Causality and its Role in Reasoning, Explainability, and Generalizability

Adèle H. Ribeiro

https://adele.github.io/ | adele.ribeiro@uni-marburg.de

Faculty of Mathematics and Computer Science
Philipps-Universität Marburg

Lisbon Machine Learning School (LxMLS)
July 20th, 2023
Recent **Breakthroughs** in AI

- We can learn models that makes **predictions** extremely well in high-dimensional settings.
- In particular, there are huge progresses in natural processing language, computer vision, and reinforcement learning.
Driverless Cars to San Jose

AI system is better than human doctors at predicting breast cancer

The New York Times

The Shift

GPT-4 Is Exciting and Scary

Today, the new language model from OpenAI may not seem all that dangerous. But the worst risks are the ones we cannot anticipate.
Current Challenges in AI

A fairer way forward for AI in healthcare: Without careful implementation, inequality.

Linda Nordling

Why artificial intelligence needs to understand consequences

A machine with a grasp of cause and effect could learn more like a human, through imagination and regret.
Judea Pearl — Causality

Director of the Cognitive Systems Laboratory at the University of California, Los Angeles.

In 2011, he won the A. M. Turing Award (the highest distinction in computer science and a $250,000 prize)

“for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.”
— Association for Computing Machinery (ACM)

“Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality.”
— The Book of Why: The New Science of Cause and Effect
Yoshua Bengio — Deep Learning

Professor at the University of Montreal, and the Founder and Scientific Director of Mila – Quebec AI Institute

In 2018, he won the A. M. Turing Award, with Geoffrey Hinton, and Yann LeCun

“for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing.”

— Association for Computing Machinery (ACM)

“Causality is very important for the next steps of progress of machine learning,” — interview with IEEE Spectrum, 2020.
Guido W. Imbens & Joshua D. Angrist

In 2021, they won the Nobel Prize in Economics (about $1 million)

“for their methodological contributions to the analysis of causal relationships”

Guido W. Imbens
Professor of Applied Econometrics in Stanford University

Joshua D. Angrist
Professor of Economics at the Massachusetts Institute of Technology
Causality is an essential component in the development of the new generation of Artificial Intelligence methods, allowing important capabilities such as

**Explainability:** provides a better understanding of the underlying mechanisms, e.g., learning directionality and confounding through causal structure learning.

**Reasoning:** can determine the effect of *unrealized* interventions rather than just predicting an outcome (i.e., can distinguish between association and causation).

**Fairness:** captures and disentangles any mechanisms of discrimination that may be present, including direct, indirect-mediated, and indirect-confounded.

**Generalizability:** allows the transportability of causal effects across different domains.

**Data Fusion:** provides language and theory to cohesively combine prior knowledge and data from multiple and heterogeneous studies.
Causal Data Science

Goal is to develop language, criteria, and algorithms for:

- **Data-Fusion**: cohesively combining heterogenous datasets,
- **Causal Inference**: inferring the effects of interventions, and
- **Decision-Making**: making robust and generalizable decisions.

Causal inference and the data-fusion problem

Elias Bareinboim\textsuperscript{a,b,1} and Judea Pearl\textsuperscript{a}

\textsuperscript{a}Department of Computer Science, University of California, Los Angeles, CA 90095; and \textsuperscript{b}Department of Computer Science, Purdue University, West Lafayette, IN 47907

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved March 15, 2016 (received for review June 29, 2015)

http://causalfusion.net
Causality Theory by Judea Pearl
Causality Theory by Judea Pearl

https://causality101.net/
Prediction vs Reasoning

Statistical Association vs Causation
Predictive Tasks

**Task:** Can I guess how serious/big is the fire by the number of firefighters in action?

Yes!

$X$: Number of firefighters in action  
$Y$: Seriousness of fire

$\rho_{XY} \neq 0 \implies X$ is a good predictor of $Y$

$P(Y = y | X = x) \neq P(Y = y)$

Observational Probability Distribution

**Conclusion:** The seriousness of fire increases with the number of firefighters.
**Conclusion:** The size of the fire increases with the number of firefighters. In other words, the fewer the firefighters, the smaller the fire.

Should we decrease the number of firefighters to reduce the fire?
Effect of Interventions

$X$: Number of firefighters in action
$Y$: Seriousness of fire

$X = f_X(Y, U_X, U_{XY})$

$Y = f_Y(U_Y, U_{XY})$

In other words, $Y$ is not a function of $X$

Y is not caused by $X$
Effect of Interventions

\[ X: \text{Number of firefighters in action} \]
\[ Y: \text{Seriousness of fire} \]

In other words, \( Y \) is not caused by \( X \).

Changing \( X \) won’t change the value of \( Y \).

\[ P(Y = y \mid do(X = x)) = P(Y = y) \]

**Interventional Probability Distribution**

The action/intervention on \( X \), \( do(X = x) \), is independent of \( Y \).

**Conclusion:** we cannot change the size of the fire by changing the number of firefighters.
Structural Causal Model (SCM)

EXPLAINABILITY AND THE DATA GENERATING MODEL
Definition: A structural causal model $\mathcal{M}$ (or, data generating model) is a tuple $\langle V, U, \mathcal{F}, P(u) \rangle$, where

- $V = \{V_1, \ldots, V_n\}$: are endogenous variables
- $U = \{U_1, \ldots, U_m\}$: are exogenous variables
- $\mathcal{F} = \{f_1, \ldots, f_n\}$: are functions determining $V$, i.e., $v_i \leftarrow f_i(pa_i, u_i)$, where $Pa_i \subseteq V$ are endogenous causes (parents) of $V_i$ and $U_i \subseteq U$ are exogenous causes of $V_i$.
- $P(U)$ is the probability distribution over $U$.

