Causal inference in degenerate systems: An impossibility result

Yue Wang

Department of Computational Medicine,
University of California, Los Angeles

Outline

- What is causal effect.
- Existing causal quantities and their problems.
- Criteria for a “good” causal quantity
- An impossibility theorem.
- Algorithms and simulations.

What is causal effect

- Heating with fire causes water to boil.
- HIV exposure causes AIDS.
- Smoking causes lung cancer.

What is causal effect

- If A and B are related, does A cause B?
- A real (indirect) causation: Driving without a seat belt can reduce the risk of cancer.
- Correlation does not imply causation.

- When the rooster crows, the sun rises. (B causes A)
- As ice cream sales increase, the rate of drowning deaths increases sharply. (A hidden factor C causes both A and B)

What is causal effect

- When A directly causes B, the causal relation can be further classified or quantified.
- Heating with fire causes water to boil. (Deterministic)
- HIV exposure causes AIDS. (Stochastic, strong effect)
- Smoking causes lung cancer. (Stochastic, weak effect)

What is causal effect

- Skip philosophical discussions (Hume et al.) of causal effect...

Purpose

- We have some random variables $X_1, X_2, \ldots, X_n, Y$.
- $X_1, \ldots, X_n$ (cause variables) are exactly all the possible direct causes of Y (result variable). We assume Y does not cause any X_i, and there is no hidden cause of Y.
- Our purpose is to quantify the effect of a direct causal relationship $X_1 \to Y$, based on the joint probability distribution of $X_1, X_2, \ldots, X_n, Y$.

Causal quantities

- There are various quantities that measure a causal relation.
- Correlation coefficient and its variations.
- Average treatment effect and its variations.
- Mutual information and its variations.
- Granger causality and transfer entropy (for stochastic processes).

- Idea: if X causes Y, then X contains information of Y. X has predict power on Y. Use information to quantify causal effect.
- Measure of information: entropy.

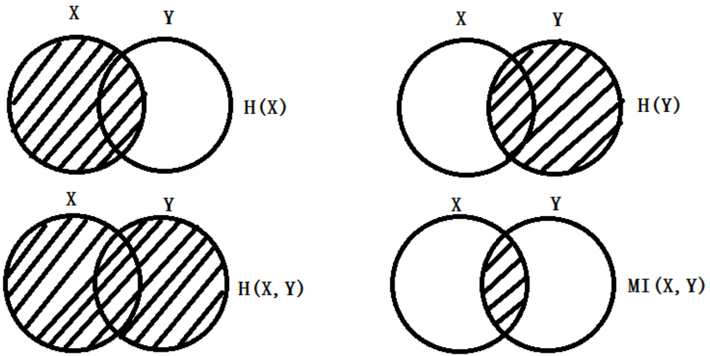
$$\mathrm{H}(X) = -\sum_i p_i \log p_i,$$

where $p_i = \mathbb{P}(X = x_i)$.

- $\mathrm{H}(X) \geq 0$. Equality holds if and only if X is deterministic.

Mutual information (MI)

$$\mathrm{MI}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y).$$



Mutual information (MI)

- Intuition: the information shared between X and Y. The information gain of Y if we know X. The predict power of X on Y.
- If X causes Y, then $\textsc{mi}(X, Y)$ can be used to describe the causal effect of $X \to Y$.
- $\textsc{mi}(X, Y) \geq 0$. Equality holds if and only if X and Y are independent.

