A short course on network causal Inference: theory and applications

Negar Kiyavash

EPFL

June 2021
Overview

1. What is Causal Inference?
2. Difficulties
3. Random Variables
4. Graphs
5. Granger Causality
6. Structural Equation Model
7. Intervention
8. Graphical models
9. Faithfulness
10. Do operation
11. Learning Causal Bayes Nets
Motivation

We often are interested in discovering causation vs correlation.

Example: [Chocolate - Nobel Prizes] Messerli [2012] reports that there is a significant correlation between a country’s chocolate consumption (per capita) and the number of Nobel prizes awarded to its citizens.
We must be careful with drawing conclusions like “Eating chocolate produces Nobel prize” or “Geniuses are more likely to eat lots of chocolate.”
We must be careful with drawing conclusions like “Eating chocolate produces Nobel prize” or “Geniuses are more likely to eat lots of chocolate.”

Correlation does not imply causation!
Application areas include:

- **Computational Neuroscience**: Advances in recording technologies have given neuroscience researchers access to large amounts of data, e.g., individual recordings of neurons in different parts of the brain.
Application areas include:

- **Computational Neuroscience:** Advances in recording technologies have given neuroscience researchers access to large amounts of data, e.g., individual recordings of neurons in different parts of the brain.
Application areas include:

- **Computational Neuroscience**: Advances in recording technologies have given neuroscience researchers access to large amounts of data, e.g., individual recordings of neurons in different parts of the brain.

- Could we understand firing of which neurons causes others to fire and hence learn the functional connectivity in the brain?
Application areas include:

- **Financial Markets:** Financial instability can lead to financial crises due to its contagion or spillover effects to other parts of the economy. Having an accurate measures of systemic risk and inter-dependencies between financial institution gives central banks and policy makers the ability to take proper actions in order to stabilize financial markets.
Application areas include:

- **Social Networks**: For networks with large numbers of nodes, such as millions of people in a social network, e.g., Twitter, having efficient algorithms that recover the graphical models is critical.

Vertical lines depict each time a message was posted by that agent. A major research goal is to infer whether, and how strongly, the news corporation influences the users by analyzing these time-series.
**Incomplete universe:** Not observing all the relevant variables may lead to false conclusion. For instance, in the chocolate-Noble prize example, the correlation stems from some hidden variables like economic strength of a country.
Difficulties

- **Incomplete universe**: Not observing all the relevant variables may lead to false conclusion. For instance, in the chocolate-Noble prize example, the correlation stems from some hidden variables like economic strength of a country.

- **Computational Issues**: Understanding the causal interaction in a large network such as social networks, requires large processing large amounts of data (think: computational power and large memory usage).
**Difficulties**

- **Incomplete universe**: Not observing all the relevant variables may lead to false conclusion. For instance, in the chocolate-Noble prize example, the correlation stems from some hidden variables like economic strength of a country.

- **Computational Issues**: Understanding the causal interaction in a large network such as social networks, requires large processing large amounts of data (think: computational power and large memory usage).

- **Simultaneous effects**: In time series analysis, inaccurate sampling rate will lead to simultaneous influences between time series. Such influences cannot be captured using, for example, Granger-causality analysis and requires finer and more complex analysis.
**Simpson’s paradox**: The table reports the success rates of two treatments for kidney stones

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Overall</th>
<th>Patients with small stones</th>
<th>Patients with large stones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment A:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open surgery</td>
<td>78% (273/350)</td>
<td>93% (81/87)</td>
<td>73% (192/263)</td>
</tr>
<tr>
<td>Treatment B:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percutaneous nephrolithotomy</td>
<td>83% (289/350)</td>
<td>87% (234/270)</td>
<td>69% (55/80)</td>
</tr>
</tbody>
</table>

- Although the overall success rate of treatment B seems better, treatment B performs worse than treatment A on both patients with small kidney stones and patients with large kidney stones.
- How do we deal with this inversion of conclusion?
Simpson’s paradox

Another example of Simpson’s paradox:

Admission data on university level:

<table>
<thead>
<tr>
<th></th>
<th>Applicants</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8442</td>
<td>44%</td>
</tr>
<tr>
<td>Women</td>
<td>4321</td>
<td>35%</td>
</tr>
</tbody>
</table>

Admission data on department level:

<table>
<thead>
<tr>
<th>Departments</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>63%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>62%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
</tbody>
</table>

Among 85 departments, there are 6 against men but only 4 against women.
Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$: probability space, where $\Omega$ is the set of all possible outcomes, $F$ is the set of events and $\mathbb{P}$ is the assignment of probabilities to the events.
Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$: probability space, where $\Omega$ is the set of all possible outcomes, $F$ is the set of events and $\mathbb{P}$ is the assignment of probabilities to the events.

- A random variable is a measurable function $X : \Omega \rightarrow E$ from a set of possible outcomes $\Omega$ to a measurable space $E$. The probability that $X$ takes on a value in a measurable set $S \subseteq E$ is written as

$$
\mathbb{P}(X \in S) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in S\})
$$
Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$: probability space, where $\Omega$ is the set of all possible outcomes, $F$ is the set of events and $\mathbb{P}$ is the assignment of probabilities to the events.

- A random variable is a measurable function $X : \Omega \rightarrow E$ from a set of possible outcomes $\Omega$ to a measurable space $E$. The probability that $X$ takes on a value in a measurable set $S \subseteq E$ is written as

$$
P(X \in S) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in S\})$$

- In many cases, $X$ is real-valued, i.e. $E = \mathbb{R}$. 
Preliminaries

Throughout the lecture we use the following notation.

- \((\Omega, F, \mathbb{P})\) : probability space, where \(\Omega\) is the set of all possible outcomes, \(F\) is the set of events and \(\mathbb{P}\) is the assignment of probabilities to the events.

- A random variable is a measurable function \(X : \Omega \rightarrow E\) from a set of possible outcomes \(\Omega\) to a measurable space \(E\). The probability that \(X\) takes on a value in a measurable set \(S \subseteq E\) is written as

\[
\mathbb{P}(X \in S) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in S\})
\]

- In many cases, \(X\) is real-valued, i.e. \(E = \mathbb{R}\).

- \(P^X\) is the distribution of the \(p\)-dimensional random vector \(X\).
Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$ : probability space, where $\Omega$ is the set of all possible outcomes, $F$ is the set of events and $\mathbb{P}$ is the assignment of probabilities to the events.

- A random variable is a measurable function $X : \Omega \rightarrow E$ from a set of possible outcomes $\Omega$ to a measurable space $E$. The probability that $X$ takes on a value in a measurable set $S \subseteq E$ is written as

$$
\mathbb{P}(X \in S) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in S\})
$$

- In many cases, $X$ is real-valued, i.e. $E = \mathbb{R}$.

- $\mathbb{P}^X$ is the distribution of the $p$-dimensional random vector $X$.

- We call $X$ independent of $Y$ and write $X \indep Y$ if and only if

$$
\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)
$$
We call $X_1, \ldots, X_p$ jointly (or mutually) independent if and only if

$$\mathbb{P}(X_1, \ldots, X_p) = \mathbb{P}(X_1) \cdots \mathbb{P}(X_p).$$
We call $X_1, \ldots, X_p$ jointly (or mutually) independent if and only if
\[
P(X_1, \ldots, X_p) = P(X_1) \cdots P(X_p).
\]

We call $X$ independent of $Y$ conditional on $Z$ and write $X \perp Y \mid Z$ if and only if
\[
P(x, y \mid z) = P(x \mid z)P(y \mid z)
\]
for all $x, y, z$ such that $p(z) > 0$. Otherwise, $X$ and $Y$ are dependent conditional on $Z$ and we write $X \not\perp Y \mid Z$. 


• We call $X_1, \ldots, X_p$ jointly (or mutually) independent if and only if

$$P(X_1, \ldots, X_p) = P(X_1) \ldots P(X_p).$$

• We call $X$ independent of $Y$ conditional on $Z$ and write $X \perp Y | Z$ if and only if

$$P(x, y | z) = P(x | z) P(y | z)$$

for all $x, y, z$ such that $p(z) > 0$. Otherwise, $X$ and $Y$ are dependent conditional on $Z$ and we write $X \not\perp Y | Z$.

