Ch 5: Forks, Chains, and Immoralities

Prashant Shekhar, PhD
Assistant Professor of Data Science
Department of Mathematics
Embry-Riddle Aeronautical University, FL, USA
Email: shekharp@erau.edu

Outline

Introduction

Introduction

- Here, we focus on the mappings between the statistical and graphical properties of a system. In a perfect world, we'd like to be able to do it in both directions: from graph independence to statistical independence and the other way around.
- It turns out that this is possible under certain assumptions. The key concept in this chapter is one of independence.

Independence

- We say that two variables, X and Y, are **independent** ($X \perp\!\!\!\perp Y$) when our knowledge about X does not change our knowledge about Y (and vice versa)

$$P(Y) = P(Y|X)$$

in other words

$$P(X, Y) = P(X)P(Y)$$

- We say that X and Y are **conditionally independent** given Z ($X \perp\!\!\!\perp Y|Z$), when X does not give us any new information about Y assuming that we observed Z. It is represented as

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- New notations:
  - Independence in the distribution: $X \perp\!\!\!\perp_P Y|Z$
  - Independence in the graph: $X \perp\!\!\!\perp_G Y|Z$

Independence in a graph (a working definition)

- Two nodes are **unconditionally (or marginally) independent** in the graph when there's no open path that connects them directly or indirectly.
- Two nodes, X and Y, are **conditionally independent** given (a set of) node(s) Z when Z blocks all open paths that connect X and Y.

Conditions for Causal Inference

- Local Markov Property
- Global Markov Property

Conditions for Causal Inference: local Markov property

- From the causal inference point of view, we need to make sure that we can map the graphical (conditional) independencies into statistical (conditional) independencies.
- In order to achieve this we need to satisfy the Local/Causal Markov condition.

Local/Causal Markov property

This condition states that the node, V_i, is independent of all its non-descendants (excluding its parents) given its parents. Therefore, formally, it can be presented as follows:

$$V_i \perp\!\!\!\perp_G V_j | PA(V_i) \tag{1}$$

$$\forall j \neq i \in G(V, E) \setminus \Big[DE(V_i) \cup PA(V_i) \Big]$$

- $G(V, E)$: graph G with nodes V and edges E
- $DE(V_i)$: descendents of node V_i
- $PA(V_i)$: parents of node V_i

Local Markov property: example1

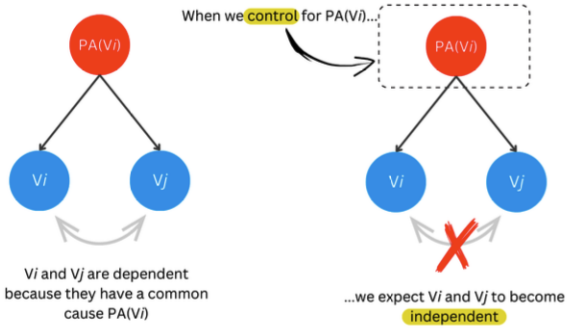


Figure: Conditional independence in a fork structure

Local Markov property: example2

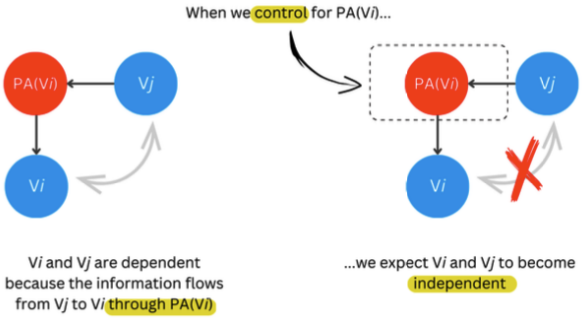


Figure: Conditional independence in a chain structure

Conditional for Causal inference: global Markov property

- Note that if there was an unobserved common cause of both nodes (V_i and V_j) in any of the scenarios, controlling for $PA(V_i)$ would not render the nodes independent, and the condition would be violated.

Global Markov property

This condition states that when the causal Markov condition holds, the following is true:

$$X \perp\!\!\!\perp_G Y|Z \implies X \perp\!\!\!\perp_P Y|Z \tag{2}$$

This means if X and Y are independent in the graph given Z, then they are also statistically independent given Z.