Conditional Mutual information (CMI)

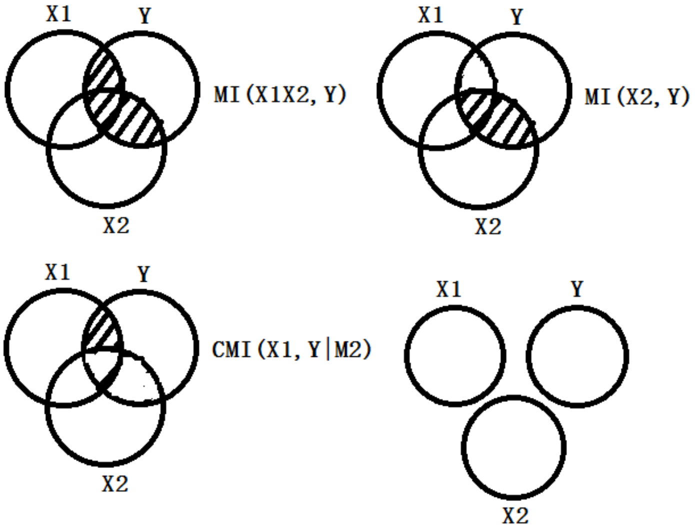
- Generalize MI for more variables.

$$\text{CMI}(X_1, Y \mid X_2) = \text{MI}(X_1 X_2, Y) - \text{MI}(X_2, Y).$$

- Conditioned on the knowledge of X_2, how much extra information of Y could X_1 provide.
- Can be used to describe the causal effect of $X_1 \to Y$ if X_1 and X_2 cause Y.
- $\text{CMI}(X_1, Y \mid X_2) \geq 0$. Equality holds if and only if X_1 and Y are independent conditioned on X_2. This means that with the knowledge of X_2, X_1 contains no new knowledge of Y.

Conditional Mutual information (CMI)

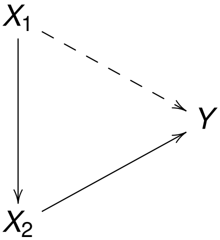
$$\mathrm{CMI}(X_1, Y \mid X_2) = \mathrm{MI}(X_1 X_2, Y) - \mathrm{MI}(X_2, Y).$$

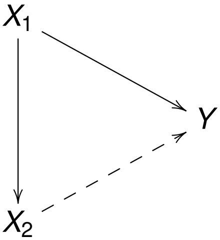


Conditional Mutual information (CMI)

- Venn diagram does not work.
- X_1, X_2 are independent variables, with equal probabilities to take 0 or 1. $Y = X_1 + X_2 \mod 2$.
- Any two of X_1, X_2, Y are independent, but these two could determine the third variable.
- $\textsc{mi}(X_1, Y) = 0$, $\textsc{cmi}(X_1, Y \mid X_2) > 0$.

Problem of CMI

- What if $\mathbb{P}(X_1 = X_2) \approx 1$?
- 



  $X_2 = X_1 + \epsilon$, $Y = X_2 + \delta$ $\quad$ $X_2 = X_1 + \epsilon$, $Y = X_1 + \delta$

  $X_1 \perp\!\!\!\perp Y \mid X_2$ $\quad$ $X_2 \perp\!\!\!\perp Y \mid X_1$

  ϵ and δ are independent, equal 0 with high probabilities.
- $\textsc{cmi}(X_1, Y \mid X_2)$ is 0 in the first case, and 0.0065 in the second case.

Problem of CMI

Table: Joint distributions of X_1, X_2, Y in two cases

X_1	X_2	Y	Case 1	Case 2
0	0	0	0.4990005	0.4990005
0	0	1	0.0004995	0.0004995
1	1	0	0.0004995	0.0004995
1	1	1	0.4990005	0.4990005
0	1	0	0.0000005	0.0004995
0	1	1	0.0004995	0.0000005
1	0	0	0.0004995	0.0000005
1	0	1	0.0000005	0.0004995

New methods

- Utilize the slight difference between X_1 and X_2.
- Causal strength (CS):
  D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. Ann. Stat., 41(5):2324–2358, 2013.
- $$\mathrm{CS}(X_1, Y) = \sum_{x_1, x_2, y} \mathbb{P}(x_1, x_2, y) \log \frac{\mathbb{P}(y \mid x_1, x_2)}{\sum_{x_1'} \mathbb{P}(y \mid x_1', x_2)\mathbb{P}(x_1')}.$$

