Causal Insights from Merging Data Sets and Merging data sets via Causal Insights

Dominik Janzing

Amazon Research Tübingen, Causality Team

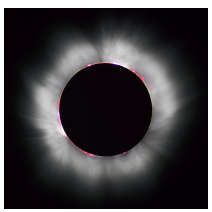
19 April 2023

How do humans learn causality?

1. interventions

source: Wikipedia, Derek Jensen

2. observations under different background conditions



sources: Luc Viatour, https://lucnix.be / Wikipedia, Tomruen

3. putting observations in the broader context of other knowledge

source: Wikipedia, Adriaen van de Venne

How ML tries to learn causal directions since the 1990s

PC-algorithm:
infer causal DAG G on n variables $X_1, \ldots, X_n$ using iid samples from $P(X_1, \ldots, X_n)$ using causal Markov condition and causal faithfulness

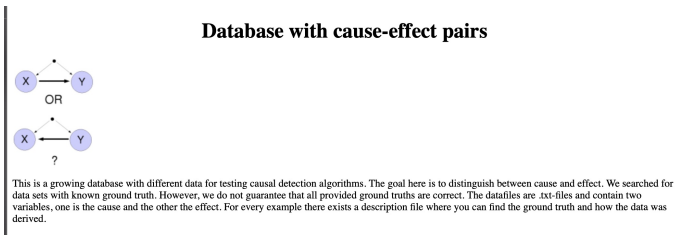
- no interventions
- no different background conditions
- no context

How ML tries to learn causal directions since the 2000s

- model assumptions beyond causal faithfulness, e.g. additive noise (linear / non-linear), post-nonlinear models

  Kano, Shimizu,.. 2003, Hoyer et al 2008, Zhang & Hyvärinnen 2009, see our book for more references

- semi-parametric models guarantee identifiability even for the cause-effect problem
- cause-effect problem is a classification task, which renders benchmarking easier



`https://webdav.tuebingen.mpg.de/cause-effect/` (but don't overfit this dataset!)

- still a difficult task

Combine information from different datasets!

Data from different environments (since the 2010s)

Given the joint distribution of two variables X, Y under different background conditions $C = 1, \ldots, k$

- observe that $P_c(X)$ changes across the conditions
- observe that $P_c(Y|X) = P(Y|X)$ remains constant
- infer $X \to Y$ because independence of mechanisms states that $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ change independently across environments

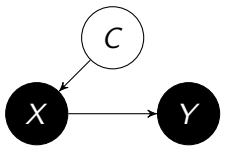
Peters, Janzing, Schölkopf: Elements of Causal Inference 2017

Schölkopf et al: On causal and anticausal learning 2012

Peters et al: Causal inference using invariant prediction: identification and confidence intervals, 2012

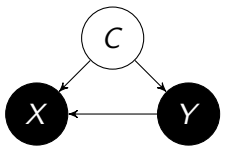
Relation to faithfulness

- let C be a variable known to be root node
- $P_c(X)$ changing across c means $X \not\perp\!\!\!\perp C$
- constant $P(Y|X)$ across c means $Y \perp\!\!\!\perp C\,|X$
- causal faithfulness implies causal structure



(also common causes of X and Y excluded by faithfulness)

- $Y \to X$ required mechanisms $P_c(X|Y)$ and $P_c(Y)$ to change together in a contrived way to keep $P(Y|X)$ constant

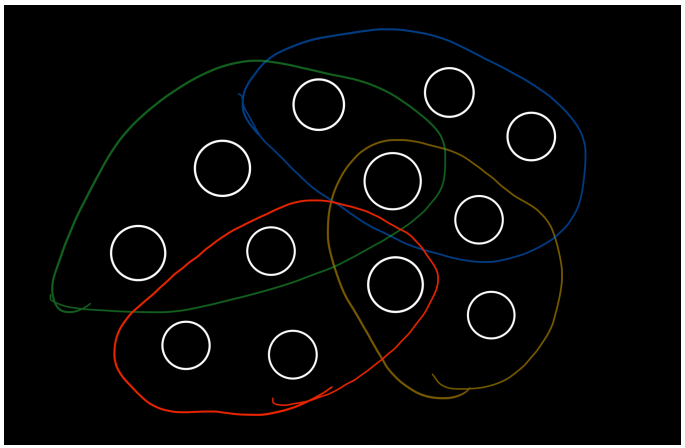


violation of faithfulness / independence of mechanisms
(perspective depending on whether C is a variable or not)

generalization to n nodes: Budhathoki et al: Why did the distribution change? AISTATS 2021

Merge data from different sets of variables

First combine statistical information from different sources, get causal insights later



"Marginal problem:" infer properties of joint distribution $P(\mathbf{X}) = P(X_1, \ldots, X_n)$ from distributions of subsets $\mathbf{X}_{S_1}, \mathbf{X}_{S_2}, \ldots, \mathbf{X}_{S_k}$ (a priori no unique solution)

Kellerer: Maßtheoretische Marginalprobleme, 1964

Example: distribution $P(X, Y, Z)$ is not determined by $P(X, Y), P(X, Z), P(Y, Z)$

Our current approach: MaxEnt

Garido Mejia, Kirschbaum, Janzing: Obtaining Causal Information by Merging Datasets with MAXENT, AISTATS 2022

- shows why merging generates causal knowledge
- shows why causal knowledge helps for merging
- yields joint distribution that shares qualitative properties with the true joint

(not clear how it could scale though)

Inferring joint distributions via MaxEnt

- qiven n variables $\mathbf{X} = (X_1, \ldots, X_n)$
- given observed expectations for subsets of variables

$$\mathbb{E}[f_1(\mathbf{X}_{S_1})] = c_1, \ldots, \mathbb{E}[f_k(\mathbf{X}_{S_k})] = c_k$$

- let $\hat{P}(X_1, \ldots, X_n)$ be the distribution maximizing entropy subject to these constraints

- $\hat{P}$ can be justified as 'best guess' for P, given the available information

  see e.g. Grünwald, Dawid: Game theory, maximum entropy,... , 2004

- we will see: even if $\hat{P}$ is not close to P, it may still capture essential properties of P

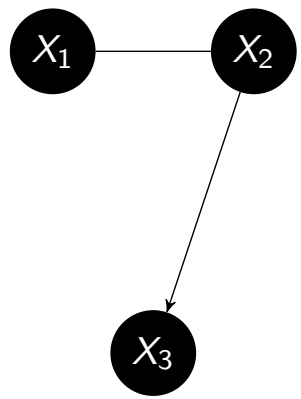
Predicting cond. indep. from bivariate observations

Lemma (Garido Mejia, Kirschbaum, DJ, 2021)

Let $\hat{P}(X_1, X_2, X_3)$ maximize entropy subject to $P(X_1, X_2), P(X_1, X_3), P(X_2, X_3)$ then any conditional independence true in P is also true in $\hat{P}$.

