Cognitive Modeling
Lecture 19: Causal Learning

Sharon Goldwater
School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk
March 8, 2010

Causal Graphical Models

In the last lecture, we introduced causal graphical models:
- they are an extension of graphical models that can deal with interventions as well as observations;
- we saw that respecting the direction of causality results in efficient representation and inference;

Today, we’ll look at modeling human learning of causal relationships using causal graphical models.

Rating Causality

**Experiment:** subjects are shown contingency data and must rate \( P(C \rightarrow E) \), the probability that an event \( C \) causes outcome \( E \).

**Example:** case studies with data from experiments in which rats are injected with a certain chemical and tested for expression of a certain gene.

- Case 1: 40 out of 100 injected rats express the gene, 0 out 100 uninjected rats express the gene (40/100, 0/100);
- Case 2: 7 out of 100 injected rats express the gene, 0 out 100 uninjected rats express the gene (7/100, 0/100);
- Case 3: 53 out of 100 injected rats express the gene, 46 out 100 uninjected rats express the gene (53/100, 46/100).

How do you rate \( P(C \rightarrow E) \) in each case?
Rating Causality

Experimental results (ratings on a 0–20 scale):

| Case  | Rating   | $P(e^+|c^+)$ |
|-------|----------|--------------|
| 1     | 14.9 ± 0.8| 0.40         |
| 2     | 8.6 ± 0.9 | 0.07         |
| 3     | 4.9 ± 0.7 | 0.53         |

So clearly, subjects are not just using conditional probability: $P(C \rightarrow E) \neq P(e^+|c^+)$. Two competing rational models have been proposed in the literature to explain these experimental results:

- $\Delta P$ model
- Causal power model

$\Delta P$ vs. Causal Power

The $\Delta P$ model assumes people estimate $P(C \rightarrow E)$ as:

$$\Delta P = P(e^+|c^+) - P(e^+|c^-)$$

- $P(e^+|c^+)$ and $P(e^+|c^-)$ are computed as relative frequencies.
- Causality is indicated by a large difference in the probability of the effect when the cause is absent or present.
- Can be shown to be equivalent to evaluating the associative strength between cause and effect.

Causal Power

The causal power model assumes people estimate $P(C \rightarrow E)$ as:

$$power = \frac{\Delta P}{1 - P(e^+|c^-)}$$

- Based on axiomatic characterization of causality (Cheng 1997).
- Normalizes $\Delta P$ by cases where $C$ could be observed to influence $E$.
  - (36/60, 30/60): $\Delta P = 0.1$, power = 0.2.
  - (60/60, 54/60): $\Delta P = 0.1$, power = 1.

Both $\Delta P$ and causal power predict some trends in experimental data (more on this later), but don’t fully account for the data.

| Case  | Rating   | $P(e^+|c^+)$ | $P(e^+|c^-)$ | $\Delta P$ | power |
|-------|----------|--------------|--------------|------------|-------|
| 1     | 14.9 ± 0.8| 0.40         | 0            | 0.40       | 0.40  |
| 2     | 8.6 ± 0.9 | 0.07         | 0            | 0.07       | 0.07  |
| 3     | 4.9 ± 0.7 | 0.53         | 0.46         | 0.07       | 0.13  |
Problematic Effects

1. Effect of $P(e^+|c^-)$ when $\Delta P = 0$:
   - **Example**: (8/8, 8/8), (4/8, 4/8), (0/8, 0/8).
   - Both $\Delta P$ and power predict $P(C \rightarrow E) = 0$ for all cases.
   - But: subjects judge $P(C \rightarrow E)$ to decrease across these cases.
   - Intuitive explanation: when $P(e^+|c^-)$ is lower, more opportunity to observe $C$ exert an effect, but still no effect.

2. Sample size effect:
   - **Example**: (2/4, 0/4), (10/20, 0/20), (25/50, 0/50).
   - Both $\Delta P$ and power predict $P(C \rightarrow E) = .5$ for all cases.
   - But: subjects judge $P(C \rightarrow E)$ to increase across cases.
   - Intuitive explanation: in small samples, effects could be just random noise.

3. Non-monotonic effects of changing $P(e^+|c^-)$:
   - **Example**: (30/30, 18/30), (24/30, 12/30), (12/30, 0/30).
   - $\Delta P$ predicts constant $P(C \rightarrow E)$, power predicts a decrease.
   - But: subjects judge $P(C \rightarrow E)$ slightly lower for middle case.
   - Previous researchers assumed this effect was just odd and ignored it.

Rethinking Causal Learning

Using Bayes nets, Tenenbaum and Griffiths (2001) provide an explanation for the failures of $\Delta P$ and causal power and suggest an alternative model.

- Both $\Delta P$ and causal power can be viewed as estimating parameters of a particular causal graphical model.
- Tenenbaum and Griffiths (2001) suggest that subjects are actually performing structure learning: choosing between two different causal graphical models.

That is, previous models assumed people are judging the strength of causation, new model assumes they are judging the existence of causation.
Analyzing $\Delta P$ and Causal Power

Given the following Bayes net:

```
B --w_B--> E --w_C--> C
```

- $C$: cause
- $E$: effect
- $B$: background (alternative cause/causes), with $B=1$ always.
- $w_B, w_C$: parameters (effect strengths) $P(E|B)$, $P(E|C)$.

