36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 31: November 18

Lecturer: Siva Balakrishnan

31.1 Review and Outline

In the last class we started discussing directed graphical models:

1. Parameter counting and independence
2. Factorization
3. Local Markov property

Today, we will first re-visit some things that were wrong in the previous lecture and then discuss d-separation.

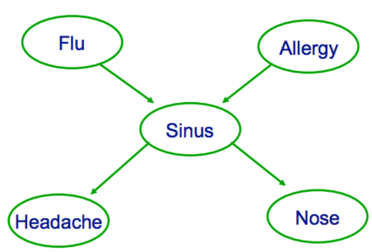
31.2 Marginal independence

The first thing we need to fix from last lecture is the rule for marginal dependence.

We first need to define an "unblocked path". An unblocked path is one that can be traced without traversing a pair of arrows that collide "head-to-head".

Now, the rule for marginal dependence is simple: two random variables are marginally independent if there is no unblocked path between them.

Going back to the graph from last class:



We can now verify that the only marginal independence is that $F \perp\!\!\!\perp A$, as there are unblocked paths between every other pair of variables.

31.3 Conditioning on colliders

The other thing that I did not really give a clean example of was the phenomenon that conditioning on a collider makes RVs dependent.

1. **Example 1:** A simple example is the following. Suppose that (like in Pittsburgh), being sunny and being warm are independent, i.e. knowing that it is sunny does not mean it is more likely to be warm for instance.

   Suppose I now denote the random variable X which is 1 if it is not sunny, and Y which is 1 if it is cold. Now, I can define a common effect Z which is 1 if it snows, which in turn happens whenever it is cold and not sunny, i.e. $Z = X$ and Y.

   X and Y are marginally independent. Now, lets see what happens if I condition on $Z = 0$, then knowing that $X = 1$ for instance tells us that $Y = 0$, i.e. we can verify that:

   $$P(Y = 0|X = 1, Z = 0) = 1 \neq P(Y = 0|Z = 0) = 2/3.$$

   So we can conclude that $X \not\perp\!\!\!\perp Y|Z$.

2. **Explaining away:** Explaining away is the phenomenon that if an effect has two independent causes, then conditioning on the effect makes the causes dependent. Additionally, it can make these causes negatively associated, i.e. knowing that one of the causes happens can make it less likely that the other cause also happened.

   Again lets see a simple example: This is an abstract version of the example we considered last time. Roughly, Z is whether the grass is wet, X is whether a fire engine came by and Y is whether it rained. The common effect $Z = X$ or Y, and again the causes X and Y are independently, Ber(1/2).

   In this case, we have:

X	Y	Z	P(X,Y,Z)
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	1	1/4

   Now, it is easy to verify that:

   $$P(X = 1|Y = 1, Z = 1) = 1/4 < P(X = 1|Z = 1) = 2/3.$$

   In words, knowing that one of the causes (Y) is true makes it less likely that the other cause is true, i.e. one cause explains away the effect. So if we know that the grass is wet and that it just rained we would be less inclined to think that a fire truck also came by.

31.4 d-separation

In the last class we discussed the basic conditional independence properties that are implied by the graph structure. These are given by the local Markov property which gives us that:

$$X \perp\!\!\!\perp \{\text{non-descendants of } X\} |\ \text{parents of } X.$$

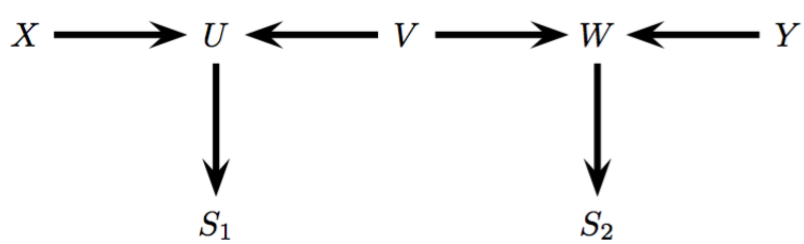
However, it turns out that there are many other independence conditions one can derive from this basic set.

Let X and Y be distinct vertices and let W be a set of vertices not containing X or Y. Then X and Y are d-separated given W if there exists no undirected path between X and Y such that

1. every collider on the path has a descendant in W, and
2. no other vertex on the path is in W.

The intuition is simple: if you have a regular path then conditioning breaks the path, if you have colliders then conditioning opens the path. Furthermore, if you condition on a descendant of a collider you open the path. Now, if there are no open paths then the conditional independence relationship holds.