• We call $X$ and $Y$ uncorrelated if $E[X^2], B[Y^2] < \infty$ and

$$E[XY] = E[X]E[Y]$$
A graph $G = (V, E)$ consists of (finitely many) nodes or vertices $V = \{1, \ldots, p\}$ and edges $E \subseteq V^2$ with $(v, v) \notin E$ for any $v \in V$. 
A graph \( G = (V, E) \) consists of (finitely many) nodes or vertices \( V = \{1, \ldots, p\} \) and edges \( E \subseteq V^2 \) with \((v, v) \notin E\) for any \( v \in V\).

A graph \( G_1 = (V_1, E_1) \) is called a subgraph of \( G \) if \( V_1 = V \) and \( E_1 \subseteq E \).
A graph $G = (V, E)$ consists of (finitely many) nodes or vertices $V = \{1, \ldots, p\}$ and edges $E \subseteq V^2$ with $(v, v) \notin E$ for any $v \in V$.

A graph $G_1 = (V_1, E_1)$ is called a subgraph of $G$ if $V_1 = V$ and $E_1 \subseteq E$.

A node $i$ is called a parent of $j$ if $(i, j) \in E$ and $(j, i) \notin E$ and a child if $(j, i) \in E$ and $(i, j) \notin E$. 


A graph $G = (V, E)$ consists of (finitely many) nodes or vertices $V = \{1, ..., p\}$ and edges $E \subseteq V^2$ with $(v, v) \notin E$ for any $v \in V$.

A graph $G_1 = (V_1, E_1)$ is called a subgraph of $G$ if $V_1 = V$ and $E_1 \subseteq E$.

A node $i$ is called a parent of $j$ if $(i, j) \in E$ and $(j, i) \notin E$ and a child if $(j, i) \in E$ and $(i, j) \notin E$.

Two nodes $i$ and $j$ are adjacent if either $(i, j) \in E$ or $(j, i) \in E$. 
A graph \( G = (V, \mathcal{E}) \) consists of (finitely many) nodes or vertices \( V = \{1, \ldots, p\} \) and edges \( \mathcal{E} \subseteq V^2 \) with \((v, v) \notin \mathcal{E}\) for any \( v \in V \).

A graph \( G_1 = (V_1, \mathcal{E}_1) \) is called a subgraph of \( G \) if \( V_1 = V \) and \( \mathcal{E}_1 \subseteq \mathcal{E} \).

A node \( i \) is called a parent of \( j \) if \((i, j) \in \mathcal{E}\) and \((j, i) \notin \mathcal{E}\) and a child if \((j, i) \in \mathcal{E}\) and \((i, j) \notin \mathcal{E}\).

Two nodes \( i \) and \( j \) are adjacent if either \((i, j) \in \mathcal{E}\) or \((j, i) \in \mathcal{E}\).

The skeleton of \( G \) does not take the directions of the edges into account: it is the graph \((V, \tilde{\mathcal{E}})\) with \((i, j) \in \tilde{\mathcal{E}}\), if \((i, j) \in \mathcal{E}\) or \((j, i) \in \mathcal{E}\).
A directed path in $G$ is a sequence of (at least two) distinct vertices $i_1, \ldots, i_n$, such that there is an edge from $i_k$ and $i_{k+1}$ for all $k = 1, \ldots, n - 1$. 
A *directed path* in $G$ is a sequence of (at least two) distinct vertices $i_1, \ldots, i_n$, such that there is an edge from $i_k$ and $i_{k+1}$ for all $k = 1, \ldots, n - 1$.

Node $i$ is an *ancestor* of node $j$, if there is a directed path from $i$ to $j$. Then, $j$ is a *descendant* of $i$. 
A directed path in $G$ is a sequence of (at least two) distinct vertices $i_1, ..., i_n$, such that there is an edge from $i_k$ and $i_{k+1}$ for all $k = 1, ..., n - 1$.

Node $i$ is an ancestor of node $j$, if there is a directed path from $i$ to $j$. Then, $j$ is a descendant of $i$.

Graph $G$ is called directed acyclic graph (DAG) if it has no directed cycle, if there is no pair $(j, k)$ with directed paths from $j$ to $k$ and from $k$ to $j$. 
A directed path in $G$ is a sequence of (at least two) distinct vertices $i_1, ..., i_n$, such that there is an edge from $i_k$ and $i_{k+1}$ for all $k = 1, ..., n - 1$.

Node $i$ is an ancestor of node $j$, if there is a directed path from $i$ to $j$. Then, $j$ is a descendant $i$.

Graph $G$ is called directed acyclic graph (DAG) if it has no directed cycle, if there is no pair $(j, k)$ with directed paths from $j$ to $k$ and from $k$ to $j$.

Adjacency matrix: We can represent a DAG $G = (V, E)$ over $p$ nodes with a binary $p \times p$ matrix $A$ (taking values 0 or 1): $A_{i,j} = 1$ iff $(i, j) \in E$. 
A joint distribution over a set of variables can be factorized using Bayes rule.

A factorization of a joint distribution can be visualized using a directed graph (Bayesian network)
A joint distribution over a set of variables can be factorized using Bayes rule.

A factorization of a joint distribution can be visualized using a directed graph (Bayesian network)

**Example:**

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)
\]

Edges represent conditional dependencies.
Clive Granger (1969): "We say that X is causing Y if we are better able to predict (the future of) Y using all available information than if the information apart from (the past of) X had been used."
Clive Granger (1969): "We say that X is causing Y if we are better able to predict (the future of) Y using all available information than if the information apart from (the past of) X had been used."

Granger’s Formulation: AR model

\[ Y_t = c + \sum_{\tau=1}^{p} a_\tau Y_{t-\tau} + b_\tau X_{t-\tau} + \epsilon_t \]

\[ Y_t = c' + \sum_{\tau=1}^{p} a'_\tau Y_{t-\tau} + \epsilon'_t \]
Clive Granger (1969): "We say that X is causing Y if we are better able to predict (the future of) Y using all available information than if the information apart from (the past of) X had been used."

Granger’s Formulation: AR model

\[
Y_t = c + \sum_{\tau=1}^{p} a_{\tau} Y_{t-\tau} + b_{\tau} X_{t-\tau} + \epsilon_t
\]

\[
Y_t = c' + \sum_{\tau=1}^{p} a'_{\tau} Y_{t-\tau} + \epsilon'_t
\]

F-test: to assess quality of prediction.
Granger Causality

- **Clive Granger** (1969): "We say that X is causing Y if we are better able to predict (the future of ) Y using all available information than if the information apart from (the past of) X had been used."

  Granger’s Formulation: AR model

  \[
  Y_t = c + \sum_{\tau=1}^{p} a_{\tau} Y_{t-\tau} + b_{\tau} X_{t-\tau} + \epsilon_t
  \]

  \[
  Y_t = c' + \sum_{\tau=1}^{p} a'_{\tau} Y_{t-\tau} + \epsilon'_t
  \]

  - F-test: to assess quality of prediction.
  - RSS: predictive sum of squared residues.

  \[
  RSS = \sum_{\tau=1}^{T} \epsilon_t^2, \quad RSS' = \sum_{\tau=1}^{T} (\epsilon')_t^2, \quad T_s = \frac{(RSS' - RSS)/p}{RSS/(T - 2p - 1)}
  \]

  - If \( T_s > \) some critical value, reject the null hypothesis
**Granger Causality**

- **Clive Granger** (1969): "We say that X is causing Y if we are better able to predict (the future of) Y using all available information than if the information apart from (the past of) X had been used."