Assumption: $\mathcal{M}$ is recursive, i.e., there are no feedback (cyclic) mechanisms.
Pre-Interventional/Observational SCM

\[\mathcal{M} = \{ \mathcal{V}, \mathcal{U}, \mathcal{F}, \mathcal{P} \} \]

\[\mathcal{V} = \{ X, Y \}\]
\[\mathcal{U} = \{ U_{XY}, U_X, U_Y \}\]
\[\mathcal{F} = \begin{cases} X = f_X(U_X, U_{XY}) \\ Y = f_Y(X, U_Y, U_{XY}) \end{cases}\]
\[\mathcal{P}(U)\]

Observational Distribution

\[P(\mathcal{V}) \doteq P_{\mathcal{M}}(\mathcal{V})\]

Can we predict better the value of \(Y\) after observing \(X = x\)?

\[P(Y = y | X = x) \neq P(Y = y) \implies X \text{ is correlated to } Y\]

Post-Interventional/Interventional SCM

\[\mathcal{M}_x = \{ \mathcal{V}, \mathcal{U}, \mathcal{F}, \mathcal{P} \} \]

\[\mathcal{V} = \{ X, Y \}\]
\[\mathcal{U} = \{ U_{XY}, U_X, U_Y \}\]
\[\mathcal{F} = \begin{cases} X = x \\ Y = f_Y(x, U_Y, U_{XY}) \end{cases}\]
\[\mathcal{P}(U)\]

Interventional Distribution

\[P(\mathcal{V} | do(X = x)) \doteq P_{\mathcal{M}_x}(\mathcal{V})\]

Can we predict better the value of \(Y\) after making an intervention \(do(X = x)\)?

\[\exists x \text{ s.t. } P_{\mathcal{M}_x}(Y = y) \neq P(Y = y) \implies X \text{ is a cause of } Y\]
Structural Equation Model (SEM)

\[ M = \begin{cases} 
V = \{X, Y, Z\} \\
U = \{\epsilon_X, \epsilon_Y, \epsilon_Z\} \\
F = \begin{cases} 
Z = \beta_{Z0} + \epsilon_Z \\
X = \beta_{X0} + \beta_{XZ}Z + \epsilon_X \\
Y = \beta_{Y0} + \beta_{YZ}Z + \beta_{YX}X + \epsilon_Y 
\end{cases} \\
U \sim N \left( 0, \Sigma = \begin{bmatrix} 
\sigma_X & 0 & 0 \\
0 & \sigma_Y & 0 \\
0 & 0 & \sigma_Z 
\end{bmatrix} \right) 
\end{cases} \]

- Linear functions
- Normal distribution
- Markovianity / Causal Sufficiency: Error terms in \( U \) are independent of each other (diagonal covariance matrix).

Full specification of an SCM requires parametric and distributional assumptions. Estimation of such models usually requires strong assumptions (e.g., Markovianity).
The knowledge required to fully specify an SCM is usually *unavailable* in practice.

Is it possible to identify the effect of interventions from *observational* data without fully specifying the SCM (i.e., in a non-parametric fashion)?

Yes, with structural knowledge encoded as a causal diagram!
Encoding Structural Causal Knowledge

Acyclic Directed Acyclic Graph (ADMG)
Causal Diagrams
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)

\( \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \)

\( \mathcal{M} = \begin{cases} 
V = \{A, B, C\} \\
U = \{U_A, U_B, U_C, U_{AB}\} \\
\mathcal{F} = \begin{cases} 
A \leftarrow f_A(U_{AB}, U_A) \\
B \leftarrow f_B(U_{AB}, U_B) \\
C \leftarrow f_C(A, B, U_C) 
\end{cases} \\
P(U) 
\end{cases} \)

Induced Causal Diagram

An SCM \( \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \) induces a causal diagram such that, for every \( V_i, V_j \in V \):

\( V_i \rightarrow V_j \), if \( V_i \) appears as argument of \( f_j \in \mathcal{F} \).
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)

\[ \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \]

\[ \mathcal{M} = \begin{cases} 
V = \{A, B, C\} \\
U = \{U_A, U_B, U_C, U_{AB}\} \\
\mathcal{F} = \begin{cases} 
A \leftarrow f_A(U_{AB}, U_A) \\
B \leftarrow f_B(U_{AB}, U_B) \\
C \leftarrow f_C(A, B, U_C) 
\end{cases} \\
P(U) 
\end{cases} \]

Induced Causal Diagram

An SCM \( \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \) induces a causal diagram such that, for every \( V_i, V_j \in V \):

\( V_i \rightarrow V_j \), if \( V_i \) appears as argument of \( f_j \in \mathcal{F} \).
Causal Diagram: Encoder of **Structural Knowledge**

### Structural Causal Model (SCM)

\[ \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \]

- \( V = \{A, B, C\} \)
- \( U = \{U_A, U_B, U_C, U_{AB}\} \)
- \( \mathcal{F} = \{A \leftarrow f_A(U_{AB}, U_A), B \leftarrow f_B(U_{AB}, U_B), C \leftarrow f_C(A, B, U_C)\} \)
- \( P(U) \)

An SCM \( \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \) induces a causal diagram such that, for every \( V_i, V_j \in V \):

- \( V_i \rightarrow V_j \), if \( V_i \) appears as argument of \( f_j \in \mathcal{F} \).
- \( V_i \leftrightarrow V_j \) if the corresponding \( U_i, U_j \in U \) are correlated or \( f_i, f_j \) share some argument \( U \in U \).
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)
\[ \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \]

\[ \begin{align*}
\mathcal{M} & = \begin{cases}
V = \{A, B, C\} \\
U = \{U_A, U_B, U_C, U_{AB}\} \\
\mathcal{F} = \begin{cases}
A \leftarrow f_A(U_{AB}, U_A) \\
B \leftarrow f_B(U_{AB}, U_B) \\
C \leftarrow f_C(A, B, U_C)
\end{cases}
\end{cases}
\end{align*} \]

An SCM \( \mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle \) induces a causal diagram such that, for every \( V_i, V_j \in V \):

- \( V_i \rightarrow V_j \), if \( V_i \) appears as argument of \( f_j \in \mathcal{F} \).
- \( V_i \leftrightarrow V_j \) if the corresponding \( U_i, U_j \in U \) are correlated or \( f_i, f_j \) share some argument \( U \in U \).
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)

\(\mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle\)

\(\mathcal{M} = \{\)

- \(V = \{A, B, C\}\)
- \(U = \{U_A, U_B, U_C, U_{AB}\}\)
- \(\mathcal{F} = \{\)
  - \(A \leftarrow f_A(U_{AB}, U_A)\)
  - \(B \leftarrow f_B(U_{AB}, U_B)\)
  - \(C \leftarrow f_C(A, B, U_C)\)
- \(P(U)\)

\(\)\n
Induced Causal Diagram

An SCM \(\mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle\) induces a causal diagram such that, for every \(V_i, V_j \in V\):

- \(V_i \rightarrow V_j\), if \(V_i\) appears as argument of \(f_j \in \mathcal{F}\).
- \(V_i \leftrightarrow V_j\) if the corresponding \(U_i, U_j \in U\) are correlated or \(f_i, f_j\) share some argument \(U \in U\).
D-Separation and Implied Conditional Independencies

**Definition (inactive):** A triplet \( \langle V_i, V_m, V_j \rangle \) is said to be **inactive** relative to a set \( Z \) if the middle node \( V_m \):

1. Is a non-collider and is in \( Z \); or
2. Is a collider and neither it nor any of its descendants in \( Z \).

**Definition (d-separation):** A path \( p \) in a causal diagram \( G \) is said to be **d-separated** (or blocked) by a set of variables \( Z \) if and only if \( p \) contains an inactive triplet in it.

