A short course on network causal Inference: theory and applications

Negar Kiyavash

EPFL

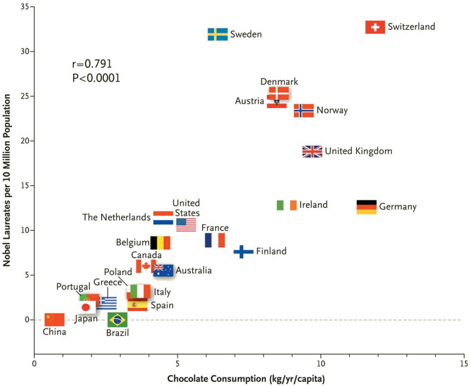
June 2021

Overview

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.”

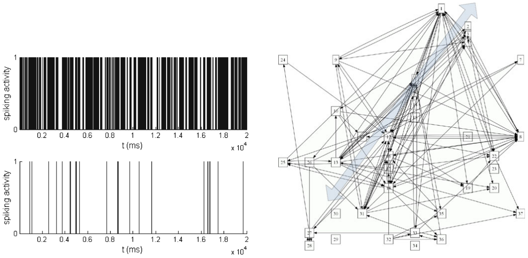
- We must be careful with drawing conclusions like “Eating chocolate produces Nobel prize” or “Geniuses are more likely to eat lots of chocolate.”
- Correlation does not imply causation!

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.

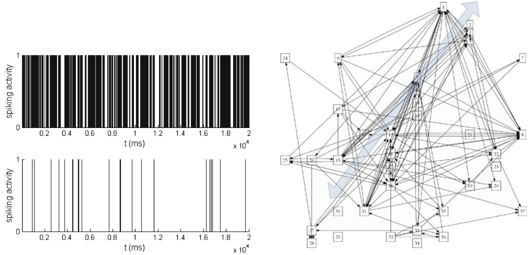
Application areas include:

- **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:

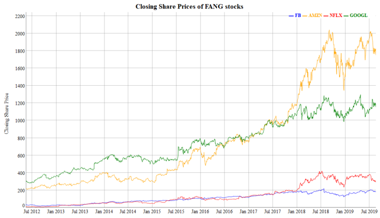
- **Computational Neuroscience:** Advances in recording technologies have given neuroscience researchers access to large amounts of data, e.g., individual recordings of neurons in different parts of the brain.

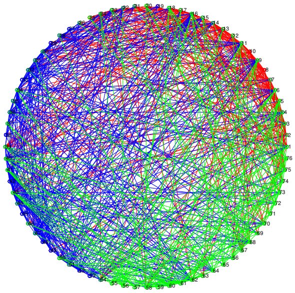


- Could we understand firing of which neurons causes others to fire and hence learn the functional connectivity in the brain?

Application areas include:

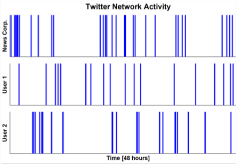
- **Financial Markets:** Financial instability can lead to financial crises due to its contagion or spillover effects to other parts of the economy. Having an accurate measures of systemic risk and inter-dependencies between financial institution gives central banks and policy makers the ability to take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

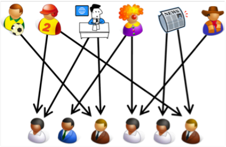




Application areas include:

- **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.

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.

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

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

- 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:

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Admission data on department level:

	Men		Women	
Departments	Applicants	Admitted	Applicants	Admitted
A	825	63%	108	82%
B	560	62%	25	68%
C	325	37%	593	39%

Among 85 departments, there are 6 against men but only 4 against women.

Preliminaries

Throughout the lecture we use the following notation.

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

Preliminaries

Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$: probability space, where Ω 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 $X : \Omega \to E$ from a set of possible outcomes Ω 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 Ω 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 $X : \Omega \to E$ from a set of possible outcomes Ω 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. $E = \mathbb{R}$.

Preliminaries

Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$: probability space, where Ω 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 $X : \Omega \to E$ from a set of possible outcomes Ω 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.

Preliminaries

Throughout the lecture we use the following notation.

