Eliciting Causal Bayesian Networks with Scoring Rules

Siddharth Namachivayam

Problem

- Scoring rules can elicit an agent's beliefs about probability distributions.
- How can we elicit an agent's beliefs about causal Bayesian networks?
- If we can:
    1. Credibly perform interventions
    2. Identify a causal Bayesian network from information about its interventional distributions

Then maybe we can combine scoring rules to discover the agent's beliefs…

Setup

- $V = \{V_1, \ldots, V_K\}$ is a set of K *discrete* random variables each with at least 2 outcomes.

- G is the causal Bayesian network over V believed by the agent.

- For any random variable R, we denote its outcomes by $[R]$.

Setup

- Given random variables X and Y, the agent's beliefs about the distribution of Y conditional on $\text{do}(X = x)$ are denoted by $b(Y \mid \text{do}(X = x))$.
    - Sometimes we will also write $b_{Y|\text{do}(x)}$ when X is clear from context.
    - Additionally, sometimes we will write $b(Y \mid \text{do}(X))$ when speaking of the function $x \mapsto b_{Y|\text{do}(x)}$.
- Finally, we let $b^* = \{b(V \mid \text{do}(X)) : X \subseteq V\}$ be the set of all the agent's interventional beliefs.

Identifiability

- Unfortunately G is not (in general) identifiable from $b*$.

- Say for variables $X, Y \in V$ that X has *zero direct effect* on Y, or equivalently, that ZDE(X, Y) holds iff:

$$b(Y \mid \text{do}(V \backslash \{X, Y\})) = b(Y \mid \text{do}(V \backslash \{X, Y\}, X))$$

, i.e. intervening on X after intervening on $V \backslash \{X, Y\}$ does not change the distribution of Y.

- Bareinboim, Brito, Pearl 2012 show that if G does not have an arrow going from X to Y then ZDE(X, Y).

- However, ZDE(X, Y) does not imply that G does not have an arrow going from X to Y.

ZDE Faithfulness

- We assume the agent's beliefs satisfy *ZDE faithfulness*, that is:

  ZDE(X, Y) holds iff G does not have an arrow going from X to Y.

- Given ZDE faithfulness, G is identifiable from $b*$!

- Just need to know $b(V_i \mid \text{do}(V \backslash \{V_i\}))$ for each $i \in K$.

Scoring Rules

- Suppose an agent's beliefs over $[R] = \{1,...,n\}$ are given by $\vec{b} \in \Delta^n$.
- A *scoring rule* is a function $s : \Delta^n \to \mathbb{R}^n$ that for all reports $r \in \Delta^n$ returns a vector $\vec{s}(r)$ of payoffs such that:

$$\vec{b} \in \text{argmax}_{r \in \Delta^n} \vec{b} \cdot \vec{s}(r).$$

- A *proper* scoring rule is a scoring rule $s : \Delta^n \to \mathbb{R}^n$ such that $\forall \vec{b} \in \Delta^n$:

$$\text{max}_{r \in \Delta^n} \vec{b} \cdot \vec{s}(r) \geq 0.$$

- A *strictly* proper scoring rule is a proper scoring rule $s : \Delta^n \to \mathbb{R}^n$ such that $\forall \vec{b} \in \Delta^n$:

$$\{\vec{b}\} = \text{argmax}_{r \in \Delta^n} \vec{b} \cdot \vec{s}(r).$$

Mechanism Design

- Using a strictly proper scoring rule $s^i : \Delta^{|[V_i]|} \to \mathbb{R}^{|[V_i]|}$, the designer can $\forall v_{-i} \in [V \backslash \{V_i\}]$ successfully elicit:

$$b(V_i \,|\, \mathrm{do}(V \backslash \{V_i\} = v_{-i}))$$

by promising to interventionally realize v_{-i}.

- However, one might worry:

1. Interventions are costly to perform

2. The agent could learn from each experiment.

Mechanism #1

- We can bypass both issues!
- The designer can promise to randomize uniformly over which of the:

$$Z = \sum_{i=1}^{K} | [V \backslash \{V_i\}] |$$

scoring rules will actually pay out and perform only its intervention.

Objection #1

- The agent might incur a cognitive cost when processing each of the counterfactual conditions in mechanism #1.
- As Z increases, the agent's expected reward per bit of information on each scoring rule will shrink while processing costs stay fixed.
- This threatens incentive compatibility as the agent will just report their non-interventional beliefs for each scoring rule.