- Granger’s Formulation: AR model

\[
Y_t = c + \sum_{\tau=1}^{p} a_\tau Y_{t-\tau} + b_\tau X_{t-\tau} + \epsilon_t
\]

\[
Y_t = c' + \sum_{\tau=1}^{p} a'_\tau Y_{t-\tau} + \epsilon'_t
\]

- F-test: to assess quality of prediction.
- RSS: predictive sum of squared residues.

\[
RSS = \sum_{\tau=1}^{T} \epsilon_t^2, \quad RSS' = \sum_{\tau=1}^{T} (\epsilon')^2_t, \quad T_s = \frac{(RSS' - RSS)/p}{RSS/(T - 2p - 1)}
\]

- If \( T_s > \) some critical value, reject the null hypothesis
- Cons: Linear assumption, stationarity, time synchronization.
Going beyond linear models

Sequential Predictors: \( w_i = g_i(Y_1, \ldots, Y_{i-1}, X_1, \ldots, X_i) \) and
\( \tilde{w}_i = \tilde{g}_i(Y_1, \ldots, Y_{i-1}) \)
Sequential Predictors: $w_i = g_i(Y_1, ..., Y_{i-1}, X_1, ..., X_i)$ and $\tilde{w}_i = \tilde{g}_i(Y_1, ..., Y_{i-1})$

Outcome $y$ is revealed, the loss incurred: $\ell(y, w)$
Going beyond linear models

- Sequential Predictors: \( w_i = g_i(Y_1, \ldots, Y_{i-1}, X_1, \ldots, X_i) \) and \( \tilde{w}_i = \tilde{g}_i(Y_1, \ldots, Y_{i-1}) \)
- Outcome \( y \) is revealed, the loss incurred: \( \ell(y, w) \)
- Reduction in loss (regret): \( \frac{1}{T} \sum_{i=1}^{T} \ell(y_i, w_i) - \ell(y_i, \tilde{w}_i) \)
Sequential Predictors: \( w_i = g_i(Y_1, ..., Y_{i-1}, X_1, ..., X_i) \) and \( \tilde{w}_i = \tilde{g}_i(Y_1, ..., Y_{i-1}) \)

Outcome \( y \) is revealed, the loss incurred: \( \ell(y, w) \)

Reduction in loss (regret): \( \frac{1}{T} \sum_{i=1}^{T} \ell(y_i, w_i) - \ell(y_i, \tilde{w}_i) \)

Case:

Logarithmic loss: \( \ell(y, w) = -\log w(y) \)

Predictors: beliefs (the optimal predictors are conditional densities)
Going beyond linear models

- Sequential Predictors: \( w_i = g_i(Y_1, ..., Y_{i-1}, X_1, ..., X_i) \) and \( \tilde{w}_i = \tilde{g}_i(Y_1, ..., Y_{i-1}) \)
- Outcome \( y \) is revealed, the loss incurred: \( \ell(y, w) \)
- Reduction in loss (regret): \( \frac{1}{T} \sum_{i=1}^{T} \ell(y_i, w_i) - \ell(y_i, \tilde{w}_i) \)

**Case:**

- Logarithmic loss: \( \ell(y, w) = -\log w(y) \)
- Predictors: beliefs (the optimal predictors are conditional densities)
- Then the regret will be:

\[
\frac{1}{T} \mathbb{E} \left[ \sum_{i=1}^{T} \log \frac{\mathbb{P}(Y_i|Y_{i-1}, X_i)}{\mathbb{P}(Y_i|Y_{i-1})} \right] := \frac{1}{T} I(X^T \rightarrow Y^T)
\]

- Entropy of random variable \( X \): \( H(X) := -\mathbb{E}[\log \mathbb{P}(X)] \)
- Mutual information between \( X \) and \( Y \): \( I(X; Y) := H(X) - H(X|Y) \)
A structural equation model (SEM) (also called a functional model) is defined as a tuple $S := (S, \mathbb{P}^N)$, where $S = (S_1, \ldots, S_p)$ is a collection of $p$ equations

$$S_j : \quad X_j = f_j(\text{PA}_j, N_j), \quad j = 1, \ldots, p,$$

$\text{PA}_j \subseteq \{X_1, \ldots, X_p\}/\{X_j\}$ are called parents of $X_j$

$\mathbb{P}^N = \mathbb{P}(N_1, \ldots, N_p)$ is the joint distribution of the noise variables and they are jointly independent.
A structural equation model (SEM) (also called a functional model) is defined as a tuple $S := (S, \mathbb{P}^N)$, where $S = (S_1, ..., S_p)$ is a collection of $p$ equations

$$S_j : \quad X_j = f_j(\text{PA}_j, N_j), \quad j = 1, ..., p,$$

$\text{PA}_j \subseteq \{X_1, ..., X_p\}/\{X_j\}$ are called parents of $X_j$

$\mathbb{P}^N = \mathbb{P}(N_1, ..., N_p)$ is the joint distribution of the noise variables and they are jointly independent.

**Example 1:**

$$X_1 = f_1(N_1), \quad X_2 = f_2(X_1, N_2), \quad X_3 = f_3(X_2, N_3)$$

$$X_1 \rightarrow X_2 \rightarrow X_3$$
A structural equation model (SEM) (also called a functional model) is defined as a tuple $S := (S, \mathbb{P}^N)$, where $S = (S_1, \ldots, S_p)$ is a collection of $p$ equations

$$S_j : \quad X_j = f_j(\text{PA}_j, N_j), \quad j = 1, \ldots, p,$$

where $\text{PA}_j \subseteq \{X_1, \ldots, X_p\}/\{X_j\}$ are called parents of $X_j$

$\mathbb{P}^N = \mathbb{P}(N_1, \ldots, N_p)$ is the joint distribution of the noise variables and they are jointly independent.

**Example 1:**

$X_1 = f_1(N_1), \quad X_2 = f_2(X_1, N_2), \quad X_3 = f_3(X_2, N_3)$

**Example 2:**

$X = N_x, \quad Y = 4X + N_y, \quad X \rightarrow Y$
**Intervention Distribution:** Consider $\mathbb{P}^X$ that has been generated from an SEM $\mathcal{S} := (S, \mathbb{P}^N)$. We can then replace one (or more) structural equations (without generating cycles in the graph) and obtain a new SEM $\tilde{\mathcal{S}}$. 
**Intervention Distribution:** Consider $\mathbb{P}^X$ that has been generated from an SEM $\mathcal{S} := (S, \mathbb{P}^N)$. We can then replace one (or more) structural equations (without generating cycles in the graph) and obtain a new SEM $\tilde{\mathcal{S}}$.

The distributions in the new SEM is intervention distributions and the variables whose structural equation have replaced have been “intervened on”.

**Intervention Distribution:** Consider $\mathbb{P}^X$ that has been generated from an SEM $S := (S, \mathbb{P}^N)$. We can then replace one (or more) structural equations (without generating cycles in the graph) and obtain a new SEM $\tilde{S}$.

The distributions in the new SEM is intervention distributions and the variables whose structural equation have replaced have been "intervened on".

**Intervention on variable $X_j$:**

$$\mathbb{P}^X_{\tilde{S}} = \mathbb{P}_S \left( X \mid do(X_j = \tilde{f}_j(\tilde{P}A_j, \tilde{N}_j)) \right)$$
**Intervention Distribution:** Consider $\mathbb{P}^X$ that has been generated from an SEM $\mathcal{S} := (S, \mathbb{P}^N)$. We can then replace one (or more) structural equations (without generating cycles in the graph) and obtain a new SEM $\tilde{\mathcal{S}}$.

The distributions in the new SEM is intervention distributions and the variables whose structural equation have replaced have been “intervened on”.

Intervention on variable $X_j$:

$$\mathbb{P}_{\tilde{\mathcal{S}}}^X = \mathbb{P}_S\left(X \mid \text{do}(X_j = \tilde{f}_j(\tilde{P}A_j, \tilde{N}_j))\right)$$

*Perfect intervention:* when $\tilde{f}_j(\tilde{P}A_j, \tilde{N}_j)$ puts a point mass on a real value $a$, we simply write $\mathbb{P}_S\left(X \mid \text{do}(X_j = a)\right)$.
**Example:** A patient with poor eyesight comes to the hospital and goes blind \((B = 1)\) after the doctor suggests the treatment \(T = 1\). Let us assume

\[
T = N_T \\
B = T \cdot N_B + (1 - T)(1 - N_B)
\]

where \(N_B \sim Ber(0.01)\).