- $(\Omega, F, \mathbb{P})$: probability space, where Ω 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 $X : \Omega \to E$ from a set of possible outcomes Ω 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 $X \perp\!\!\!\perp Y$ if and only if

$$\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$$

- We call $X_1, ..., X_p$ jointly (or mutually) independent if and only if

$$\mathbb{P}(X_1, ..., X_p) = \mathbb{P}(X_1)...\mathbb{P}(X_p).$$

- We call $X_1, ..., X_p$ jointly (or mutually) independent if and only if

$$\mathbb{P}(X_1, ..., X_p) = \mathbb{P}(X_1)...\mathbb{P}(X_p).$$

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

$$\mathbb{P}(x, y|z) = \mathbb{P}(x|z)\mathbb{P}(y|z)$$

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

- We call $X_1, ..., X_p$ jointly (or mutually) independent if and only if

$$\mathbb{P}(X_1, ..., X_p) = \mathbb{P}(X_1)...\mathbb{P}(X_p).$$

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

$$\mathbb{P}(x, y|z) = \mathbb{P}(x|z)\mathbb{P}(y|z)$$

  for all x, y, z such that $p(z) > 0$. Otherwise, X and Y are dependent conditional on Z and we write $X \not\perp\!\!\!\perp Y|Z$.
- We call X and Y *uncorrelated* if $\mathbb{E}[X^2], \mathbb{B}[Y^2] < \infty$ and

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Graphs

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

Graphs

- A graph $G = (V, \mathcal{E})$ consists of (finitely many) nodes or vertices $V = \{1, ..., p\}$ 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 $\mathcal{E}_1 \subseteq \mathcal{E}$.

Graphs

- A graph $G = (V, \mathcal{E})$ consists of (finitely many) nodes or vertices $V = \{1, ..., p\}$ 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 $(i, j) \notin \mathcal{E}$.

Graphs

- A graph $G = (V, \mathcal{E})$ consists of (finitely many) nodes or vertices $V = \{1, ..., p\}$ 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 $(j, i) \in \mathcal{E}$.

Graphs

- A graph $G = (V, \mathcal{E})$ consists of (finitely many) nodes or vertices $V = \{1, ..., p\}$ 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 $i_1, ..., i_n$, such that there is an edge from i_k and i_{k+1} for all $k = 1, ..., n-1$.

- 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.

- 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.

- 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 $(i, j) \in \mathcal{E}$.

Graphical representation

- 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)

Graphical representation

- 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:**

$$\mathbb{P}(X_1, X_2, X_3, X_4) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_1, X_2)\mathbb{P}(X_4|X_1, X_2, X_3)$$

$$\begin{tikzcd}
 & X_1 \arrow[dl] \arrow[dr] \arrow[dd] & \\
X_2 \arrow[rr] \arrow[dr] & & X_3 \arrow[dl] \\
 & X_4 &
\end{tikzcd}$$

- Edges represent conditional dependencies.

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 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

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

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

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.

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.

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

- If $T_s >$ some critical value, reject the null hypothesis
- Cons: Linear assumption, stationarity, time synchronization.

Going beyond linear models

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

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: $\ell(y, w)$

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

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)

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:

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

- Entropy of random variable X: $H(X) := -\mathbb{E}[\log \mathbb{P}(X)]$
- Mutual information between X and Y : $I(X;Y) := H(X) - H(X|Y)$

Structural Equation Model (SEM)

- A *structural equation model* (SEM) (also called a functional model) is defined as a tuple $\mathcal{S} := (S, \mathbb{P}^N)$, where $S = (S_1, ..., S_p)$ is a collection of p equations

$$S_j : \quad X_j = f_j(PA_j, N_j), \quad j = 1, ..., p,$$

- $PA_j \subseteq \{X_1, ..., X_p\}/\{X_j\}$ 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.