- Part mutual information (PMI):
  J. Zhao, Y. Zhou, X. Zhang, and L. Chen. Part mutual information for quantifying direct associations in networks. Proc. Natl. Acad. Sci., 113(18):5130–5135, 2016.
-

$$\mathrm{PMI}(X_1, Y \mid X_2) =$$

$$\sum_{x_1, x_2, y} \mathbb{P}(x_1, x_2, y) \log \frac{\mathbb{P}(x_1, y \mid x_2)}{\sum_{x_1'} \mathbb{P}(y \mid x_1', x_2)\mathbb{P}(x_1') \sum_{y'} \mathbb{P}(x_1 \mid x_2, y')\mathbb{P}(y')}.$$

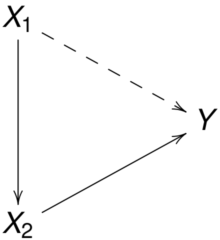
$$\begin{tikzcd} X_1 \arrow[dd] \arrow[dr, dashed] & \\ & Y \\ X_2 \arrow[ur] & \end{tikzcd} \qquad \begin{tikzcd} X_1 \arrow[dd] \arrow[dr] & \\ & Y \\ X_2 \arrow[ur, dashed] & \end{tikzcd}$$

$X_2 = X_1 + \epsilon$, $Y = X_2 + \delta$ $\qquad$ $X_2 = X_1 + \epsilon$, $Y = X_1 + \delta$

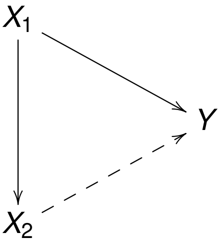
$X_1 \perp\!\!\!\perp Y \mid X_2$ $\qquad$ $X_2 \perp\!\!\!\perp Y \mid X_1$

In the first case, $\textsc{cmi}(X_1, Y \mid X_2)$, $\textsc{cs}(X_1, Y)$ and $\textsc{pmi}(X_1, Y \mid X_2)$ are 0. In the second case, $\textsc{cmi}(X_1, Y \mid X_2)$, $\textsc{cs}(X_1, Y)$ and $\textsc{pmi}(X_1, Y \mid X_2)$ are 0.0065, 0.6852, 0.9661.

Problem of new methods

- 

  $X_2 = X_1 + \epsilon,\ Y = X_2 + \delta$
  $X_1 \perp\!\!\!\perp Y \mid X_2$

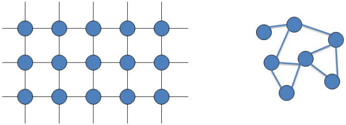
  

  $X_2 = X_1 + \epsilon,\ Y = X_1 + \delta$
  $X_2 \perp\!\!\!\perp Y \mid X_1$
- These two joint distributions are almost the same, but the resulting causal effects are very different.
- CS and PMI may not be continuous with joint distribution (under total variation distance) when both $\{X_1\}$ and $\{X_2\}$ have all the information of Y contained in $\{X_1, X_2\}$.

Markov boundary (MB)

- Markov chain $Z(t-2) \to Z(t-1) \to Z(t) \to Z(t+1)$. Use $\mathcal{S} = \{Z(t-2), Z(t-1), Z(t)\}$ to predict $Y = Z(t+1)$: $\{Z(t)\}$ is enough.
- Markov random field: neighbors are enough to predict one variable.