In this example, we have

\[
\mathbb{P}_S(B = 0 | do(T = 1)) = 0.99 \\
\mathbb{P}_S(B = 0 | do(T = 0)) = 0.01
\]
**Another Example:** Suppose that $\mathbb{P}(X, Y)$ is induced by a structural equation model $\mathcal{S}$

$$X = N_x, \quad Y = 3X + N_y, \quad \Rightarrow \quad X \rightarrow Y$$

with $N_x, N_y \sim \mathcal{N}(0, 1)$. The

- $\mathbb{P}(Y) = \mathcal{N}(0, 10)$
- $\mathbb{P}(Y|\text{do}(X = 2)) = \mathcal{N}(6, 1)$,  $\mathbb{P}(Y|\text{do}(X = 1.2)) = \mathcal{N}(3.6, 1)$
- $\mathbb{P}(X|\text{do}(Y = 2)) = \mathbb{P}(X|\text{do}(Y = 1.2)) = \mathcal{N}(0, 1) = \mathbb{P}(X)$
Another Example: Suppose that \( P(X, Y) \) is induced by a structural equation model \( S \)

\[
X = N_x, \quad Y = 3X + N_y, \quad \Rightarrow \quad X \rightarrow Y
\]

with \( N_x, N_y \sim \mathcal{N}(0, 1) \). The

\[
\begin{align*}
P(Y) &= \mathcal{N}(0, 10) \\
P(Y|do(X = 2)) &= \mathcal{N}(6, 1), \quad P(Y|do(X = 1.2)) = \mathcal{N}(3.6, 1) \\
P(X|do(Y = 2)) &= P(X|do(Y = 1.2)) = \mathcal{N}(0, 1) = P(X)
\end{align*}
\]

Intervening on \( X \) changes the distribution of \( Y \) but not the other way around.
**Total causal effect:** Given an SEM $S$, there is a (total) causal effect from $X_i$ to $X_j$ iff

$$X_i \perp X_j \text{ in } \mathbb{P}_S(X_1, \ldots, X_p | do(X_i = \tilde{N}_x))$$

for some variable $\tilde{N}_x$. 
Total causal effect

- **Total causal effect**: Given an SEM $\mathcal{S}$, there is a (total) causal effect from $X_i$ to $X_j$ iff

$$X_i \perp\!\!\!\!\perp X_j \text{ in } \mathbb{P}_S(X_1, \ldots, X_p|do(X_i = \tilde{N}_x))$$

for some variable $\tilde{N}_x$.

- **Example**: Consider the following SEM,

$$X = N_x, \quad Y = 3X + N_y.$$  

When we replace the structural equation for $X$ with $X = \tilde{N}_x$, the dependency between $X$ and $Y$ does not vanish. Thus, there is a causal effect from $X$ to $Y$.  


Proposition:
If there is no directed path from $X$ to $Y$, then there is no causal effect.

Example: Consider the following SEM

$$A = N_a, \quad B = A \oplus N_b, \quad C = B \oplus N_c,$$

where $N_a \sim \text{Ber}(1/2)$, $N_b \sim \text{Ber}(1/3)$ and $N_c \sim \text{Ber}(1/20)$ are independent. $\oplus$ denotes addition modulo 2 (i.e. $1 \oplus 1 = 0$)

- $\mathbb{P}_S(B \mid \text{do}(C = 1)) = \mathbb{P}(B)$
- $\mathbb{P}_S(B \mid \text{do}(A = 1)) = \text{Ber}(2/3) \neq \mathbb{P}(B)$
- There are causal effects from $A$ to $B$ and $A$ to $C$. 
Counterfactual SEM: Consider an SEM $S := (S, \mathbb{P}^N)$ over nodes $X$. Given some observations $x$, we define a counterfactual SEM by replacing the distribution of noise variables:

$$S_{X=x} := (S, \mathbb{P}^N_{S,X=x})$$

- $\mathbb{P}^N_{S,X=x} = \mathbb{P}^N_{S|X=x}$
- The new set of noises need not be independent.


- **Counterfactual SEM**: Consider an SEM $\mathcal{S} := (S, \mathbb{P}^N)$ over nodes $X$. Given some observations $x$, we define a counterfactual SEM by replacing the distribution of noise variables:

$$S_{X=x} := (S, \mathbb{P}^N_{S,X=x})$$

- $\mathbb{P}^N_{S,X=x} = \mathbb{P}^N_{S | X=x}$

- The new set of noises need not be independent.

- *Counterfactual statements* can be seen as do-statements in the new counterfactual SEM.
Counterfactual

- **Example:** Consider the following SEM

\[ X = N_x, \quad Y = X^2 + N_y, \quad Z = 2Y + X + N_z \]

where \( N_x, N_y, N_z \sim \mathcal{N}(0, 1) \).

- Suppose we observe \((x, y, z) = (1, 2, 4)\)
**Example:** Consider the following SEM

\[ X = N_x, \quad Y = X^2 + N_y, \quad Z = 2Y + X + N_z \]

where \( N_x, N_y, N_z \sim \mathcal{N}(0, 1) \).

- Suppose we observe \((x, y, z) = (1, 2, 4)\)
- \( \mathbb{P}_{S}^{N}(1,2,4) \) puts a point mass on \((N_x, N_y, N_z) = (1, 1, -1)\).
**Example:** Consider the following SEM

\[ X = N_x, \quad Y = X^2 + N_y, \quad Z = 2Y + X + N_z \]

where \( N_x, N_y, N_z \sim \mathcal{N}(0, 1) \).

Suppose we observe \((x, y, z) = (1, 2, 4)\)

\( \mathbb{P}^{N|(1,2,4)}_S \) puts a point mass on \((N_x, N_y, N_z) = (1, 1, -1)\).

Counterfactual statement: “Z would have been 11, had X been 2.”

means \( \mathbb{P}^{Z|do(X=2)}_S, (1,2,4) \) is a point mass on 11.
**Example:** Consider the following SEM

\[ X = N_x, \quad Y = X^2 + N_y, \quad Z = 2Y + X + N_z \]

where \( N_x, N_y, N_z \sim \mathcal{N}(0, 1) \).

Suppose we observe \((x, y, z) = (1, 2, 4)\)

\[ \mathbb{P}^N_{S(1,2,4)} \] puts a point mass on \((N_x, N_y, N_z) = (1, 1, -1)\).

counterfactual statement: “Z would have been 11, had X been 2.”

means \[ \mathbb{P}^Z|_{do(X=2)}_{S,(1,2,4)} \] is a point mass on 11.

“Y would have been 5, had X been 2.”

“Z would have been 11, had Y been 5.”
Graphical models can encode a set of conditional dependence and independence of variables.

**Markov property** enables us to read off CI from a graph.

**Faithfulness** allows us to read off graphical property (d-separation) from CI.
Graphical models can encode a set of conditional dependence and independence of variables.

**Markov property** enables us to read off CI from a graph.

**Faithfulness** allows us to read off graphical property (d-separation) from CI.

**Example:** \( P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_3) \)
Graphical models can encode a set of conditional dependence and independence of variables.

**Markov property** enables us to read off CI from a graph.

**Faithfulness** allows us to read off graphical property (d-separation) from CI.

**Example:**
\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_3)
\]

CI: \(X_4 \perp X_1|X_3, \ X_4 \perp X_2|X_3\).

From the graph: \(X_1\) and \(X_4\) are “d-separated” by \(X_3\).
Three nodes are called an *immorality* or a *v-structure* if one node is a child of the two others that themselves are not adjacent.

\[ i \rightarrow j \leftarrow k \]

j is called a **collider**.
Three nodes are called an *immorality* or a *v-structure* if one node is a child of the two others that themselves are not adjacent.

\[ i \rightarrow j \leftarrow k \]

\(j\) is called a *collider*.

**Blocked path:** In a DAG, a path between \(i_1\) and \(i_n\) is blocked by a set \(S\) (with neither \(i_1\) nor \(i_n\) in \(S\)) whenever there is a node \(i_k\), such that one of the following happens:

1. \(i_k \in S\) and \(i_{k-1} \rightarrow i_k \rightarrow i_{k+1}\) or \(i_{k-1} \leftarrow i_k \leftarrow i_{k+1}\) or \(i_{k-1} \leftarrow i_k \rightarrow i_{k+1}\).
2. \(i_{k-1} \rightarrow i_k \leftarrow i_{k+1}\) and neither \(i_k\) nor any of its descendants is in \(S\).
D-separation

- Three nodes are called an **immorality** or a **v-structure** if one node is a child of the two others that themselves are not adjacent.

\[ i \rightarrow j \leftarrow k \]

\( j \) is called a **collider**.