A set \( Z \) d-separates \( X \) and \( Y \) if and only if \( Z \) blocks every path between a node in \( X \) and a node in \( Y \).

\[
\begin{align*}
X & \xrightarrow{\text{d-separated}} Y \\
B & \xrightarrow{\text{d-separated}} W
\end{align*}
\]

Does \( Z \) d-separates \( X \) and \( Y \)?
\[
Z: \quad \checkmark \{\} \quad \checkmark \{B\} \quad \times \{W\} \quad \times \{B, W\}
\]

We have that \((X \perp Y)_G\), \((X \perp Y \mid B)_G\), and \((X \perp Y \mid W)_G\), but \((X \perp Y \mid B, W)_G\)

**Global Markov property:** \((X \perp Y \mid Z)_G \Rightarrow (X \perp Y \mid Z)_P\)

D-separations in \( G \) imply conditional independencies in \( P \)
Markov Blanket (MB) of $V$: the bidirected connected component (district) of $V$ (excluding $V$ itself) and the parents of the district of $V$, i.e.:

$$\text{mb}_G(V) = \text{dis}_G(V) \cup \text{Pa}_G(\text{dis}_G(V)) \setminus \{V\}$$
Randomized Experiments

Randomized Experiments / Control Trials (e.g. RCT) allow the identification of causal effects by leveraging randomization of the treatment assignment.
Pearl’s Inferential Hierarchy

Associational vs Interventional vs Counterfactual
What is induced by the SCM?

**Observational SCM**

\[ M = \{ \mathcal{V}, \mathcal{U}, \mathcal{F}, P(\mathcal{U}) \} \]

\[ \mathcal{V} = \{ X, Y \} \]
\[ \mathcal{U} = \{ U_{XY}, U_X, U_Y \} \]
\[ \mathcal{F} = \{ X = f_X(U_X, U_{XY}), Y = f_Y(X, U_Y, U_{XY}) \} \]

**Interventional SCM**

\[ M_x = \{ \mathcal{V}, \mathcal{U}, \mathcal{F}, P(\mathcal{U}) \} \]

\[ \mathcal{V} = \{ X, Y \} \]
\[ \mathcal{U} = \{ U_{XY}, U_X, U_Y \} \]
\[ \mathcal{F} = \{ X = x, Y = f_Y(x, U_Y, U_{XY}) \} \]

**Observational Causal Diagram**

Observational Distribution

**Interventional Causal Diagram**

Interventional Distribution

Loss of Information
### Reality

Structural Causal Model (SCM)

$$\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$$

- **Observational**
  - \( \mathbf{V} = \{X, Y\} \)
  - \( \mathbf{U} = \{U_{XY}, U_X, U_Y\} \)
  - \( \mathcal{F} = \{X \leftarrow f_X(U_X, U_{XY}), Y \leftarrow f_Y(X, U_Y, U_{XY}), P(\mathbf{U})\} \)

- **Interventional**
  - \( \mathcal{M}_x = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle \)
    - \( \mathbf{V} = \{X, Y\} \)
    - \( \mathbf{U} = \{U_{XY}, U_X, U_Y\} \)
    - \( \mathcal{F} = \{X \leftarrow x, Y \leftarrow f_Y(x, U_Y, U_{XY}), P(\mathbf{U})\} \)

### Structural Knowledge

Causal Diagram

- \( G \)

- \( \mathcal{F} \)

### Data

- \( \hat{P}(Y | X = x) \)
  - \( \mathcal{F} \)

- \( \hat{P}(Y | do(X = x)) = ? \)
  - \( \mathcal{F} \)
Markovian Parametrization

Data

Potential Causal Diagrams

Potential SCMs

\[ \mathcal{M}_{11} = \langle V, U_1, \mathcal{F}_{11}, P_{11}(u_1) \rangle \]

\[ \cdots \]

\[ \mathcal{M}_{1k_1} = \langle V, U_1, \mathcal{F}_{1k_1}, P_{1k_1}(u_1) \rangle \]

\[ \mathcal{M}_{21} = \langle V, U_2, \mathcal{F}_{21}, P_{21}(u_2) \rangle \]

\[ \cdots \]

\[ \mathcal{M}_{2k_2} = \langle V, U_2, \mathcal{F}_{2k_2}, P_{2k_2}(u_2) \rangle \]

\[ \mathcal{M}_{31} = \langle V, U_3, \mathcal{F}_{31}, P_{31}(u_3) \rangle \]

\[ \cdots \]

\[ \mathcal{M}_{3k_3} = \langle V, U_3, \mathcal{F}_{3k_3}, P_{3k_3}(u_3) \rangle \]

\[ \cdots \]

\[ \mathcal{M}_{41} = \langle V, U_4, \mathcal{F}_{41}, P_{41}(u_4) \rangle \]

\[ \cdots \]

\[ \mathcal{M}_{4k_4} = \langle V, U_4, \mathcal{F}_{4k_4}, P_{4k_4}(u_4) \rangle \]

\[ \mathcal{M}_{51} = \langle V, U_5, \mathcal{F}_{51}, P_{51}(u_5) \rangle \]

\[ \cdots \]

\[ \mathcal{M}_{5k_5} = \langle V, U_5, \mathcal{F}_{5k_5}, P_{5k_5}(u_5) \rangle \]

\[ \cdots \]

Encoded Knowledge / Assumptions

True Model

Markovian Parametrization

Observed Data

\[ P(Y|X = x) \]
Multiple neural nets fit the data equally well, leading to different causal explanations!
Ladder of Causation

<table>
<thead>
<tr>
<th>Layer</th>
<th>Task / Language</th>
<th>Typical Question</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Countertfactual</td>
<td>Structural Causal Model</td>
<td>What if I had acted differently?</td>
<td>Was it the aspirin that stopped my headache?</td>
</tr>
<tr>
<td>Interventional</td>
<td>ML- Reinforcement (Causal Bayes Net)</td>
<td>What if I do X? What would Y be if I intervene on X?</td>
<td>Will my headache be cured if I take aspirin?</td>
</tr>
<tr>
<td>Associational</td>
<td>ML- (Un)Supervised (Decision trees, Deep nets, …)</td>
<td>What if I see? How would seeing X change my belief in Y?</td>
<td>What does a symptom tell us about the disease?</td>
</tr>
</tbody>
</table>

Ladder of Causation

<table>
<thead>
<tr>
<th>Layer</th>
<th>Task / Language</th>
<th>Typical Question</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Imaginining</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Doing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Seeing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cross-layer inferences:

Doing
- most of the inferences are about causal effects (policies, treatments, decisions)
- Will my headache be cured if I take aspirin?

