Causal Inference in Machine Learning

Ricardo Silva
Department of Statistical Science and Centre for Computational Statistics and Machine Learning
ricardo@stats.ucl.ac.uk
Part I:
Did you have breakfast today?
Researchers reviewed 47 nutrition studies and concluded that children and adolescents who ate breakfast had better mental function and better school attendance records than those who did not.

They suggested several possible reasons. For example, eating breakfast may modulate short-term metabolic responses to fasting, cause changes in neurotransmitter concentrations or simply eliminate the distracting physiological effects of hunger.

http://www.nytimes.com/2005/05/17/health/17nutr.html?_r=0
Spurious causality (?)

- Eating makes you faithful
  - Will he cheat? How to tell. Ladies, you probably think that it's just in his nature. He can't help it - he HAS to cheat. But here's the sad truth: you're not feeding him enough. If you're worried your guy might cheat, try checking out his waistline. A new study says the size of his belly may reveal whether he'll stray.

- Relaxing makes you die
  - In a prospective cohort study of thousands of employees who worked at Shell Oil, the investigators found that embarking on the Golden Years at age 55 doubled the risk for death before reaching age 65, compared with those who toiled beyond age 60.

http://www.medpagetoday.com/PrimaryCare/PreventiveCare/1980
What is a cause, after all?

- A causes B:
  \[ P(B \mid A \text{ is manipulated to } a_1) \neq P(B \mid A \text{ is manipulated to } a_2) \]

- Next the concept of an *external agent*

- Examples of manipulations:
  - Medical interventions (treatments)
  - Public policies (tax cuts for the rich)
  - Private policies (50% off! Everything must go!)

- A manipulation (intervention, policy, treatment, etc.) changes the data generation mechanism. It sets a new *regime*
But what exactly is a manipulation?

- Some intervention T on A can only be "effective" if T is a cause of A
- ??!??
- Don’t be afraid of circularities
  - Or come up with something better, if you can

Bart: What is "the mind"? Is it just a system of impulses or is it...something tangible?

Simpsons, The (1987)
An axiomatic system

- When you can’t define something, axiomatize it:
  - From points to lines and beyond
- We will describe languages that have causal concepts as primitives
- The goal: use such languages to
  - Express causal assumptions
  - Compute answers to causal queries that are entailed by such assumptions
Causal queries: hypothetical causation vs. counterfactual causation

- I have a headache. If I take an aspirin now, will it go away?

- I had a headache, but it passed. Was it because I took an aspirin two hours ago? Had I not taken such an aspirin, would I still have a headache?
Prediction vs. explanation

- The first case is a typical “predictive” question
  - You are calculating the effect of a hypothetical intervention
  - Pretty much within decision theory
    - Think well before offering the 50% discount!

- The second case is a typical “explanatory” question
  - You are calculating the effect of a counterfactual intervention
    - Have things been different…
  - Ex.: law

- What about scientific/medical explanation?
This talk will focus solely on prediction.

Explanation is fascinating, but too messy, and not particularly useful (at least as far as Science goes)…
Preparing axioms: Seeing vs. doing

- Observe again the notation
  \[ P(B \mid A \text{ is manipulated to } a_1) \]

- Why not...
  \[ P(B \mid A = a_1) \]

...?
Seeing vs. doing: an example

- The reading in a barometer is useful to predict rain
  \[ P(\text{rain} \mid \text{barometer reading} = \text{high}) > P(\text{rain} \mid \text{barometer reading} = \text{low}) \]

- But hacking a barometer won’t cause rain
  \[ P(\text{rain} \mid \text{barometer hacked to high}) = P(\text{rain} \mid \text{barometer hacked to low}) \]

- (Sometimes this is called intervening vs. conditioning. You should see this as a way of indexing regimes.)
Why is seeing different from doing?

- **Issue #1: directionality**

  ![Diagram](Drinking -> Car accidents)
Why is seeing different from doing?