Structural Equation Model (SEM)

- A *structural equation model* (SEM) (also called a functional model) is defined as a tuple $\mathcal{S} := (S, \mathbb{P}^N)$, where $S = (S_1, ..., S_p)$ is a collection of p equations

$$S_j : \quad X_j = f_j(PA_j, N_j), \quad j = 1, ..., p,$$

- $PA_j \subseteq \{X_1, ..., X_p\}/\{X_j\}$ 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:**

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

$$X_1 \longrightarrow X_2 \longrightarrow X_3$$

Structural Equation Model (SEM)

- A *structural equation model* (SEM) (also called a functional model) is defined as a tuple $\mathcal{S} := (S, \mathbb{P}^N)$, where $S = (S_1, ..., S_p)$ is a collection of p equations

$$S_j : \quad X_j = f_j(PA_j, N_j), \quad j = 1, ..., p,$$

- $PA_j \subseteq \{X_1, ..., X_p\}/\{X_j\}$ 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:**

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

$$X_1 \longrightarrow X_2 \longrightarrow X_3$$

- **Example 2:**

$$X = N_x, \quad Y = 4X + N_y, \quad X \to Y$$

Intervention

- **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

- **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

- **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:

$$\mathbb{P}^X_{\tilde{\mathcal{S}}} = \mathbb{P}_{\mathcal{S}}\Big(X | do(X_j = \tilde{f}_j(\tilde{PA}_j, \tilde{N}_j))\Big)$$

Intervention

- **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:

$$\mathbb{P}^X_{\tilde{\mathcal{S}}} = \mathbb{P}_{\mathcal{S}}\Big(X | do(X_j = \tilde{f}_j(\tilde{PA}_j, \tilde{N}_j))\Big)$$

- *Perfect intervention*: when $\tilde{f}_j(\tilde{PA}_j, \tilde{N}_j)$ puts a point mass on a real value a, we simply write $\mathbb{P}_{\mathcal{S}}\big(X | do(X_j = a)\big)$.

- **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

$$\begin{aligned} T &= N_T \\ B &= T.N_B + (1 - T)(1 - N_B) \end{aligned}$$

where $N_B \sim Ber(0.01)$.

- In this example, we have

$$\mathbb{P}_{\mathcal{S}}(B = 0 | do(T = 1)) = 0.99$$

$$\mathbb{P}_{\mathcal{S}}(B = 0 | do(T = 0)) = 0.01$$

- **Another Example:** Suppose that $\mathbb{P}(X, Y)$ is induced by a structural equation model $\mathcal{S}$

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

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

$$\begin{aligned}
&\mathbb{P}(Y) = \mathcal{N}(0, 10) \\
&\mathbb{P}(Y|do(X = 2)) = \mathcal{N}(6, 1), \quad \mathbb{P}(Y|do(X = 1.2)) = \mathcal{N}(3.6, 1) \\
&\mathbb{P}(X|do(Y = 2)) = \mathbb{P}(X|do(Y = 1.2)) = \mathcal{N}(0, 1) = \mathbb{P}(X)
\end{aligned}$$

- **Another Example:** Suppose that $\mathbb{P}(X, Y)$ is induced by a structural equation model $\mathcal{S}$

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

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

$$\mathbb{P}(Y) = \mathcal{N}(0, 10)$$
$$\mathbb{P}(Y|do(X = 2)) = \mathcal{N}(6, 1), \quad \mathbb{P}(Y|do(X = 1.2)) = \mathcal{N}(3.6, 1)$$
$$\mathbb{P}(X|do(Y = 2)) = \mathbb{P}(X|do(Y = 1.2)) = \mathcal{N}(0, 1) = \mathbb{P}(X)$$

- Intervening on X changes the distribution of Y but not the other way around.

Total causal effect

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

$$X_i \not\perp\!\!\!\perp X_j \text{ in } \mathbb{P}_{\mathcal{S}}(X_1, ..., X_p | do(X_i = \tilde{N}_x))$$

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

$$X_i \not\perp\!\!\!\perp X_j \text{ in } \mathbb{P}_{\mathcal{S}}(X_1, ..., X_p | do(X_i = \tilde{N}_x))$$

  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

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

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

- $\mathbb{P}_{\mathcal{S}}(B|do(C=1)) = \mathbb{P}(B)$
- $\mathbb{P}_{\mathcal{S}}(B|do(A=1)) = Ber(2/3) \neq \mathbb{P}(B)$
- There are causal effects from A to B and A to C.