We can analyze the $\Delta P$ and Causal Power models as two different parameterizations (i.e., ways of defining $P(E|B, C)$).

**Linear parameterization:** the effect strengths of $B$ and $C$ are additive.

\[
\begin{align*}
P(e^+|c^-, b^+) &= w_B \\
P(e^+|c^+, b^+) &= w_B + w_C
\end{align*}
\]

**Noisy-OR parameterization:** $C$ and $B$ act as independent causes.

\[
\begin{align*}
P(e^+|c^-, b^+) &= w_B \\
P(e^+|c^+, b^+) &= w_B + w_C - w_B w_C
\end{align*}
\]

Reduces to standard OR if $w_B = w_C = 1$.

---

Structure Learning

Tenenbaum and Griffiths (2001) show that:
- $\Delta P$ corresponds Bayes net with linear parameterization;
- causal power corresponds to Bayes net with noisy-OR parameterization

where parameters $w_B$ and $w_C$ are estimated using maximum likelihood estimation.

**Key insight:** causal inference is a judgment of whether a causal link exists, not how strong the effect is. So, subjects are really doing structure learning for Bayes nets.
Hypothesis: subjects are deciding between the following two Bayes nets:

\[ h_C = 1 \]

\[ h_C = 0 \]

Does cause C have an influence on effect E?

Tenenbaum and Griffiths (2001) use Bayesian inference over model structures to make this decision.

\[ \text{support} = \log \frac{P(h_C = 1 | X)}{P(h_C = 0 | X)} \]

Assuming the prior probability of each graph is 0.5,

\[ \text{support} = \log \frac{P(X | h_C = 1)}{P(X | h_C = 0)} \]

Compute \( P(X | h_C = 1) \) by summing over possible parameter values (Bayesian inference):

\[ P(X | h_C = 1) = \int_0^1 \int_0^1 P(X | w_B, w_C, h_C = 1)p(w_B, w_C | h_C = 1)dw_B dw_C \]

Similarly for \( P(X | h_C = 0) \).

Actual computation requires a computer program.

Can also compute other values from this model, e.g. \( p(w_C | X) \).

Causal Support is high when \( p(w_C | X) \) has most of its mass on non-zero values.
Comparison of the Models

Comparison of the three models:

| Model   | Form of $P(E | B, C)$ | $P(C \rightarrow E)$ |
|---------|-----------------------|-----------------------|
| $\Delta P$ | Linear               | $w_C$                |
| Power    | Noisy-OR             | $w_C$                |
| Support  | Noisy-OR             | $\log \frac{P(h_C = 1)}{P(h_C = 0)}$ |

Comparison with Experimental Data

Comparison of model performance with Buehner and Cheng’s (1997) experimental data:

- subjects judged $P(C \rightarrow E)$ for hypothetical medical studies (similar to gene expression example);
- each subjects saw eight cases in which $C$ occurred and eight cases in which $C$ didn’t occur;
- compare predictions of all three models to human judgments.

Overall, Causal Support has highest correlation with human data for this and other experimental data.
**Sample Size Effect**

- Left: $p(w_C | X)$. Right: Causal Support.
- More data $\Rightarrow$ more certainty in non-zero value of $w_C$.

**Non-monotonic Effect**

- Top: $E$ occurs with $C$ in all cases where it can $\Rightarrow$ high certainty in high value of $w_C$.
- Bottom: $E$ never occurs without $C$ $\Rightarrow$ lower value of $w_C$, but high certainty in non-zero value.
- Middle: Neither extreme $\Rightarrow$ most probable value of $w_C$ is high, but lower certainty in non-zero value.

**Discussion: results**

- Causal Support correlates better with human data than previous models in a range of experiments.
- Captures several trends other models do not:
  - effects when $\Delta P = 0$;
  - non-monotonic effects;
  - sample size effects.
- Predictions stem from the assumption that humans are learning causal structure rather than estimating its strength.
- Also able to draw inferences based on very few observations (this was tested in subsequent experiments).

**Discussion: methods**

Causal Support model uses Bayesian inference to compare probabilities of different Bayes net structures.

- Previous models ask: what is the best (maximum-likelihood) estimate of $w_C$?
  
  Estimates further from zero $\Rightarrow$ greater $P(C \rightarrow E)$

- Causal Support asks: what is the most probable causal structure?
  
  More mass of $w_C$ away from zero $\Rightarrow$ greater $P(C \rightarrow E)$
Summary

- Two standard models of causal inference exist:
  - $\Delta P$: prob. of positive cause minus prob. of negative cause;
  - causal power: $\Delta P$ normalized by one minus probability of negative cause;
- these models can be analyzed as Bayes nets with linear parameterization and noisy-OR parameterization;
- but: more plausible to assume that the structure of the Bayes net is also learned;
- the causal support model achieves this by using Bayesian inference over the structure of the net;
- it accounts for patterns in the experimental data that other models fail to capture.

References