The following example illustrates this: consider the following graph:

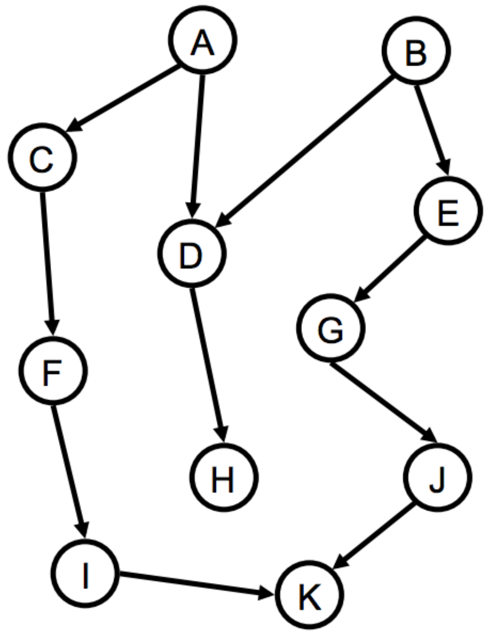


The following statements follow from the d-separation rules:

1. X and Y are d-separated (given the empty set);
2. X and Y are d-connected given $\{S_1, S_2\}$;
3. X and Y are d-separated given $\{S_1, S_2, V\}$.

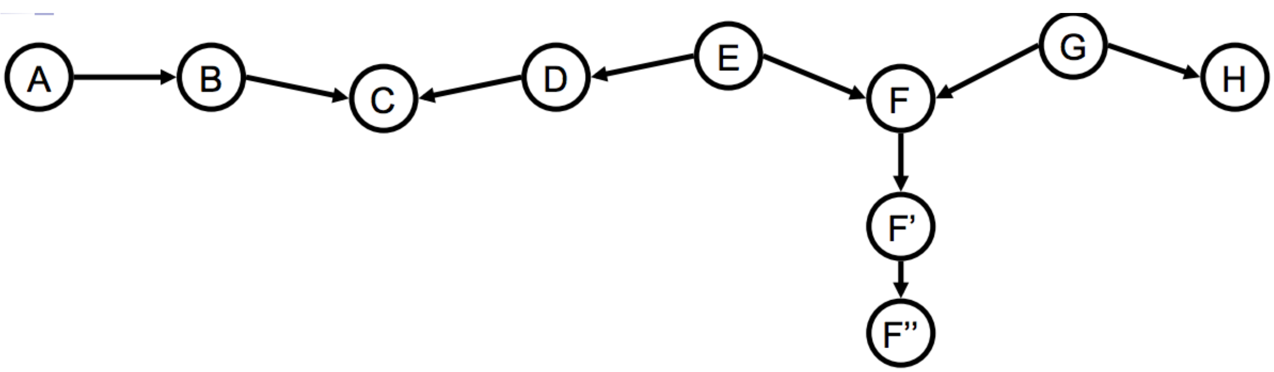
Lets look at another example:

Which of the following are correct?



1. $A \perp\!\!\!\perp B$
2. $E \perp\!\!\!\perp F | K$
3. $E \perp\!\!\!\perp F | K, I$

And one more:



In this Bayes net, when are A and H independent?

31.5 Directed graphs: advanced topics

So far, what we have seen is that directed graphical models can be a useful way to compactly represent a joint probability distribution, and in particular the graph implies a certain factorization of the joint distribution. Furthermore, it can be used to graphically determine various independence and conditional independences that are much less obvious to determine from the joint distribution directly.

A graphical model (say the RVs are all discrete) is completely specified by the conditional probabilities $P(X|\text{parents}(X))$.

There are many questions that of a more statistical nature that we will not have time to cover in this class:

1. **Inference:** Given a fully specified directed graphical model, answer questions of the form: "What is the probability that $X = 1|Y = 0, Z = 1$?"

   Inference in a graphical model is basically asking conditional/marginal probability queries.

   Answering these questions uses the graphical model structure in an intricate way, and the class of algorithms are known as message-passing algorithms.

2. **Estimation:** Estimation is the question: given data $X_1, \ldots, X_n$, where each $X_i \in \mathbb{R}^d$ and a graph G, estimate the conditional probabilities $P(X|\text{parents}(X))$.

3. **Graph-structure estimation:** In a lot of applications, we are just given a collection of data, and we want to estimate the structure of the data that underlies the RVs that we measure.

   The brute force way, tests every possible conditional independence assumption, and then creates the graph that respects the valid conditional independence assumptions. This is both computationally and statistically impossible to do for large graphs, so more often many different heuristics are used to determine conditional independence relationships, and the graph structure.