Counterfactual

- **conterfactual 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:

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

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

Counterfactual

- **conterfactual 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:

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

- $\mathbb{P}^N_{\mathcal{S},X=x} = \mathbb{P}^{N|X=x}_{\mathcal{S}}$
- The new set of noises need not be independent.
- *Counterfactual statements* can be seen as do-statements in the new conterfactual 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 $(x, y, z) = (1, 2, 4)$

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 $(x, y, z) = (1, 2, 4)$
- $\mathbb{P}_{\mathcal{S}}^{N|(1,2,4)}$ puts a point mass on $(N_x, N_y, N_z) = (1, 1, -1)$.

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 $(x, y, z) = (1, 2, 4)$
- $\mathbb{P}_{\mathcal{S}}^{N|(1,2,4)}$ 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 $\mathbb{P}_{\mathcal{S},(1,2,4)}^{Z|do(X=2)}$ is a point mass on 11.

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 $(x, y, z) = (1, 2, 4)$
- $\mathbb{P}_{\mathcal{S}}^{N|(1,2,4)}$ 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 $\mathbb{P}_{\mathcal{S},(1,2,4)}^{Z|do(X=2)}$ 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 model

- 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 model

- 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:** $\mathbb{P}(X_1, X_2, X_3, X_4) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_1, X_2)\mathbb{P}(X_4|X_3)$

$$\begin{tikzcd}
 & X_1 \arrow[dl] \arrow[dr] & \\
X_2 \arrow[rr] & & X_3 \arrow[dl] \\
 & X_4 &
\end{tikzcd}$$

Graphical model

- 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:** $\mathbb{P}(X_1, X_2, X_3, X_4) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_1, X_2)\mathbb{P}(X_4|X_3)$

$$
\begin{tikzcd}
 & X_1 \arrow[dl] \arrow[dr] & \\
X_2 \arrow[rr] & & X_3 \arrow[dl] \\
 & X_4 &
\end{tikzcd}
$$

- CI: $X_4 \perp\!\!\!\perp X_1 | X_3$, $X_4 \perp\!\!\!\perp X_2 | X_3$.
- From the graph: X_1 and X_4 are "d-separated" by X_3.

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.

$$i \to j \leftarrow k$$

  j is called a **collider**.

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

- 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.

  $$i \to j \leftarrow k$$

  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 $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$.
  2. $i_{k-1} \to i_k \leftarrow i_{k+1}$ 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^{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

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

    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

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

    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

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

    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
  - the Markov factorization property with respect to the DAG G if

    $$\mathbb{P}(X_1, ..., X_p) = \prod_{i=1}^{p} \mathbb{P}(X_i | X_{PA_i})$$

    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

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

    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

    $$\mathbb{P}(X_1, ..., X_p) = \prod_{i=1}^{p} \mathbb{P}(X_i | X_{PA_i})$$

    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

$$\begin{tikzcd}
& X_1 \arrow[dl] & \\
X_2 \arrow[dr] & & X_3 \arrow[ul] \arrow[dl] \\
& X_4 &
\end{tikzcd}$$

- X_2 and X_3 are d-separated by X_1, $\Rightarrow X_2 \perp\!\!\!\perp X_3 | X_1$
- X_1 and X_4 are d-separated by $\{X_2, X_3\}$, $\Rightarrow X_1 \perp\!\!\!\perp X_4 | X_2, X_3$
- $\mathbb{P}(X) = \mathbb{P}(X_3)\mathbb{P}(X_1|X_3)\mathbb{P}(X_2|X_1)\mathbb{P}(X_4|X_2, X_3)$

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.

$$
\begin{tikzcd}
 & X_1 \arrow[dl] \arrow[dr] & & X_4 & \\
X_2 \arrow[rr] & & X_3 \arrow[ur] & & X_5 \arrow[ul]
\end{tikzcd}
$$

$$
\begin{tikzcd}
 & X_1 \arrow[dl] \arrow[dr] & & X_4 & \\
X_2 & & X_3 \arrow[ll] \arrow[ur] & & X_5 \arrow[ul]
\end{tikzcd}
$$

