An Impossibility Theorem of Quantifying Causal Effect

Yue Wang

Department of Applied Mathematics
University of Washington
yuewang@uw.edu
Institut des Hautes Études Scientifiques (IHÉS)

June 25, 2018

Collaborator

- This work is collaborated with Dr. Linbo Wang at Department of Biostatistics, Harvard University.
- Full paper can be found at https://arxiv.org/abs/1711.04466

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. (Deterministic)
- HIV exposure causes AIDS. (Stochastic, strong effect)
- Smoking causes lung cancer. (Stochastic, weak effect)

What is causal effect

- Skip 100 pages of philosophical discussions of causal effect...

Purpose

- We have some random variables $X_1, X_2, \cdots, X_n, Y$.
- $X_1, \cdots, X_n$ (cause variables) are exactly all the possible direct causes of Y (result variable). We assume there is no hidden cause of Y.
- Our purpose is to quantify the effect of a causal relationship $X_1 \to Y$, based on the joint probability distribution of $X_1, X_2, \cdots, X_n, Y$.

Information theory

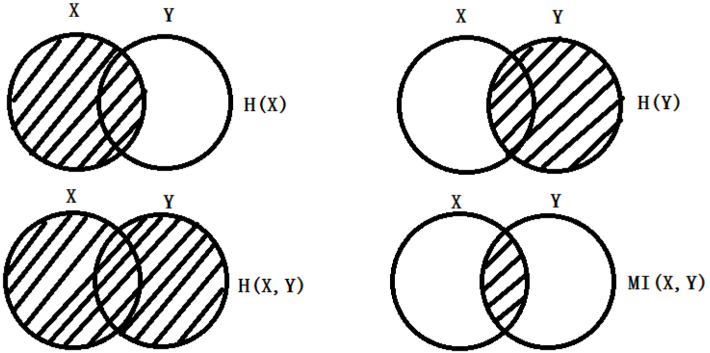
- Idea: if X causes Y, then X contains information of Y. X has predict power on Y. Use information to quantify causal effect.
- Measure of information: entropy.

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

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 $\mathrm{MI}(X, Y)$ can be used to describe the causal effect of $X \to Y$.
- $\mathrm{MI}(X, Y) \geq 0$. Equality holds if and only if X and Y are independent.

Conditional Mutual information (CMI)

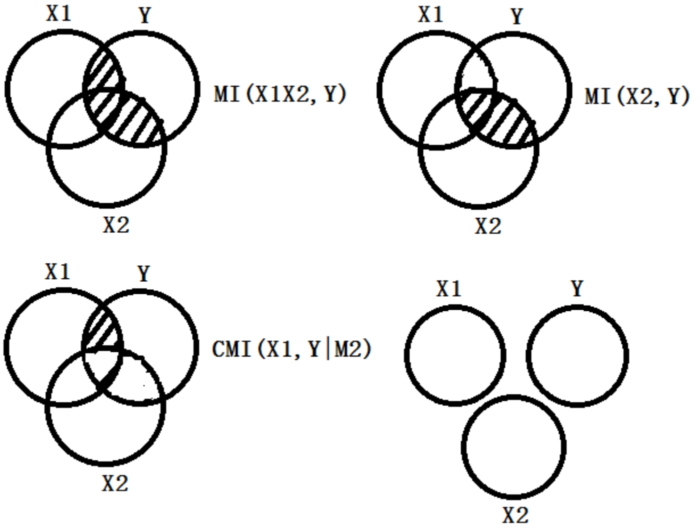
- Generalize MI for more variables.

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

- 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.
- $\mathrm{CMI}(X_1, Y \mid X_2) \geq 0$. 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.
- $\mathrm{MI}(X_1, Y) = 0$, $\mathrm{CMI}(X_1, Y \mid X_2) > 0$.

