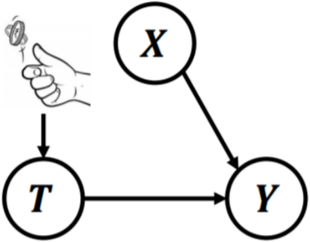
Beyond Neyman-Rubin model

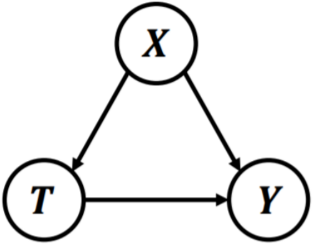
Credit : J.Josse, B. Neal, S. Wager, J. Peters

Refresher

We have considered two settings

- RCTs [$T \perp\!\!\!\perp (Y(0), Y(1))$]
- Neyman Rubin model [$T \perp\!\!\!\perp (Y(0), Y(1)) | X$]





Refresher

Under these assumptions, we have addressed several questions

- Estimate the Average Treatment Effect (ATE)

$$ATE := \mathbb{E}[Y_i(1) - Y_i(0)]$$

- Taking into account Heteregeneous Treatment Effects and estimate the Conditional Average Treatment Effect (CATE)

$$CATE(x) := \mathbb{E}[Y_i(1) - Y_i(0) | X = x]$$

Refresher

Underlying assumptions in Course 1-Course 3

- The graph is assumed to be known
- It involves only three variables

Some natural questions

- What about more complex situations?
  Some answers in Course 4
- How can we learn the graph?
  Some answers in Course 5

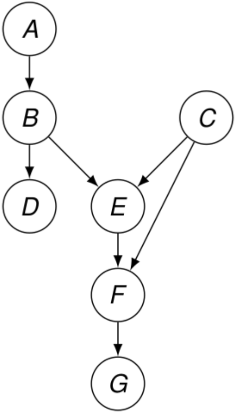
Refresher

Roadmap

- Graph terminology
- Basic building blocks
- Why should we care?
- Identifying the causal effect : some criteria

Vocabulary

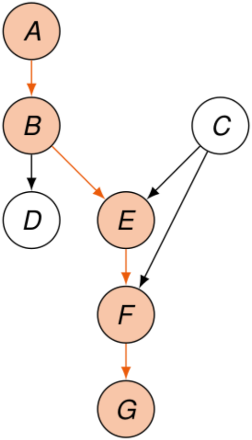
Let us consider the following graph $G = (V, E)$