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

Markov Blanket

- 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

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

Markov Blanket

- 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

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

Markov Blanket

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:**

$$
\begin{tikzcd}
& Z \arrow[dl] \arrow[dr] & & Y \arrow[dl] \arrow[dd] \\
T \arrow[dr] & & R \arrow[dl] \arrow[dr] & \\
& X & & W
\end{tikzcd}
$$

- 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?

Faithfulness

- **Definition:** $\mathbb{P}^X$ is said to be *faithful* to the DAG G if

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

for all disjoint sets A, B, C.

Faithfulness

- **Definition:** $\mathbb{P}^X$ is said to be *faithful* to the DAG G if

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

  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].

[1] $p \Rightarrow q \;\equiv\; \neg q \Rightarrow \neg p$

Faithfulness

- **Definition:** $\mathbb{P}^X$ is said to be *faithful* to the DAG G if

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

  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 $N_Z \sim \mathcal{N}(0, \sigma_z^2)$.

$$G_1 : \qquad \begin{tikzcd} & X \arrow[dl, "a"'] \arrow[dr, "c"] & \\ Y \arrow[rr, "b"'] & & Z \end{tikzcd}$$

- **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 $N_Z \sim \mathcal{N}(0, \sigma_z^2)$.

$$G_1 : \quad \begin{tikzcd} & X \arrow[dl, "a"'] \arrow[dr, "c"] & \\ Y \arrow[rr, "b"'] & & Z \end{tikzcd}$$

- 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 $\tilde{N}_X \sim \mathcal{N}(0, \delta_x^2)$, $\tilde{N}_Y \sim \mathcal{N}(0, \delta_y^2)$, and $\tilde{N}_Z \sim \mathcal{N}(0, \delta_z^2)$.

$$G_2: \quad \begin{tikzcd} & X \arrow[dl, "\tilde{a}"'] & \\ Y & & Z \arrow[ll, "\tilde{b}"'] \end{tikzcd}$$

- **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 $\tilde{N}_Z \sim \mathcal{N}(0, \delta_z^2)$.

$$G_2 : \quad \begin{tikzcd} & X \arrow[dl, "\tilde{a}"'] & \\ Y & & Z \arrow[ll, "\tilde{b}"'] \end{tikzcd}$$

- If we choose

$$\begin{aligned} &\delta_X^2 = \sigma_X^2,\ \tilde{a} = a,\ \delta_Z^2 = b^2\sigma_Y^2 + \sigma_Z^2 \\ &\tilde{b} = (b\sigma_Y^2)/(b^2\sigma_Y^2 + \sigma_Z^2),\ \delta_Y^2 = \sigma_Y^2 - (b^2\sigma_Y^4)/(b^2\sigma_Y^2 + \sigma_Z^2) \end{aligned}$$

- **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 $\tilde{N}_Z \sim \mathcal{N}(0, \delta_z^2)$.

$$G_2 : \quad \begin{tikzcd} & & X \arrow[dll, "\tilde{a}"'] \\ Y & & & Z \arrow[lll, "\tilde{b}"'] \end{tikzcd}$$

- If we choose

$$\begin{aligned} &\delta_X^2 = \sigma_X^2,\ \tilde{a} = a,\ \delta_Z^2 = b^2\sigma_Y^2 + \sigma_Z^2 \\ &\tilde{b} = (b\sigma_Y^2)/(b^2\sigma_Y^2 + \sigma_Z^2),\ \delta_Y^2 = \sigma_Y^2 - (b^2\sigma_Y^4)/(b^2\sigma_Y^2 + \sigma_Z^2) \end{aligned}$$

- 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 $\tilde{N}_Z \sim \mathcal{N}(0, \delta_z^2)$.

$$G_2 : \quad \begin{tikzcd} & X \arrow[dl, "\tilde{a}"'] & \\ Y & & Z \arrow[ll, "\tilde{b}"'] \end{tikzcd}$$