Seeing
- most of the available data is observational, passively collected
- What does a symptom tell us about the disease?

Causal Hierarchy Theorem: to answer questions in layer $i$, we need information from layer $i$ or higher.
Association vs Causation

Will we be able to decide the true relationship just by “seeing” more data?

https://xkcd.com/925/ - Creative Commons Attribution-NonCommercial 2.5 License.
Causal Effect Identification

Graphical Criteria, Do-Calculus, and ID-Algorithm
Causal Effect

The causal effect of a (set of) treatment variable(s) $X$ on a (set of) outcome variable(s) $Y$ is a quantity derived from $P(Y \mid do(X))$ that tells us how much $Y$ changes due to an intervention $do(X = x)$.

Examples:

• *Average Treatment Effect (ATE)* for discrete treatments:
  \[
  \mathbb{E}[Y \mid do(X = x')] - \mathbb{E}[Y \mid do(X = x)],
  \]
  defined for two treatment levels $x'$ and $x$ of $X$.

• *Average Treatment Effect (ATE)* for continuous treatments,
  \[
  \frac{\partial \mathbb{E}[Y_i \mid do(X_j = x_j)]}{\partial x_j}, \text{ for all } Y_i \in Y, \text{ and } X_j \in X.
  \]
  The derivative shows the rate of change of $Y$ w.r.t. $do(X = x)$

  where $\mathbb{E}[Y \mid do(X = x)] = \sum_{y \in \Omega_Y} yP(y \mid do(x))$
Classical Causal Effect Identification

1. Query
   \[ P(y \mid do(x)) \]

2. Causal Constraints
   \[ X \rightarrow M \rightarrow Y \]

3. Probability Distributions
   \[ P(x, m, y) \]
   **Observational Distribution**

\[ P(y \mid do(x)) = \sum_m P(m \mid x) \sum_{x'} P(y \mid m, x') P(x') \]

- Interventional Distribution
- Available Distributions
- Structural knowledge available

Causal Effect Identifiability: The causal effect of a (set of) treatment variable(s) $X$ on a (set of) outcome variable(s) $Y$ is said to be identifiable from a causal diagram $G$ and the probability of the observed variables $P(V)$ if the interventional distribution $P(Y \mid do(X))$ is uniquely computable, i.e., if for every pair of SCMs $\mathcal{M}_1$ and $\mathcal{M}_2$ that induce $G$ and $P^{\mathcal{M}_1}(V) = P^{\mathcal{M}_2}(V) = P(V) > 0$, $P^{\mathcal{M}_1}(Y \mid do(X)) = P^{\mathcal{M}_2}(Y \mid do(X)) = P(Y \mid do(X))$.

In words, causal effect identifiability means that, no matter the form of true SCM, for all models $\mathcal{M}$ agreeing with $\langle G, P(V) \rangle$, they also agree in $P(y \mid do(x))$. 
The Effect Identification Problem

**Causal Effect Identifiability:** The causal effect of a (set of) treatment variable(s) $X$ on a (set of) outcome variable(s) $Y$ is said to be identifiable from a causal diagram $G$ and the probability of the observed variables $P(V)$ if the interventional distribution $P(Y \mid do(X))$ is **uniquely computable**, i.e., if for every pair of SCMs $\mathcal{M}_1$ and $\mathcal{M}_2$ that induce $G$ and $P_{\mathcal{M}_1}(V) = P_{\mathcal{M}_2}(V) = P(V) > 0$, $P_{\mathcal{M}_1}(Y \mid do(X)) = P_{\mathcal{M}_2}(Y \mid do(X)) = P(Y \mid do(X))$.

In words, causal effect identifiability means that, no matter the form of true SCM, for all models $\mathcal{M}$ agreeing with $\langle G, P(V) \rangle$, they also agree in $P(y \mid do(x))$. 
Tools for Causal Identification

1. Truncated Factorization / G-computation formula
2. Graphical criteria
   1. Parent adjustment
   2. Backdoor Adjustment
   3. Front-door Adjustment
3. Do-Calculus (a.k.a. Causal Calculus)
4. Identify Algorithm (a.k.a. ID algorithm)

Markovian Models
A few interesting (albeit still constrained) scenarios
General Semi-Markovian Scenarios


Identification in Markovian Models

**Truncated Factorization — Markovian:** Let $G$ be a causal diagram for the collection $P_*$ of all interventional distributions $P_x(V)$, for any $X \subseteq V$. It follows that $P_x(V)$ factorizes as:

$$P_x(v) = P(v | do(x)) = \prod_{v_i \in V \setminus X} P_x(v_i | pa_i) \Bigg|_{x=x}$$

Follows from $P_x(v) = P(v | do(x))$ being Markov relative to $G_X$

$$= \prod_{v_j \in V \setminus X} P(v_i | pa_i) \Bigg|_{x=x}$$

Markovian SCMs have the modularity property, i.e., $P_x(v_i | pa_i) = P(v_i | pa_i)$

Causal Effect of $X$ on $Y$: $P(y | do(x)) = \sum_{V \setminus (Y \cup X)} \prod_{v_j \in V \setminus X} P(v_i | pa_i) \Bigg|_{x=x}$

- In Markovian Models, the joint interventional distribution (and hence any causal effect) is always identifiable.
- This factorization is a.k.a “manipulation theorem” (Spirtes et al. 1993) or G-computation (Robins 1986, p. 1423).
Example: Identifiable Effect

Causal Effect of $X$ on $Y$: \[
P(y \mid do(x)) = \sum_{V \setminus (Y \cup X)} \prod_{V \setminus X} P_x(v_i \mid pa_i) \bigg|_{X=x}\]

$P(x, y, z) = P(z)P(x \mid z)P(y \mid x, z)$

$P(y, z \mid do(x)) = P(z)P(y \mid x, z)$

$\implies P(y \mid do(x)) = \sum_z P(z)P(y \mid x, z)$
Identification in Semi-Markovian Models

Adjustment over parents:

Let $G$ be a causal graph with no unmeasured parents.