Proof: basic information theory, conditional independence results in larger entropy

Application: let X_1, X_2 be potential causes of X_3 and the set be causally sufficient.



If there is no direct link $X_1 \to X_3$, we observe $X_1 \perp\!\!\!\perp X_3 \,|X_2$ also in $\hat{P}$

Causal structure of n nodes from bivariates?

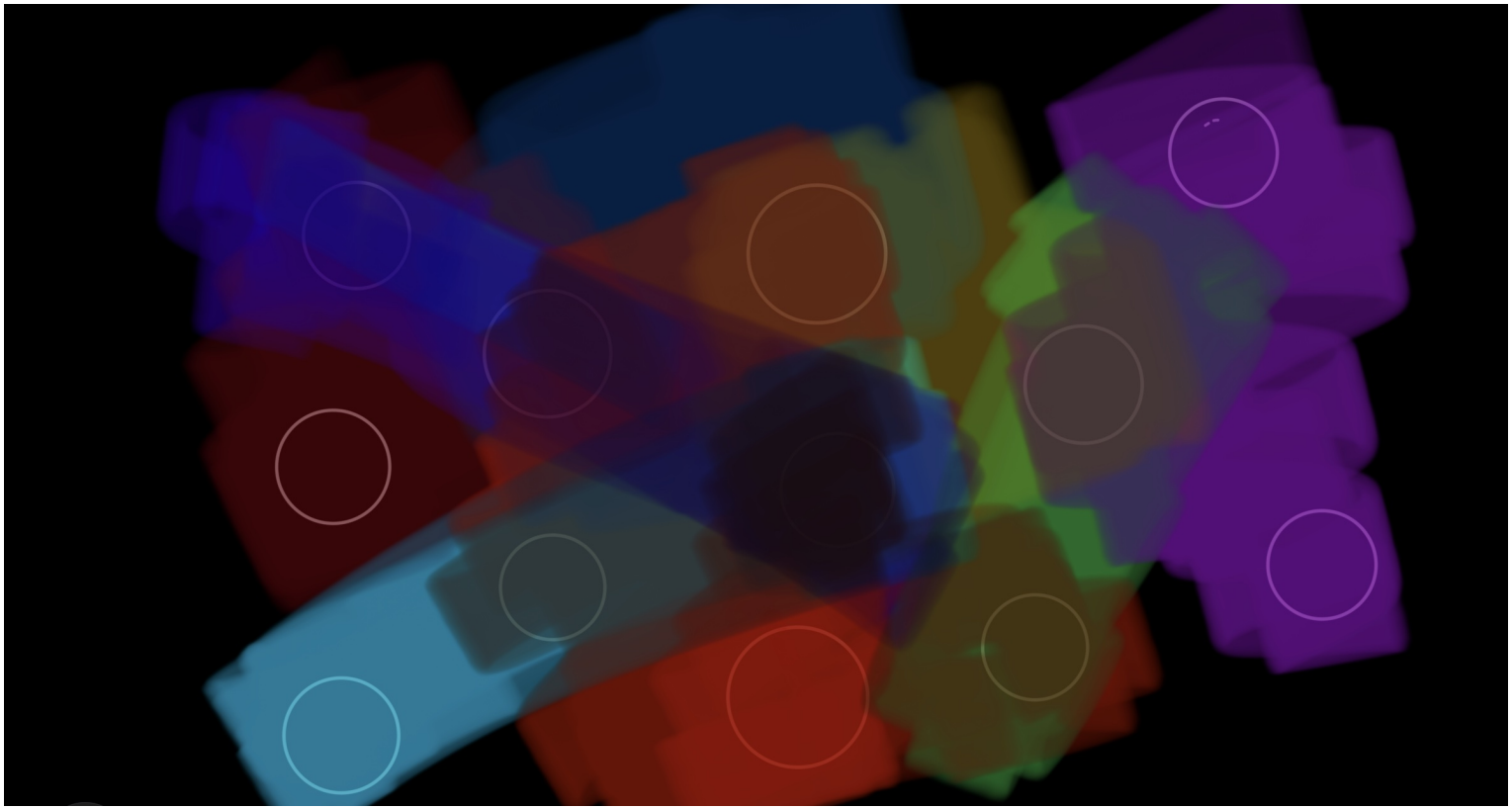
MaxEnt subject to bivariates yields

$$\hat{p}(x_1, \ldots, x_n) \sim e^{\sum_{ij} g_{ij}(x_i, x_j)}$$

- **information geometric view:** projection on manifold of pair-interactions
- **defines undirected graph:** draw an edge iff the interaction term is non-zero
- **relation to DAG G non-trivial:** contains, subject to a genericity condition, at least the 'moral graph' corresponding to G

see Lauritzen: Graphical models, 1996, for relation between directed and undirected graphical models

Vision



Combine all available datasets from a domain to get the best guess for the joint distribution, then try to infer causality

We can do better merging if we know causal directions!

Replacing MaxEnt with Causal MaxEnt

... if we know causal directions:

Let $X_1, \ldots, X_d$ be variables in causal order

- Step 1: obtain $\hat{P}(X_1)$ by maximizing $H(X_1)$ subject to all constraints that restrict $P(X_1)$
- Step 2: then obtain $\hat{P}(X_1|X_1)$ by maximizing $H(X_2|X_1)$ subject to all constraints that restrict $P(X_1, X_2)$
- Step 3: then obtain $\hat{P}(X_3|X_2, X_1)$ by ...
- ...

results more plausible distributions than standard MaxEnt

Sun, Janzing, Schölkopf 2006, Janzing 2021

Interrupt: Motivating Causal MaxEnt

discuss difference between MaxEnt and Causal MaxEnt for a simple case

Principle of Insufficient Reason (PIR)