- If we choose

$$\begin{aligned} &\delta_X^2 = \sigma_X^2,\ \tilde{a} = a,\ \delta_Z^2 = b^2\sigma_Y^2 + \sigma_Z^2 \\ &\tilde{b} = (b\sigma_Y^2)/(b^2\sigma_Y^2 + \sigma_Z^2),\ \delta_Y^2 = \sigma_Y^2 - (b^2\sigma_Y^4)/(b^2\sigma_Y^2 + \sigma_Z^2) \end{aligned}$$

- 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 $\delta_\cdot^2 > 0$.

Do operation

- Consider an SEM S,

$$X_j = f_j(PA_j, N_j)$$

- Due to the Markov property, we have

$$\mathbb{P}_S(X) = \prod_{i=1}^{p} \mathbb{P}(X_j | X_{PA_j})$$

Do operation

- Consider an SEM S,
$$X_j = f_j(PA_j, N_j)$$
- Due to the Markov property, we have
$$\mathbb{P}_S(X) = \prod_{i=1}^{p} \mathbb{P}(X_j | X_{PA_j})$$
- Now consider the SEM $\tilde{S}$ after $do(X_k = \tilde{N}_k)$ with $\tilde{N}_k \sim \tilde{p}(X_k)$,
$$\mathbb{P}_{S,do(X_k=\tilde{N}_k)}(X) = \tilde{p}(X_k) \prod_{j \neq k} \mathbb{P}(X_j | X_{PA_j})$$

Do operation

- Consider an SEM S,

$$X_j = f_j(PA_j, N_j)$$

- Due to the Markov property, we have

$$\mathbb{P}_S(X) = \prod_{i=1}^{p} \mathbb{P}(X_j | X_{PA_j})$$

- Now consider the SEM $\tilde{S}$ after $do(X_k = \tilde{N}_k)$ with $\tilde{N}_k \sim \tilde{p}(X_k)$,

$$\mathbb{P}_{S,do(X_k=\tilde{N}_k)}(X) = \tilde{p}(X_k) \prod_{j \neq k} \mathbb{P}(X_j | X_{PA_j})$$

- Perfect intervention: $do(X_k = a)$

$$\mathbb{P}_{S,do(X_k=a)}(X) = \begin{cases} \prod_{j \neq k} \mathbb{P}(X_j | X_{PA_j}) & X_k = a \\ 0 & \text{Otherwise} \end{cases} \tag{1}$$

- **Example:** Consider the Simpson's paradox in which all variables are binary.
- T: Treatment, Z: size of stone, R: recovery

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

$$\begin{tikzcd} & Z \arrow[dl] \arrow[dr] & \\ T \arrow[rr] & & R \end{tikzcd}$$

- We are interested in

$$\mathbb{P}_S(R = 1 | do(T = A))$$
$$\mathbb{P}_S(R = 1 | do(T = B))$$

- We have

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

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

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

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

- 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

$$
\begin{aligned}
\mathbb{P}_S(R=1|do(T=A)) &= \sum_{z=0}^{1} \mathbb{P}_{S,do(T=A)}(R=1, Z=z, T=A) \\
&= \sum_{z=0}^{1} \mathbb{P}_{S,do(T=A)}(R=1|Z=z, T=A)\mathbb{P}_{S,do(T=A)}(Z=z, T=A) \\
&= \sum_{z=0}^{1} \mathbb{P}_S(R=1|T=A, Z=z)\mathbb{P}_S(Z=z|do(T=A)) \\
&= \sum_{z=0}^{1} \mathbb{P}_S(R=1|T=A, Z=z)\mathbb{P}_S(Z=z)
\end{aligned}
$$

- The last step uses the perfect intervention formula in (1)
- Using the values in the table, we obtain

$$
\begin{aligned}
\mathbb{P}_S(R=1|do(T=A)) &\approx 0.93.\frac{357}{700} + 0.73.\frac{343}{700} = 0.832 \\
\mathbb{P}_S(R=1|do(T=B)) &\approx 0.87.\frac{357}{700} + 0.69.\frac{343}{700} = 0.782
\end{aligned}
$$

- But $\mathbb{P}_S(R=1|T=A) = 0.78$,

Instrumental variable

- A major complication is the possibility of inconsistent parameter estimation due to the existence of hidden confounders.