Markov Random Fields



Can be generalized to any **undirected** graphs (nodes, edges)
Neighborhood system: each node is connected to its neighbors
neighbors are reciprocal
Markov property: each node only depends on its neighbors

Note: the black lines on the left graph are illustrating the 2D grid for the image pixels
they are not edges in the graph as the blue lines on the right

Markov boundary (MB)

Markov boundary of Y within $\mathcal{S} = \{X_1, \ldots, X_n\}$: $\mathcal{S}_1 \subset \mathcal{S}$, which is minimal, and has the same information on Y. (Remove all redundant variables with no predict power.)

$$\text{CMI}(\mathcal{S} \backslash \mathcal{S}_1, Y \mid \mathcal{S}_1) = 0,$$

$$\forall \mathcal{S}_2 \subsetneqq \mathcal{S}_1, \quad \text{CMI}(\mathcal{S} \backslash \mathcal{S}_2, Y \mid \mathcal{S}_2) > 0.$$

This means $Y \perp\!\!\!\perp \mathcal{S} \backslash \mathcal{S}_1 \mid \mathcal{S}_1$.

Markov boundary (MB)

- MB may not be unique. (Set $X_1 = X_2$ in the above example, then both $\{X_1\}$ and $\{X_2\}$ are MB of Y within $\{X_1, X_2\}$.)
- Statnikov et al. (2013) studied 13 benchmark data sets from reality, and found that five of them have multiple MB.
- Assume MB is unique. A cause variable inside MB has non-zero unique information of Y, and a cause variable outside MB has zero unique information of Y. Therefore the unique MB should be exactly all the cause variables with non-zero causal effect.

Problem of new methods

- When there are multiple MB (called “degenerate”), both CS and PMI are not directly defined. (Contains $0/0$ in the expression.)
- Try to use continuation: choose a sequence of distributions (for which CS and PMI are defined) converging to the original distribution, and check whether the corresponding CS and PMI converge.

Problem of new methods

Consider a probability distribution $\mathfrak{p}$ on $\mathcal{S} \cup Y$, under which Y has multiple MB in $\mathcal{S}$, and X_1 is in at least one, but not all of such MB. Now $\textsc{cmi}(X_1, Y \mid \mathcal{S} \backslash \{X_1\}) = 0$.

Theorem (Wang & Wang, 2020)

There exist constants $c_1 < c_2$, such that in any arbitrarily small neighborhood of $\mathfrak{p}$, $\textsc{cs}(X_1 \to Y)$ takes any value between c_1 and c_2. The same applies to PMI.

Problem of new methods

- When there are multiple MB, both CS and PMI cannot be well-defined.
- Similar to the behavior of a complex analytical function near an essential singularity (Picard's great theorem).
- It is not numerically feasible to calculate CS and PMI in such cases.

Sketch of the proof

Lemma (Wang & Wang, 2020)

Assume Y has multiple MB, X_1 is inside some MB but not all MB, and set X_2 to be all other variables. Then there exist x_1', x_2 such that $\mathbb{P}(x_1') > 0$, $\mathbb{P}(x_2) > 0$, $\mathbb{P}(x_1', x_2) = 0$.

Under small perturbations, $\mathbb{P}(x_1', x_2)$ is very small but positive, so that CS and PMI can be defined, but $\mathbb{P}(y \mid x_1', x_2)$ can change significantly.

Sketch of the proof

- $$\text{CS}(X_1, Y) = \sum_{x_1, x_2, y} \mathbb{P}(x_1, x_2, y) \log \frac{\mathbb{P}(y \mid x_1, x_2)}{\sum_{x_1'} \mathbb{P}(y \mid x_1', x_2)\mathbb{P}(x_1')}.$$

- $$\text{PMI}(X_1, Y \mid X_2) =$$
$$\sum_{x_1, x_2, y} \mathbb{P}(x_1, x_2, y) \log \frac{\mathbb{P}(x_1, y \mid x_2)}{\sum_{x_1'} \mathbb{P}(y \mid x_1', x_2)\mathbb{P}(x_1') \sum_{y'} \mathbb{P}(x_1 \mid x_2, y')\mathbb{P}(y')}.$$

- If $\mathbb{P}(x_1')$ and $\mathbb{P}(x_2)$ are not small, but $\mathbb{P}(x_1', x_2)$ is very small, then changing $\mathbb{P}(y \mid x_1', x_2)$ has a significant impact on CS and PMI, but the whole distribution is perturbed slightly.