- **Blocked path**: In a DAG, a path between \( i_1 \) and \( i_n \) is blocked by a set \( S \) (with neither \( i_1 \) nor \( i_n \) in \( S \)) whenever there is a node \( i_k \), such that one of the following happens:

1. \( i_k \in S \) and \( i_{k-1} \rightarrow i_k \rightarrow i_{k+1} \) or \( i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \) or \( i_{k-1} \leftarrow i_k \rightarrow i_{k+1} \).
2. \( i_{k-1} \rightarrow i_k \leftarrow i_{k+1} \) and neither \( i_k \) nor any of its descendants is in \( S \).

- **D-separation**: Two disjoint subsets of vertices \( A \) and \( B \) are d-separated by a third (also disjoint) subset \( S \) if every path between nodes in \( A \) and \( B \) is blocked by \( S \).
D-separation

- Three nodes are called an *immorality* or a *v-structure* if one node is a child of the two others that themselves are not adjacent.

  \[ i \rightarrow j \leftarrow k \]

  \( j \) is called a *collider*.

- **Blocked path**: In a DAG, a path between \( i_1 \) and \( i_n \) is blocked by a set \( S \) (with neither \( i_1 \) nor \( i_n \) in \( S \)) whenever there is a node \( i_k \), such that one of the following happens:

  1. \( i_k \in S \) and \( i_{k-1} \rightarrow i_k \rightarrow i_{k+1} \) or \( i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \) or \( i_{k-1} \leftarrow i_k \rightarrow i_{k+1} \).

  2. \( i_{k-1} \rightarrow i_k \leftarrow i_{k+1} \) and neither \( i_k \) nor any of its descendants is in \( S \).

- **D-separation**: Two disjoint subsets of vertices \( A \) and \( B \) are d-separated by a third (also disjoint) subset \( S \) if every path between nodes in \( A \) and \( B \) is blocked by \( S \).

- **Unblocked path**: a path can be traced without traversing colliding (head to head) arrows.
D-separation

- Three nodes are called an *immorality* or a *v-structure* if one node is a child of the two others that themselves are not adjacent.
  \[ i \rightarrow j \leftarrow k \]

  \( j \) is called a **collider**.

- **Blocked path**: In a DAG, a path between \( i_1 \) and \( i_n \) is blocked by a set \( S \) (with neither \( i_1 \) nor \( i_n \) in \( S \)) whenever there is a node \( i_k \), such that one of the following happens:
  1. \( i_k \in S \) and \( i_{k-1} \rightarrow i_k \rightarrow i_{k+1} \) or \( i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \) or \( i_{k-1} \leftarrow i_k \rightarrow i_{k+1} \).
  2. \( i_{k-1} \rightarrow i_k \leftarrow i_{k+1} \) and neither \( i_k \) nor any of its descendants is in \( S \).

- **D-separation**: Two disjoint subsets of vertices \( A \) and \( B \) are d-separated by a third (also disjoint) subset \( S \) if every path between nodes in \( A \) and \( B \) is blocked by \( S \).

- **Unblocked path**: A path can be traced without traversing colliding (head to head) arrows.

- Given a DAG \( G \), we obtain the undirected **moralized graph** \( G^{\text{mor}} \) of \( G \) by connecting the parents of each node and removing the directions of the edges.
Markov Properties

- **Markov property** Given a DAG $G$ and a joint distribution $\mathbb{P}^X$, this distribution is said to satisfy
  - the *global Markov* property with respect to the DAG $G$ if
    \[ A, B \text{ d-separated by } C \Rightarrow A \perp B \mid C \]
    for all disjoint sets $A, B, C$. 

Markov Properties

- **Markov property** Given a DAG $G$ and a joint distribution $\mathbb{P}^X$, this distribution is said to satisfy the **global Markov** property with respect to the DAG $G$ if

$$A, B \text{ d-separated by } C \Rightarrow A \perp\!
\!
\!
\perp B | C$$

for all disjoint sets $A, B, C$.

- the **local Markov** property with respect to the DAG $G$ if each variable is independent of its non-descendants given its parents.
Markov Properties

- **Markov property** Given a DAG $G$ and a joint distribution $\mathbb{P}^X$, this distribution is said to satisfy
  - the *global Markov* property with respect to the DAG $G$ if
    \[ A, B \text{ d-separated by } C \implies A \perp\perp B | C \]
    for all disjoint sets $A, B, C$.
  - the *local Markov* property with respect to the DAG $G$ if each variable is independent of its non-descendants given its parents.
  - the Markov factorization property with respect to the DAG $G$ if
    \[ \mathbb{P}(X_1, \ldots, X_p) = \prod_{i=1}^p \mathbb{P}(X_i | X_{PA_i}) \]
    where $X_{PA_i}$ denotes the parents of node $i$ in DAG $G$. 

Markov Properties

- **Markov property** Given a DAG $G$ and a joint distribution $\mathbb{P}^X$, this distribution is said to satisfy
  
  - the *global Markov* property with respect to the DAG $G$ if
    
    $$A, B \text{ d-separated by } C \Rightarrow A \perp B | C$$
    
    for all disjoint sets $A, B, C$.
  
  - the *local Markov* property with respect to the DAG $G$ if each variable is independent of its non-descendants given its parents
  
  - the Markov factorization property with respect to the DAG $G$ if
    
    $$\mathbb{P}(X_1, \ldots, X_p) = \prod_{i=1}^{p} \mathbb{P}(X_i | X_{PA_i})$$
    
    where $X_{PA_i}$ denotes the parents of node $i$ in DAG $G$.
  
- For distributions with positive continuous densities, the global and local property are equivalent.
Example: In the following DAG, we have

\[ \begin{array}{c}
X_1 \\
X_2 \\
X_3 \\
X_4 \\
\end{array} \]

- \( X_2 \) and \( X_3 \) are d-separated by \( X_1 \), \( \Rightarrow X_2 \perp X_3 | X_1 \)
- \( X_1 \) and \( X_4 \) are d-separated by \( \{X_2, X_3\} \), \( \Rightarrow X_1 \perp X_4 | X_2, X_3 \)
- \( P(X) = P(X_3)P(X_1|X_3)P(X_2|X_1)P(X_4|X_2, X_3) \)
Markov Equivalence Class

- **Markov equivalence class of graphs** We denote by \( \mathcal{M}(G) \) the set of distributions that are Markov with respect to \( G \):

  \[ \mathcal{M}(G) := \{ \mathbb{P} : \text{satisfies the global (or local) Markov property w.r.t. } G \} \]

- Two DAGs \( G_1 \) and \( G_2 \) are Markov equivalent if \( \mathcal{M}(G_1) = \mathcal{M}(G_2) \).
Markov Equivalence Class

- **Markov equivalence class of graphs** We denote by $\mathcal{M}(G)$ the set of distributions that are Markov with respect to $G$:

  $$\mathcal{M}(G) := \{\mathbb{P} : \text{satisfies the global (or local) Markov property w.r.t. } G\}$$

- Two DAGs $G_1$ and $G_2$ are Markov equivalent if $\mathcal{M}(G_1) = \mathcal{M}(G_2)$.

**Theorem [Verma and Pearl, 1991]**

Two DAGs are Markov equivalent if and only if they have the same skeleton and the same immoralities.
**Example:** Next two DAGs are Markov equivalent.
<table>
<thead>
<tr>
<th>( p )</th>
<th>number of DAGs with ( p ) nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>543</td>
</tr>
<tr>
<td>5</td>
<td>29281</td>
</tr>
<tr>
<td>6</td>
<td>3781503</td>
</tr>
<tr>
<td>7</td>
<td>1138779265</td>
</tr>
<tr>
<td>8</td>
<td>783702329343</td>
</tr>
<tr>
<td>9</td>
<td>1213442454842881</td>
</tr>
<tr>
<td>10</td>
<td>4175098976430598143</td>
</tr>
<tr>
<td>11</td>
<td>3160345936418917607425</td>
</tr>
<tr>
<td>12</td>
<td>521939651343829405020504063</td>
</tr>
<tr>
<td>13</td>
<td>18676600744432035186664816926721</td>
</tr>
<tr>
<td>14</td>
<td>1439428141044398334941790719839535103</td>
</tr>
<tr>
<td>15</td>
<td>237725265553410354992180218286376719253505</td>
</tr>
<tr>
<td>16</td>
<td>837566707737333320287699303047996412235223138303</td>
</tr>
<tr>
<td>17</td>
<td>6270792119692388989446452602494921906963551482675201</td>
</tr>
<tr>
<td>18</td>
<td>99421195322159515895228914592354524516555026878588305014783</td>
</tr>
<tr>
<td>19</td>
<td>327719012271075917361775733311261125883583076258421902583546773505</td>
</tr>
<tr>
<td>20</td>
<td>2344880451051088988152559855229099188899081192234291298795803236068491263</td>
</tr>
</tbody>
</table>
Consider a graph $G = (V, E)$ and a target node $Y$. The Markov blanket of $Y$ is the smallest set $M$ such that

