A Brief Introduction to Causal Discovery and Causal inference
Correlation vs. Causation
Correlation Is Not Causation

The golden rule of causal analysis: no causal claim can be established purely by a statistical method.
Statistical Implications of Causality

• Better to talk of (in)dependence other than correlation.
• Most statisticians would agree that causality does tell us something about dependence.
• But dependence does tell us something about causality too.
(Conditional) Independence

• Two random variables $X$ and $Y$ are called independent if for each values of $(X, Y)$ denoted by $(x, y)$,
  
  • $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$
  
  • Denoted by $X \perp Y$ or $(X \perp Y)_D$
  
  • Otherwise they are dependent

• Two random variables $X$ and $Y$ are called conditionally independent given $Z$, if for each values of $(X, Y, Z)$ denoted by $(x, y, z)$,
  
  • $P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z)$
  
  • Denoted by $X \perp Y|Z$ or $(X \perp Y|Z)_D$
  
  • Otherwise they are conditionally dependent
Statistical Implications of Causality

• Reichenbach’s *Common Cause Principle* (1956) links causality and (in)dependence.

> It seems that a dependence between events $A$ and $B$ indicates either that $A$ causes $B$, or that $B$ causes $A$, or that $A$ and $B$ have a common cause.

> If $A$ and $B$ have a common cause $C$ (only), then conditioning on $C$ would make $A$ and $B$ independent. In this case, $C$ is said to ‘screen off’ the dependence between $A$ and $B$. 
The Bridge: DAG

• Use the Directed Acyclic Graph (DAG) to represent the cause-effect relations
  • Nodes as variables
  • Edges as direct causal connections

• If a DAG represents the true causal relationship, then the DAG encodes all the conditional independence relations in the true distribution which can be read-off from the graph using $d$-separation.
Make Use of DAG for Causal Discovery and Causal Inference

DAG links causality with data

‘Structural causal model’, the more symbolic extension to the DAG

Build DAG (causal graph) from observational data

Infer causal effects from given DAG (causal graph)

Causal Discovery  Build DAG (causal graph) from observational data  Infer causal effects from given DAG (causal graph)  Causal Inference
Structural Equation/Causal Model
Structural Causal Model

• A causal model is triple $\mathcal{M} = < U, V, F >$, where
  
  • $U$ is a set of exogenous (hidden) variables whose values are determined by factors outside the model;
  
  • $V = \{X_1, \ldots, X_i, \ldots\}$ is a set of endogenous (observed) variables whose values are determined by factors within the model;
  
  • $F = \{f_1, \ldots, f_i, \ldots\}$ is a set of deterministic functions where each $f_i$ is a mapping from $U \times (V \setminus X_i)$ to $X_i$. Symbolically, $f_i$ can be written as $x_i = f_i(pa_i, u_i)$

where $pa_i$ is a realization of $X_i$’s parents in $V$, i.e., $Pa_i \subseteq V$, and $u_i$ is a realization of $X_i$’s parents in $U$, i.e., $U_i \subseteq U$. 

Each causal model $\mathcal{M}$ is associated with a direct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where

- $\mathcal{V}$ is the set of nodes representing the variables $U \cup V$ in $\mathcal{M}$;
- $\mathcal{E}$ is the set of edges determined by the structural equations in $\mathcal{M}$: for $X_i$, there is an edge pointing from each of its parents $Pa_i \cup U_i$ to it.
  - Each direct edge represents the potential direct causal relationship.
  - Absence of a direct edge represents zero direct causal relationship.

Assuming the acyclicity of causality, $\mathcal{G}$ is a directed acyclic graph (DAG).

Standard terminology
- parent, child, ancestor, descendent, path, direct path
A Causal Model and Its Graph

Observed Variables $V = \{I, H, W, E\}$

Hidden Variables $U = \{U_I, U_H, U_W, U_E\}$

Model ($M$)

$$i = f_I(u_I)$$

$$h = f_H(i, u_H)$$

$$w = f_W(h, u_W)$$

$$e = f_E(i, h, w, u_E)$$

Assume $U_I$ and $U_H$ are correlated.

Graph ($G$)

$H$ (Hour of Study)

$I$ (Interest)

$U_I$

$U_H$

$E$ (Exam Grade)

$W$ (Working Strategy)
A Markovian Model and Its Graph

With causal sufficiency assumption

Model \((M)\)

\[ i = f_I(u_I) \]
\[ h = f_H(i, u_H) \]
\[ w = f_W(h, u_W) \]
\[ e = f_E(i, h, w, u_E) \]

Assume \(U_I, U_H, U_W, U_E\) are mutually independent.