Objection #1

- We can formalize this objection in the context of when each scoring rule $s^i : \Delta^{|[V_i]|} \to \mathbb{R}^{|[V_i]|}$ is of the form:

$$s_v^i(r) = B \cdot \log_2\left(\frac{r_v}{1/|[V_i]|}\right)$$

, i.e. a log scoring rule. Here, the designer's worst case loss is:

$$B \cdot \log_2(\max_{i \in K}|[V_i]|)$$

Objection #1

- We can cash out the cost of processing the counterfactual condition in the scoring rule designed to elicit $b(V_i \,|\, \text{do}(V \backslash \{V_i\} = v_{-i}))$ as:

$$C \cdot \text{KL}(b_{V_i|\text{do}(v_{-i})} \,||\, b(V_i))$$

- If the agent chooses to not process the counterfactual condition $b(V_i)$ is reported instead.

Objection #1

- One can show the expected benefit of processing the counterfactual condition for each scoring rule is:

$$\frac{B}{Z} \cdot \mathsf{KL}(b_{V_i|\mathsf{do}(v_{-i})} || b(V_i)).$$

- Hence, mechanism #1 will be incentive compatible iff $B/Z > C$.

- Furthermore, the infimum worst case loss for mechanism #1 to be incentive compatible is:

$$(C \cdot Z) \cdot \log_2(\max_{i \in K} |[V_i]|)$$

Objection #1

- To illustrate the issue, consider when each $[V_i] = \{0,1\}$.
- The infimum worst case loss is:

$$C \cdot K \cdot 2^{K-1}.$$

Lesson

- We need a mechanism that randomizes over fewer interventions.
- The Peter-Clark algorithm reconstructs a *partially oriented causal skeleton* from the joint distribution.
- This drastically reduces the number of interventions needed to identify G.
- Idea: Assume the agent's beliefs satisfy some kind of causal faithfulness!

Causal Faithfulness

- $\forall A \subseteq V$ let do(A_r) randomize uniformly over all $a \in A$.
- Let G_A be the DAG obtained by removing all edges going into A from G.
- We say an agent's beliefs satisfy *causal faithfulness* if $\forall A \subseteq V$ the intervention do(A_r) induces a distribution $b(V \mid \text{do}(A_r))$ where:

X and Y are dependent conditional on every subset $V' \subseteq V \backslash \{X, Y\}$

iff

X and Y are adjacent in G_A.

Tests

- An *adjacency test* for $X, Y \in V$ is an intervention do(A_r) such that $X, Y \notin A$.
- A *directional test* for $X, Y \in V$ is an intervention do(A_r) such that $X \in A$ or $Y \in A$ but not both.
- Eberhardt, Glymour, Scheines 2005 note that as long as each pair of variables is subject to one adjacency test and one directional test we can orient all edges in G.

Tests

- Towards that end, they form a sequence of subsets $A \subseteq V$ corresponding to interventions do(A_r) as follows:

seq(V) :=

1. If $|V| = 1$ return $\langle \emptyset \rangle$.
2. Else, partition V into sets V^1 and V^2 where $V^1 = \lfloor |V|/2 \rfloor$ and $V^2 = \lceil |V|/2 \rceil$.
3. Return $\langle V^1 \rangle$ + (seq(V^1) ∪ seq(V^2))

where ∪ first pads the shorter list with $\emptyset$ at the end to make it equal in length to the longer list and takes their component-wise union.

Mechanism #2

- Using a strictly proper scoring rule $s : \Delta^{|[V]|} \to \mathbb{R}^{|[V]|}$, the designer can $\forall A \in \text{seq}(V)$ successfully elicit:

$$b(V\,|\,\text{do}(A_r)).$$

- Thus the designer can promise to randomize uniformly over which of the $|\,\text{seq}(V)\,| = \lceil 1 + \log_2(K) \rceil$ scoring rules will actually payout and perform only its intervention to elicit G.

Mechanism #2

- Assuming each $s : \Delta^{|[V]|} \to \mathbb{R}^{|[V]|}$ is a log scoring rule:

$$s_v(r) = B \cdot \mathrm{log}_2\left(\frac{r_v}{1/|[V]|}\right)$$

we can show the infimum worst case loss for mechanism #2 to be incentive compatible is:

$$(C \cdot \lceil 1 + \log_2(K) \rceil) \cdot \mathrm{log}_2(|[V]|)$$

Mechanism #2

- To see why this answers objection #1 consider when each $[V_i] = \{0,1\}$.
- The infimum worst case loss is now:

$$C \cdot K \cdot \lceil 1 + \log_2(K) \rceil$$

- Significant improvement from earlier:

$$C \cdot K \cdot 2^{K-1}$$

- In fact, no matter how many outcomes each V_i has we can always show:

$$\frac{\text{infimum worst case loss for mechanism \#2 to be IC}}{\text{infimum worst case loss for mechanism \#1 to be IC}} \leq \frac{\lceil 1 + \log_2(K) \rceil}{2^{K-1}}$$

Objection #2

- The PC algorithm can reconstruct a partially oriented skeleton of S from a given joint distribution.