Problem of CMI

- What if $\mathbb{P}(X_1 = X_2) \approx 1$?
- $$\begin{tikzcd} X_1 \arrow[dd] \arrow[dr, dashed] & \\ & Y \\ X_2 \arrow[ur] & \end{tikzcd} \qquad \begin{tikzcd} X_1 \arrow[dd] \arrow[dr] & \\ & Y \\ X_2 \arrow[ur, dashed] & \end{tikzcd}$$

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

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

  ϵ and δ are independent, equal 0 with high probabilities.
- $\mathrm{CMI}(X_1, Y \mid X_2)$ 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

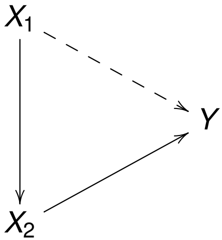
New methods

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

New methods

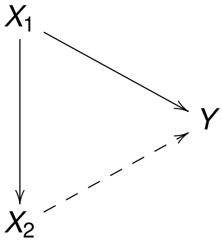
- Part mutual information (PMI):
  J. Zhao, Y. Zhou, X. Zhang, and L. Chen. Part mutual information for quantifying direct associations in networks. Proc. Natl. Acad. Sci., 113(18):5130–5135, 2016.
-
$$\mathrm{PMI}(X_1, Y \mid X_2) = \sum_{x_1, x_2, y} \mathbb{P}(x_1, x_2, y) \log \frac{\mathbb{P}(x_1, y \mid x_2)}{\sum_{x_1'} \mathbb{P}(y \mid x_1', x_2)\mathbb{P}(x_1') \sum_{y'} \mathbb{P}(x_1 \mid x_2, y')\mathbb{P}(y')}.$$

New methods



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

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



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

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

In the first case, $\mathrm{CMI}(X_1, Y \mid X_2)$, $\mathrm{CS}(X_1, Y)$ and $\mathrm{PMI}(X_1, Y \mid X_2)$ are 0. In the second case, $\mathrm{CMI}(X_1, Y \mid X_2)$, $\mathrm{CS}(X_1, Y)$ and $\mathrm{PMI}(X_1, Y \mid X_2)$ are 0.0065, 0.6852, 0.9661.

Problem of new methods

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

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

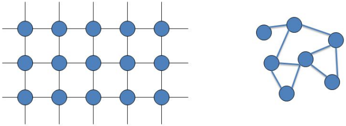
  $X_1 \perp\!\!\!\perp Y \mid X_2$ $\qquad$ $X_2 \perp\!\!\!\perp Y \mid X_1$

- 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 $\mathcal{S} = \{X_1, \cdots, X_n\}$: $\mathcal{S}_1 \subset \mathcal{S}$, which is minimal, and has the same predict power with $\mathcal{S}$. (Remove all redundant variables with no predict power.)

$$\mathrm{MI}(\mathcal{S}_1, Y) = \mathrm{MI}(\mathcal{S}, Y),$$

$$\forall \mathcal{S}_2 \subsetneq \mathcal{S}_1, \quad \mathrm{MI}(\mathcal{S}_2, Y) < \mathrm{MI}(\mathcal{S}, Y).$$

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.)
- Assume MB is unique. A cause variable inside MB has positive irreplaceable predict power of Y, and a cause variable outside MB has zero irreplaceable predict power of Y. Therefore the unique MB should be exactly all the cause variables with positive causal effect.

Problem of new methods

- When there are multiple MB, 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

- Y. Wang and L. Wang. On the boundary between qualitative and quantitative methods for causal inference. arXiv:1711.04466.

Theorem (Wang & Wang, 2017)

In any arbitrarily small neighborhood of a distribution with multiple MB, CS (also PMI) takes any value in an interval with positive length.

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).
- Calculating CS and PMI in such case is not numerically feasible.

Sketch of the proof

Lemma (Wang & Wang, 2017)

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

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

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

- 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.

Purpose

- We have cause variables $\mathcal{S} = \{X_1, X_2, \cdots, X_n\}$ and result variable Y.
- Our purpose is to quantify the effect of a causal relationship $X_1 \rightarrow Y$.
- Current 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 positive causal effect.