Graph \((G)\)

- \(I\) (Interest)
- \(H\) (Hour of Study)
- \(W\) (Working Strategy)
- \(E\) (Exam Grade)
Causal Graph of Markovian Model

Each node is associated with an observable conditional probability table (CPT) $P(x_i|p_{a_i})$
Causal Discovery
\( d \)-Separation

\( d \)-separation

**Definition of \( d \)-separation**

- A path \( q \) is said to be blocked by conditioning on a set \( Z \) if:
  - \( q \) contains a chain \( i \rightarrow m \rightarrow j \) or a fork \( i \leftarrow m \rightarrow j \) such that the middle node \( m \) is in \( Z \), or
  - \( q \) contains a collider \( i \rightarrow m \leftarrow j \) such that the middle node \( m \) is not in \( Z \) and such that no descendant of \( m \) is in \( Z \).
- \( Z \) is said to \( d \)-separate \( X \) and \( Y \) if \( Z \) blocks every path from \( X \) to \( Y \), denoted by \( (X \perp Y|Z)_G \)

**If the DAG represents the true causal relationship**

\[(X \perp Y|Z)_G \iff (X \perp Y|Z)_D\]
**d-Separation**

- Example (blocking of paths)

  - Path from $X$ to $Y$ is blocked by conditioning on $\{U\}$ or $\{Z\}$ or both $\{U, Z\}$

- Example (unblocking of paths)

  - Path from $X$ to $Y$ is blocked by $\emptyset$ or $\{U\}$
  - Unblocked by conditioning on $\{Z\}$ or $\{W\}$ or both $\{Z, W\}$
$d$-Separation

• Examples ($d$-separation)

![Graph Diagram]

• We have following $d$-separation relations
  • $(X \perp Y|Z)_G, (X \perp Y|U)_G, (X \perp Y|ZU)_G$
  • $(X \perp Y|ZW)_G, (X \perp Y|UW)_G, (X \perp Y|ZUW)_G$
  • $(X \perp Y|VZUW)_G$

• However we do NOT have
  • $(X \perp Y|VZU)_G$
PC Algorithm (Peter Spirtes & Clark Glymour)

• Faithfulness assumption
• Causal sufficiency (no hidden common cause) assumption
• The **BEST** we can do without further assumptions (or knowledge).
• Usually **CANNOT** identity the unique causal graph (up to the Markov equivalent class)

\[
\begin{align*}
X & \rightarrow Z \rightarrow Y \\
X & \leftarrow Z \rightarrow Y \\
X & \leftarrow Z \leftarrow Y \\
X & \rightarrow Z \leftarrow Y
\end{align*}
\]
PC Algorithm: The Sketch

1. Construct the skeleton
   1. Start with a fully connected undirected graph
   2. Remove all edges $X - Y$ with $X \perp Y$
   3. Remove all edges $X - Y$ for which there is a neighbor $Z \neq Y$, $X$ with $X \perp Y | Z$
   4. Remove all edges $X - Y$ for which there are two neighbors $Z_1, Z_2 \neq Y, X$ with $X \perp Y | Z_1, Z_2$
   5. ...

2. Orient the arrows by finding v-structures $X \rightarrow Z \leftarrow Y$
Open-source Software - Tetrad

• [http://www.phil.cmu.edu/tetrad/](http://www.phil.cmu.edu/tetrad/)

• Implement a large set of Constraint-Based and Score-Based causal discovery algorithms.
We Can Do Better Than PC Algorithm

• Given $X, Y$, can we distinguish $X \rightarrow Y$ and $X \leftarrow Y$?

• If some additional assumptions are made about the functional and/or parametric forms of the underlying true data-generating structure, then one can exploit asymmetries in order to identify the direction of a structural relationship.
Additive Noise

• Given the linear structural equations
  \[ X = U_X \text{ and } Y = X + U_Y \text{ such that } U_Y \perp U_X \]
• If \( U_X \) or \( U_Y \) is non-Gaussian
• Then the causal direction \( X \to Y \) is identifiable
Independent Causal Mechanisms

• The causal generative process of a system’s variables is composed of autonomous modules that do not inform or influence each other.

• Suppose $X \rightarrow Y$, then $P(x)$ and $P(y|x)$ should be independent.

• In other words, unsupervised learning on $X$ should not improve supervised learning $X \rightarrow Y$.