- Issue #2: confounding (i.e., common causes)
**Why is seeing different from doing?**

- *Most important lesson:* unmeasured confounding (i.e., hidden common causes) is perhaps the most complicating factor of all.

- (but see also: measurement error and sampling selection bias)
The \textit{do} operator (Pearl’s notation)

- **A shorter notation**

- $P(A | B = b)$: the probability of $A$ being true given an observation of $B = b$
  - That is, no external intervention
  - This is sometimes called the distribution under the \textit{natural state} of $A$

- $P(A | \text{do}(B = b))$: the probability of $A$ given an intervention that sets $B$ to $b$
  - $P(A | \text{do}(B))$: some shorter notation for $\text{do}(B) = \text{true}$
Different do’s

- $P(A \mid \text{do}(B), C)$
  - Intervening on $B$, seeing $C$

- $P(A \mid \text{do}(B), \text{do}(C))$
  - Multiple interventions

- $P(A \mid \text{do}(P(B) = P'))$
  - A change on the distribution of $B$ (not only a point mass distribution)
Causal models

- A causal model is defined by a set of
  \[ P(A_1, A_2, \ldots, A_N \mid \text{do}(B_1), \text{do}(B_2), \ldots, \text{do}(B_M), B_{M+1}, B_{M+2}, \ldots, B_O) \]

- How to estimate this? Which data can I use?

- The Radical Empiricist says:
  
  Every do is a change of regime. Anything can happen. In general, there is no logical connection between states!

Every different set of do’s specify a brave new World.

(or does it?)
Learning causal models

- The gold standard*: randomized experiments

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Patient ID</th>
<th>Age</th>
<th>Heart Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine</td>
<td>1</td>
<td>32</td>
<td>+</td>
</tr>
<tr>
<td>Medicine</td>
<td>2</td>
<td>41</td>
<td>+</td>
</tr>
<tr>
<td>Placebo</td>
<td>3</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Placebo</td>
<td>4</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>Medicine</td>
<td>5</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

*and a recipe for knighthood
The role of randomization

- Breaking the hidden common causes
- Example: gender may cause both self-selection of treatment, and heart condition
The role of randomization

- The randomized assignment overrides the original causal mechanisms

- Notice: placebo is a surrogate for no-treatment

- With blind/double-blind assignments, its role is to avoid psychological effects
Causal models

- A causal model is defined by a set of
  \( P(A_1, A_2, \ldots, A_N \mid \text{do}(B_1), \text{do}(B_2), \ldots, \text{do}(B_M), B_{M+1}, B_{M+2}, \ldots, B_O) \)

- Do I always have to perform an experiment?
Observational studies

- The art and science of inferring causation without experiments
- This can only be accomplished if extras assumptions are added
- Most notable case: inferring the link between smoking and lung cancer
- This tutorial will focus on observational studies
Observational studies

- If you can do a randomized experiment, you should do it.
- Observational studies have important roles, though:
  - When experiments are impossible for unethical/practical reasons:
    - The case for smoking/lung cancer link
  - When there are many experiments to perform:
    - A type of exploratory data analysis/active learning tool
    - E.g., biological systems
Observational studies

- It *is* certainly true that correlation is not causation
  - And as statisticians know, it may well be the case correlation-hat is not even correlation
- But it is also lazy to stop there
Observational studies
John Snow’ Soho
Observational studies

- But in the end, don’t we always have a testable condition?
Observational studies

- Appropriate interventions are much more subtle than you might think…

```
Nasty pictures in cigarette packages → Smoke

Smoke → Lung cancer

Smoke ⊥⊥ Lung cancer | do(Smoke)
```

“Gullibility trait” expression level
(Sort of) Observational studies

- But I’m Facebook and I have 1 googol-pounds of money for experiments. I’m covered, right?
(Sort of) Observational studies

Facebook shows ad to Anna

Anna “likes” it

Bob “likes” it
(Sort of) Observational studies

Facebook shows ad to Anna
Anna “likes” it
Bob “likes” it

Radio show advert

Facebook shows ad to Anna
Anna “likes” it
Bob “likes” it
(Sort of) Observational studies

