z), I(X;Y|Z) \leftarrow q(X,Y,Z)$
  $\mathcal{S} \leftarrow \mathcal{S} \cup (H(Z), I(X;Y|Z))$
- $H^*(Z) = \min\{H : (H,I) \in \mathcal{S} \text{ AND } I < I_t\}$
- If $H^*(Z) < \min\{H(X), H(Y)\}$ output 

, otherwise output 

If common entropy is large



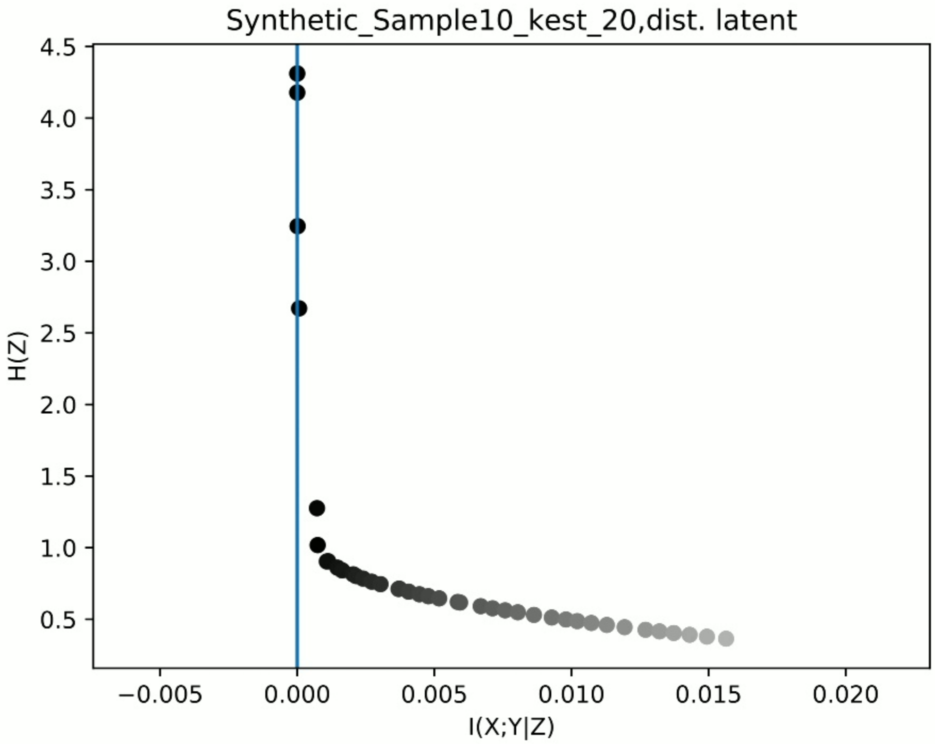
otherwise

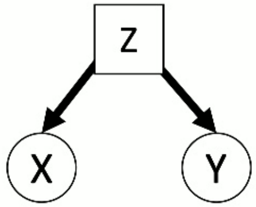