Instrumental variable

- 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.

Instrumental variable

- 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.

Instrumental variable

- 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.

$$\begin{tikzcd} & \color{red}{A} \arrow[dl, "\delta"'] \arrow[dr, "\gamma"] & \\ S \arrow[rr, "\rho"'] & & Y \end{tikzcd}$$

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

Instrumental variable

- 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.

$$\begin{tikzcd} & A \arrow[dl, "\delta"'] \arrow[dr, "\gamma"] & \\ S \arrow[rr, "\rho"'] & & Y \end{tikzcd}$$

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

- Interested in finding the influence of S on Y, i.e., finding ρ.
- No unbiased estimator exists for ρ:
$\widehat{\rho} := \frac{Cov(Y,S)}{Var(S)} == \frac{Cov(Y=\alpha+\rho S+\gamma A+N_Y,S)}{Var(S)} = \rho + \gamma \frac{Cov(A,S)}{Var(S)}$.

- Now, consider the following graph in which
  - A is a hidden confounder.
  - Z is a variable such that

$$Cov(Z, S) \neq 0 (\textit{First stage restriction criterion})$$

$$Cov(Z, \gamma A + N_Y) = 0 (\textit{Exclusion Restriction}).$$

$$\begin{tikzcd}
 & & \color{red}{A} \arrow[dl, "\delta"'] \arrow[dr, "\gamma"] & \\
Z \arrow[r, "\beta"'] & S \arrow[rr, "\rho"'] & & Y
\end{tikzcd}$$

- Now, consider the following graph in which
  - A is a hidden confounder.
  - Z is a variable such that

$$Cov(Z, S) \neq 0 (\textit{First stage restriction criterion})$$

$$Cov(Z, \gamma A + N_Y) = 0 (\textit{Exclusion Restriction}).$$

$$\begin{tikzcd}
 & & \color{red}{A} \arrow[dl, "\delta"'] \arrow[dr, "\gamma"] & \\
Z \arrow[r, "\beta"'] & S \arrow[rr, "\rho"'] & & Y
\end{tikzcd}$$

- Interested in finding ρ.

- Now, consider the following graph in which
  - A is a hidden confounder.
  - Z is a variable such that

$$Cov(Z,S) \neq 0 (\textit{First stage restriction criterion})$$

$$Cov(Z, \gamma A + N_Y) = 0 (\textit{Exclusion Restriction}).$$

$$\begin{tikzcd} & & \color{red}{A} \arrow[dl, "\delta"'] \arrow[dr, "\gamma"] & \\ Z \arrow[r, "\beta"'] & S \arrow[rr, "\rho"'] & & Y \end{tikzcd}$$

- Interested in finding ρ.
- In this model, variable Z (also called *Instrumental Variable*) can help to estimate ρ.

- 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 $\rho\beta$: $\widehat{\rho\beta} := \frac{Cov(Y,Z)}{Var(Z)}$.

- In this model, we have

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

- First, estimate $\rho\beta$: $\widehat{\rho\beta} := \frac{Cov(Y,Z)}{Var(Z)}$.
- Second, estimate β: $\widehat{\beta} := \frac{Cov(S,Z)}{Var(Z)}$.