Criteria for quantifying causal effect

- C0. The effect of $X_1 \to Y$ is identifiable from the joint distribution of cause variables and result variable. (No extra information needed.)
- C1. If there is unique MB $\mathcal{M}$, and $X_1 \notin \mathcal{M}$, then the effect of $X_1 \to Y$ is 0. (X_1 is totally redundant.)
- C2. If there is unique MB $\mathcal{M}$, and $X_1 \in \mathcal{M}$, then the effect of $X_1 \to Y$ is at least CMI$(X_1, Y \mid \mathcal{M} \backslash \{X_1\})$.
- C3. The effect of $X_1 \to Y$ is a continuous function of the joint distribution.
- CMI fails in C2. CS and PMI fail in C3.

An impossibility theorem

Theorem (Wang & Wang, 2017)

Assume in a distribution, Y has multiple MB. X_1 belongs to at least one MB, but not all MB. Then in any neighborhood of this distribution, the effect of $X_1 \to Y$ cannot be defined while satisfying criteria C0–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, 2017)

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

Lemma (Wang & Wang, 2017)

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 C1 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 C2 shows that the effect of $X_1 \to Y$ is at least $\mathrm{CMI}(X_1, Y \mid \mathcal{M}_0 \backslash \{X_1\}) > 0$.
- Let $\epsilon \to 0$. Criterion C3 shows that the effect of $X_1 \to Y$ should be at least $\mathrm{CMI}(X_1, Y \mid \mathcal{M}_0 \backslash X)$, 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 is either very small (CMI) or fluctuate violently (CS, PMI). Therefore in such case, quantitative method is not feasible.
- A practical problem: detecting whether MB is unique from data.

Uniqueness of Markov boundary

- We need theoretical and practical methods to determine the uniqueness of MB.
- Construct the "essential set" $\mathcal{E}$: $X_i \in \mathcal{E}$ if and only if $\mathrm{MI}(\mathcal{S} \backslash \{X_i\}, Y) < \mathrm{MI}(\mathcal{S}, Y)$.

Lemma (Wang & Wang, 2017)

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

Theorem (Wang & Wang, 2017)

MB is unique if and only if $MI(\mathcal{E}, Y) = MI(\mathcal{S}, Y)$.

- An assumption-free algorithm of determining the uniqueness of MB.

Algorithm 1: An assumption-free algorithm for determining the uniqueness of MB

(1) **Input**

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

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

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

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

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

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

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

output: Y has a unique MB

Else

output: Y has multiple MB

Uniqueness of Markov boundary

- Disadvantage: all independence tests concern all variables, which requires larger amount of observations.
- Consider other methods. First we need algorithms to find one MB.
- Known methods: IAMB, KIAMB, Semi-Interleaved HITON-PC, MBOR, BLCD, PCMB, GLL-PC. All of them have additional requirements on data (to exclude morbid cases).
- An assumption-free algorithm of detecting one MB: discard redundant variables one by one.

Algorithm 2: An assumption-free algorithm for producing one MB

(1) **Input**

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

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

(3) **Repeat**

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

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 Markov boundary

- TIE* algorithm: detect all MB. Requires another algorithm of detecting one MB.
- Execute TIE* until finding the second MB, or finishing with unique MB.

Algorithm 3: A general algorithm for determining the uniqueness of MB

(1) **Input**

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

Algorithm Ω which can correctly produce one MB

(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 $\mathcal{S} \mid \{X_i\}$

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 Markov boundary

- The idea of Algorithm 3: first find one MB, then determine whether each variable X_i inside MB is essential, through finding one MB in $\mathcal{S}\backslash\{X_i\}$, and then comparing these two MB.
- We can directly determine whether X_i is essential.

Algorithm 4: An assumption-free algorithm for determining the uniqueness of MB

(1) **Input**
Observations 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 2 on $\mathcal{S}$
(3) **For** $i = 1, \ldots, m$,
If $Y \perp\!\!\!\perp X_i \mid \mathcal{S} \setminus \{X_i\}$
If $X_i \not\!\perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$
Output Y has multiple MB
Terminate
(4) **Output** Y has a unique MB

Simulation setup

- Implement Algorithm 1, 3 and 4. In algorithm 3, the algorithm of finding one MB is set to KIAMB (Alg. 3-KI) and Algorithm 2 (Alg. 3-AF).
- Test on three artificial cases. In Case C, the assumption of KIAMB is failed. $\alpha = 0.001$.