(Bernoulli, Laplace), also called "Principle of Indifference" by Jaynes

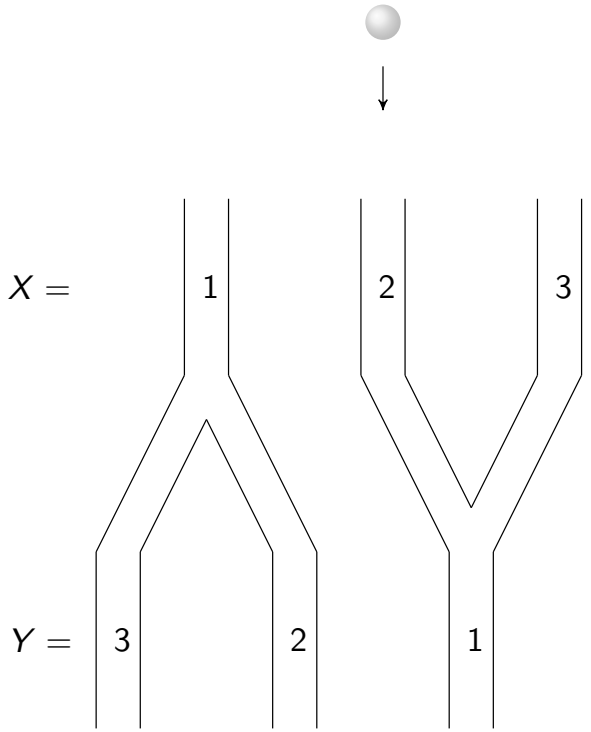
Assign equal probabilities to all possible outcomes under consideration if there is no evidence for preference to any of them.

- knowing only that there is one ball in one of the n urns, one assigns probability $1/n$ to each case (maximizes entropy)
- this subjective probability reflects the symmetry of the problem (Jaynes, Jeffrey,...), 'uninformative prior' in Bayesian inference

Why we need a **Causal** Principle of Insufficient Reason

A ball entering the tube from above can take 4 different paths:

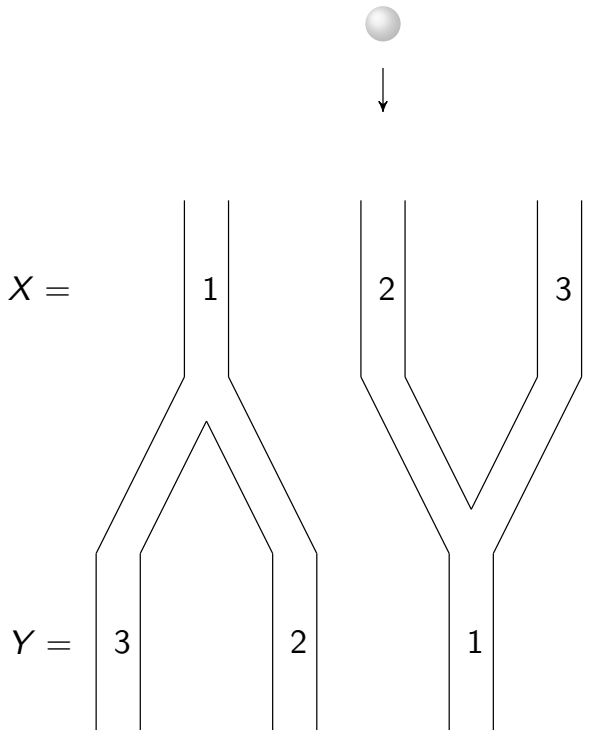
$$
\begin{aligned}
X = 1 \quad &\wedge \quad Y = 3 \\
X = 1 \quad &\wedge \quad Y = 2 \\
X = 2 \quad &\wedge \quad Y = 1 \\
X = 3 \quad &\wedge \quad Y = 1
\end{aligned}
$$



Should we really consider all 4 paths equally likely? (amounts to assigning higher probabilities to entries that enable more options for the exits)

Why we need a **Causal** Principle of Insufficient Reason

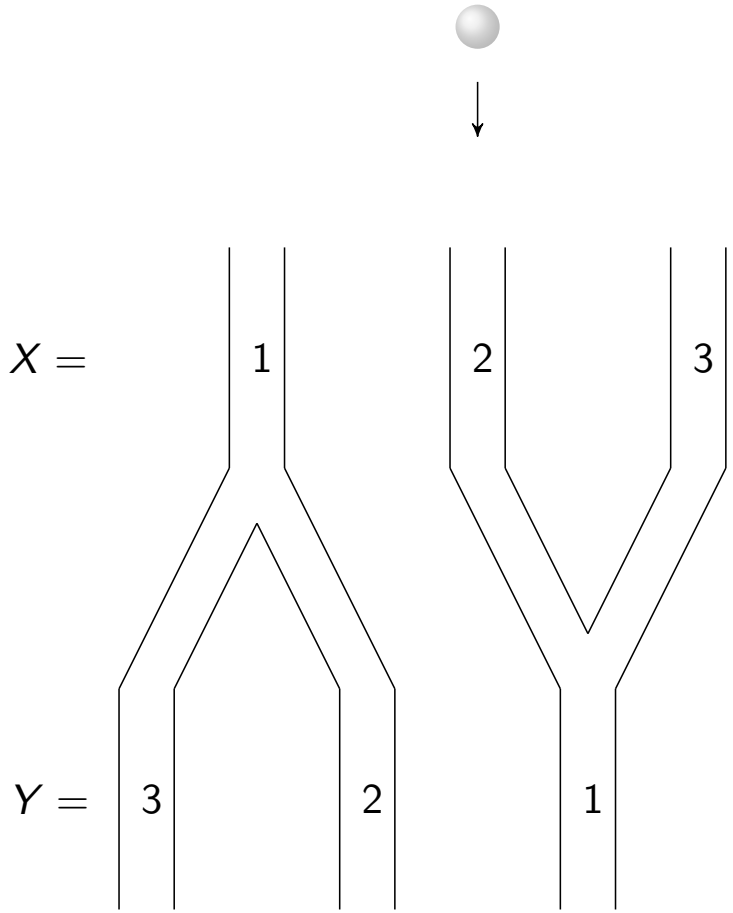
- more plausible to consider **all entrances** equally likely
- given that entrance $X = 1$ is taken, both options at the bifurcation point are equally likely
- exits $Y = 1, 2, 3$ obtain probabilities $2/3, 1/6, 1/6$



device is symmetric w.r.t. X and Y, asymmetry of $P_{X,Y}$ based on causal direction

Causal Principle of Insufficient Reason (CPIR)