Topological ordering: for any edge $X_i \rightarrow X_j$, $i < j$

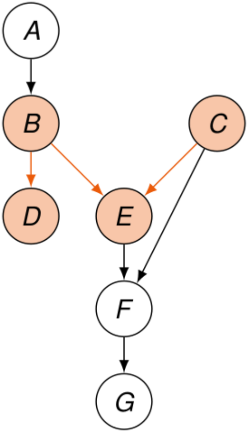
Vocabulary

Directed path: $A \rightarrow B \rightarrow E \rightarrow F \rightarrow G$



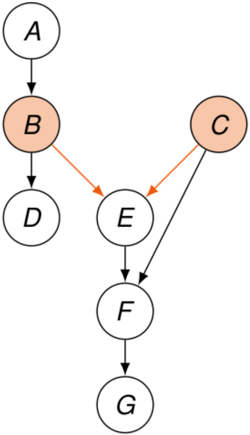
Vocabulary

Path : $D \leftarrow B \rightarrow E \leftarrow C$



Vocabulary

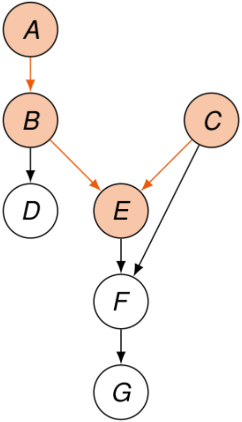
Parents, ancestors: $Pa(E) = \{B, C\}$



A node without parents is called a source node

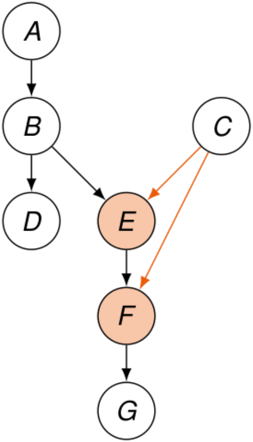
Vocabulary

Parents, ancestors: $Pa(E) = \{B, C\}$, $An(E) = \{A, B, C, E\}$



Vocabulary

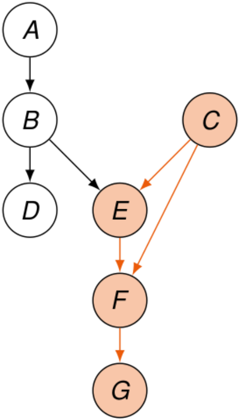
Children, descendants: $Ch(C) = \{E, F\}$



A node without children is called a sink node

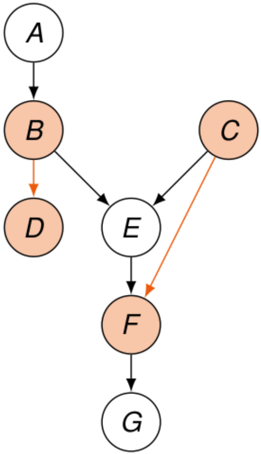
Vocabulary

Children, descendants: $Ch(C) = \{E, F\}$, $De(C) = \{C, E, F, G\}$



Vocabulary

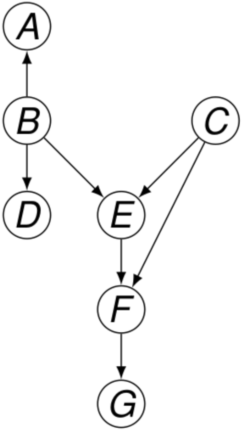
Induced subgraph $G[S]$: $G[\{B, C, D, F\}]$

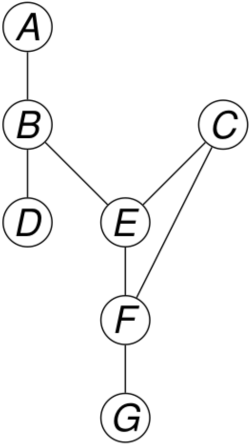


Vocabulary

- A graph G is called a partially directed acyclic graph (PDAG) if there is no directed cycle, that is, if there is no pair (j, k) with directed paths from j to k and from k to j.
- G is called a directed acyclic graph (DAG) if it is a PDAG and all edges are directed.

Vocabulary





a DAG and its corresponding skeleton

Local Markov Property

- Aim : find a relationship between the topology of a DAG and the properties of the corresponding joint distribution of the variables of the graph
- A first tool : the local Markov property

Local Markov Property

- Local Markov property with respect to the DAG G if each variable is independent of its non-descendants given its parents
- Markov factorization property with respect to the DAG G if

$$p(x) = p(x_1, \cdots, x_d) = \prod_j p(x_j | pa_j^G).$$

For this last property, need to assume that P_X has a density p.

Markov Property

As long as the joint distribution has a density, these two definitions are equivalent.

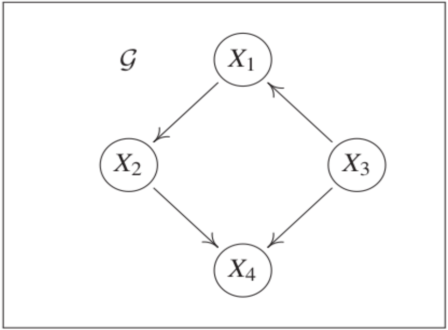
Equivalence of Markov properties

If P_X has a density p, then the two Markov properties in the previous definition are equivalent.

Local Markov Property

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(X_2, X_3, N_4)$$

- $N_1, \ldots, N_4$ jointly independent
- $\mathcal{G}$ is acyclic



Extracted from Peters et al. Chapter 6

Is the distribution $\mathbb{P}_{X_1, X_2, X_3, X_4}$ is Markovian with respect to the graph G?