- Let S^* denote the set of vertices which belong to an unoriented edge in the skeleton.

- It suffices now to subject every pair of variables in S^* to a directional test.

- Wouldn't this allow us to randomize over even fewer scoring rules?

Mechanism #3

- Step #1- The agent reports to a scoring rule $s^{t_1} : \Delta^{|[V]|} \to \mathbb{R}^{|[V]|}$ which pays out in case the designer performs no intervention.

- Step #2- The designer:

    - Reconstructs a partially oriented skeleton S from the report in step #1.

    - Computes seq(S^*) and $\forall A \in$ seq$(S^*)\backslash\{\emptyset\}$, allows the agent to report to a scoring rule $s^{t_2} : \Delta^{|[V]|} \to \mathbb{R}^{|[V]|}$ which pays out in case the designers performs the intervention do(A_r).

- Step #3- The designer randomizes uniformly over which of the $|$seq$^*|$ interventions to performs and pays out the agent.

Objection #3

- While $|\text{seq}^*| \leq |\text{seq}|$, IC cannot be established.

- Agents can leverage their knowledge that the PC algorithm determines which interventions the designer will randomize over.

- Namely, if they know more about certain interventions than others, they may prefer the PC algorithm to output a skeleton different from the one they truly believe.

- So, mechanism #3 must assume agents are *myopic* to guarantee IC, i.e. agents ignore profits from future actions at each stage of the game.

Possible Solution

- Maybe we can mitigate the incentive to lie in step #1 by making the reward sizes for the scoring rules in step #2 small in comparison.

- Formally, suppose the scoring rule in step 1, $s^{t_1} : \Delta^{|[V]|} \to \mathbb{R}^{|[V]|}$, is of the form:

$$s_v^{t_1}(r) = B_1 \cdot \log_2\left(\frac{r_v}{1/|[V]|}\right)$$

while the scoring rules in step 2, $s^{t_2} : \Delta^{|[V]|} \to \mathbb{R}^{|[V]|}$, are of the form:

$$s_v^{t_2}(r) = B_2 \cdot \log_2\left(\frac{r_v}{1/|[V]|}\right)$$

Possible Solution

- $B_2 \cdot \log_2(|[V]|)$ is the maximum profit an agent can obtain from a scoring rule used in step #2.

- $B_1 \cdot \mathrm{KL}(b(V)||r)$ is the cost of reporting r instead of $b(V)$ to the scoring rule in step #1.

- In equilibrium then:

$$\mathrm{KL}(b(V)||r) \leq \frac{B_2}{B_1} \cdot \log_2(|[V]|).$$

- Since making $B_1 >> B_2$ implies $r \approx b(V)$, we might hope for *approximate* IC.

Possible Solution

- Without additional assumptions, this may be no more than a pipe dream.
- The causal skeleton output by the PC algorithm could be highly sensitive to the input joint distribution.
- But even if we could find suitable assumptions, the presence of processing costs prevents us from decreasing B_2 arbitrarily.
- Therefore, such an approach would likely increase B_1 so high that the designer's worst case loss is greater than in mechanism #2.

Summary

Mechanism	**Identification Assumptions**	**Myopic IC**	**IC**	**Worst-Case Loss (Log Scoring Rules)**
Mechanism #1	ZDE faithfulness	Yes	Yes	$(C \cdot Z) \cdot \log_2(\max_{i \in K} \lvert [V_i] \rvert)$
Mechanism #2	Causal faithfulness	Yes	Yes	$(C \cdot \lceil 1 + \log_2(K) \rceil) \cdot \log_2(\lvert [V] \rvert)$
Mechanism #3	Causal faithfulness	Yes	No	$(C \cdot \lvert \mathrm{seq}^* \rvert) \cdot \log_2(\lvert [V] \rvert)$

References

- Bareinboim, Brito, Pearl: https://link.springer.com/chapter/10.1007/978-3-642-29449-5_1
- Spirtes, Glymour, Scheines: https://direct.mit.edu/books/monograph/2057/Causation-Prediction-and-Search
- Eberhardt, Glymour, Scheines: https://arxiv.org/pdf/1207.1389