- We could show that the global Markov property, the local Markov property, and another property called the Markov factorization property are equivalent

Assumptions for Causal Discovery

- Faithfulness assumption
- Causal minimality condition
- No hidden confounding

Assumptions for Causal discovery

- So far, we have discussed the importance of mapping between graphical and distributional independence structures.
- Now, let's reverse the direction. Causal discovery aims at discovering (or learning) the true causal graph from observational and/or (partially) interventional data.
- This task is possible when certain conditions are met.

Assumption for Causal Discovery: faithfulness assumption

- Causal discovery has a couple of different flavors. In this section, we'll focus on a family of methods called **constraint-based causal discovery** (sometimes also called **independence-based causal discovery**).
- These methods are designed to find statistical independencies in the observational data and try to recreate the true causal graph from these independencies.
- One of the main assumptions behind this family of methods is called the faithfulness assumption.

Faithfulness assumption

Its simplest formulation is the following (converse of global Markov property):

$$X \perp\!\!\!\perp_P Y|Z \implies X \perp\!\!\!\perp_G Y|Z \tag{3}$$

The formula says that if X and Y are independent in the distribution given Z, they will also be independent in the graph given Z.

Problems with faithfulness assumption

- Faithfulness assumption might be difficult to fulfill sometimes.
- The most critical reason for this is the estimation error when testing for conditional independence in the finite sample size regime (falsely seem conditionally independent).
- Moreover, it's not very difficult to find situations where the assumption is violated

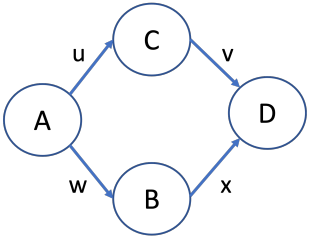


Figure: uu

With the structural equations

- $C := uA$
- $B := wA$
- $D := vC + xB$, i,e $D := (vu + xw)A$
- Hence Association of A and D cancels out at $vu = -xw$, i.e., $A \perp\!\!\!\perp D$
- However, that doesnt mean independence in graph as we can clearly see. Hence:

$$A \perp\!\!\!\perp_P D \not\Longrightarrow A \perp\!\!\!\perp_G D$$

- The probability of such situations in the real world is very small though.

Assumption for Causal Discovery: causal minimality condition

- There might be more than one graph that entails the same distribution! That's problematic when we want to recover causal structure (represented as a causal graph) because the mapping between the graph and the distribution is ambiguous.
- To address this issue, we use the causal minimality condition.

Causal Minimality assumption

This assumption states that DAG G is minimal to distribution, P, if and only if G induces P, but no proper sub-graph of G induces P. In other words, if graph G induces P, removing any edge from G should result in a distribution that is different than P.

- The assumption is usually perceived as a form of Ockham's razor (simplest solution is usually the correct solution), its implications have practical significance for constraint-based causal discovery methods and their ability to recover correct causal structures.

Assumptions for Causal Discovery

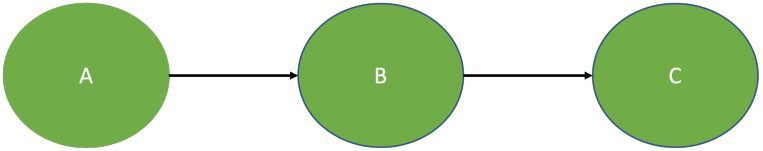
- The assumption of **no hidden confounding** (sometimes also referred to as **causal sufficiency**) is another very commonly assumption used in causal discovery and causal inference.
- Although meeting this assumption is not necessary for all causal methods, it's pretty common.
- Note that causal sufficiency and the causal Markov condition are related (and have some practical overlap), but they are not identical.
- In many real-world scenarios, we might find it challenging to verify whether hidden confounding exists in a system of interest.

Chains, forks and colliders/immoralities

Chain (of events)

- We have a DAG representing a collision warning system in an autonomous vehicle.
- A system like this consists of a detector (or detector module) and an alerting system (sometimes also an automatic driving assistance system).
- When there's an obstacle on the collision course detected by the detector, it sends a signal that activates the alerting system. **Let's say that the detector is in state 1 when it detects an obstacle and it's in state 0 when it detects no obstacles**.
- Please note that the existence of the obstacle does not give us any new information about the alert when we know the detector state.
- Hence, if detector is in state 1, the alert goes off irrespective of if there is an obstacle or not. **The obstacle and the alert are independent, given the detector state**.