• Will be different if decompose the distribution to $P(y)$ and $P(x|y)$.
Causal Inference
WHAT Kind OF QUESTIONS SHOULD THE CAUSAL MODEL ANSWER
THE CAUSAL HIERARCHY

• Observational Questions: (What is?) \( P(y | A) \)
  • “What if we see A”

• Action Questions: (What if?) \( P(y \ | \ do(A)) \)
  • “What if we do A?”

• Counterfactuals Questions: (Why?) \( P(y_A' \ | \ A) \)
  • “What if we did things differently?”

• Options:
  • “With what probability?”
Ladder of Causality

Intervention

• Physical intervention
Intervention and *do*-Operation

• The basic operation of manipulating a causal model.
  • Simulate the physical intervention.
  • Forces some observed variables $X \in V$ to take certain constants $x$.

• Mathematically formulated as $do(X = x)$ or simply $do(x)$.

• The **effect of intervention** on all other observed variables $Y = V \setminus X$ is represented by the **post-intervention distribution** of $Y$.
  • Denoted by $P(Y = y | do(X = x))$ or simply $P(y | do(x))$;
Intervention and \textit{do}-Operation

• In causal model $\mathcal{M}$, intervention \textit{do}(\(x^*\)) is defined as the substitution of structural equation \(x = f_X(pa_X, u_X)\) with value \(x^*\). The causal model after performing \textit{do}(\(x^*\)) is denoted by $\mathcal{M}_{x^*}$.

\[
\mathcal{M}: \quad x = f_X(pa_X, u_X) \quad \xrightarrow{\text{do}(x^*)} \quad \mathcal{M}_{x^*}: \quad x = x^*
\]

• From the point of view of the causal graph, performing \textit{do}(\(x^*\)) is equivalent to setting the node $X$ to value $x^*$ and removing all the incoming edges in $X$. 

\[
\begin{array}{c}
\text{\(X\)} \quad \xrightarrow{\text{do}(x^*)} \quad \text{\(X^*\)}
\end{array}
\]
Intervention in Markovian Model

• In the Markovian model, the post-intervention distribution $P(y|do(x))$ can be calculated from the CPTs, known as the truncated factorization:

$$P(y|do(x)) = \prod_{Y \in Y} P(y|Pa_Y) \delta_{X \leftarrow x}$$

• where $\delta_{X \leftarrow x}$ means assigning attributes in $X$ involved in the term ahead with the corresponding values in $x$.

• Specifically, for a single attribute $Y$ given an intervention on a single attribute $X$,

$$P(y|do(x)) = \sum_{V \setminus \{X,Y\}} \prod_{v \in V \setminus \{X\}} P(v|Pa_v) \delta_{X \leftarrow x}$$
Intervention Example

• What is the probability of getting grade A if we change the study hour to 2?

Graph ($G$)

Model ($M$)

$$i = f_I(u_i)$$

$$h = f_H(i, u_H)$$

$$w = f_W(h, u_W)$$

$$e = f_E(i, h, w, u_E)$$
Intervention Example

• What is the probability of getting grade A if we change the study hour to 2, i.e., $do(H = 2)$?

Graph ($G'$)

Model ($M'$)

- $i = f_I(u_I)$
- $h = 2$
- $w = f_W(h, u_W)$
- $e = f_E(i, h, w, u_E)$

• Find $P(E = 'A'|do(H = 2))$
Intervention Example

Graph (\(G'\))

- \(I\) (Interest)
- \(W\) (Working Strategy)
- \(E\) (Exam Grade)

\[ do(H = 2) \]
(Hour of Study)

Model (\(M'\))

\[ i = f_i(u_I) \]
\[ h = 2 \]
\[ w = f_W(h, u_W) \]
\[ e = f_E(i, h, k, u_E) \]

\[ P(y|do(x)) = \sum_{v \in \mathcal{V}\setminus\{x, y\}} \prod_{v \in \mathcal{V}\setminus\{x\}} P(v|Pa_v) \delta_{x\leftarrow x} \]

\[ P(E = 'A'|do(H = 2)) = \sum_{I, W} P(i)P(w|H = 2)P(E = 'A'|i, H = 2, w) \]
Applications of CI in ML

- Fair machine learning
- Reinforcement learning
- Transfer learning and multitask learning
Open-source packages for causal inference

• Microsoft/DoWhy:
  • https://github.com/microsoft/dowhy

• IBM/causallib
  • https://github.com/IBM/causallib
Useful Resources

• Four books

• Website: [http://bayes.cs.ucla.edu/](http://bayes.cs.ucla.edu/)