Then, the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_{pa_x} P \left( y \mid x, pa_x \right) P \left( pa_x \right)$$

Proof follows from the truncated factorization for Markovian models!

$$X = \{X\}$$
$$Y = \{Y\}$$
$$Pa_X = \{Z_1, Z_2\}$$

$$P(y \mid do(x)) = \sum_{z_1, z_2} P \left( y \mid x, z_1, z_2 \right) P \left( z_1, z_2 \right)$$
Adjustment over parents:

Let $G$ be a causal graph with no unmeasured parents.

Then, the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_{pa_x} P \left( y \mid x, pa_x \right) P \left( pa_x \right)$$

Proof follows from the truncated factorization for Markovian models!

$$X = \{X\}$$
$$Y = \{Y\}$$
$$Pa_X = \{Z_1, Z_2\}$$

$$P(y \mid do(x)) = \sum_{z_1, z_2} P \left( y \mid x, z_1, z_2 \right) P \left( z_1, z_2 \right)$$

After conditioning on the parents, the association between $X$ and $Y$ is only due to the direct path.
Adjustment over parents:

Let $G$ be a causal graph with no unmeasured parents.

Then, the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_{pa_x} P(y \mid x, pa_x) P(pa_x)$$

$$Pa_x = \{Z_2\}$$

$$U_x = \{U_{X,Z_2}\}$$
Adjustment over parents:

Let $G$ be a causal graph with no unmeasured parents.

Then, the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_{p_a x} P\left( y \mid x, p_a x \right) P\left( p_a x \right)$$

Adjustment over parents:

$$P(y \mid do(x)) = \sum_{z_1, z_2} P\left( y \mid x, z_1, z_2 \right) P\left( z_1, z_2 \right)$$

$Pa_x = \{Z_2\}$

$U_x = \{U_{X,Z_2}\}$
Identification in Semi-Markovian Models

Adjustment over parents:

Let $G$ be a causal graph with no unmeasured parents.

Then, the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_{p_{ax}} P\left(y \mid x, p_{ax}\right) P\left(p_{ax}\right)$$

Adjustment over parents:

$Pa_x = \{Z_2\}$

$U_x = \{U_{X,Z2}\}$

After conditioning on the $\{Z_1, Z_2\}$, the association between $X$ and $Y$ is also due to a spurious / confounding path.
Backdoor Adjustment

Let $X$ be a set of treatment variables and $Y$ a set of outcome variables in the causal graph $G$.

If there exists a set $Z$ such that:

1. for every $X \in X$ and $Y \in Y$, $Z$ blocks every path between $X$ and $Y$ that has an arrow into $X$, and
2. no node in $Z$ is a descendant of a variable $X \in X$ (all variables in $Z$ are pre-treatment)

Then, $Z$ satisfies the backdoor criterion for $(X, Y)$ and, then the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_z P(y \mid x, z) P(z)$$

Also known as *confounding paths*, or *backdoor paths*.

Backdoor Adjustment

Let $X$ be a set of treatment variables and $Y$ a set of outcome variables in the causal graph $G$.

If there exists a set $Z$ such that:

1. for every $X \in X$ and $Y \in Y$, $Z$ blocks every path between $X$ and $Y$ that has an arrow into $X$, and

2. no node in $Z$ is a descendant of a variable $X \in X$ (all variables in $Z$ are pre-treatment)

Then, $Z$ satisfies the backdoor criterion for $(X, Y)$ and, then the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_z P(y \mid x, z) P(z)$$

Also known as confounding paths, or backdoor paths.

Backdoor Adjustment

Let $\mathbf{X}$ be a set of treatment variables and $\mathbf{Y}$ a set of outcome variables in the causal graph $G$.

If there exists a set $\mathbf{Z}$ such that:

1. for every $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\mathbf{Z}$ blocks every path between $X$ and $Y$ that has an arrow into $X$, and

2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ (all variables in $\mathbf{Z}$ are pre-treatment)

Then, $\mathbf{Z}$ satisfies the \textbf{backdoor criterion} for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

$$P(\mathbf{y} | do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} | \mathbf{x}, \mathbf{z}) P(\mathbf{z})$$
Estimation via Propensity Scores

**Theorem:** If the set \(Z\) satisfies the parent / backdoor criterion w.r.t. the ordered pair \((X, Y)\) in the causal graph \(G\), then the causal effect of \(X\) on \(Y\) is identifiable (uniquely computable) and given by:

\[
P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z)
\]

\[
= \sum_z \frac{P(y \mid x, z)P(x \mid z)P(z)}{P(x \mid z)}
\]

\[
= \sum_z \frac{P(y, x, z)}{P(x \mid z)}
\]

Only if \(Z\) satisfies the BD criterion, Inverse Probability Weighting/Propensity Score can be used to estimate \(P(y \mid do(x))\).
What if backdoor adjustment does not work?

Identification via Front-Door Adjustment

Let $X$ be a set of treatment variables and $Y$ a set of outcome variables in the causal graph $G$.

If there exists a set $M$ such that:

1. $M$ intercepts all directed paths from any vertex $X \in X$ to any vertex $Y \in Y$;
2. There is no unblocked back-door path from any vertex $X \in X$ to vertex $M \in M$; and
3. All back-door paths from any vertex $M \in M$ to any vertex $Y \in Y$ are blocked by $X$.

Then, $M$ satisfies the **front-door criterion** and, then the effect of $X$ on $Y$ is given by:

$$P(y \mid do(x)) = \sum_m P(m \mid x) \sum_{x'} P(y \mid m, x')P(x')$$

$X = \{X\}$

$Y = \{Y\}$

![Diagram showing successful and unsuccessful applications of front-door adjustment](image)

$M = \{M\}$

$M = \{M_1, M_2\}$
Many scenarios beyond back-door and front-door!