Sketch of the proof

- Construct two sequences of distributions, both of which converge to the original distribution.
- CS (or PMI) of two sequences always exist, but converge to different values.
- Distribution of one sequence can continuously transform into distribution of the other sequence, during which CS (or PMI) is always defined.
- When there are multiple MB, CS and PMI are not well defined, and CMI is zero.
- Questions?

Purpose

- We have cause variables $\mathcal{S} = \{X_1, X_2, \ldots, X_n\}$ and result variable Y.
- Our purpose is to quantify the effect of a causal relationship $X_1 \to Y$, based on the joint distribution of $X_1, X_2, \ldots, X_n, Y$
- Existing causal quantities have different problems.
- We propose several criteria for a "good" causal quantity.
- We focus on the case where MB is unique. In such case, the unique MB should be exactly all the variables with non-zero causal effect.

Criteria for quantifying causal effect

- C1. The strength of $X \to Y$ is a continuous function of the joint distribution of Y and $\mathcal{S}$, under the total variation distance.
- C2. If there is a unique MB $\mathcal{M}$ of Y within $\mathcal{S}$, and $X \notin \mathcal{M}$, then the strength of $X \to Y$ is 0.
- C3. If there is a unique MB $\mathcal{M}$ of Y within $\mathcal{S}$, and $X \in \mathcal{M}$, then the absolute value of the strength of $X \to Y$ is at least $\text{CMI}(X, Y \mid \mathcal{M} \setminus \{X\})$.
- In C3, $\text{CMI}(X, Y \mid \mathcal{M} \setminus \{X\})$ can be replaced by any positive constant, which only depends on $X, Y, \mathcal{M} \setminus \{X\}$.

Criteria for quantifying causal effect

- CMI satisfies C1 and C2.
- CS and PMI satisfies C2 and C3.
- A naive causal effect measure that takes a large positive constant value (such as 2H(Y)) satisfies C1 and C3.

An impossibility theorem

Consider a probability distribution $\mathfrak{p}$ on $\mathcal{S} \cup Y$, under which Y has multiple MB in $\mathcal{S}$, and X is in at least one, but not all of such MB. We want to define a quantity for $X \to Y$ that can be calculated from $\mathfrak{p}$.

Theorem (Wang & Wang, 2020)

In any neighborhood $\mathfrak{N}$ of $\mathfrak{p}$, all identifiable measures of the strength of $X \to Y$ must violate at least one of the criteria in C1 – C3.

Sketch of proof

For variable X_1, we define X_1 with ϵ-noise to be X_1^ϵ, which equals X_1 with probability $1-\epsilon$, and equals an independent noise with probability ϵ. Denote all cause variables by $\mathcal{S}$.

Lemma (Strict Data Processing Inequality, Wang & Wang, 2020)

$\mathcal{S}_1$ is a group of variables without X_1, Y. If we add ϵ-noise on X_1 to get X_1^ϵ, then $\mathrm{CMI}(X_1^\epsilon, Y \mid \mathcal{S}_1) \leq \mathrm{CMI}(X_1, Y \mid \mathcal{S}_1)$, and the equality holds if and only if $\mathrm{CMI}(X_1, Y \mid \mathcal{S}_1) = 0$.

Lemma (Wang & Wang, 2020)

Assume Y has multiple MB. For one MB $\mathcal{M}_0$, if we add ϵ noise on all variables of $\mathcal{S} \backslash \mathcal{M}_0$, then in the new distribution, $\mathcal{M}_0$ is the unique MB.