Drug/placebo assignment → Patient complies → Health
(Sort of) Observational studies

Drug/placebo assignment → Patient complies → Health

Here be dragons
Observational studies: starting from natural state models

- How are full joint/conditional distributions specified?

\[ P(A_1, A_2, \ldots, A_N \mid B_1, B_2, \ldots, B_M, B_{M+1}, B_{M+2}, \ldots, B_O) \]

- There is a notion of modularity in the natural state. Why wouldn’t we have some *stable modularity across “Worlds”*?
Definitions and axioms of causal modularity: DAGs

- Directed acyclic graphs
- Start with a “reference system”, a set of events/random variables $V$
- Each element of $V$ is a vertex in causal graph $G$
- A causes B is causal graph $G$ only if A is an ancestor of B
- DAGs with such an assumption are causal graphs
Definitions and axioms of causal modularity

- A is a *direct cause* of B wrt V if and only if A causes B for some choice of intervention in $V \setminus \{A, B\}$
- “A is a direct cause of B” implies the edge
The Causal Markov Condition

- Let $G$ be a DAG representing a causal system over $V$, and $P$ a distribution over $V$
- $(G, P)$ satisfy the Causal Markov Condition if and only if:

$$A \perp \{\text{All of its (non-parental) non-descendants}\} \mid A\text{'s parents}$$

where A’s parents are its direct causes in G

(Spirtes et al, 2000)
The Causal Markov Condition

\[ \begin{align*}
D & \perp \!\!\! \perp \{E, G, H\} \mid \{A, B, C\} \\
G & \perp \!\!\! \perp \text{everybody else} \mid E
\end{align*} \]
Limitations of the Causal Markov condition?

“The Interactive Fork”

\[ P(\text{Picture} \mid \text{Switch}) < P(\text{Picture} \mid \text{Switch, Sound}) \]

*Where did the independence go?*

(Spiritse et al, 2000)
Instead of an exhaustive “table of interventional distributions”:

- $G = (V, E)$, a causal graph with vertices $V$ and edges $E$
- $P(\theta)$, a probability over the “natural state” of $V$, parameterized by $\theta$
- $(G, \theta)$ is a causal model if pair $(G, P)$ satisfies the Causal Markov condition
- We will show how to compute the effect of interventions
To summarize: what’s different?

- As you probably know, DAG models can be non-causal
- What makes causal?

![Graph with directed edge from A to B]

Answer: because I said so!
To summarize

- A causal graph is a way of encoding causal assumptions
- Graphical models allow for the evaluation of the consequences of said assumptions
- Typical criticism:
  - “this does not advance the ‘understanding’ of causality”
- However, it is sufficient for predictions
- And no useful non-equivalent alternatives are offered
Example of axioms in action: Simpson’s paradox

The “paradox”:

P(E | F, C) < P(E | F, ~C)  
P(E | ~F, C) < P(E | ~F, ~C)  

P(E | C) > P(E | ~C)

Which table to use? (i.e., condition on gender or not?)

(Pearl, 2000)
To condition or not to condition: some possible causal graphs
Dissolving a “paradox” using the *do* operator

- Let our population have some subpopulations
  - Say, F and ~F
- Let our treatment C not cause changes in the distribution of the subpopulations
  - \( P(F \mid \text{do}(C)) = P(F \mid \text{do}(\neg C)) = P(F) \)
- Then for outcome E it is impossible that we have, simultaneously,
  - \( P(E \mid \text{do}(C), F) < P(E \mid \text{do}(\neg C), F) \)
  - \( P(E \mid \text{do}(C), \neg F) < P(E \mid \text{do}(\neg C), \neg F) \)
  - \( P(E \mid \text{do}(C)) > P(E \mid \text{do}(\neg C)) \)
\begin{proof}

\begin{align*}
P(E|do(C)) &= P(E|do(C), F)P(F|do(C)) \\
&\quad + P(E|do(C), \neg F)P(\neg F|do(C)) \\
&= P(E|do(C), F)P(F) + P(E|do(C), \neg F)P(\neg F).
\end{align*}

\begin{align*}
P(E|do(\neg C)) &= P(E|do(\neg C), F)P(F) \\
&\quad + P(E|do(\neg C))P(\neg F)
\end{align*}