Conditional Front-Door

\[ P(y \mid do(x)) = \sum_{m,z} P(m \mid x, z) \]

\[ \sum_{x'} P(y \mid m, x', z)P(x', z) \]

Napkin

\[ P(y \mid do(x)) = \frac{\sum_{z_2} P(x, y \mid z_1, z_2)P(z_2)}{\sum_{z_2} P(x \mid z_1, z_2)P(z_2)} \]

Unnamed

\[ P(y \mid do(x)) = \sum_{z_2, z_3} P(y \mid x, z_1, z_2, z_3)P(z_2) \]

\[ \sum_{z_1} P(z_3 \mid x, z_1)P(z_1) \]

And many others....
Do-Calculus (a.k.a. Causal Calculus)

Graphical conditions implying invariances between observational ($\mathcal{L}_1$) and interventional ($\mathcal{L}_2$) distributions

**Theorem:** Let $X, Y, Z, W$ be any disjoint subjects of variables.

**Rule 1** (Insertion/Deletion of Observations)

$$P(y \mid do(w), x, z) = P(y \mid do(w), z), \text{ if } (Y \perp\!
\perp X \mid Z, W)_{G_W}$$

**Rule 2** (Exchange of Actions and Observations)

$$P(y \mid do(w), do(x), z) = P(y \mid do(w), x, z), \text{ if } (Y \perp\!
\perp X \mid Z, W)_{G_{WX}}$$

**Rule 3** (Insertion/Deletion of Actions)

$$P(y \mid do(w), do(x), z) = P(y \mid do(w), z), \text{ if } (Y \perp\!
\perp X \mid Z, W)_{G_{WX(Z)}}$$

$G_{WX}$: graph $G$ after removing the incoming arrows into $W$ and the outgoing arrows from $X$;

$X(Z)$: set of $X$-nodes that are not ancestors of any $Z$-node in $G_W$.

Pearl, 1995
Identification in Non-Markovian Models

\[ P(y \mid do(x)) = \sum_m P(y \mid do(x), m)P(m \mid do(x)) \]
\[ = \sum_m P(y \mid do(x), do(m))P(m \mid do(x)) \]
\[ = \sum_m P(y \mid do(x), do(m))P(m \mid x) \]
\[ = \sum_m P(y \mid do(m))P(m \mid x) \]
\[ = \sum_m \sum_{x'} P(y \mid m, x')P(x' \mid do(m))P(m \mid x) \]

Probability Axioms
Rule 2
Rule 3
Probability Axioms
Rule 2
Rule 3
The Identify (ID) Algorithm

Algorithm 1 ID(x, y) given Causal Diagram G

Input: two disjoint sets X, Y ⊆ V
Output: Expression for P_x(y) or FAIL

1: Let D = An(Y)_{v \notin X}
2: Let the c-components of G_D be D_i, i = 1, ..., k
3: P_x(y) = \sum_{d \setminus Y} \prod_i \text{IDENTIFY}(D_i, V, P)

4: function IDENTIFY(C, T, Q = Q[T])
5: if C = T then return Q[T]

/* Let S^B denote the c-component of {B} in G_T */
6: if \exists B \in T \setminus C such that S^B \cap \text{ch}(B) = \emptyset then
7: Compute Q[T \setminus \{B\}] from Q; \triangleright Lemma 1
8: return IDENTIFY(C, T \setminus \{B\}, Q[T \setminus \{B\}])
9: else
10: throw FAIL

Lemma 1. Given a causal diagram D over V, X ∈ T ⊆ V, and P_{V \setminus t}, i.e., an expression for Q[T]. If X is not in the same c-component with a child in D_T, then Q[T \setminus \{X\}] is identifiable and given by

\[
Q[T \setminus \{X\}] = \frac{P_{V \setminus t}}{Q[S^X]} \times \sum_x Q[S^X]
\]

(2)

where S^X is the c-component of X in D_T and Q[S^X] is computable from P_{V \setminus t} by [Tian, 2002, Lemma 11].
Causal Effect Identification

1. Query
   \[ P(y \mid do(x)) \]

2. Causal Contraints
   \[ X \rightarrow M \rightarrow Y \]

3. Probability Distributions
   \[ P(x, m, y) \]
   Observational Distribution

ID-Algorithm and many recent generalizations.

Solution
   \[ P(y \mid do(x)) = \sum_m P(m \mid x) \sum_{x'} P(y \mid m, x') P(x') \]

Available Distributions

More on Causal Effect Identification

Identification from observational and experimental data:


Identification of stochastic/soft (and possibly imperfect) interventions:


Identification and Estimation via Deep Neural Networks:

Identification and Estimation via Deep Neural Networks

Trained Model: 
**$\mathcal{G}$-NCM $\widehat{M}$**

- **$V$**: Endogenous variables
- **$\widehat{U}$**: Create one for every bidirected clique
- **$\widehat{F}$**: Feedforward neural network for each variable in $V$ with parents from the graph
- **$P(\widehat{U})$**: All Unif(0,1)

**Inductive bias based on the causal diagram**: the enforced constraints empower the NCM with the ability to solve causal inference tasks.
Expressiveness of NCMs

True Model: SCM $\mathcal{M}^*$
- $V$: Endogenous variables
- $U$: Exogenous variables
- $\mathcal{F}$: Set of functions for variables in $V$
- $P(U)$: Probability distribution over $U$

Trained Model: NCM $\hat{\mathcal{M}}$
- $V$: Endogenous variables
- $\hat{U}$: Exogenous variables
- $\hat{\mathcal{F}}$: Feedforward neural network for each variable in $V$
- $P(\hat{U})$: All Unif(0,1)

**Thm:** For any SCM $\mathcal{M}^*$, there exists an NCM $\hat{\mathcal{M}}$ such that $\hat{\mathcal{M}}$ matches $\mathcal{M}^*$ on all three PCH layers!

This does not imply that the estimated NCM $\hat{\mathcal{M}}$ matches the true SCM $\mathcal{M}^*$!
Solution: A Neural Algorithm for Identification

**Algorithm 1:** Identifying/estimating queries with NCMs.

**Input:** causal query \( Q = P(y \mid do(x)) \), \( L_1 \) data \( P(v) \), and causal diagram \( G \)

**Output:** \( P^{\text{MC}}(y \mid do(x)) \) if identifiable, \text{FAIL} otherwise.

1. \( \hat{M} \leftarrow \text{NCM}(V, G) \) // from Def. 7
2. \( \theta_{\min} \leftarrow \arg \min_{\theta} P^{\hat{M}(\theta)}(y \mid do(x)) \) s.t. \( L_1(\hat{M}(\theta)) = P(v) \)
3. \( \theta_{\max} \leftarrow \arg \max_{\theta} P^{\hat{M}(\theta)}(y \mid do(x)) \) s.t. \( L_1(\hat{M}(\theta)) = P(v) \)
4. if \( P^{\hat{M}(\theta_{\min}^*)}(y \mid do(x)) \neq P^{\hat{M}(\theta_{\max}^*)}(y \mid do(x)) \) then
5. return \text{FAIL}
6. else
7. return \( P^{\hat{M}(\theta_{\min}^*)}(y \mid do(x)) \) // choose min or max arbitrarily

Maximize and minimize the induced causal query \( Q \) while maintaining \( L_1 \)-consistency (can be done with likelihood estimation).