- assign equal probabilities to every option for the cause (knowledge about the mechanism relating cause to effect is irrelevant at this point)
- assign equal probabilities to every remaining option for the effect, given the cause



Causal PIR (abstract version)

- given two variables X, Y with finite range $\mathcal{X}, \mathcal{Y}$
- given the knowledge that $X \to Y$ and that the causal mechanism only allows $(x, y) \in R$ where $R \subset \mathcal{X} \times \mathcal{Y}$
- let P_X be the uniform distribution over all x allowed by R
- for any x, let $P_{Y|X=x}$ be the uniform distribution over all y with $(x, y) \in R$
- Then the Causal PIR prior reads $P_{X \to Y} := P_{Y|X} P_X$
  not uniform over R

Properties of Causal MaxEnt vs MaxEnt

- Causal MaxEnt distribution has at most the same entropy as MaxEnt

  (maximizing $H(X) + H(Y|X)$ cannot result in less entropy than first maximizing $H(X)$ and then $H(Y|X)$)

- hence, Causal MaxEnt uses causal direction to provide a more specific prediction

- in the spirit of 'On causal and anticausal learning' by Schölkopf et al, 2012: knowing causal directions entails inductive bias that helps for better predictions

Use Causal MaxEnt for the merging problem

Bivariates overlapping at X_1

Given $P(X_1, X_2)$ and $P(X_1, X_3)$ with causal order X_1, X_2, X_3

- step 1 sets $\hat{P}(X_1) := P(X_1)$ given by the marginals $P(X_1, X_2)$ and $P(X_1, X_3)$
- step 2 sets $\hat{P}(X_2|X_1) := P(X_2|X_1)$ given by $P(X_1, X_2)$
- step 3 maximizes $H(X_3|X_2, X_1)$ subject to the given $P(X_3, X_1)$ which results in $X_3 \perp\!\!\!\perp X_2 \,|X_1$
- corresponds to the fork $X_2 \leftarrow X_1 \rightarrow X_3$

Bivariates overlapping at X_2

Given $P(X_1, X_2)$ and $P(X_2, X_3)$ with causal order X_1, X_2, X_3

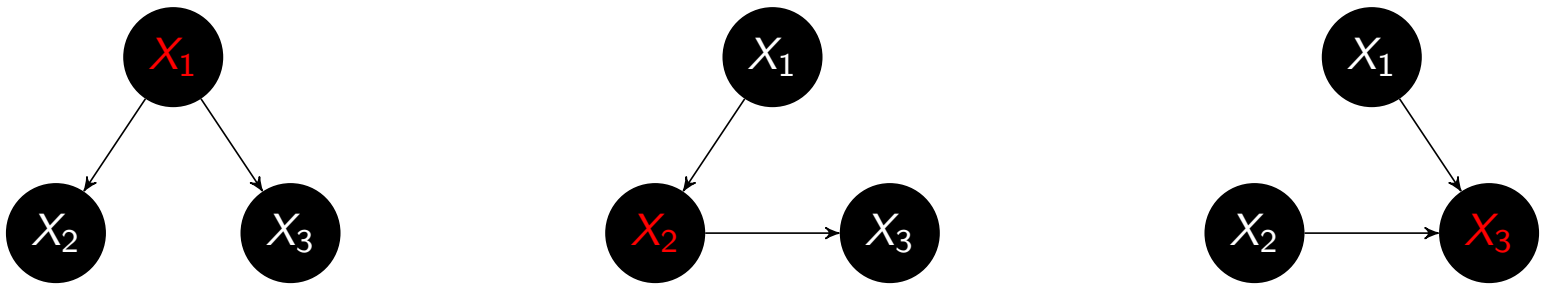
- skip steps 1 and 2 yield $\hat{P}(X_1, X_2) := P(X_1, X_2)$ is given
- step 3 maximizes $H(X_3|X_2, X_1)$ subject to the given $P(X_3, X_2)$
- results in $X_1 \perp\!\!\!\perp X_3 \,|X_2$
- corresponds to the chain $X_1 \to X_2 \to X_3$

Bivariates overlapping at X_3

Given $P(X_1, X_3)$ and $P(X_2, X_3)$ with causal order X_1, X_2, X_3

- step 1 sets $\hat{P}(X_1) := P(X_1)$ as given by $P(X_1, X_2)$
- step 2 sets $\hat{P}(X_2) := P(X_2)$ taken independently from $P(X_2, X_3)$, hence $X_1 \perp\!\!\!\perp X_2$
- corresponds to the collider $X_1 \to X_3 \leftarrow X_2$
- step 3 obtains $\hat{P}(X_3|X_2, X_1)$ by maximizing $H(X_3|X_2, X_1)$

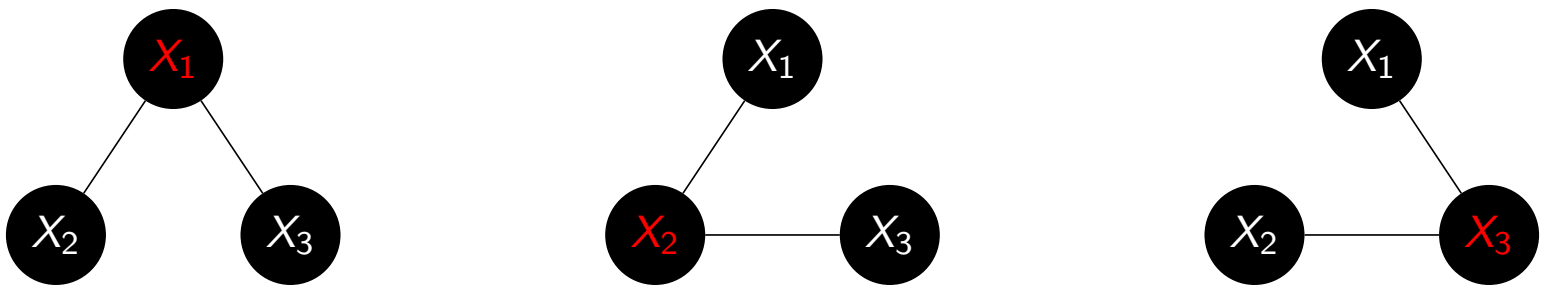
Common rule



In each of the three cases, Causal MaxEnt dropped the edge whose marginal was missing