\end{proof}

\begin{align*}
P(E|do(C)) < P(E|do(\neg C)),
\end{align*}
Part II:
Predictions with observational data
Goals and methods

- Given: a causal graph, observational data
- Task: estimate $P(E \mid do(C))$
- Approach:
  - Perform a series of modifications on $P(E \mid do(C))$, as allowed by the causal assumptions, until no $do$ operators appear
  - Estimate quantity using observational data
  - That is, reduce the causal query to a probabilistic query

(Spirtes et al, 2000 – Chapter 7; Pearl, 2000 – Chapter 3)
The trivial case

- Graph:

- A representation of a $do(A)$ intervention
The trivial case

- B is independent of T given A
  - $P(B \mid \text{do}(A)) = P(B \mid A, T) = P(B \mid A)$

- Term on the right is identifiable from observational data
  - *do*-free

- That is, $P(B \mid \text{do}(A))$ can be estimated as $P(B \mid A)$
A less trivial case

- Knowledge:

- Query: $P(B \mid \text{do}(A))$
A less trivial case

- With intervention

- B and T are not independent given A anymore…
A less trivial case

- Solution: conditioning

- Now, B is independent of T given A and F
A less trivial case

\[ P(B \mid \text{do}(A)) = P(B \mid \text{do}(A), F)P(F \mid \text{do}(A)) + P(B \mid \text{do}(A), \neg F)P(\neg F \mid \text{do}(A)) = \]

\[ = P(B \mid A, \neg F)P(\neg F) + P(B \mid A, F)P(F) = \]

“F-independent” intervention
Simplified operation for independent point interventions

Before intervention:

\[ P(A, B, F) = P(B | A, F)P(A | F)P(F) \]

After intervention:

\[ P(A, B, F | do(A)) = P(B | A, F)P(A | F)P(F) \]

\[ = P(B | A, F) \delta(A = true)P(F) \]

A “mechanism substitution” system
Those “back-doors”…

- Any common ancestor of A and B in the graph is a confounder
- Confounders originate “back-door” paths that need to be blocked by conditioning
In general, one should condition on and marginalize minimal sets, since this reduces statistical variability.
Unobserved confounding

- If some variables are hidden, then there is no data for conditioning

- Ultimately, some questions cannot be answered without extra assumptions

- But there are other methods beside back-door adjustment
Interestingly enough, $P(Y \mid \text{do}(X))$ is identifiable in this case.

- Even though we will be conditioning on a variable $Z$ that is in the causal path!
The front-door criterion

\[
P(Y \mid \text{do}(X)) = \frac{P(Z \mid X) P(Y \mid Z, U) P(U)}{\sum_Z \sum_u P(Z \mid X) P(Y \mid Z, U) P(U)}
\]

\[
P(X, Y, Z, U) = P(U) P(X \mid U) P(Z \mid X) P(Y \mid Z, U)
\]

\[
P(Y, Z, U \mid \text{do}(X)) = P(Y \mid Z, U) P(Z \mid X) P(U)
\]

\[
P(Y \mid \text{do}(X)) = \sum_Z P(Z \mid X) \sum_u P(Y \mid Z, U) P(U)
\]
The front-door criterion

\[
P(U | X) = P(U | Z, X) \\
P(Y | Z, U) = P(Y | X, Z, U)
\]

\[
\sum_u P(Y | Z, U)P(U) = \sum_x \sum_u P(Y | X, Z, U)P(U | X)P(X)
\]

\[
= \sum_x \sum_u P(Y | X, Z, U)P(U | X, Z)P(X)
\]

\[
= \sum_x P(Y | X, Z)P(X) \quad \text{U free!}
\]
Back-door and front-door criteria combined result in a set of reduction rules.