**Thm.:** \( Q \) is identifiable if and only if they match!

**Corol.:** If \( Q \) is identifiable, then we can compute it by performing the mutilation procedure on \( \hat{M} \)!

The approach is equivalent to established symbolic approaches (Thm. 4), and in identifiable cases, the result is an NCM that can serve as a proxy model for estimating the query (Corol. 2).
Can we relax some causal assumptions?
Causal Effect Identification

1. **Query**
   \[ P(y \mid do(x)) \]

2. **Causal Constraints**
   \( X \rightarrow M \rightarrow Y \)

3. **Probability Distributions**
   \[ P(x, m, y) \]

   **Observational Distribution**

   \[ P(y \mid do(x)) = \sum_m P(m \mid x) \sum_{x'} P(y \mid m, x') P(x') \]

   **Interventional Distribution**

   **Available Distributions**

   **Solution**
   yes / no

   Usually hard to be fully specified.

---

Causal diagrams are powerful tools that allow for inferences based on weaker knowledge (structural invariances) than the encoded in the true, underlying SCM.

Still, structural knowledge for every pair of variables may not be available in many real-world, complex, high-dimensional systems.

Question:

Is it possible to relax the assumption of having a fully specified causal diagram and still be able to identify a causal effect?
A causal diagram cannot be specified given the existing knowledge!

How can we identify $P(y \mid do(x))$ in this case?
Cluster DAGs (C-DAGs)

A) Age
(B) Blood pressure
(C) Comorbidities
(D) Medication history
(X) Lisinopril
(S) Sleep Quality
(Y) Stroke

A cluster DAG $G_C$ over a given partition $C = \{ C_1, \ldots, C_k \}$ of $V$ is compatible with a causal diagram $G$ over $V$ if for every $C_i, C_j \in C$:

- $C_i \rightarrow C_j$ if $\exists V_i \in C_i$ and $V_j \in C_j$ such that $V_i \rightarrow V_j$
- $C_i \leftrightarrow C_j$ if $\exists V_i \in C_i$ and $V_j \in C_j$ such that $V_i \leftrightarrow V_j$

and $G_C$ contains no cycles.
Many causal diagrams are compatible with the current knowledge!

Can be seen as an equivalence class of causal diagrams, where any relationships are allowed among the variables within each cluster.

Can we infer causal effects without deciding on any one particular causal diagram?
C-DAG: Flexible Encoder of Model Assumptions

- **N clusters of size one (full knowledge - DAG)**
- **... (partial knowledge - C-DAG)**
- **One cluster of size N (no knowledge)**
C-DAG: Flexible Encoder of Model Assumptions

Clusters are manually created by domain experts:
- due to lack of knowledge, consensus, or interest on the internal causal structure;
- to communicate relationships among semantically meaningful entities.
Identification of Causal Effects from C-DAGs

1. Query
   \( P(y \mid do(x)) \)

2. C-DAG
   \( X \rightarrow M_{1,2,3} \rightarrow Y \)

3. Data
   \( P(x, m_1, m_2, m_3, y) \)

Inference Engine

Solution
   \( \text{yes} / \text{no} \)

\[ P(y \mid do(x)) = \sum_{m_{123}} P(m_{123} \mid x) \sum_{x'} P(y \mid m_{123}, x') P(x') \]

Inferred (Interventional) Distribution

Available (Observational) Distribution

Effect Identifiability given a C-DAG

An effect identifiable in a C-DAG $G_C$ is identifiable in all compatible causal diagrams $G$ using the same identification formula!
Effect Non-Identifiability given a C-DAG

An effect is not identifiable in a C-DAG $G_C$ if there exists at least one compatible causal diagrams $G$ in which the effect is not identifiable.
Beyond Backdoor Adjustment

Again, an effect identifiable in a C-DAG $G_C$ is identifiable in all compatible causal diagrams $G$ using the same identification formula!
What if no knowledge is available?

Can we learn a causal diagram $\mathcal{G}$ from observational data?

Causal Discovery:

In non-parametric settings, we can’t learn the true causal diagram, but algorithms such as the Fast Causal Inference (FCI) can learn a graphical representation of its Markov equivalence class!

Causal Discovery

Fast Causal Inference (FCI)

A constraint-based causal discovery algorithms that accounts for unobserved confounders
Causal Discovery

**Goal:** Learn a graphical representation of the Markov Equivalence Class from observational data.

**Assumptions:** the observed distribution is the marginal of a distribution $P$ that satisfies the following conditions for the true causal diagram $G$ (an ADMG):

1) **I-Map / Semi-Markov Condition:** for any disjoint subsets $X$, $Y$ and $Z$:

   $$(X \perp Y \mid Z)_G \Rightarrow (X \perp Y \mid Z)_P.$$  

   $G$ is an **I-Map of $P$**

   $P$ is semi-Markov **relative** to $G$.

2) **Faithfulness Condition:** for any disjoint subsets $X$, $Y$ and $Z$:

   $$(X \perp Y \mid Z)_P \Rightarrow (X \perp Y \mid Z)_G.$$  

   $P$ is **faithful** to $G$

**Note:** Estimation of the marginal distribution from limited data requires and **additional assumption**:

3) An adequate *conditional independence test* is available.
Fast Causal Inference (FCI) Algorithm

**FCI:** Learn a PAG $\mathcal{P}$ representing the **Markov Equivalence Class (MEC)** from $P$, i.e.:

$$(X \perp \perp Y \mid Z)_{\mathcal{P}} \Leftrightarrow (X \perp \perp Y \mid Z)_{P,G}$$

Evaluated through m-separation

Every non-circle edge mark represents an invariance in the MEC in terms of ancestral and non-ancestral relationships

- **Arrowhead** $\implies$ **non-ancesterrality**
- **Tail** $\implies$ **ancest rallly**
- **Circle** $\implies$ **non-invariance**

| $A \rightarrow B$ | $\implies$ anecstrally |
| $A \leftarrow B$ | $\implies$ spurious association |
| $A \leftrightarrow B$ | $\implies$ selection bias |


Conditional Independence Tests

Gaussian errors and independent observations: partial correlation test

R package: https://cran.r-project.org/web/packages/pcalg/

Kernel-based non-parametric test:

R package: https://cran.r-project.org/web/packages/CondIndTests

Continuous (conditional Gaussian) or Discrete (Binary, Ordinal, Multinomial) - Linear Regression

- R package: https://cran.r-project.org/web/packages/MXM/

Gaussian errors and correlated observations (family data):