Comparison to usual MaxEnt

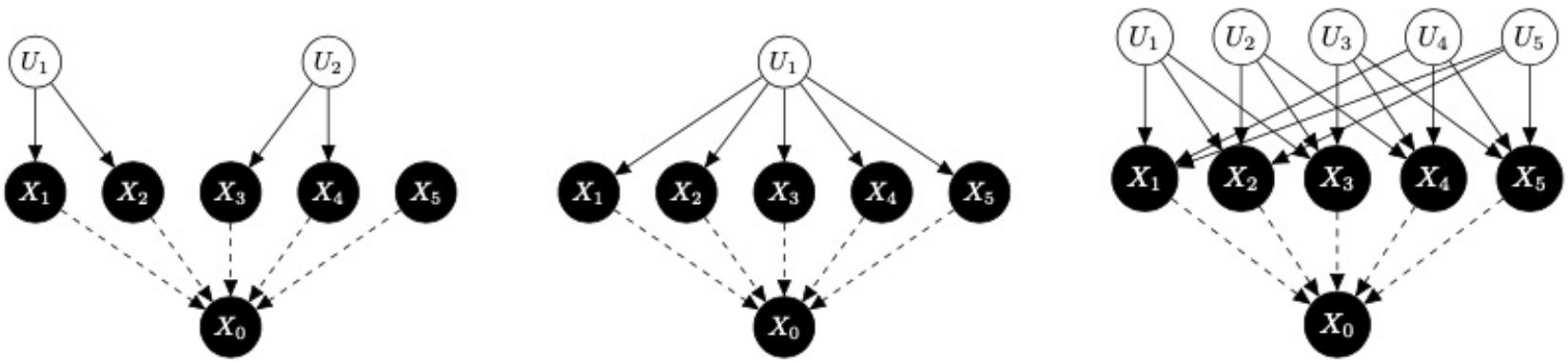
corresponds to undirected graphical models



yields always $X_i \perp\!\!\!\perp X_j \,|X_k$ when $P(X_i, X_j)$ is missing

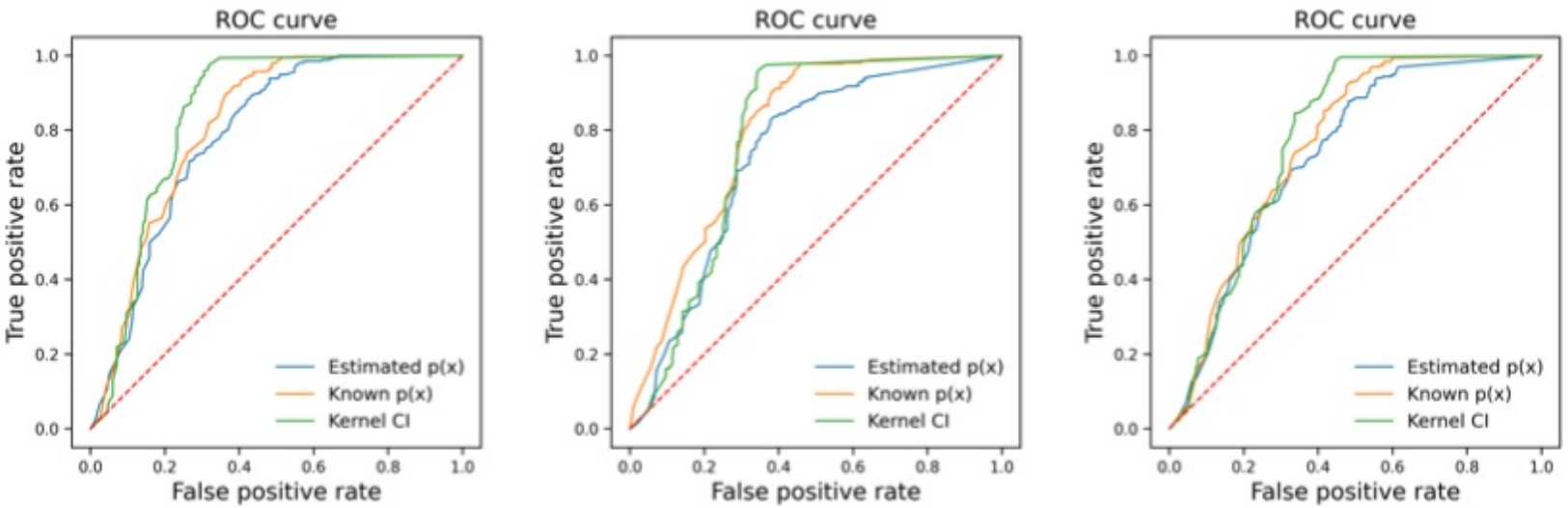
Experiments with Causal MaxEnt on simulated data

- potential causes: confounded binary variables $X_1, ..., X_5$, target X_0
- randomly generate links between some X_j and X_0 and structural equations
- observe all bivariate distributions
- infer presence of links from Causal MaxEnt distribution $\hat{P}(X_1, \ldots, X_5, X_0)$



(3 different confounding structures)

Results



- identification of edges performs almost as good as kernel conditional independence tests (green curve) on the true joint distribution
- Causal MaxEnt (blue curve) only slightly worse than if joint distribution of potential causes is given (orange curve)

Experiments with real data and known joint samples

`www.gapminder.org/data/` provides country-level data of social, economic, and environmental factors (each country is a data point)

Variables include, e.g.

- children per woman (FER)
- GDP per capita (GDP)
- Human Development Index (HDI)
- life expectancy (LE)
- CO2 emissions (CO2)

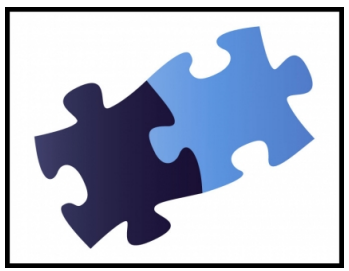
(ongoing work) all 4 triple-wise conditional independences we checked coincided with kernel CI tests

Merging as a new type of prediction problem

Different view on the above

- causal models help inferring joint statistics of variables that have not been observed together (regardless of Causal MaxEnt)

- causal information helps putting distributions together



(asymmetry of puzzle pieces makes it easier)

- test causal models by testing the inferred joint distribution without interventions

Exploit causal discovery for merging

- assume a causal discovery algorithm (e.g. LiNGAM) infers from $P(X, Y)$ that $X \to Y$ is an unconfounded sense
- assume it also infers that $Y \to Z$ in an unconfounded sense from $P(Y, Z)$
- then the joint causal model reads $X \to Y \to Z$, hence

$$P(Z, X) = \sum_y P(Z|y)P(y, X).$$

- can be tested on data from $P(Z, X)$

Generalization: New type of learning scenario

- given the variables $\mathbf{X} := \{X_1, \ldots, X_n\}$
- given an algorithm that infers a causal graph G on $\mathbf{X}$ from marginal distributions

$$P(\mathbf{X}_{S_1}), P(\mathbf{X}_{S_k}), \ldots, P(\mathbf{X}_{S_k}) \quad \text{(“training data“)}$$

for $S_i \subset \{1, \ldots, n\}$

- use G to predict properties of previously unseen $P(\mathbf{X}_{S_{k+1}})$ (“test data”)

Why should we consider such a scenario?