Local Markov Property

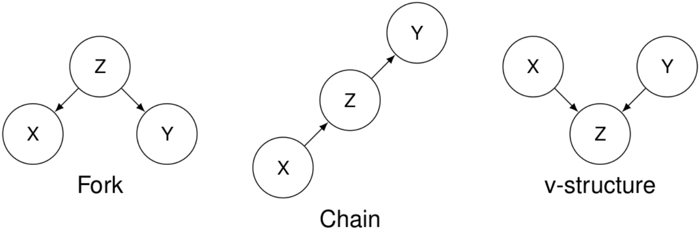
- According to (i) or (ii) of the the theorem, it is the case if

$$X_2 \perp\!\!\!\perp X_3 | X_1 \text{ and } X_1 \perp\!\!\!\perp X_4 | X_2, X_3$$

- If we use (iii), it is the case if

$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1|x_3)p(x_2|x_1)p(x_4|x_2, x_3)$$

Basic building blocks



Fork, chains and v-structures

Basic building blocks

Fork

In whole generality $p(X_1, X_2, X_3) = p(X_1|X_2, X_3)p(X_3|X_2)p(X_2)$.
Observe now that by graphical properties

$$p(X_1|X_2, X_3) = p(X_1|X_2)$$

Combining these two equations yields the required result.

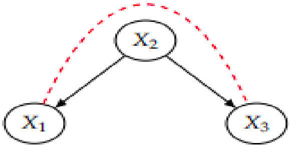


Figure 3.13: Fork with flow of association drawn as a dashed red arc.

Extracted from N. Bradley, Chapter 3

Basic building blocks

Chain

Let us now consider the case of the chain

$$p(X_1, X_2, X_3) = p(X_3|X_1, X_2)p(X_1|X_2)p(X_2)$$

which simplifies since $p(X_3|X_1, X_2) = p(X_3|X_2)$.

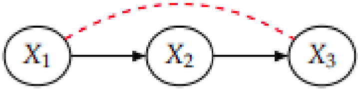


Figure 3.12: Chain with flow of association drawn as a dashed red arc.

Extracted from N. Bradley, Chapter 3

Basic building blocks

- Chains and forks share the same set of independencies.
- When we condition on X_2 in both graphs, it blocks the flow of association from X_1 to X_3.
- We now consider another building block : v-structures or immoralities

Basic building blocks

v-structure

- In the case of *v* structures, one has $X_1 \perp\!\!\!\perp X_3$
- But X_1 and X_3 are not independent w.r.t. X_2

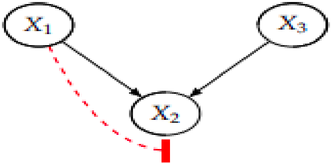


Figure 3.16: Immorality with association blocked by a collider.

Extracted from N. Bradley, Chapter 3

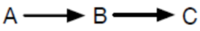
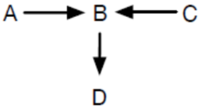
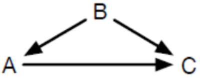
Basic building blocks

v-structure

Proof of independency of X_1 and X_3

$$
\begin{aligned}
p(x_1, x_3) &= \sum_{x_2} p(x_1, x_2, x_3) \\
&= \sum_{x_2} p(x_2 | x_1, x_3) p(x_1 | x_3) p(x_3) \\
&= \sum_{x_2} p(x_2 | x_1, x_3) p(x_1) p(x_3) \\
&= p(x_1) p(x_3) \left[\sum_{x_2} p(x_2 | x_1, x_3) \right] \\
&= p(x_1) p(x_3)
\end{aligned}
$$

Why should we care?

Graph:	Regression:	Implications:
	$\mathbb{E}[C\|\cancel{A},\boxed{B}]$	$A \perp\!\!\!\perp C\|B$
	$\mathbb{E}[C\|\cancel{A}]$	$A \perp\!\!\!\perp C$
	$\mathbb{E}[C\|\boxed{A},\boxed{D}]$	$A \not\perp\!\!\!\perp C\|D$
	$\mathbb{E}[C\|\boxed{A},\boxed{B}]$	$A \perp\!\!\!\perp B$

Extracted from https://arxiv.org/abs/2202.09875

Why should we care?