R package: https://github.com/adele/FamilyBasedPGMs
Learning Structural Invariances

\[ P(v) \]

\[ X \perp Y \]

\[ Z \perp Y | Z \]

\[ P(x, y, z) = P(z | x, y)P(x | y)P(y) \]

\[ = P(z | x, y)P(x)P(y) \]

Markov Equivalence Class (MEC)
Learning Structural Invariances

\[ \mathcal{M}_1 = \begin{cases} V = \{X, Y, Z\} \\ U = \{U_x, U_y, U_z\} \\ \mathcal{F} = \begin{cases} X \leftarrow f_X(U_x) \\ Z \leftarrow f_Z(X, Y, U_x) \\ Y \leftarrow f_Y(U_y) \\ P(U) \end{cases} \end{cases} \]

\[ \mathcal{M}_{N-1} = \begin{cases} V = \{X, Y, Z\} \\ U = \{U_{XZ}, U_{YZ}, U_x, U_y, U_z\} \\ \mathcal{F} = \begin{cases} X \leftarrow f_X(U_{XZ}, U_x) \\ Z \leftarrow f_Z(Y, U_{XZ}, U_x) \\ Y \leftarrow f_Y(U_y) \\ P(U) \end{cases} \end{cases} \]

\[ \mathcal{M}_N = \begin{cases} V = \{X, Y, Z\} \\ U = \{U_{XZ}, U_{YZ}, U_x, U_y, U_z\} \\ \mathcal{F} = \begin{cases} X \leftarrow f_X(U_{XZ}, U_x) \\ Z \leftarrow f_Z(U_{XZ}, U_{YZ}, U_z) \\ Y \leftarrow f_Y(U_y) \\ P(U) \end{cases} \end{cases} \]

Data

\[ P(x, y, z) = P(z | x, y)P(x | y)P(y) = P(z | x, y)P(x)P(y) \]

Conditional (in)dependencies

\[ P(v) \]

\[ X \perp Y \]

\[ X \perp Z \]

\[ Z \perp Y \]

\[ X \perp Y | Z \]

Markov Equivalence Class (MEC)

Partial Ancestral Graph (PAG)

Causal Discovery

FCI algorithm

\[ X \rightarrow Z \rightarrow Y \]

Other examples

Fast Causal Inference (FCI) Algorithm

True (unknown) causal diagram

True (unknown) causal diagram

Conditional Independence Tests

Complete Graph

X ⊥ W
X ⊥ Y|Z, W

observed in the data

X ⊥ W
X ⊥ Y|Z, W

Skeleton

X ⊥ W
X ⊥ Y|Z, W

Partial Ancestral Graph (PAG)

Implied by the PAG using m-separation

Implied by the ADMG using d-separation

Implied by the ADMG using d-separation

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.
PAG represents the Markov Equivalence Class

\[ X \rightarrow Z \rightarrow W \rightarrow Y \]

Partial Ancestral Graph (PAG)

- \( Z \) is not an ancestor of \( X \) or \( W \).
- \( Z \) and \( W \) are ancestors of \( Y \).
- \( Z \) is not confounded with \( Y \).

\[ X \perp \perp W \]
\[ X \perp \perp Y | Z, W \]
Causal discovery from observational and experimental data:


Can we identify causal effects from the equivalence class?

**Effect Identification:**

For Covariate Adjustment, we can use the Generalized Adjustment Criterion.

Recently, we proposed complete calculus and algorithms for the identification of marginal and conditional causal effect in PAGs!


General Identification in Markov Equivalence Classes

1. Query
   \[ P(y \mid do(x)) \]

2. PAG
   \[ W \rightarrow X \rightarrow Y \rightarrow Z \]

3. Data
   \[ P(x, m, y) \]

Solution

IDP / CIDP

\[ P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z) \]

Available Distributions

Interventional Distribution

The CIDP and IDP algorithms are available at the PAGId R package: https://github.com/adele/PAGId

An effect identifiable in a PAG $\mathcal{P}$ is identifiable in all causal diagrams $G$ in the Markov Equivalence Class using the same identification formula!
Effect Non-Identifiability given a PAG

An effect not identifiable in a PAG $\mathcal{P}$ is not identifiable in at least one causal diagrams $G$ in the Markov Equivalence Class

$$P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z)$$
Continuous Process of Scientific Discovery and Causal Hypothesis Refinement

SCM $M^*$
(Unobserved Nature)

Z $\leftarrow f_Z(U_z)$
X $\leftarrow f_X(Z, U_x)$
Y $\leftarrow f_Y(X, Z, U_y)$

$P(U_z, U_x, U_y)$

knowledge(t)

data(t) $\rightarrow$ distributions(t)

new discoveries(t+1)

queries(t)

Experimental validation(t)

new insights

new challenges

perform new observations and/or experiments(t+1)

A Statistical Learning
B Causal Learning
C Causal Inference
D Causal Exp. Design

Causal Inference Workflow
Causal Inference Workflow

Continuous Process of Scientific Discovery and Causal Hypothesis Refinement

SCM M* (Unobserved Nature)

Z ← f_Z(U_z)
X ← f_X(Z, U_x)
Y ← f_Y(X, Z, U_y)
P(U_z, U_x, U_y)

new discoveries(t+1)
data(t) A distributions(t) B knowledge(t)

causal hypothesis (t)

queries(t)

not answerable

answerable

Experimental validation(t)

new insights

new challenges

perform new observations and/or experiments(t+1)

A Statistical Learning  B Causal Learning  C Causal Inference  D Causal Exp. Design
1. Generalized Effect Identification:
   - For when multiple observational and experimental datasets are available, possible under partial observability.

2. Partial Identification:
   - For when the effect is not point-identifiable, but an interval for it can be derived.

3. Effect Transportability:
   - For when the target is a different population/domain.

4. Counterfactual Identification:
   - Identification of $\mathcal{L}_3$ quantities, such as $P(y_x | x')$ and $P(y_x | x', z)$.

5. Fairness Evaluation:
   - To identify path-specify effects related to protective variables.

6. Effect Estimation beyond backdoor scenarios:
   - Via doubly robust machine learning and different plug-in density estimators.
Conclusions

Causal inference can help overcome critical challenges in Artificial Intelligence, including robustness, generalizability, explainability, and fairness.

Causal Data Science: principled way of combining data and substantive knowledge about the phenomenon under investigation to generate causal explanations and better decision-making.

Recent developments for causal inference when knowledge is largely unavailable and coarse are expected to help the practice of causal data analysis and meet the growing demand in the Empirical Sciences for sound causal explanations and more robust and generalizable decision-making.
Thank you! :)  

Feel free to reach out to me if you have any questions: 

adele.ribeiro@uni-marburg.de