Sketch of proof

- Assume $X_1 \in \mathcal{M}_0$, $X_1 \notin \mathcal{M}_1$ for MB $\mathcal{M}_0, \mathcal{M}_1$.
- We can add ϵ-noise on $\mathcal{S} \backslash \mathcal{M}_1$, such that $\mathcal{M}_1$ is the unique MB. Criterion C2 shows that the effect of $X_1^\epsilon \to Y$ is 0.
- We can add ϵ-noise on $\mathcal{S} \backslash \mathcal{M}_0$, such that $\mathcal{M}_0$ is the unique MB. Criterion C3 shows that the absolute value of the effect of $X_1 \to Y$ is at least $\text{CMI}(X_1, Y \mid \mathcal{M}_0 \backslash \{X_1\}) > 0$.
- Let $\epsilon \to 0$. Criterion C1 shows that the absolute value of the effect of $X_1 \to Y$ should be at least $\text{CMI}(X_1, Y \mid \mathcal{M}_0 \backslash X)$, and should be 0.

Summary

- Quantifying causal effect with multiple MB is an essentially ill-posed problem.
- When a distribution with unique MB is close to another distribution with multiple MB, a reasonable causal quantity can be very small (CMI) or fluctuate violently (CS, PMI). Therefore in such case, quantitative method is not feasible.
- Questions?

Uniqueness of MB

- In practice, whether should we apply quantitative methods?
- We need practical methods to determine the uniqueness of MB.
- First step: algorithms that can produce one MB.
- Known methods: IAMB, KIAMB, Semi-Interleaved HITON-PC, MBOR, BLCD, PCMB, GLL-PC. All of them have additional requirements on the joint distribution (to exclude morbid cases).
- We propose Algorithm 1, using the idea of discarding redundant variables one by one.

Algorithm 1: An assumption-free algorithm for producing one MB

(1) **Input**

Joint distribution of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y

(2) **Set** $\mathcal{M}_0 = \mathcal{S}$

(3) **Repeat**

Set $X_0 = \arg\min_{X \in \mathcal{M}_0} \mathrm{CMI}(X, Y \mid \mathcal{M}_0 \setminus \{X\})$

If $X_0 \perp\!\!\!\perp Y \mid \mathcal{M}_0 \setminus \{X_0\}$

Set $\mathcal{M}_0 = \mathcal{M}_0 \setminus \{X_0\}$

Until $X_0 \not\perp\!\!\!\perp Y \mid \mathcal{M}_0 \setminus \{X_0\}$

(4) **Output** $\mathcal{M}_0$ is a MB

Uniqueness of MB

- Theoretical results on the uniqueness of MB.
- Construct the "essential set" $\mathcal{E}$: $X_i \in \mathcal{E}$ if and only if $\text{MI}(\mathcal{S} \backslash \{X_i\}, Y) < \text{MI}(\mathcal{S}, Y)$.

Lemma (Wang & Wang, 2020)

$\mathcal{E}$ is the intersection of all MB.

Theorem (Wang & Wang, 2020)

MB is unique if and only if $\mathcal{E}$ is an MB.

- Based on such results, we have several algorithms of determining the uniqueness of MB.

Uniqueness of MB

- Algorithm 2: use another algorithm to produce one MB, and check whether each variable in this MB is essential.
- The checking is not very straightforward, since we want to avoid testing conditional independence for too many variables.
- This is equivalent to executing TIE* algorithm until producing the second MB, or finishing with the unique MB.

Uniqueness of MB

Algorithm 2: A general algorithm for determining uniqueness of MB

(1) **Input**

Joint distribution of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y

An algorithm Ω which could produce one MB correctly

(2) **Set** $\mathcal{M}_0 = \{X_1, \ldots, X_m\}$ to be the result of Algorithm Ω on $\mathcal{S}$

(3) **For** $i = 1, \ldots, m$,

Set $\mathcal{M}_i$ to be the result of Algorithm Ω on $\mathcal{S} \setminus \{X_i\}$

If $Y \perp\!\!\!\perp \mathcal{M}_0 \mid \mathcal{M}_i$

Output Y has multiple MB

Terminate

(4) **Output** Y has a unique MB

Uniqueness of MB

- Use Algorithm 1 as "Ω" in Algorithm 2, to get "Alg. 2-AF", an assumption free algorithm for determining MB uniqueness.
- Use KIAMB as "Ω" in Algorithm 2, to get "Alg. 2-KI", an algorithm for determining MB uniqueness (requires composition property).
- We propose two more algorithms for better comparison.