$$Y \text{ d-sep. } V \setminus (\{Y\} \cup M) \text{ given } M.$$

If $\mathbb{P}^X$ is Markov w.r.t. $G$, then

$$Y \perp\!
\!
\!
\!
\perp V \setminus (\{Y\} \cup M) | M$$
Consider a graph $G = (V, E)$ and a target node $Y$. The Markov blanket of $Y$ is the smallest set $M$ such that

$$Y \text{ d-sep. } V \setminus (\{Y\} \cup M) \text{ given } M.$$ 

If $P^X$ is Markov w.r.t. $G$, then

$$Y \independent V \setminus (\{Y\} \cup M)|M$$

The Markov blanket of a node is the set of nodes consisting of its parents, its children, and any other parents of its children.
Example:

Markov blanket of $Z$ is $M_Z := \{T, R, Y\}$, because $Z$ is d-sep. from $\{X, W\}$ by $M_Z$.

What is the Markov blanket of $R$?
Definition: \( \mathbb{P}^X \) is said to be faithful to the DAG \( G \) if

\[
A \perp B | C \Rightarrow A, B \text{ d-sep. by } C
\]

for all disjoint sets \( A, B, C \).
**Definition:** $\mathbb{P}^X$ is said to be *faithful* to the DAG $G$ if

$$A \perp \! \! \! \perp B \mid C \Rightarrow A, B \text{ d-sep. by } C$$

for all disjoint sets $A, B, C$.

- Markov assumption enables us to read off independence from a graph. Faithfulness allows us to infer dependencies from the graph.  

---

\[ p \Rightarrow q \equiv \neg q \Rightarrow \neg p \]
**Definition:** $\mathbb{P}^X$ is said to be *faithful* to the DAG $G$ if

$$A \perp B \mid C \Rightarrow A, B \text{ d-sep. by } C$$

for all disjoint sets $A, B, C$.

- Markov assumption enables us to read off independence from a graph. Faithfulness allows us to infer dependencies from the graph\(^1\).

- A distribution satisfies *causal minimality* with respect to $G$ if it is Markov with respect to $G$, but not to any proper subgraph of $G$. 

\(^1 p \Rightarrow q \equiv \neg q \Rightarrow \neg p \)
**Example:** Consider the following SEM,

\[
X = N_X, \quad Y = aX + N_Y, \quad Z = bY + cX + N_Z,
\]

where \( N_X \sim \mathcal{N}(0, \sigma_X^2) \), \( N_Y \sim \mathcal{N}(0, \sigma_Y^2) \), and \( N_Z \sim \mathcal{N}(0, \sigma_Z^2) \).

\[ G_1 : \quad Y \xrightarrow{a} X \quad \xrightarrow{c} Z \]

\[ \xrightarrow{b} \]

\[
Y \xrightarrow{b} Z
\]
Example: Consider the following SEM,

\[ X = N_X, \quad Y = aX + N_Y, \quad Z = bY + cX + N_Z, \]

where \( N_X \sim \mathcal{N}(0, \sigma_X^2) \), \( N_Y \sim \mathcal{N}(0, \sigma_Y^2) \), and \( N_Z \sim \mathcal{N}(0, \sigma_Z^2) \).

\[ G_1 : \quad \begin{array}{c}
X \\
\downarrow^a
\end{array} \quad Y \quad \begin{array}{c}
\downarrow^b
\end{array} \quad Z \quad \begin{array}{c}
\downarrow^c
\end{array} \]

If \( ab + c = 0 \), the distribution is not faithful with respect to \( G_1 \) since we obtain \( X \perp\!\!\!\perp Z \)

\[ \mathbb{E}[(X - \mu_X)(Z - \mu_Z)] = \mathbb{E}[XZ] = (ac + b)\mathbb{E}[X^2] = 0 \]
Example: Consider the following SEM,

\[ X = \tilde{N}_X, \quad Y = \tilde{a}X + \tilde{b}Z + \tilde{N}_Y, \quad Z = \tilde{N}_Z, \]

where \( \tilde{N}_X \sim \mathcal{N}(0, \delta^2_X) \), \( \tilde{N}_Y \sim \mathcal{N}(0, \delta^2_Y) \), and \( \tilde{N}_Z \sim \mathcal{N}(0, \delta^2_Z) \).

\[ G_2 : \ \egin{array}{c}
X \\
\tilde{a}
\end{array} \leftarrow \begin{array}{c}
Y \\
\tilde{b}
\end{array} \rightarrow \begin{array}{c}
Z
\end{array} \]
• **Example:** Consider the following SEM,

\[ X = \tilde{N}_X, \quad Y = \tilde{a}X + \tilde{b}Z + \tilde{N}_Y, \quad Z = \tilde{N}_Z, \]

where \( \tilde{N}_X \sim \mathcal{N}(0, \delta_X^2) \), \( \tilde{N}_Y \sim \mathcal{N}(0, \delta_Y^2) \), and \( \tilde{N}_Z \sim \mathcal{N}(0, \delta_Z^2) \).

![](image)

If we choose

\[
\delta_X^2 = \sigma_X^2, \quad \tilde{a} = a, \quad \delta_Z^2 = b^2 \sigma_Y^2 + \sigma_Z^2
\]

\[
\tilde{b} = (b \sigma_Y^2)/(b^2 \sigma_Y^2 + \sigma_Z^2), \quad \delta_Y^2 = \sigma_Y^2 - (b^2 \sigma_Y^4)/(b^2 \sigma_Y^2 + \sigma_Z^2)
\]
Example: Consider the following SEM,

\[ X = \tilde{\mathcal{N}}_X, \quad Y = \tilde{a}X + \tilde{b}Z + \tilde{\mathcal{N}}_Y, \quad Z = \tilde{\mathcal{N}}_Z, \]

where \( \tilde{\mathcal{N}}_X \sim \mathcal{N}(0, \delta_X^2), \) \( \tilde{\mathcal{N}}_Y \sim \mathcal{N}(0, \delta_Y^2), \) and \( \tilde{\mathcal{N}}_Z \sim \mathcal{N}(0, \delta_Z^2). \)

If we choose

\[ \tilde{\mathcal{N}}_X \sim \mathcal{N}(0, \delta_X^2), \quad \tilde{\mathcal{N}}_Y \sim \mathcal{N}(0, \delta_Y^2), \quad \tilde{\mathcal{N}}_Z \sim \mathcal{N}(0, \delta_Z^2). \]

then both SEMs will lead to the same covariance matrix and the same observational distribution.

\[ \Sigma = 
\begin{pmatrix}
\sigma_X^2 & a\sigma_X^2 & 0 \\
 a\sigma_X^2 & a^2\sigma_X^2 + \sigma_Y^2 & b\sigma_Y^2 \\
 0 & b\sigma_Y^2 & b^2\sigma_Y^2 + \sigma_Z^2
\end{pmatrix} \]
Example: Consider the following SEM,

\[
X = \tilde{N}_X, \quad Y = \tilde{a}X + \tilde{b}Z + \tilde{N}_Y, \quad Z = \tilde{N}_Z,
\]

where \( \tilde{N}_X \sim \mathcal{N}(0, \delta_X^2) \), \( \tilde{N}_Y \sim \mathcal{N}(0, \delta_Y^2) \), and \( \tilde{N}_Z \sim \mathcal{N}(0, \delta_Z^2) \).

\[ \begin{array}{ccc}
X \\
G_2: & Y & \leftarrow \tilde{a} \\
& \tilde{b} & Z
\end{array} \]

