A CONVERSATION WITH

NIHAT AY

@KyleCranmer
New York University
Department of Physics
Center for Data Science
CILVR Lab
Outline

1. The Structure of Causal Networks
2. Markov Properties & Markov Equivalence
3. Intervention & Causal Effects
4. Identifiability of Causal Effects
5. Application to Embodied Intelligence
6. The Common Cause Principle
7. Knockout Interventions for System Identification
PREDICTION

INFERENCE

\( \theta \)

Theory

Data

x
The Scientific Method as an Ongoing Process

1. Make Observations
   - What do I see in nature?
   - This can be from one's own experiences, thoughts, or reading.

2. Think of Interesting Questions
   - Why does that pattern occur?

3. Formulate Hypotheses
   - What are the general causes of the phenomenon I am wondering about?

4. Gather Data to Test Predictions
   - Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.

5. Refine, Alter, Expand, or Reject Hypotheses

6. Develop Testable Predictions
   - If my hypothesis is correct, then I expect a, b, c, ...

7. Develop General Theories
   - General theories must be consistent with most or all available data and with other current theories.
THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE
All models are wrong, and increasingly you can succeed without them.

— Peter Norvig
The ladder of Causation

Figure 1.2. The Ladder of Causation, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first rung, learning from association. Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge. Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena. (Source: Drawing by Maayan Harel.)
The morale of the story

The morale of this story is summed up in the following picture:

- **data**
- **data + graph assumptions**

[Link to the original source](https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/)
1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: What if I see ...?
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?
2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: What if I do ...? How?
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?
3. COUNTERFACTUALS

**ACTIVITY:** Imagining, Retrospection, Understanding

**QUESTIONS:** *What if I had done ...? Why?*  
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

**EXAMPLES:** Was it the aspirin that stopped my headache?  
Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?
CLAIM: Under the hypothesis of independent mechanisms and small changes across different distributions:

- smaller sample complexity to recover from a distribution change

  - E.g. for transfer learning, agent learning, domain adaptation, etc.

What if we had the right causal structure?
Causally interpreted Bayesian networks

Definition: A DAG $G = (V, E)$ together with a family $\kappa_v : \mathbb{X}_{\text{pa}(v)} \times \mathbb{X}_v \to [0, 1]$, $v \in V$, of Markov kernels is called Bayesian network. The interpretation of the Markov kernels $\kappa_v$ as mechanisms of the nodes implies a causal nature of the Bayesian network.

Consider two distinct nodes $v, w \in V$.
- If $v \to w$ we call $v$ a (pot.) direct cause of $w$ and $w$ a (pot.) direct effect of $v$.
- If $v \sim w$ we call $v$ a (pot.) cause of $w$ and $w$ an (pot.) effect of $v$.

Markov Equivalence

The data generating process

\[ p(u, a, b, c) = \varphi(u) \alpha(u; a) \beta(a; b) \gamma(u, b; c) \]

\[ B \perp U \mid A, \quad C \perp A \mid B, U \]
A toy example

Ferenc Huszár

Intervention and the *do*-operation

\[
\hat{\alpha}(a) := \begin{cases} 
1, & \text{if } a = \hat{a} \\
0, & \text{if } a \neq \hat{a}
\end{cases}
\]

\[
p(u, a, b, c) = \varphi(u) \alpha(u; a) \beta(u; b) \gamma(a, b; c)
\]

\[
p(u, a, b, c|\text{do}(\hat{a})) := \varphi(u) \hat{\alpha}(a) \beta(u; b) \gamma(a, b; c)
\]
Intervention and the \textit{do}-operation

\[ p(b \mid \text{do}(\hat{a})) = \sum_{u,v,w,c} p(u,a,b,c \mid \text{do}(\hat{a})) \]
\[ = \sum_{u,v,w,c} \varphi(u) \hat{a}(a) \beta(u;b) \gamma(a,b;c) \]
\[ = \sum_{v,w,c} \varphi(u) \beta(u;b) \gamma(\hat{a},b;c) \]
\[ = \sum_{u,v,w} p(u) \beta(u;b) \sum_{c} \gamma(\hat{a},b;c) \]
\[ = p(b) \neq p(b \mid \hat{a}) \]