Uniqueness of MB

Algorithm S1: An assumption-free algorithm for determining the uniqueness of MB

Using Algorithm 1 to find one MB, and directly check whether each variable in this MB is essential.

(1) **Input**

Joint distribution of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y

(2) **Set** $\mathcal{M}_0 = \{X_1, \ldots, X_m\}$ to be the result of Algorithm 1 on $\mathcal{S}$

(3) **For** $i = 1, \ldots, m$,

If $X_i \perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$

Output Y has multiple MB

Terminate

(4) **Output** Y has a unique MB

Algorithms

Algorithm S2: An assumption-free algorithm for determining the uniqueness of MB

Directly build the essential set, and check if it is an MB.

(1) **Input**

Joint distribution of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y

(2) **Set** $\hat{\mathcal{E}} = \emptyset$

(3) **For** $i = 1, \ldots, k$,

If $\quad X_i \not\perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$

$\hat{\mathcal{E}} = \hat{\mathcal{E}} \cup \{X_i\}$

(4) **If** $Y \perp\!\!\!\perp \mathcal{S} \mid \hat{\mathcal{E}}$

output: Y has a unique MB

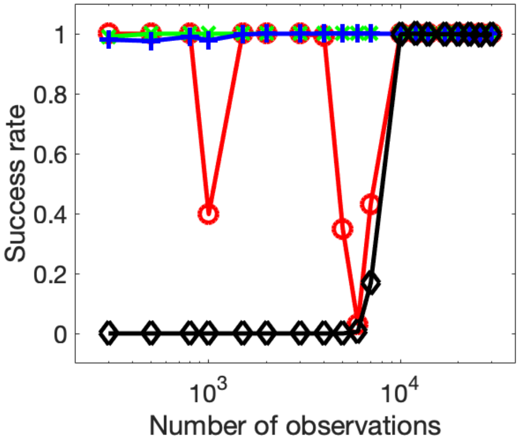
Else

output: Y has multiple MB

Simulation setup

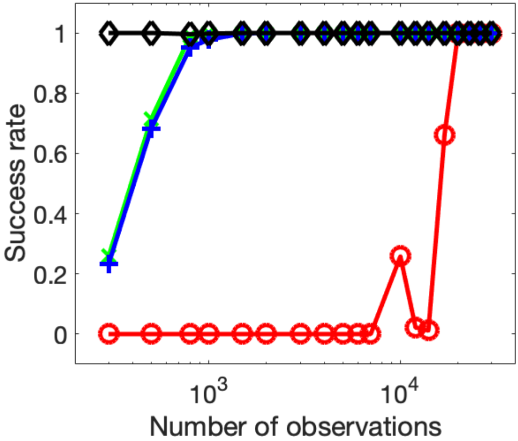
- Implement Alg. 2-AF, Alg. 2-KI, Alg. S1 and Alg. S2.
- Test on four artificial cases.
- In Case 1 and Case 3, MB is unique. In Case 2 and Case 4, MB is not unique.
- In Case 3 and Case 4, the assumption of KIAMB, i.e. the composition property, is failed (thus Alg. 2-KI is failed).

Algorithms performances



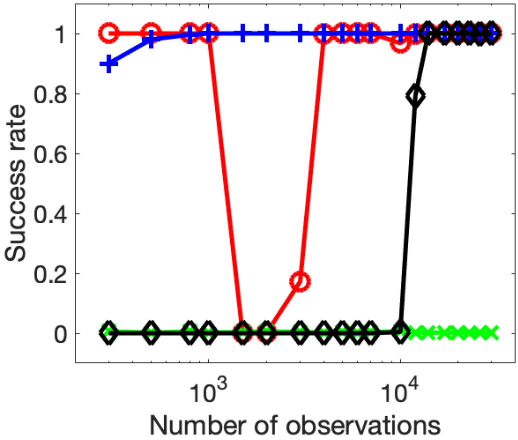
Case 1. Success rates of Alg. 2-AF (blue '+'); Alg. 2-KI (green '×'); Alg. S1 (black '◇'); Alg. S2 (red '○') with different numbers of observations. Number of observations is in logarithm.