Chains in DAGs

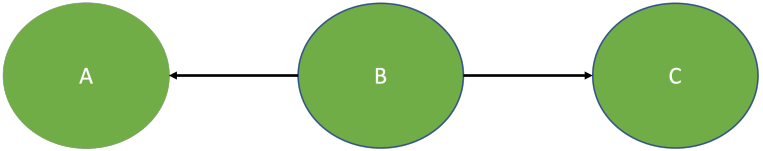


- We can generalize the independence property from our collision warning example to all chain structures.
- It means that in every structure that can be encoded as $A \rightarrow B \rightarrow C$, A and C are independent in the graph given B. Therefore, formally:

$$A \perp\!\!\!\perp_G C | B$$

- Intuitively, controlling for B closes the only open path that exists between A and C. Note that A and C become dependent when we do not control for B.

Forks



- A fork is a structure where the edge between nodes A and B is reversed compared to the chain structure.
- In the fork, node B becomes what we usually call a common cause of nodes A and C.
- **An example**: suppose you are driving a car and you see a person on the road (event B). On the car side this leads to emergency braking (event A), and on your side you get a adrenaline boost into your bloodstream (event C) due to the possibility of an accident. In this way, the presence of the person on the road is a joint cause of 2 separate events.

Forks: Independence structure

Consider the structural equations for the fork model:

- $U_A, U_B, U_C \sim N(0,1)$ *each*
- $B := U_B$
- $A := B + U_A$
- $C := B + U_C$

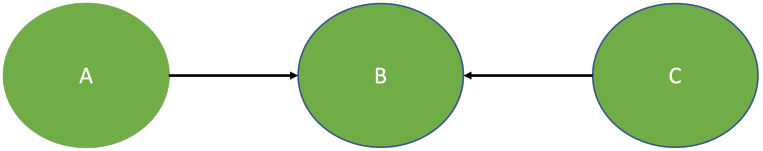
Now imagine we look at $B = 0$. Then A and C will only be influenced by their respective noise terms that are independent by definition. Hence:

- A and C are conditionally independent given B.

$$A \perp\!\!\!\perp_G C | B$$

- In case we do not control for B, A and C are dependent.
- **This independence pattern is identical to the one that we've obtained from the chain structure**.

Colliders/immoralities/v-structures



- The most important characteristic of a collider is that its independence pattern is reversed compared to chains and forks.
- **In colliders, A and C are unconditionally independent**.

$$A \perp\!\!\!\perp C$$

- If we control for B, A and C become dependent (opposite of chains and forks).
- Thanks to the unique properties of colliders, they can be immensely helpful when we're trying to recover graph structures from observational data.

Collider example 1: *LLama* $\rightarrow$ *Brakes* $\leftarrow$ *Sun*

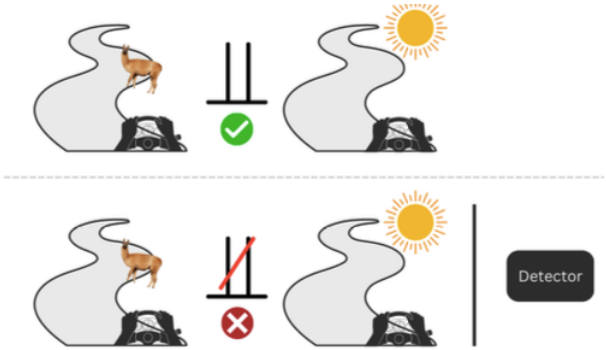


Figure: **Top panel**: Presence of llama on the road is independent of glare from the sun. **Bottom panel**: Given that detector applied the brakes, the probability of a llama becomes inversely correlated (i.e. dependent) with probability of glare from the sun.