\[ p(c \mid \text{do}(\hat{u})) = \sum_{u,a,b} p(u,a,b,c \mid \text{do}(\hat{u})) \]
\[ = \sum_{u,a,b} \varphi(u) a(u;a) \beta(u;b) \gamma(a,b,c) \]
\[ = \sum_{a,b} \beta(\hat{u};b) \gamma(a,b,c) \]
\[ = \sum_{a,b} \varphi(\hat{u}) a(\hat{u};a) \beta(\hat{u};b) \gamma(a,b,c) \]
\[ = p(c \mid \hat{u}) \]
A toy example

\[ P(y|do(X)) = p(y|x) \]

\[ z = \text{randn()} \]
\[ x = 3 \]
\[ y = (y-1)/4 + \sqrt{3}\cdot\text{randn()}/2 \]
\[ x = 3 \]

\[ P(y|do(X)) = p(y) \]

\[ x = \text{randn()} \]
\[ x = 3 \]
\[ y = 1 + 2\cdot\text{randn()} \]
\[ x = 3 \]
\[ x = z \]
\[ x = 3 \]
\[ y = z + 1 + \sqrt{3}\cdot\text{randn()} \]
\[ x = 3 \]

\[ P(y|do(X)) = p(y) \]


Identifiability of causal effects

**Theorem:** Let $\mathcal{B}$ be a Bayesian network with DAG $G = (V, E)$. For two distinct nodes $v$ and $w$ the following holds:

$$p(x_w \mid do(x_v)) = \sum_{x_{pa(v)}} p(x_{pa(v)}) p(x_w \mid x_v, x_{pa(v)}).$$

$$p(x_w \mid do(x_v)) = \sum_{x_i, x_j} p(x_i, x_j) p(x_w \mid x_v, x_i, x_j).$$
3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: What if I had done ...? Why?
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?
Example 1: David Blei’s election example

This is an example David brought up during the Causality Panel and I referred back to this in my talk. I’m including it here for the benefit of those who attended my MLSS talk:

*Given that Hilary Clinton did not win the 2016 presidential election, and given that she did not visit Michigan 3 days before the election, and given everything else we know about the circumstances of the election, what can we say about the probability of Hilary Clinton winning the election, had she visited Michigan 3 days before the election?*

Let’s try to unpack this. We are interested in the probability that:

- she *hypothetically* wins the election

conditioned on four sets of things:

- she lost the election
- she did not visit Michigan
- any other relevant observable facts
- she *hypothetically* visits Michigan

It’s a weird beast: you’re simultaneously conditioning on her visiting Michigan and not visiting Michigan. And you’re interested in the probability of her winning the election given that she did not. WHAT?

Why would quantifying this probability be useful? Mainly for credit assignment. We want to know why she lost the election, and to what degree the loss can be attributed to her failure to visit Michigan three days before the election. Quantifying this is useful, it can help political advisors make better decisions next time.
Identifiability of causal effects (front-door example)

\[ p(c \mid do(a)) = \sum_{u,b} \varphi(u) \beta(a; b) \gamma(u, b; c) \]
\[ = \sum_{u,b} p(u) p(b \mid a) p(c \mid u, b) \]
\[ = \sum_{u,b} \left( \sum_{a'} p(a') p(u \mid a') \right) p(b \mid a) p(c \mid u, b) \]
\[ = \sum_{b} p(b \mid a) \sum_{a'} p(a') \left( \sum_{u} p(u \mid a', b) p(c \mid u, a', b) \right) \]
\[ = \sum_{b} p(b \mid a) \sum_{a'} p(a') p(c \mid a', b) \]

Diagram: A -> B -> C
In robotics

The sensorimotor loop

by Keyan-Ghazi Zahedi

Causal effects in the sensorimotor loop

\[ p(s \mid do(a)) = \sum_c p(s \mid c, a) p(c) \]

Figure 3. Graphical representations of the environment, the agent, and the various models. Circles are stochastic nodes, rectangles are deterministic nodes. (a) Agent interacting with the environment, generating a trajectory \( \{y_t, a_t\}_{t=0} \). These trajectories are the training data for the models. (b) Same as (a) but also including the backdoor \( z_t \) in the generated trajectory. The red arrows indicate the locations of the interventions. (c) Standard autoregressive generative model of observations. The model predicts the observation \( y_t \), which it then feeds into \( h_{t+1} \). (d) Example of a Non-Causal Partial Model (NCPM) that predicts the observation \( y_t \) without feeding it into \( h_{t+1} \). (e) Proposed Causal Partial Model (CPM), with a backdoor \( z_t \) for the actions.