Causal Inference Algorithm

InferGraph

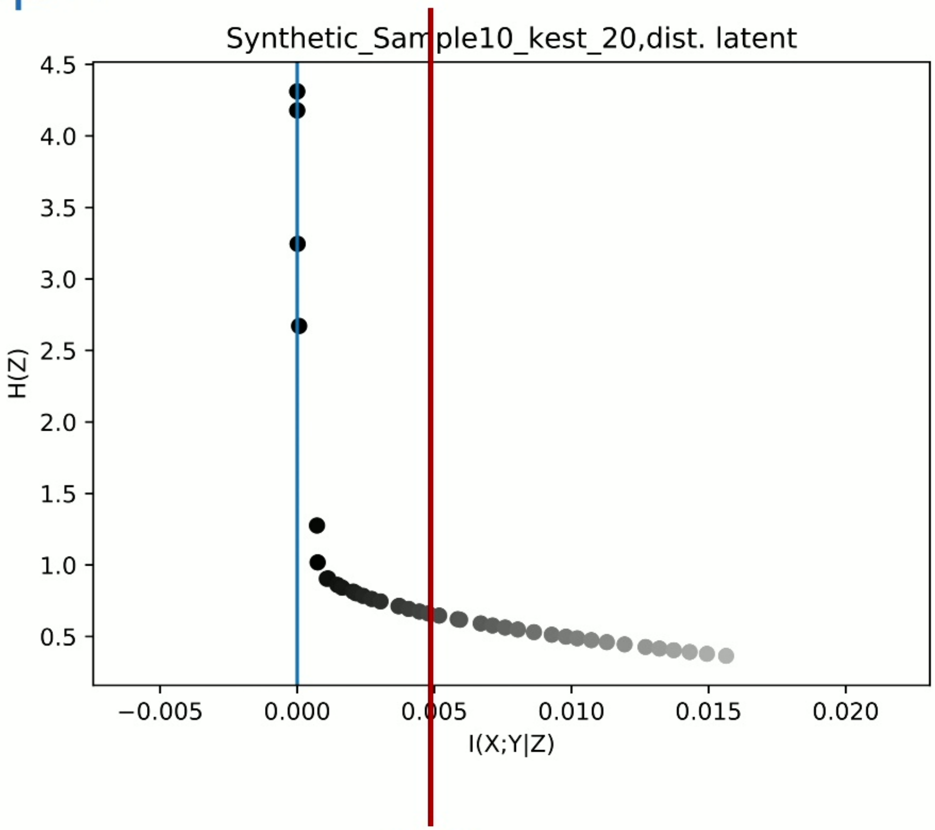
$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$





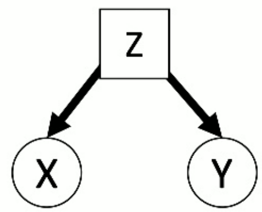
ation

Causal Inference Algorithm
InferGraph



Indep.
threshold

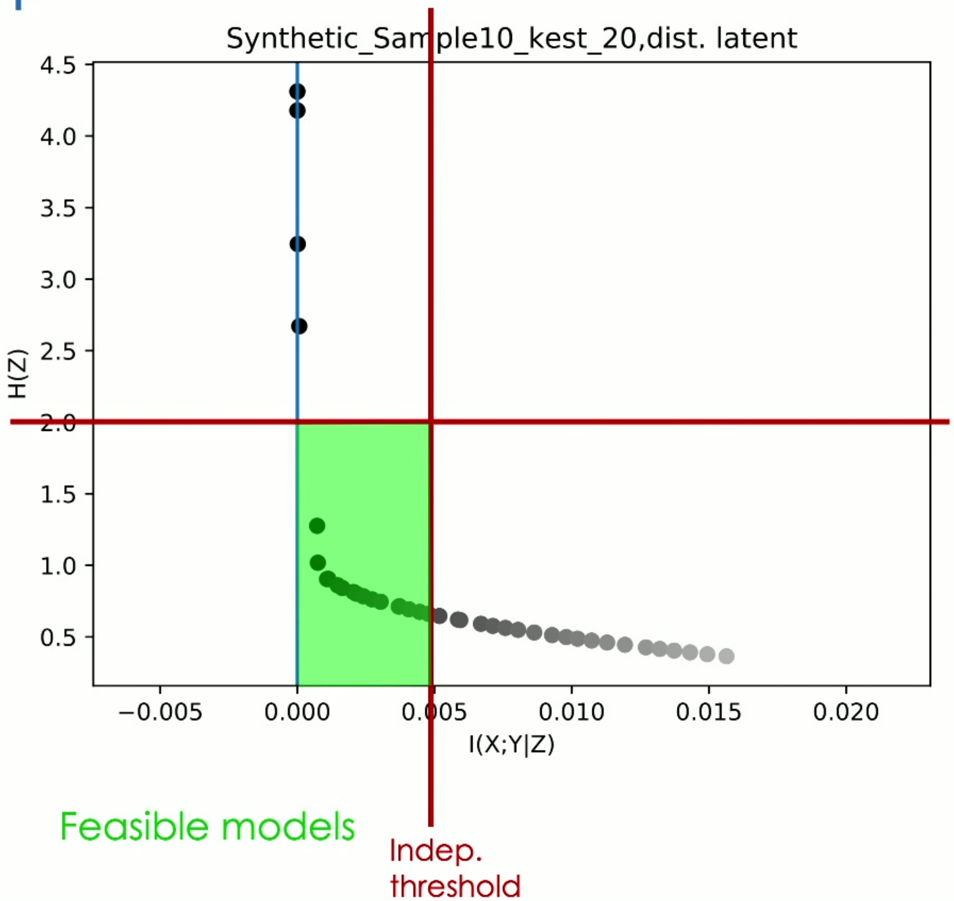
$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