If we choose

\[
\delta_X^2 = \sigma_X^2, \quad \tilde{a} = a, \quad \delta_Z^2 = b^2 \sigma_Y^2 + \sigma_Z^2 \\
\tilde{b} = (b \sigma_Y^2)/(b^2 \sigma_Y^2 + \sigma_Z^2), \quad \delta_Y^2 = \sigma_Y^2 - (b^2 \sigma_Y^4)/(b^2 \sigma_Y^2 + \sigma_Z^2)
\]

then both SEMs will lead to the same covariance matrix and the same observational distribution.

\[
\Sigma = \begin{pmatrix}
\sigma_X^2 & a \sigma_X^2 & 0 \\
a \sigma_X^2 & a^2 \sigma_X^2 + \sigma_Y^2 & b \sigma_Y^2 \\
0 & b \sigma_Y^2 & b^2 \sigma_Y^2 + \sigma_Z^2
\end{pmatrix}
\]

However, the distribution is faithful with respect to \( G_2 \) if \( \tilde{a}, \tilde{b} \neq 0 \) and all \( \delta_\cdot^2 > 0 \).
Consider an SEM $S$,

$$X_j = f_j(\text{PA}_j, N_j)$$

Due to the Markov property, we have

$$\mathbb{P}_S(X) = \prod_{i=1}^{p} \mathbb{P}(X_j | X_{\text{PA}_j})$$
Do operation

- Consider an SEM $S$,
  \[ X_j = f_j(\text{PA}_j, N_j) \]

- Due to the Markov property, we have
  \[ P_S(X) = \prod_{i=1}^{p} P(X_j|X_{\text{PA}_j}) \]

- Now consider the SEM $\tilde{S}$ after $do(X_k = \tilde{N}_k)$ with $\tilde{N}_k \sim \tilde{p}(X_k)$,
  \[ P_{S,do(X_k=\tilde{N}_k)}(X) = \tilde{p}(X_k) \prod_{j \neq k} P(X_j|X_{\text{PA}_j}) \]
Do operation

- Consider an SEM $S$,
  \[ X_j = f_j(PA_j, N_j) \]

- Due to the Markov property, we have
  \[
  P_S(X) = \prod_{i=1}^{p} P(X_j | X_{PA_j})
  \]

- Now consider the SEM $\tilde{S}$ after $do(X_k = \tilde{N}_k)$ with $\tilde{N}_k \sim \tilde{p}(X_k)$,
  \[
  \mathbb{P}_{S, do(X_k=\tilde{N}_k)}(X) = \tilde{p}(X_k) \prod_{j \neq k} P(X_j | X_{PA_j})
  \]

- Perfect intervention: $do(X_k = a)$
  \[
  \mathbb{P}_{S, do(X_k=a)}(X) = \begin{cases} 
  \prod_{j \neq k} P(X_j | X_{PA_j}) & X_k = a \\
  0 & \text{Otherwise}
  \end{cases}
  \] (1)
**Example:** Consider the Simpson’s paradox in which all variables are binary.

**T:** Treatment, **Z:** size of stone, **R:** recovery

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Patients with small stones</th>
<th>Patients with large stones</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment A:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open surgery</td>
<td>78% (273/350)</td>
<td><strong>93%</strong> (81/87)</td>
<td>73% (192/263)</td>
</tr>
<tr>
<td><strong>Treatment B:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percutaneous nephrolithotomy</td>
<td><strong>83%</strong> (289/350)</td>
<td>87% (234/270)</td>
<td>69% (55/80)</td>
</tr>
</tbody>
</table>

We are interested in

$$P_S(R = 1|do(T = A))$$

$$P_S(R = 1|do(T = B))$$
We have

$$\mathbb{P}_S(R = 1 \mid do(T = A)) = \sum_{z=0}^{1} \mathbb{P}_{S, do(T=A)}(R = 1, Z = z, T = A)$$

$$= \sum_{z=0}^{1} \mathbb{P}_{S, do(T=A)}(R = 1 \mid Z = z, T = A) \mathbb{P}_{S, do(T=A)}(Z = z, T = A)$$

$$= \sum_{z=0}^{1} \mathbb{P}_S(R = 1 \mid T = A, Z = z) \mathbb{P}_S(Z = z \mid do(T = A))$$

$$= \sum_{z=0}^{1} \mathbb{P}_S(R = 1 \mid T = A, Z = z) \mathbb{P}_S(Z = z)$$
We have

\[ \mathbb{P}_S(R = 1|do(T = A)) = \sum_{z=0}^{1} \mathbb{P}_{S, do(T=A)}(R = 1, Z = z, T = A) \]

\[ = \sum_{z=0}^{1} \mathbb{P}_{S, do(T=A)}(R = 1|Z = z, T = A) \mathbb{P}_{S, do(T=A)}(Z = z, T = A) \]

\[ = \sum_{z=0}^{1} \mathbb{P}_S(R = 1|T = A, Z = z) \mathbb{P}_S(Z = z|do(T = A)) \]

\[ = \sum_{z=0}^{1} \mathbb{P}_S(R = 1|T = A, Z = z) \mathbb{P}_S(Z = z) \]

The last step uses the perfect intervention formula in (1)
We have

\[ \mathbb{P}_S(R = 1|do(T = A)) = \sum_{z=0}^{1} \mathbb{P}_{S, do(T = A)}(R = 1, Z = z, T = A) \]

\[ = \sum_{z=0}^{1} \mathbb{P}_{S, do(T = A)}(R = 1|Z = z, T = A)\mathbb{P}_{S, do(T = A)}(Z = z, T = A) \]

\[ = \sum_{z=0}^{1} \mathbb{P}_S(R = 1|T = A, Z = z)\mathbb{P}_S(Z = z|do(T = A)) \]

\[ = \sum_{z=0}^{1} \mathbb{P}_S(R = 1|T = A, Z = z)\mathbb{P}_S(Z = z) \]

The last step uses the perfect intervention formula in (1)

Using the values in the table, we obtain

\[ \mathbb{P}_S(R = 1|do(T = A)) \approx 0.93 \cdot \frac{357}{700} + 0.73 \cdot \frac{343}{700} = 0.832 \]

\[ \mathbb{P}_S(R = 1|do(T = B)) \approx 0.87 \cdot \frac{357}{700} + 0.69 \cdot \frac{343}{700} = 0.782 \]

But \( \mathbb{P}_S(R = 1|T = A) = 0.78 \),
A major complication is the possibility of inconsistent parameter estimation due to the existence of hidden confounders.
A major complication is the possibility of inconsistent parameter estimation due to the existence of hidden confounders.

The *instrumental variables* method provides a way to nonetheless obtain consistent parameter estimates.
A major complication is the possibility of inconsistent parameter estimation due to the existence of hidden confounders.

The *instrumental variables* method provides a way to nonetheless obtain consistent parameter estimates.

We explain this method through a simple example.
A major complication is the possibility of inconsistent parameter estimation due to the existence of hidden confounders.

The *instrumental variables* method provides a way to nonetheless obtain consistent parameter estimates.

We explain this method through a simple example.

**Example:** Consider the following causal structure in which $A$ is a hidden confounder.

\[
\begin{align*}
Y(\text{log earning}) &= \alpha + \rho S(\text{Schooling years}) + \gamma A(\text{Individual ability}) + N_Y(\text{noise}) \\
S &\rightarrow A \\
&\downarrow \delta \\
A &\rightarrow Y \\
&\downarrow \gamma \\
&\downarrow \rho \\
S &\rightarrow Y
\end{align*}
\]
A major complication is the possibility of inconsistent parameter estimation due to the existence of hidden confounders.

The *instrumental variables* method provides a way to nonetheless obtain consistent parameter estimates.

We explain this method through a simple example.

**Example:** Consider the following causal structure in which $A$ is a hidden confounder.

$$Y(\text{log earning}) = \alpha + \rho S(\text{Schooling years}) + \gamma A(\text{Individual ability}) + N_Y(\text{noise})$$

Interested in finding the influence of $S$ on $Y$, i.e., finding $\rho$.