- In complex situations, we should revisit our concept of interpretability!
- Classical tools for interpretability of ML models : importance measures
- Some examples :
  - Coefficients of white box models as Linear Regresssion or Logistic Regression
  - Importance measure of RFs based on decreasing of Gini Impurity
  - Permutation importance
  - Shapley indices

For a review :

https://christophm.github.io/interpretable-ml-book/

Why should we care?

A motivating example

See notebook of **Lecture 4**

The concept of d separation

Blocked paths

A path is said to be blocked by a set of vertices Z if:

- it contains a chain $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ and $B \in Z$, or
- it contains a collider $A \rightarrow B \leftarrow C$ such that no descendant of B is in Z

The concept of d separation

Definition

Two (sets of) nodes X and Y are d-separated by a set of nodes Z if all of the paths between (any node in) X and (any node in) Y are blocked by Z. We denote $X \perp\!\!\!\perp_G Y|Z$

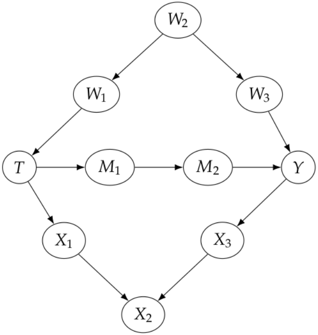
Theorem

Assume that $\mathbb{P}$ satisfies the local Markov assumption with respct to G. If X and Y are d-separated in G conditioned on Z , then X and Y are independent in $\mathbb{P}$ conditioned on Z. We can write this succinctly as follows:

$$X \perp\!\!\!\perp_G Y|Z \Rightarrow X \perp\!\!\!\perp_{\mathbb{P}} Y|Z$$

which is named the Global Markov property

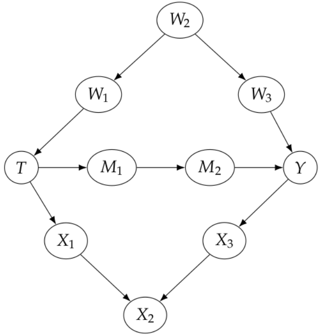
The concept of *d* separation



Are T and Y d separated by

- the empty set?
- $\{W_2\}$?
- $\{W_2, M_1\}$?
- $\{W_1, M_2\}$?
- $\{W_1, M_2, X_2\}$?

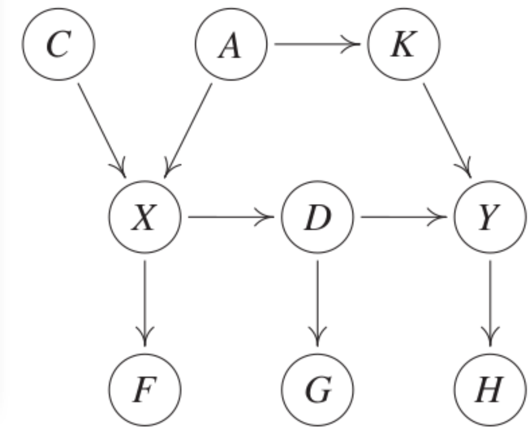
The concept of d separation



Are T and Y d separated by

- the empty set?
- $\{W\}$?
- $\{W, X_2\}$?

The concept of *d* separation



For this DAG : $C \perp\!\!\!\perp_G G | \{X\}$ and $C \not\perp\!\!\!\perp_G G | \{X, H\}$

Backdoor criterion

The backdoor criterion

Consider a causal graph G and a causal effect $P(y|do(t))$. A set of variables Z satisfies the backdoor criterion iff:

- no node in Z is a descendant of T
- Z blocks every path between T and Y that contains an arrow into T.