Notation:
Examples of do-calculus inference rules

- Insertion/deletion of observations:

\[ P(Y \mid \text{do}(X), Z, W) = P(Y \mid \text{do}(X), W), \text{ if } (Y \perp\!\!\!\!\!\!\perp Z \mid X, W) \text{ in } G_X \]

- Action/observation exchange:

\[ P(Y \mid \text{do}(X), \text{do}(Z), W) = P(Y \mid \text{do}(X), Z, W), \text{ if } (Y \perp\!\!\!\!\!\!\perp Z \mid X, W) \text{ in } G_{XZ} \]

- Sound and complete algorithms that use these rules exist (Huang and Valtorta, 2006)
A more complex example...

\[ P(Y \mid \text{do}(X), \text{do}(Z_2)) = \sum_{z_1} P(Y \mid Z_1, \text{do}(X), \text{do}(Z_2)) \times P(Z_1 \mid \text{do}(X), \text{do}(Z_2)) \]

(Now, Rule 2, for interchanging observation/intervention)

\[ = \sum_{z_1} P(Y \mid Z_1, X, Z_2) P(Z_1 \mid X) \]

Notice: \( P(Y \mid \text{do}(X)) \) is NOT identifiable!
... and even more complex examples

\[ P(Y \mid \text{do}(X)) \text{ is identifiable} \]
(I’ll leave it as an exercise)
Planning

- Sequential decision problems:
  - More than one intervention, at different times
  - Intervention at one time depends on previous interventions and outcomes

- Example: sequential AIDS treatment (Robins, 1999)
A definition of causal effect: ACE

\[ \text{ACE}(x, x', Y) = E(Y \mid \text{do}(X = x')) - E(Y \mid \text{do}(X = x)) \]

Controlled direct effects in terms of do(.):

\[ \text{DE}_a(pcp_1, pcp_2, HIV) = \]
\[ E(HIV \mid \text{do}(AZT) = a, \text{do}(PCP = pcp_1)) - E(HIV \mid \text{do}(AZT) = a, \text{do}(PCP = pcp_2)) \]
Standardized and natural direct effects

- Controlling intermediate variables can also be done in a randomized way
  - E.g., controlled according to the age of the patient
- This notion is known as standardized effect

Natural direct effects:
- Intermediate variables arise from natural state
- E.g., adjusting for intermediate psychological effects by using placebos

(Didelez, Dawid and Geneletti, 2006)
Dealing with unidentifiability

- We saw techniques that identify causal effects, if possible
- What if it is not possible?
- The dreaded "bow-pattern":

![Diagram showing a cycle between X and Y with an arrow from X to Y and another from Y to X]
Instrumental variables

- One solution: explore parametric assumptions and other variables
- Classical case: the linear instrumental variable

\[
X = aZ + \varepsilon_X
\]
\[
Y = bX + \varepsilon_Y
\]
Instrumental variables

Let $Z$ be a standard Gaussian:

- $\sigma_{YZ} = ab$, $\sigma_{xz} = a$
- That is, $b = \frac{\sigma_{YZ}}{\sigma_{xz}}$

Bounds can be generated for non-linear systems

- Advertising: see my incoming NIPS paper for an example and references
Bayesian analysis of confounding

- Priors over confounding factors
- Buyer Beware: priors have to have a convincing empirical basis
  - not a small issue
- Example: epidemiological studies of occupational hazards
  - Are industrial sand workers more likely to suffer from lung cancer?
    - Since if so, they should receive compensations

(Steenland and Greenland, 2004)
Bayesian analysis of confounding

- Evidence for:
  - Observational evidence of higher proportion of cancer incidence in said population
  - Exposure to silica is likely to damage lungs