No unbiased estimator exists for $\rho$:

$$\hat{\rho} := \frac{\text{Cov}(Y,S)}{\text{Var}(S)} = \frac{\text{Cov}(Y = \alpha + \rho S + \gamma A + N_Y, S)}{\text{Var}(S)} = \rho + \gamma \frac{\text{Cov}(A,S)}{\text{Var}(S)}.$$
Now, consider the following graph in which

- A is a hidden confounder.
- Z is a variable such that

\[ \text{Cov}(Z, S) \neq 0 \text{(First stage restriction criterion)} \]

\[ \text{Cov}(Z, \gamma A + N_Y) = 0 \text{(Exclusion Restriction)} \]

\[ Z \xrightarrow{\beta} S \xrightarrow{\rho} Y \]

\[ A \xrightarrow{\delta} \]

\[ A \xrightarrow{\gamma} \]
Now, consider the following graph in which

- $A$ is a hidden confounder.
- $Z$ is a variable such that

\[ \text{Cov}(Z, S) \neq 0 \text{(First stage restriction criterion)} \]

\[ \text{Cov}(Z, \gamma A + N_Y) = 0 \text{(Exclusion Restriction)} \]

Interested in finding $\rho$. 
Now, consider the following graph in which

- \( A \) is a hidden confounder.
- \( Z \) is a variable such that

\[
\text{Cov}(Z, S) \neq 0 (\text{First stage restriction criterion})
\]

\[
\text{Cov}(Z, \gamma A + N_Y) = 0 (\text{Exclusion Restriction}).
\]

Interested in finding \( \rho \).

In this model, variable \( Z \) (also called \textit{Instrumental Variable} ) can help to estimate \( \rho \).
In this model, we have

\[ S = \beta Z + \delta A + N_S, \]
\[ Y = \alpha + \rho S + \gamma A + N_Y. \]
In this model, we have

\[ S = \beta Z + \delta A + N_S, \]
\[ Y = \alpha + \rho S + \gamma A + N_Y. \]

First, estimate \( \rho \beta \):

\[ \hat{\rho \beta} := \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}. \]
In this model, we have

\[ S = \beta Z + \delta A + N_S, \]
\[ Y = \alpha + \rho S + \gamma A + N_Y. \]

First, estimate \( \rho \beta \):

\[ \hat{\rho} \beta := \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}. \]

Second, estimate \( \beta \):

\[ \hat{\beta} := \frac{\text{Cov}(S, Z)}{\text{Var}(Z)}. \]
In this model, we have

\[ S = \beta Z + \delta A + N_S, \]

\[ Y = \alpha + \rho S + \gamma A + N_Y. \]

First, estimate \( \rho \beta \):

\[ \hat{\rho} \beta := \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}. \]

Second, estimate \( \beta \):

\[ \hat{\beta} := \frac{\text{Cov}(S, Z)}{\text{Var}(Z)}. \]

Finally, estimate \( \rho \):

\[ \hat{\rho} := \frac{\hat{\rho} \hat{\beta}}{\hat{\beta}} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(S, Z)}. \]

This is an unbiased estimator because

\[ \hat{\rho} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(S, Z)} = \frac{\text{Cov}(\alpha + \rho S + \gamma A + N_Y, Z)}{\text{Cov}(S, Z)} = \rho + \frac{\text{Cov}(\gamma A + N_Y, Z)}{\text{Cov}(S, Z)}. \]
In this model, we have

\[ S = \beta Z + \delta A + N_S, \]
\[ Y = \alpha + \rho S + \gamma A + N_Y. \]

First, estimate \( \rho \beta \):
\[ \hat{\rho}_\beta := \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}. \]

Second, estimate \( \beta \):
\[ \hat{\beta} := \frac{\text{Cov}(S, Z)}{\text{Var}(Z)}. \]

Finally, estimate \( \rho \):
\[ \hat{\rho} := \frac{\rho \beta}{\beta} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(S, Z)}. \]

This is an unbiased estimator because

\[ \hat{\rho} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(S, Z)} = \frac{\text{Cov}(\alpha + \rho S + \gamma A + N_Y, Z)}{\text{Cov}(S, Z)} = \rho + \frac{\text{Cov}(\gamma A + N_Y, Z)}{\text{Cov}(S, Z)}. \]

According to the restrictions:

\[ \frac{\text{Cov}(\gamma A + N_Y, Z)}{\text{Cov}(S, Z)} = 0 \]

Therefore, \( \hat{\rho} = \rho \).
The goal is to infer the graph given a set of conditional dependence and independence tests.

SGS Algorithm: developed by Sprites, Glymour and Scheives.
The goal is to infer the graph given a set of conditional dependence and independence tests.

SGS Algorithm: developed by Sprites, Glymour and Scheives.

Consists of two phases
1. Learning the skeleton
2. Learning the orientations
Learning Causal Bayes Nets

- The goal is to infer the graph given a set of conditional dependence and independence tests.
- SGS Algorithm: developed by Sprites, Glymour and Scheives.
- Consists of two phases
  1. Learning the skeleton
  2. Learning the orientations
- SGS is based on two main assumptions
  - No hidden confounders
  - Graph is a DAG
It is based on the following result.

**Lemma**

- Two nodes $X$, $Y$ in a DAG $(V, E)$ are adjacent iff they cannot be d-separated by any subset $S \subseteq V \setminus \{X, Y\}$.
- If two nodes $X$, $Y$ in a DAG $(V, E)$ are not adjacent, then they are d-separated by either $PA_X$ or $PA_Y$. 
Estimation of skeleton

- It is based on the following result.

**Lemma**
- Two nodes $X$, $Y$ in a DAG $(V, E)$ are adjacent iff they cannot be d-separated by any subset $S \subseteq V \setminus \{X, Y\}$.
- If two nodes $X$, $Y$ in a DAG $(V, E)$ are not adjacent, then they are d-separated by either $PA_X$ or $PA_Y$.

**Steps:**
1. Begin with a complete graph.
2. Use conditional dependence and independence test to eliminate edges (edge elimination).
Estimation of skeleton

- It is based on the following result.

Lemma

- Two nodes X, Y in a DAG \((V, E)\) are adjacent iff they cannot be d-separated by any subset \(S \subseteq V \setminus \{X, Y\}\).
- If two nodes X, Y in a DAG \((V, E)\) are not adjacent, then they are d-separated by either \(PA_X\) or \(PA_Y\).

Steps:

1. Begin with a complete graph.
2. Use conditional dependence and independence test to eliminate edges (edge elimination).

There are different methods to perform CI tests, e.g., empirical methods, kernel-based methods. In general, CI tests are difficult to perform in practice.
Orientation of edges

Steps

1. Orient the immoralities (or v-structures) in the graph.
   - For structure $X \rightarrow Y \rightarrow Z$ with no direct edge between $X$ and $Z$.
   - Let $S$ denotes the corresponding d-separation set for $X$ and $Z$.
   - The structure $X \rightarrow Y \rightarrow Z$ is an immorality and can be oriented as $X \rightarrow Y \leftarrow Z$ if and only if $Y \notin S$.

2. We may be able to orient some further edges using e.g., Meek’s orientation rules.
   - If there exist a pair $A, C$ not directly connected and exists node $B$ such that $A \rightarrow B \leftarrow C$, then, we can orient the 2nd arrow from $B$ to $C$.
   - Avoid cycle.
   - ...
Example:

True graph

1. Complete Graph

Edge elimination

v-structures

Logical Orientation
**Edge Elimination**

- (zero Orders) Edge $AE$ removed due to unshielded collider.
- (1st Orders) $ABDF$: A d-sep. F by D, Edge $AF$ eliminated.
  - $BDF$: B d-sep. F by D, Edge $BF$ eliminated.
  - $CBDF$: C d-sep. F by D, Edge $CF$ eliminated.
  - $ECBDF$: E d-sep. F by D, Edge $EF$ eliminated.
- $DBCE$: D d-sep. E by B, Edge $ED$ eliminated.
- (2nd Orders) $BACE$: B d-sep. E by $\{A, C\}$, Edge $BE$ eliminated.
Edge Orientation

- (Statistical Orientation) $A \perp E$, $A \perp C$, $E \perp C$, $A \perp E | C \Rightarrow$ ACE is a v-structure.
- (Logical Orientation) BCE: $C \rightarrow B$, otherwise unshielded collider.
- ABC: $A \rightarrow B$, otherwise cycle.
- ABD: $B \rightarrow D$, otherwise unshielded collider.
- BDF: $D \rightarrow F$, otherwise unshielded collider.
The End