"All models are wrong, but some are useful"

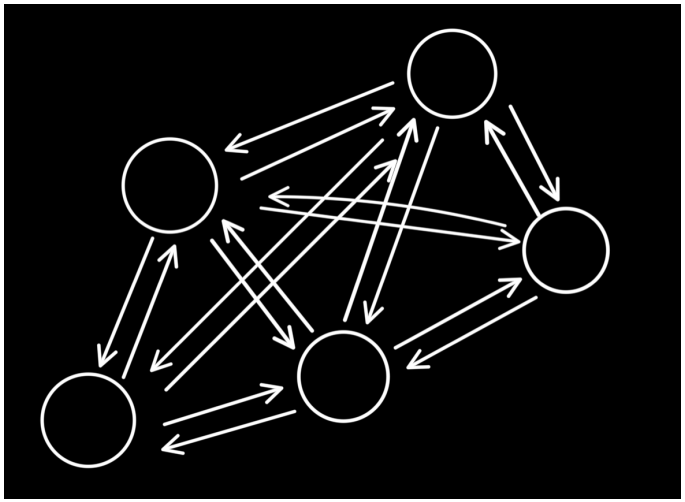
(statistician George Box, 1976)

- **Hypothesis:** causal models are helpful if they perform good at the above prediction task of merging distributions
- **Flaw in current discussions:** researcher in causal discovery seem to be believe in "the true DAG"

Intuitive conditions for causal models to be useful

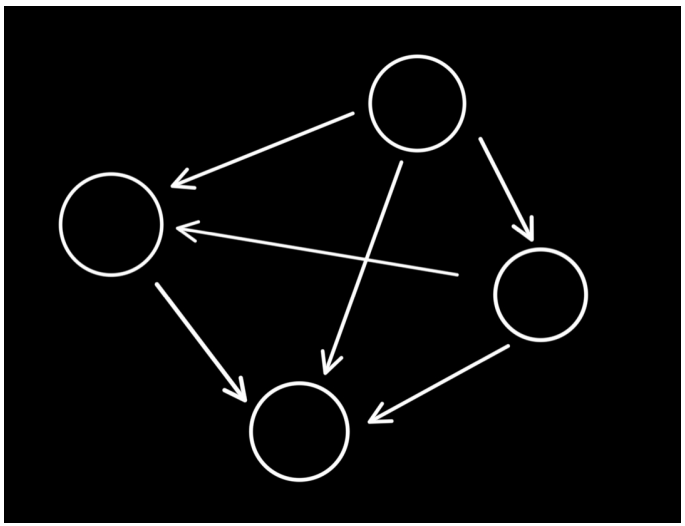
- no causal direction should be wrong
- sufficiently sparse graph to get actionable insights
- falsifiable predictions

Example for a "true" but useless causal model



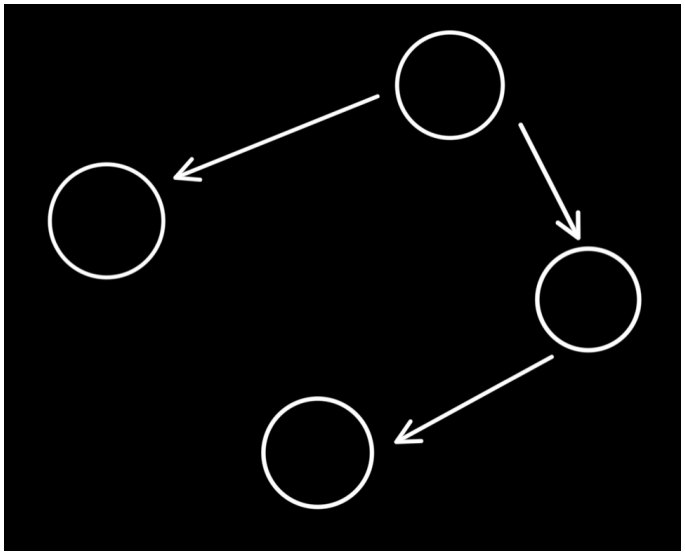
- no actionable insights except for "don't touch the system, you never know what happens"
- impossible to falsify even by interventions

Slightly more helpful causal model: complete DAG



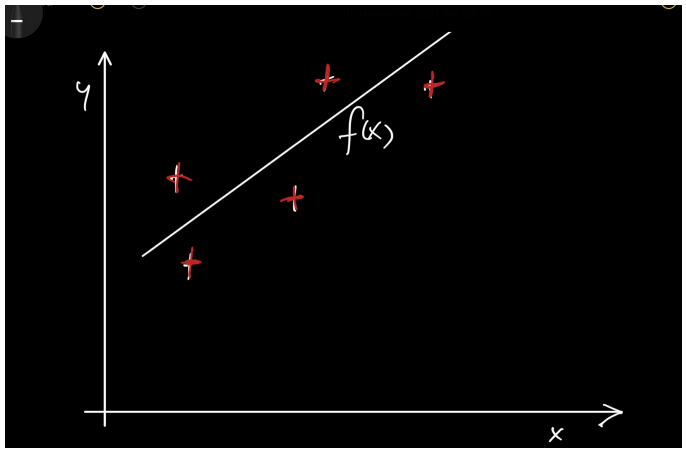
- falsifiable by interventions (affect only descendants)
- not falsifiable from passive observations (no conditional independences)