Backdoor criterion

Theorem (back-door adjustment)

If Z satisfies the backdoor criterion relative to (T, Y) and if $Pr(t, z) > 0$, then the causal effect of T on Y is identifiable and is given by

$$\mathbb{P}[y|do(t)] = \sum_{z} \mathbb{P}[y|t, z]\mathbb{P}(z)$$

Backdoor criterion

Consequence of the backdoor criterion

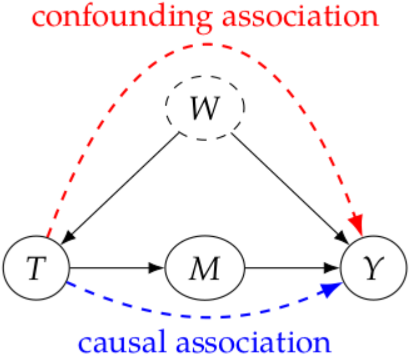
$$\begin{aligned}\mathbb{E}[Y|do(t)] &= \sum_{z} \mathbb{E}[Y|t,z]\mathbb{P}(z) \\ &= \mathbb{E}_Z[\mathbb{E}[Y|t,Z]]\end{aligned}$$

Hence we can identify the ACE

$$\mathbb{E}[Y(1) - \mathbb{E}[Y(0)] = \mathbb{E}_Z[\mathbb{E}[Y|T=1,Z]] - \mathbb{E}_Z[\mathbb{E}[Y|T=0,Z]]$$

Frontdoor criterion

In some cases whre we have latent variables, the backdoor criterion is not possible to apply



Extracted from B. Neal

Frontdoor criterion

In this example, one should use another criterion : the frontdoor criterion. One has

- $\mathbb{P}[m|do(t)] = \mathbb{P}[m|t]$
- $\mathbb{P}[y|do(m)] = \sum_t \mathbb{P}(y|m,t)\mathbb{P}(t)$ [T blocks the backdoor]
- $\mathbb{P}[y|do(t)] = \sum_m \mathbb{P}[y|do(m)]\mathbb{P}[m|do(t)]$

Hence

$$\mathbb{P}[y|do(t)] = \sum_m \mathbb{P}[m|t]\left[\sum_{t'} \mathbb{P}(y|m,t')\mathbb{P}(t')\right]$$

Frontdoor criterion

Frontdoor criterion

Consider a causal graph G and a causal effect $P(ydo(t))$. A set of variables M satisfies the frontdoor criterion iff:

- M intercepts all directed paths from T to Y ;
- There is no backdoor path from T to Y
- All backdoor paths from M to Y are blocked by T.

Frontdoor criterion

Theorem (frontdoor adjustment)

if M satisfies the frontdoor criterion relative to (T, Y) and if $Pr(t, m) > 0$, then the causal effect of T on Y is identifiable and is given by

$$\mathbb{P}[y|do(t)] = \sum_{m} \mathbb{P}[m|t] \left[\sum_{t'} \mathbb{P}(y|m, t')\mathbb{P}(t') \right]$$

Beyond backdoor and frontdoor

Beyond backdoor and frontdoor

- Other methods : use of instrumental variables [see Chapter 9 of B. Neal]
- More general : use of the three rules of Do Calculus edicted by Pearl

A case Study: Hotel Booking Cancellations

`Example extracted from the DoWhy library`

- Scientific question : estimate the impact of assigning a different room as compared to what the customer had reserved on Booking Cancellations.
- Gold standard : use experiments such as Randomized Controlled Trials wherein each customer is randomly assigned to one of the two categories i.e. each customer is either assigned a different room or the same room as he had booked before
- What if we cannot intervene or its too costly too perform such an experiment?
- Can we somehow answer our query using only observational data or data that has been collected in the past?

We shall answer these questions using causality and the `dowhy` library!

A case Study: Hotel Booking Cancellations

Description of the Dataset

- Booking information for a city hotel and a resort hotel taken from a real hotel in Portugal
- Includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things
- All personally identifying information has been removed from the data

Reference of the study case[1]

[1] https://www.sciencedirect.com/science/article/pii/S23523409183