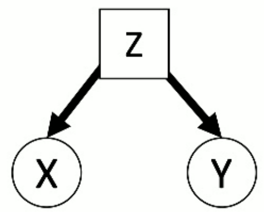


Causal Inference Algorithm
InferGraph

$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$

$$\min\{H(X), H(Y)\}$$



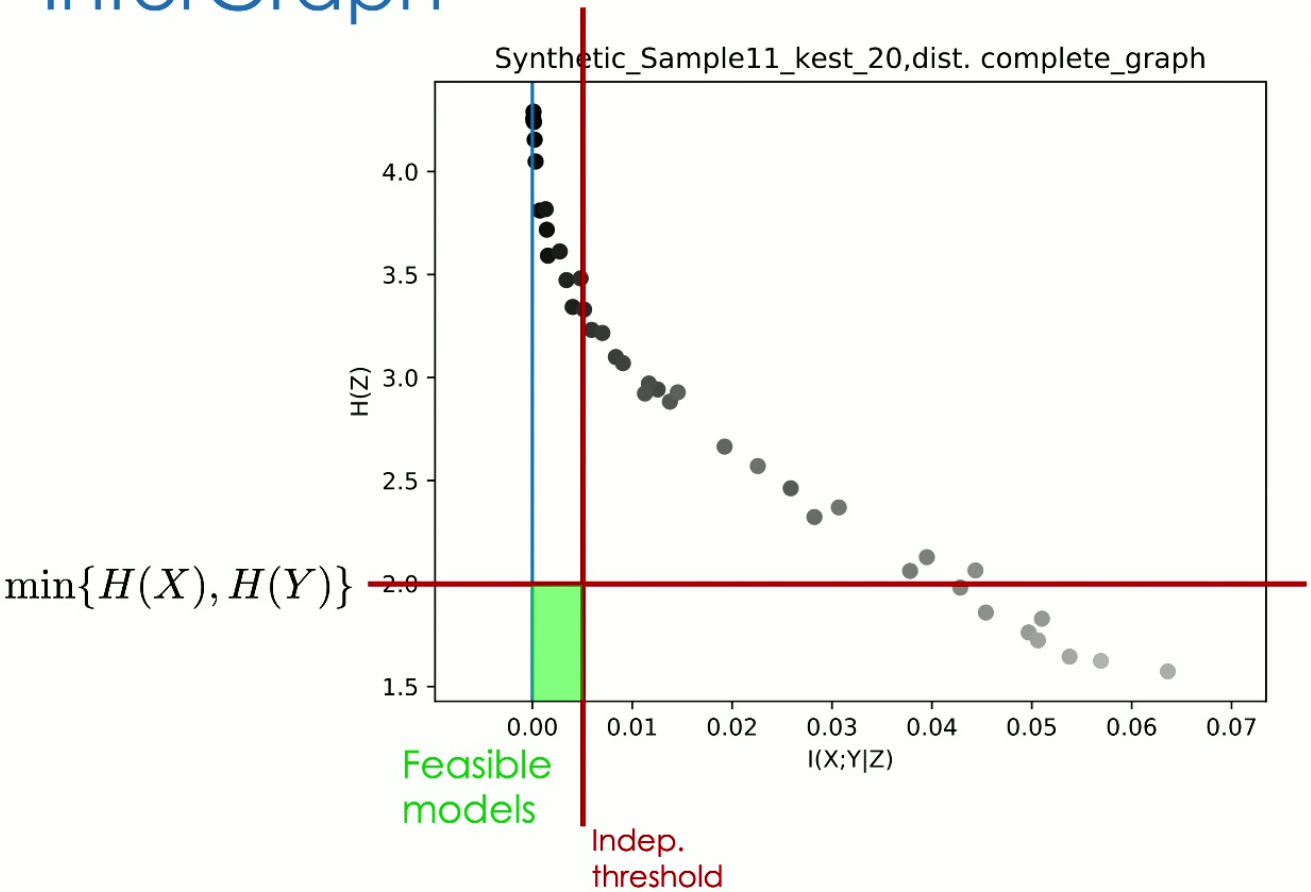


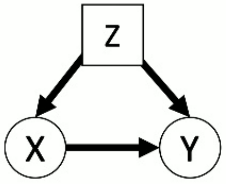
48

Causal Inference Algorithm

InferGraph

$$\mathcal{L} = I(X;Y|Z) + \beta H(Z)$$





ation

Another Application of Common Entropy

- Improve constraint-based causal discovery alg. in the small sample regime.