- Evidence against:
  - Blue-collar workers tend to smoke more than general population

(Steenland and Greenland, 2004)
Quantitative study

- Sample of 4,626 U.S. workers, 1950s-1996
  - Smoking not recorded: becomes unmeasured confounder
  - Prior: empirical priors pulled from population in general
    - Assumes relations between subpopulations are analogous

(Steenland and Greenland, 2004)
Quantitative study

(Steenland and Greenland, 2004)
Part III: Learning causal structure
From association to causation

- We require a causal model to compute predictions
- Where do you get the model?
  - Standard answer: prior knowledge
- Yet one of the goals is to use observational data
- Can observational data be used to infer a causal model?
  - or at least parts of it?
From association to causation

- This will require going beyond the Causal Markov condition...
  - independence in the causal graph $\Rightarrow$ independence in probability

- ...into the Faithfulness Condition
  - independence in the causal graph $\iff$ independence in probability

- Notice: semiparametric constraints also relevant, but not discussed here

(Spirtes et al., 2000; Pearl, 2000)
Why do we need the Faithfulness Condition?
Why would we accept the Faithfulness Condition?

- Many statisticians don’t
  - Putting the Radical Empiricist hat: “anything goes”
  - Yet many of these don’t see much of a problem with the Causal Markov condition

- But then unfaithful distributions are equivalent to accidental cancellations between paths
  - How likely is that?
Arguments for Faithfulness

- The measure-theoretical argument:
  - probability one in multinomial and Gaussian families (Spirtes et al., 2000)
- The experimental analysis argument:
  - Not spared of faithfulness issues (in a less dramatic sense)
  - How often do you see zero-effect strong causes?

![Diagram of causal relationships between Coffee-Cola, Exercise, and Heart Attack]
Arguments against Faithfulness (serious and non-serious ones)

- In practice, one only needs a distribution “close” to unfaithful for things to fail
  - Honest concern: this is possible on any sample size
- The anti-model argument:
  - “there is no such a thing as independence”
  - but accepting an independence from data is also a matter of prior. There is no such a thing called “prior-free” learning
    - What exactly does “failing to reject a null hypothesis” mean?
    - All models are null hypotheses. Mankind’s knowledge (i.e. model) of the Universe is one big null hypothesis.
- The Luddite argument:
  - “Never trust a machine to do a man’s job”
  - This is no excuse: competing models are out there and you ought to know of their existence
In practice

- There is plenty of justification for deriving what data + faithfulness entail
  - Other models can explain the data. Never trust blindly an “expert” model
    - Fear of competition for pet-theory can be a hidden reason against “automatic” causality discovery
  - No reason why use a single model: e.g. sample graphs from posterior
  - No reason to throw skepticism away
  - No reason to forget the GIGO principle
- Prior knowledge can (and should) always be added
Algorithms: principles

- Markov equivalence classes:
  - Limitations on what can be identifiable with conditional independence constraints

\[ X \perp Z \mid Y \]
The goal:

- Learn a Markov equivalence class
- Some predictions still identifiable (Spirtes et al., 2000)
- A few pieces of prior knowledge (e.g., time order) can greatly improve identifiability results
- Provides a roadmap for experimental analysis
- Side note: Markov equivalence class is not the only one
Initial case: no hidden common causes

- Little motivation for that, but easier to explain
- “Pattern”: a graphical representation of equivalence classes
More on equivalence classes

- Adjacencies are always the same in all members of a Markov equivalence class

Never equivalent, since on the left we have $X \perp \!
\!
\!\perp Y \mid \text{some set } S$
More on equivalence classes

- Unshielded colliders: always identifiable

Unshielded collider

Not a unshielded collider
More on equivalence classes

- “Propagating” unshielded colliders

Why? Different unshielded colliders

Why? Different unshielded colliders
Algorithms: two main families

- Piecewise (constraint-satisfaction) algorithms
  - Evaluate each conditional independence statement individually, put pieces together

- Global (score-based) algorithms
  - Evaluate “all” models that entail different conditional independencies, pick the “best”
    - “Best” in a statistical sense
    - “All” in a computationally convenient sense

- Two endpoints of a same continuum
A constraint-satisfaction algorithm: the PC algorithm

- Start by testing marginal independencies
  - Is $X_1$ independent of $X_2$?
  - Is $X_1$ independent of $X_3$?
  - ...
  - Is $X_{N-1}$ independent of $X_N$?