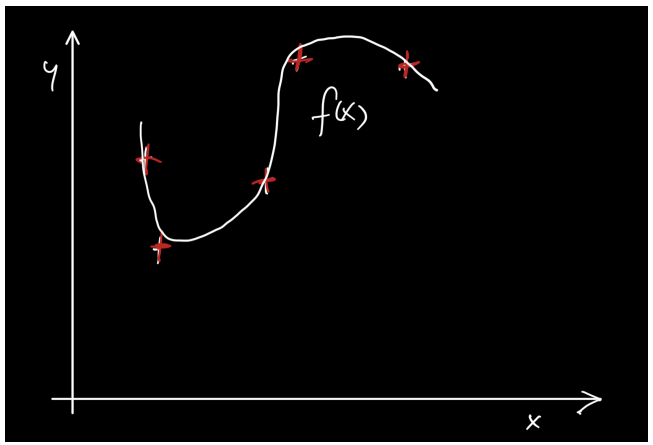
More useful after sparsification



- falsifiable from passive observations
- more actionable insights (explaining a disease by *one* cause is more helpful than attributing it to the entire lifestyle)

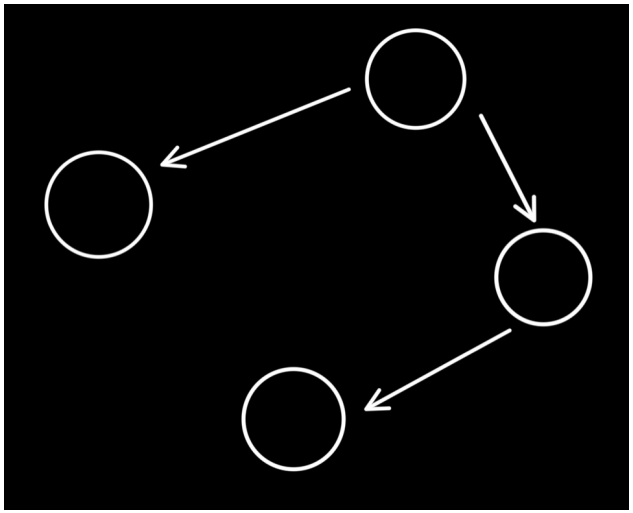
Recall: standard iid prediction tasks

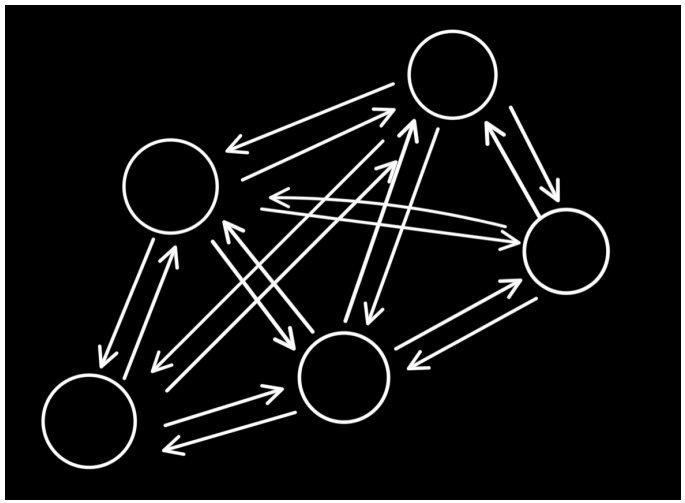




- only sufficiently simple models (that don't capture the full truth) are useful
- generalize from training to test data (guarantees from statistical learning theory for a given loss function, sample size, model class capacity)

"Overfitting" in causal discovery



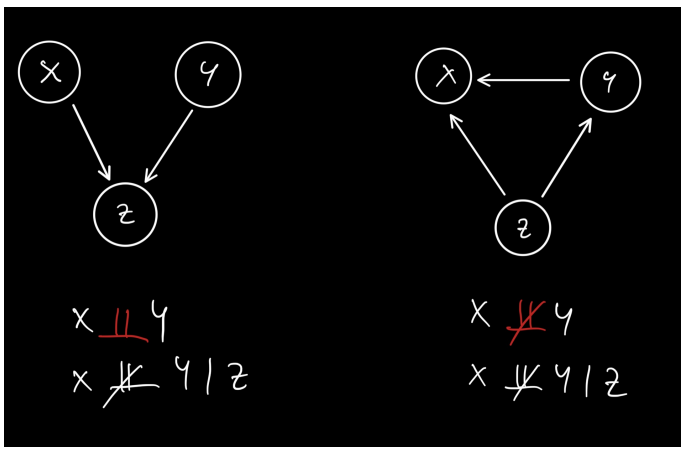


open question: in what sense does the left graph provide better predictions?

- loss function for causal models unclear
- capacity measure of model classes unclear
- prediction scenario unclear (what does 'generalization' mean? Learning causality is not primarily a finite sample problem)

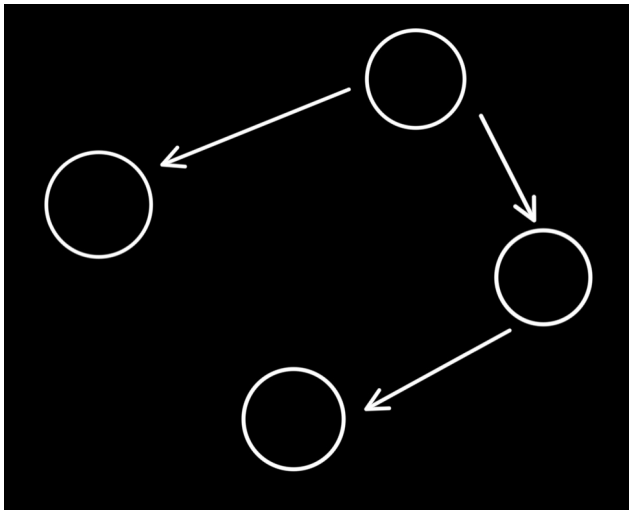
Finite sample issues of causal discovery on top

- relies on joint statistical properties that are hard to estimate from finite data (e.g. conditional independences)
- no clear advice for setting significance thresholds (inferred arrows can flip their direction)



- robustness against violating assumptions hardly understood
- joint iid data from n-variables rarely given, only observations from subsets of variables

Two different reasons for sparsification of DAGs



1. interpretable results
2. small sample sizes

lowering threshold for p-values with increasing sample size to achieve 1 obscures the difference between both reasons

Goal: statistical learning theory of causality

(generalization bounds that guide us in finding the right model complexity with respect to causal graphs)

- given some class $\mathcal{M}$ of causal graphs (e.g. DAGs)
- assume a model $M \in \mathcal{M}$ is consistent with a set of "training" marginals $P(\mathbf{X}_{S_1}), \ldots, P(\mathbf{X}_{S_k})$
- if the capacity of $\mathcal{M}$ is small enough, does M then correctly predict properties of $P(\mathbf{X}_{S_{k+1}})$ with high confidence, i.e. generalize to "test" marginals?