Algorithms performances

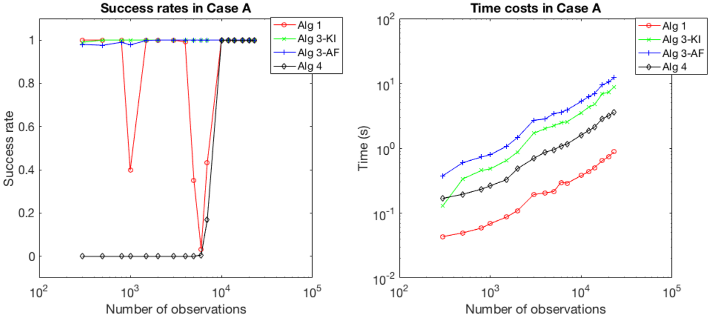


Figure: Success rates and average time costs per execution (in seconds) of Algorithms 1 (red circle), 3-KI (green 'x'), 3-AF (blue '+'), 4 (black diamond) with different numbers of observations in Case A. Number of observations and time costs are in logarithm.

Algorithms performances

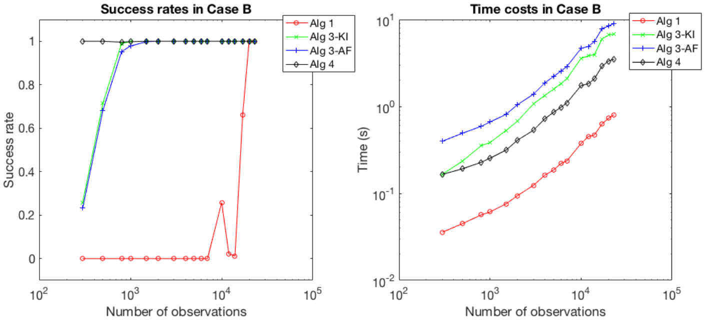


Figure: Success rates and average time costs per execution (in seconds) of Algorithms 1 (red circle), 3-KI (green ‘x’), 3-AF (blue ‘+’), 4 (black diamond) with different numbers of observations in Case B. Number of observations and time costs are in logarithm.

Algorithms performances

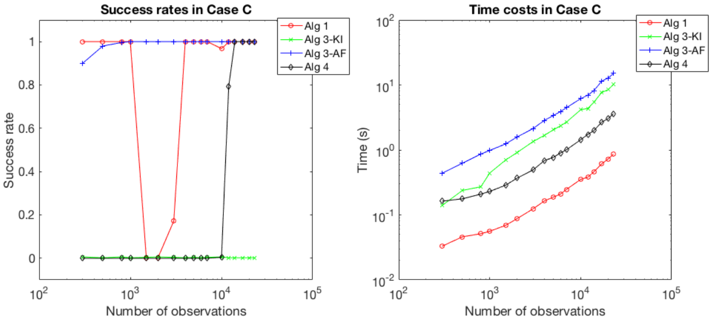


Figure: Success rates and average time costs per execution (in seconds) of Algorithms 1 (red circle), 3-KI (green 'x'), 3-AF (blue '+'), 4 (black diamond) with different numbers of observations in Case C. Number of observations and time costs are in logarithm.

Algorithms performances

- When the amount of observation is large enough, every algorithm is valid (except 3-KI in Case C).
- Success rates: 3-AF$\approx$ 3-KI$>>$1,4 if KIAMB is valid.
- Time cost: 3-AF$>$3-KI$>$4$>$1.

References

- CMI: DOBRUSHIN, R. L. (1963). General formulation of Shannon's main theorem in information theory. Amer. Math. Soc. Trans. 33, 323–438.
- CS: JANZING, D., BALDUZZI, D., GROSSE-WENTRUP, M. & SCHÖLKOPF, B. (2013). Quantifying causal influences. Ann. Stat. 41, 2324–2358.
- PMI: ZHAO, J., ZHOU, Y., ZHANG, X. & CHEN, L. (2016). Part mutual information for quantifying direct associations in networks. Proc. Natl. Acad. Sci. 113, 5130–5135.
- MB: PEARL, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.

Thank you!