- Such tests are usually frequentist hypothesis tests of independence
  - Not essential: could be Bayes factors too
The PC algorithm

- Next step: conditional independencies tests of “size” 1
  - Is $X_1$ independent of $X_2$ given $X_3$?
  - Is $X_1$ independent of $X_2$ given $X_4$?
  - ...
  - (In practice only a few of these tests are performed, as we will illustrate)
- Continue then with tests of size 2, 3, … etc. until no tests of a given size pass
- Orient edges according to which tests passed
Assume the model on the left is the real model

Observable: samples from the observational distribution

Goal: recover the pattern (equivalence class representation)
PC, Step 1: find adjacencies
PC, Step 2: collider orientation

- X and Y are independent given T
  - Therefore, $X \rightarrow T \leftarrow Y$ is not possible
  - At the same time,
    - $X \leftarrow Z \leftarrow Y$
    - $X \rightarrow Z \rightarrow Y$
    - $X \leftarrow Z \rightarrow Y$
    are not possible, or otherwise X and Y would not be independent given T
  - Therefore, it has to be the case that $X \rightarrow Z \leftarrow Y$

- Check all unshielded triples
PC, Step 3: orientation propagation

- Since $X \rightarrow Z \leftarrow W$ is not a collider, only option left is $X \rightarrow Z \rightarrow W$

- Pattern:
Advantages and shortcomings

- **Fast**
  - Only submodels are compared
  - Prunes search space very effectively

- **Consistent**
  - On the limit on infinite data

- **But brittle**
  - Only submodels are compared: very prone to statistical mistakes
  - Doesn’t enforce global constraint of acyclicity
    - Might generate graphs with cycles
      - (which is actually good and bad)
Simple application: evolutionary biology

- Using a variation of PC + bootstrapping in biological domain:

  ![Diagram showing relationships between seed weight, fruit diameter, canopy projection, number of fruit, and number of seeds dispersed](image)

(Shipley, 1999)
Simple application: botanic

- Very small sample size (35):

  ![Diagram](Shipley, 1999)
Simple application: botanic

- Forcing blue edge by background knowledge

(Shipley, 1999)
Global methods for structure learning

- Compares whole graphs against whole graphs
- Typical comparison criterion (score function): posterior distribution
  - $P(G_1 \mid \text{Data}) > P(G_2 \mid \text{Data})$, or the opposite?
- Classical algorithms: greedy search
  - Compares nested models: one model differs from the other by an adjacency
  - Some algorithms search over DAGs, others over patterns
Greedy search over DAGs

- From the current point, evaluate all edge insertions, deletions and reversals.

Current
Greedy search over patterns

- Evaluate all patterns that differ by one adjacency from the current one
- Unlike DAG-search, consistent (starting point doesn’t matter)
- But the problem is NP-hard…

![Diagram showing patterns and their adjacencies]
Combining observational and experimental data

- Model selection scores are usually decomposable:
  - Remember DAG factorization:
    \[ \prod_i P(X_i \mid \text{Parents}(X_i)) \]
  - Score factorization (such as log-posterior):
    \[ \text{Score}(G) = \sum_i S(X_i, \text{Parents}(X_i)) \]

(Cooper and Yoo, 1999)
Combining observational and experimental data

- Experimental data follows from a local probability substitution
- Apply the “mechanism substitution” principle:

(Cooper and Yoo, 1999)
Combining observational and experimental data

- For data point \( j \), natural state:
  \[
  \text{Score}(G; j) = \log P(X_j) + \log P(Y_j | X_j) + \log P(Z_j | Y_j)
  \]