Algorithms performances



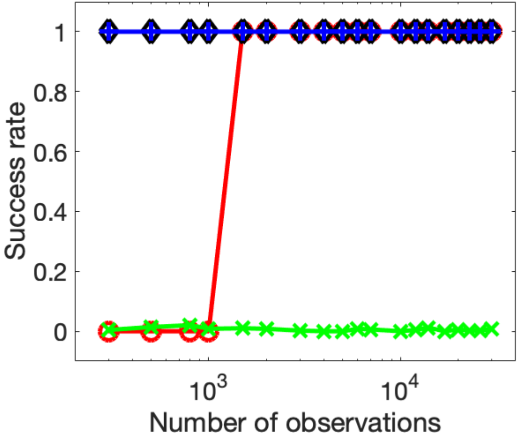
Case 2. Success rates of Alg. 2-AF (blue '+'); Alg. 2-KI (green '×'); Alg. S1 (black '◇'); Alg. S2 (red '○') with different numbers of observations. Number of observations is in logarithm.

Algorithms performances



Case 3. Success rates of Alg. 2-AF (blue '+'); Alg. 2-KI (green '×'); Alg. S1 (black '◇'); Alg. S2 (red '○') with different numbers of observations. Number of observations is in logarithm.

Algorithms performances



Case 4. Success rates of Alg. 2-AF (blue '+'); Alg. 2-KI (green '×'); Alg. S1 (black '◇'); Alg. S2 (red '○') with different numbers of observations. Number of observations is in logarithm.

Algorithms performances

- Performance: In Case 1 and Case 2, where the composition property holds (KIAMB is valid), Alg. 2-KI is slightly better than Alg. 2-AF, and both are much better than Alg. S1 and Alg. S2.
- Performance: In Case 3 and Case 4, where the composition property fails, Alg. 2-KI fails to produce correct results, while Alg. 2-AF exhibits the best performance.
- In practice, if one has a strong belief in the composition property, then we recommend Alg. 2-KI. Otherwise Alg. 2-AF is preferable.

- Questions?

Summary

- When there are multiple MB, CS and PMI are not well-defined.
- When there are multiple MB, we cannot define a “good” causal quantity. (Use qualitative methods instead.)
- Theoretical and practical methods for determining the uniqueness of MB.

- Be aware of multiple MB!

References

- Wang, Y., & Wang, L. (2020, June). Causal inference in degenerate systems: An impossibility result. In International Conference on Artificial Intelligence and Statistics (pp. 3383-3392). PMLR.
- Dobrushin, R. L. (1963). General formulation of Shannon's main theorem in information theory. American mathematical society translations, 33, 323-438.
- Janzing, D., Balduzzi, D., Grosse-Wentrup, M., & Schölkopf, B. (2013). Quantifying causal influences.

References

- Zhao, J., Zhou, Y., Zhang, X., & Chen, L. (2016). Part mutual information for quantifying direct associations in networks. Proceedings of the National Academy of Sciences, 113(18), 5130-5135.
- Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann.
- Peña, J. M., Nilsson, R., Björkegren, J., & Tegnér, J. (2007). Towards scalable and data efficient learning of Markov boundaries. International Journal of Approximate Reasoning, 45(2), 211-232.
- Statnikov, A., Lytkin, N. I., Lemeire, J., & Aliferis, C. F. (2013). Algorithms for discovery of multiple Markov boundaries. Journal of Machine Learning Research, 14(Feb), 499-566.

Thank you!