- Finitely many samples $\Rightarrow$ Incorrect CI statements

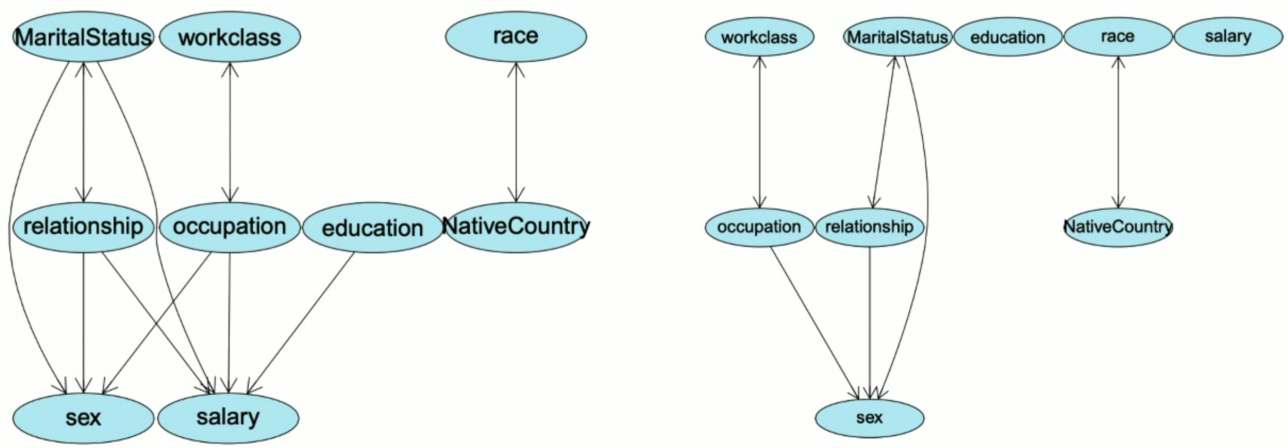
- Can be used to reject small separating sets:

We observe $X \perp\!\!\!\perp Y \,|Z$ but $G(X;Y) > H(Z)$

ation

Another Application of Common Entropy

- EntropicPC rejects separating sets using common entropy.



EntropicPC Standard PC

ation

Conclusion

- Information-theoretic measures can enable causal discovery
  - Minimum entropy couplings
  - Nonnegative Rank
  - Common Entropy

- More information-theory research needed to improve entropic causality
  (e.g., approximate common entropy)

- General case of larger graphs open.

ation

References

1. S. Compton, D. Katz, B. Qi, K. Greenewald, M. Kocaoglu, “Minimum-Entropy Coupling Approximation Guarantees Beyond the Majorization Barrier,” in Proc. of AISTATS’23, Valencia, Spain, April 2023.

2. S. Compton, K. Greenewald, D. Katz, M. Kocaoglu, “Entropic Causal Inference: Graph Identifiability”, in Proc. of ICML’22, July 2022.

3. M. Kocaoglu, S. Shakkottai, A. G. Dimakis, C. Caramanis, S. Vishwanath, “Applications of Common Entropy for Causal Inference,” in Proc. of NeurIPS’20, Online, Dec. 2020.

4. S. Compton, M. Kocaoglu, Kristjan Greenewald, Dmitriy Katz, “Entropic Causal Inference: Identifiability and Finite Sample Results,” in Proc. of NeurIPS’20, Online, Dec. 2020.

5. M. Kocaoglu, A. G. Dimakis, S. Vishwanath, B. Hassibi, "Entropic Causality and Greedy Minimum Entropy Coupling," in Proc. of ISIT’17, Aachen, Germany, June 2017.

6. M. Kocaoglu, A. G. Dimakis, S. Vishwanath, B. Hassibi, "Entropic Causal Inference," in Proc. of AAAI 2017, San Francisco, USA, Feb. 2017.

ation

Questions?

ation