- In this model, we have

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

- First, estimate $\rho\beta$: $\widehat{\rho\beta} := \frac{Cov(Y,Z)}{Var(Z)}$.
- Second, estimate β: $\widehat{\beta} := \frac{Cov(S,Z)}{Var(Z)}$.
- Finally, estimate ρ: $\widehat{\rho} := \frac{\widehat{\rho\beta}}{\widehat{\beta}} = \frac{Cov(Y,Z)}{Cov(S,Z)}$.
- This is an unbiased estimator because

$$\widehat{\rho} = \frac{Cov(Y,Z)}{Cov(S,Z)} = \frac{Cov(\alpha + \rho S + \gamma A + N_Y, Z)}{Cov(S,Z)} = \rho + \frac{Cov(\gamma A + N_Y, Z)}{Cov(S,Z)}.$$

- In this model, we have

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

- First, estimate $\rho\beta$: $\widehat{\rho\beta} := \frac{Cov(Y,Z)}{Var(Z)}$.
- Second, estimate β: $\widehat{\beta} := \frac{Cov(S,Z)}{Var(Z)}$.
- Finally, estimate ρ: $\widehat{\rho} := \frac{\widehat{\rho\beta}}{\widehat{\beta}} = \frac{Cov(Y,Z)}{Cov(S,Z)}$.
- This is an unbiased estimator because

$$\widehat{\rho} = \frac{Cov(Y,Z)}{Cov(S,Z)} = \frac{Cov(\alpha + \rho S + \gamma A + N_Y, Z)}{Cov(S,Z)} = \rho + \frac{Cov(\gamma A + N_Y, Z)}{Cov(S,Z)}.$$

- According to the restrictions:

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

- Therefore, $\widehat{\rho} = \rho$.

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.

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

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

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.

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

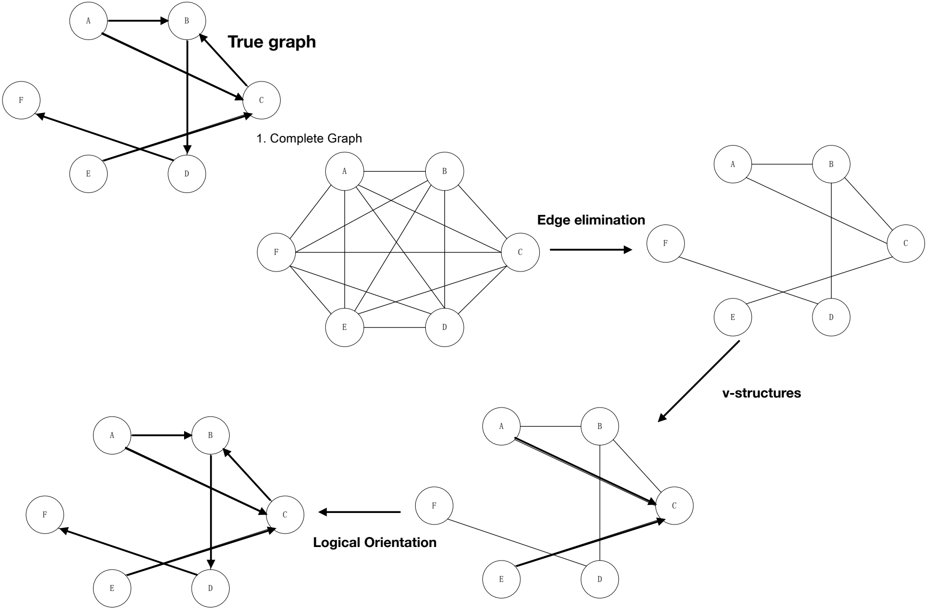
- Two nodes X, Y in a DAG (V, E) are adjacent iff they cannot be d-separated by any subset $S \subseteq V \setminus \{X, Y\}$.
- If two nodes X, Y in a DAG (V, E) are not adjacent, then they are d-separated by either PA_X or PA_Y.

- 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 $X - Y - Z$ with no direct edge between X and Z.
     - Let S denotes the corresponding d-separation set for X and Z.
     - The structure $X - Y - Z$ 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 $A \rightarrow B - C$, then, we can orient the 2nd arrow from B to C.
     - Avoid cycle.
     - ...

- **Example:**


A
B
C
D
E
F
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.
  - DBA: A d-sep. D by B, Edge DA eliminated.
  - DBC: A d-sep. C by B, Edge DC 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) $A \perp\!\!\!\perp E,\ A \not\perp\!\!\!\perp C,\ E \not\perp\!\!\!\perp C, A \not\perp\!\!\!\perp E | C \Rightarrow$ 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