- For data point \( k \), random intervention on \( Y \)
  \[
  \text{Score}(G; k) = \log P(X_k) + \log P(Y_k | T_k) + \log P(Z_k | Y_k)
  \]
  
  e.g., \( \text{Score}(G; k) = \log P(X_k) + \log \frac{1}{2} + \log P(Z_k | Y_k) \)

(Cooper and Yoo, 1999)
Notice: greedy algorithms typically return the maximum a posteriori (MAP) graph
- Or some local maxima of the posterior

Posterior distributions
- Practical impossibility for whole graphs
  - MCMC methods should be seeing as stochastic search methods, mixing by the end of the universe
  - Still: 2 graphs are more useful than 1
- Doable for (really) small subgraphs: edges, short paths (Friedman and Koller, 2000)
Computing structure posteriors: a practical approach

- Generate a few high probability graphs
  - E.g.: use (stochastic) beam-search instead of greedy search
- Compute and plot marginal edge posteriors
A word of warning

- Uniform consistency: impossible with faithfulness only (Robins et al., 2003)
  - Considering the case with unmeasured confounding

- Rigorously speaking, standard Bayesian posteriors reflect independence models, not causal models

- There is an implicit assumption that the distribution is not “close” to unfaithfulness
  - A lot of work has yet to be done to formalize this (Zhang and Spirtes, 2003)
Methods robust to hidden common causes

- What happens to these algorithms when there are hidden common causes?
Methods robust to hidden common causes

- Even if directionality is correct:
  - they don’t tell you correct direct effects
  - which directions are unconfounded
Partial ancestral graphs (PAGs)

- New representation of equivalence classes

(Spirtes et al., 2000)
Partial ancestral graphs (PAGs)

- Type of edge:
Discovery algorithms

- Discovers and partially orients *inducing paths*:
  - Sequences of edges between nodes that can’t be blocked

- Notice
  - Can’t tell if A is a direct or indirect cause of F
  - Can’t tell if B is a cause of F
Algorithms

- The “Fast” Causal Inference algorithm (FCI, Spirtes et al., 2000):
  - “Fast” because it has a clever way of avoiding exhaustive search (e.g., as in Pearl, 2000)
- Sound and complete algorithms are fairly recent: Zhang, 2005
- Bayesian algorithms are largely underdeveloped
  - Discrete model parameterization still a challenge
Conclusion
Summary and other practical issues

- There is no magic:
  - It’s assumptions + data + inference systems
  - Emphasis on assumptions
- Still not many empirical studies
  - Requires expertise, ultimately requires experiments for validation
  - Lots of work in fixed back-door designs
    - Graphical models not that useful (more so in longitudinal studies)
The future

- Biological systems might be a great domain
  - That’s how it all started after all (Wright, 1921)
  - High-dimensional: make default back-door adjustments dull
    - Lots of direct and indirect effects of interest
  - Domains of testable assumptions
    - Observational studies with graphical models can be a great aid for experimental design
  - But beware of all sampling issues: measurement error, small samples, dynamical systems, etc.
What I haven’t talked about

- Dynamical systems ("continuous-time" models)
- Other models for (Bayesian) analysis of confounding
  - Structural equations, mixed graphs et al.
  - Potential outcomes (Rosenbaum, 2002)
- Detailed discovery algorithms
  - Including latent variable models/non-independence constraints
- Active learning
- Measurement error, sampling selection bias
- Lack of overlap under conditioning
- Formalizing non-ideal interventions
  - Non-compliance, etc.
Thank you
Textbooks

Other references

To know more

- A short article by me:

- Hernan and Robins’ incoming textbook

- Pearl’s “Causality”

- Spirtes/Glymour/Scheine’s “Causation, Prediction and Search”

- Morgan and Winship’s “Counterfactuals and Causal Inference” (2nd edition out this weekend)