Collider example 2

- Consider the case where both integers A and C are randomly generated between 1 and 3.
- Let's also say that B is a sum of A and C. Hence we have a collider situation of the form: $A \rightarrow B \leftarrow C$.
- If we dont condition on B (dont observe B) then A and C are certainly independent. i.e., $A \perp\!\!\!\perp C$
- However if we observed B, and it is 4 (for example), then possibilities of A and C are as follows:
  - A = 1; C = 3
  - A = 2; C = 2
  - A = 3; C = 1

  Here its clear, that after observing B, A and C become inversely correlated, i.e., dependent.

Colliders: ambiguous cases

- As we've seen earlier, various graphical configurations might lead to the same statistical independence structure (for example chains and forks).
- In some cases, we might get lucky and have enough colliders in the graph to make up for it. In reality, though, we might often not be that fortunate.
- However, even in cases where some edges cannot be oriented using constraint-based methods, we can introduce the following concept:

Markov equivalence Classes (MEC)

A set of DAGs, $\mathcal{D} = \{G_0(V, E_0), G_1(V, E_1), .., G_n(V, E_n)\}$ is *Markov equivalent* if and only if all DAGs in $\mathcal{D}$ have the same skeleton (undirected version of DAG) and the same set of colliders. If we add the edges for all the collider structures that we've found, we will obtain a **complete partially-directed acyclic graph (CPDAG)**.

Colliders: MEC example

- If we take the CPDAG and generate a set of all possible DAGs from it, we'll obtain a MEC.
- MECs can be pretty useful. Even if we cannot recover a full DAG, a MEC can significantly reduce our uncertainty about the causal structure for a given dataset.

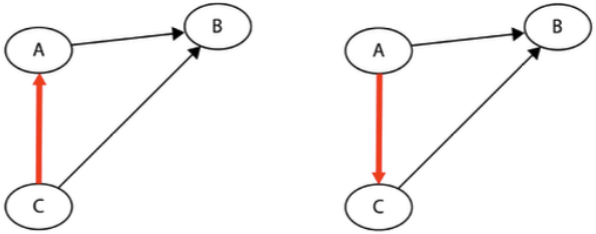
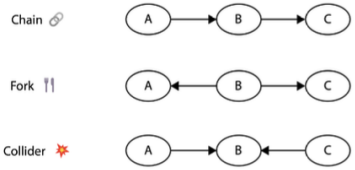


Figure: An example of a MEC. The graphs have the same set of edges. If we removed the arrows and left the edges undirected, we would obtain two identical graphs, which is an indicator that both graphs have the same skeleton. The collider ($A \rightarrow B \leftarrow C$) is present in both graphs. The only difference between the two graphs is the direction of the edge between nodes A and C. Thus, these graphs constitute a MEC.

Chains, forks, colliders, and regression

Introduction

- Here we will see how the properties of chains, forks, and colliders manifest themselves in regression analysis.
- The type of analysis that we'll conduct in this section is actually at the heart of some of the most classic methods of causal inference and causal discovery. Our plan is as follows:
  - We're going to generate three datasets, each with three variables, A, B, and C.
  - Each dataset will be based on a graph representing one of the three structures: a chain, a fork, or a collider.



  - Next, we'll fit one regression model per dataset, regressing C on the remaining two variables, and analyze the results.
  - We'll plot pairwise scatterplots for each dataset to strengthen our intuitive understanding of a link between graphical structures, statistical models, and visual data representations.

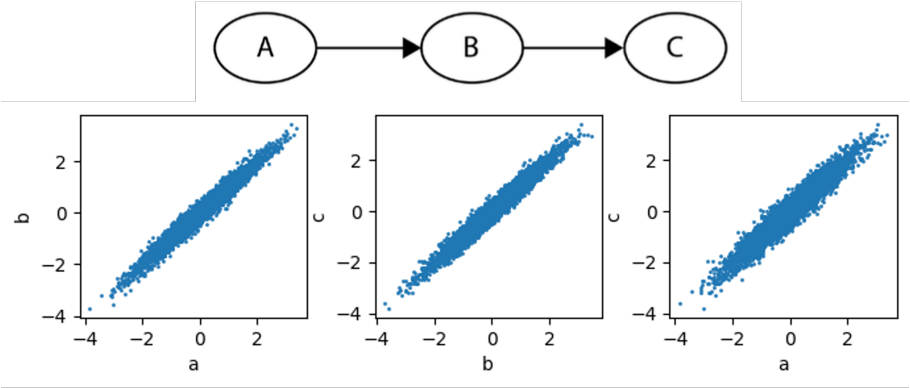
CODE

Refer to the [Python notebook] for data generation from:

- Chain based DAG
- Fork based DAG
- Collider based DAG

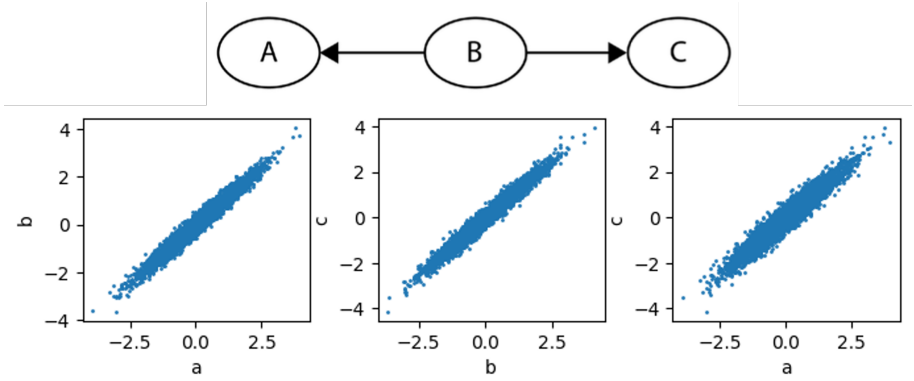
Additionally, the Jupyter notebook also provides regression analysis for predicting variable C when we control for A and B. The idea here is to note the different in behavior of the model for different data generating processes.

Chain based DAG: Fitting C = f(A,B)



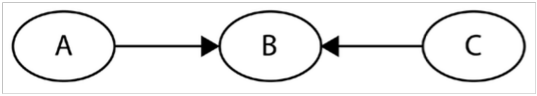
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0042	0.003	-1.452	0.147	-0.010	0.001
A	-0.0044	0.015	-0.302	0.762	-0.033	0.024
B	1.0053	0.014	69.751	0.000	0.977	1.034

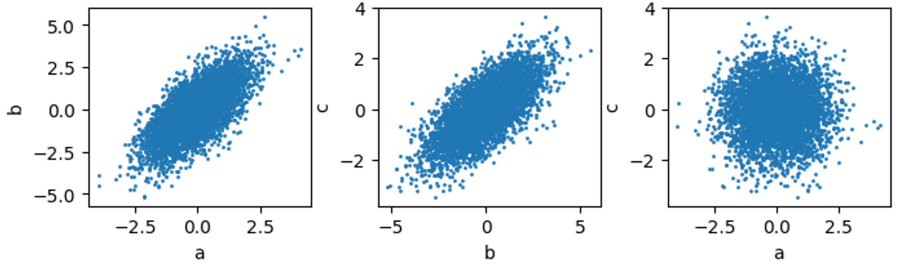
Fork based DAG: Fitting C = f(A,B)



	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0053	0.003	-1.904	0.057	-0.011	0.000
A	-0.0222	0.014	-1.578	0.115	-0.050	0.005
B	1.0287	0.014	72.095	0.000	1.001	1.057

Collider based DAG: Fitting C = f(A,B)





	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0028	0.003	-1.000	0.317	-0.008	0.003
A	-0.9621	0.004	-245.413	0.000	-0.970	-0.954
B	0.9617	0.003	350.750	0.000	0.956	0.967

Conclusion

- For Chain $A \rightarrow B \rightarrow C$:
  - In the scatterplot, every pair seems correlated (as expected)
  - However, when we control for B, based on the probability numbers, the coefficient of A becomes insignificant (as expected)
- For Fork $A \leftarrow B \rightarrow C$:
  - In the scatterplot, every pair seems correlated (as expected)
  - However, when we control for B, based on the probability numbers, the coefficient of A becomes insignificant (as expected)
- For Collider $A \rightarrow B \leftarrow C$:
  - In the scatterplot pair A, B are correlated and pair B, C are correlated. However A and C are independent (as expected)
  - However, when we control for B, based on the probability numbers, the coefficients of both A and B are significant. This is because when B is given, A and C become dependent (as expected).