Example for a simple task: DAGs as binary classifiers

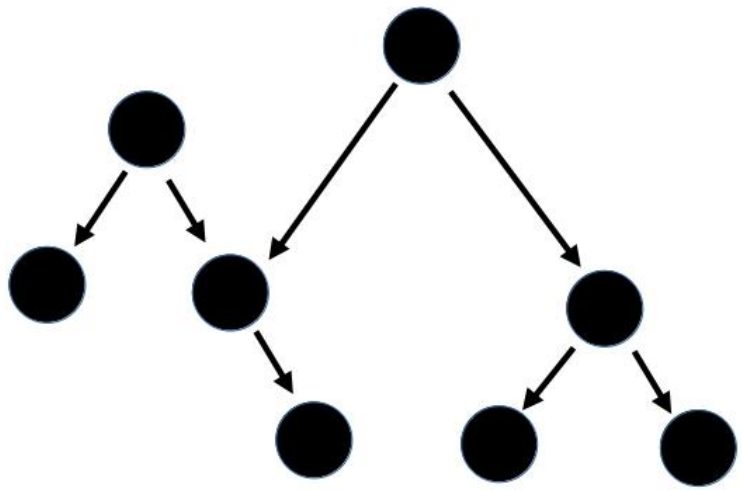
any DAG G with nodes $\mathbf{X} := (X_1, \ldots, X_n)$ can be seen as classifier on triples of sets of nodes:

$$G(\mathbf{X}_A, \mathbf{X}_B, \mathbf{X}_C) = 0 \quad \text{iff } \mathbf{X}_C \text{ d-separates } \mathbf{X}_B \text{ from } \mathbf{X}_A$$

Methodological justification of faithfulness: Markov condition alone can only predict independences, no dependences
(believe in faithfulness not because it is true but because the belief helps)

Example for a "small" causal model class

polytree DAGs: DAGs without undirected cycles (every two nodes are connected by at most one path)

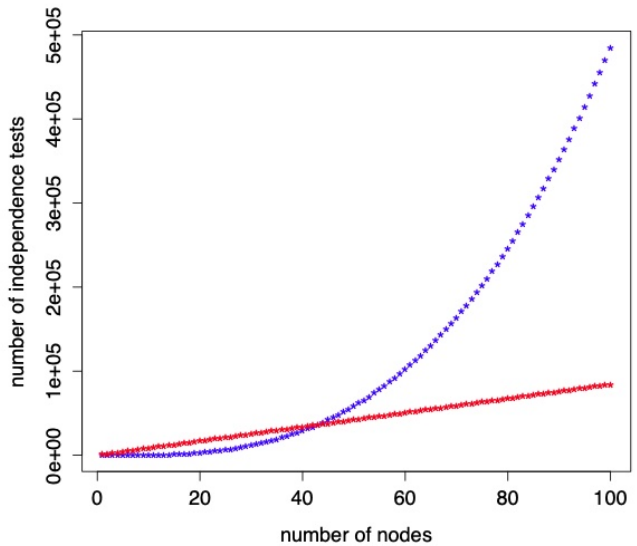


Lemma (DJ, 2018)

The set of polytree-DAGs is a class of classifiers on triples of random variables (via conditional independence) whose VC dimension is at most $n(\log_2 n + 1)$*, if* n *is the number of nodes.*

Janzing: Merging joint distributions via causal model classes with low VC dimension, arXiv:1804.03206

Generalization bound for predicting cond. independences



- **blue:** number of triples of variables
- **red:** number of triple conditional independence tests required to fit a polytree DAG such that Vapnik Chervonencis bounds from conventional statistical learning theory guarantee generalization to unobserved triples

Janzing: Merging joint distributions via causal model classes with low VC dimension, arXiv:1804.03206

Abandon the idea of absolute truth...

- irrelevant whether the inferred conditional independences are true (what does this mean btw?)

- irrelevant whether the inferred DAG is 'true' (what does this mean btw?)

VC bounds guarantee prediction of the outcome of the particular CI test chosen for the training triples
(regardless of quality of test and chosen p-value)

'Relative' causality

A DAG can be valuable for good predictions

- with respect to a particular level of complexity of dependences
  (e.g. partial correlations instead of conditional independences
  to capture linear relations only)

- with respect to a particular accuracy
  (e.g. reject independence only for very low p-values)

Did we forget about our goal of causal discovery?

is causality just about predicting statistical relations between variables that have not been observed together?

isn't causality actually defined via interventions?

Sure, but ...

assume we ask for the **causal** relation between X and Y

- add a variable I describing a randomized intervention on X and draw the arrow

- observing $I \perp\!\!\!\perp Y \,|X$ provides strong evidence for I causing Y only via X
- hence, causal structure is likely to be

- together with the causal consensus $I \to X$ and $I \not\to Y$, getting conditional independences right is almost sufficient to capture causality

Working hypothesis (with a grain of salt)

- an algorithm that correctly predicts statistical relations to variables from a sufficiently large consensus causal substructure is likely to capture causality in an interventional sense

- benchmarking causal discovery algorithm by just testing their predictions of joint statistical properties may be a good proxy for benchmarking causal results

- avoids the problem that interventions are sometimes hard to define (e.g. interventions with unknown intervention target, see also my arxiv preprint "Phenomenological Causality")

Thank you for your attention!

Some references:

1. S. Garido Mejia, E. Kirschbaum, D. Janzing: **Obtaining Causal Information by Merging Datasets with MAXENT**, AISTATS 2022.
2. D. Janzing: **Merging joint distributions via causal model classes with low VC dimension**, arXiv:1804.03206.
3. J. Peters, D. Janzing, B. Schölkopf: **Elements of Causal Inference**, 2017.
4. D. Janzing: **Causal versions of Maximum Entropy and Principle of Insufficient Reason**, JCI 2021
5. J. Peters, P. Bühlmann, N. Meinshausen: **Causal inference using invariant prediction: identification and confidence intervals**, 2016.
6. L. Gresele, J. von Kügelgen, J. Kübler, E. Kirschbaum, B. Schölkopf, D. Janzing: **Causal inference through the causal marginal problem**, ICML 2022.