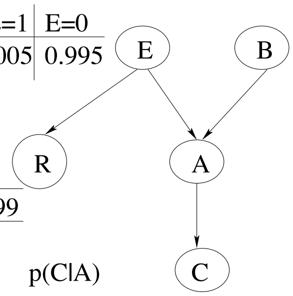
Lecture 3: Conditional independence and graph structure

- Independence maps (I-maps)
- Factorization theorem
- The Bayes ball algorithm and d-separation

Recall from last time

- Bayesian networks are a graphical model representing conditional independence relations
- The nodes of the graphs represent r.v.'s
- Each node has associated with it a conditional probability distribution (CPD) for the corresponding r.v., given its parents
- The joint probability distribution can be computed by multiplying the local CPDs

Example: A Bayesian (belief) network



p(E)

E=1	E=0
0.005	0.995

p(B)

B=1	B=0
0.01	0.99

p(R|E)

	R=1	R=0
E=0	0.0001	0.9999
E=1	0.65	0.35

p(A|B,E)

	A=1	A=0
B=0,E=0	0.001	0.999
B=0,E=1	0.3	0.7
B=1,E=0	0.8	0.2
B=1,E=1	0.95	0.05

p(C|A)

	C=1	C=0
A=0	0.05	0.95
A=1	0.7	0.3

- The nodes represent random variables
- The arcs represent "influences"
- At each node, we have a conditional probability distribution (CPD) for the corresponding variable *given its parents*

Recall: I-Maps

A directed acyclic graph (DAG) G whose nodes represent random variables $X_1, \ldots, X_n$ is an **I-map (independence map)** of a distribution p if p satisfies the independence assumptions:

$$X_i \perp\!\!\!\perp \text{Nondescendents}(X_i) | X_{\pi_i}, \forall i = 1, \ldots n$$

where X_{π_i} are the parents of X_i

Example

Consider all possible DAG structures over 2 variables. Which graph is an I-map for the following distribution?

x	y	$p(x, y)$
0	0	0.08
0	1	0.32
1	0	0.32
1	1	0.28

What about the following distribution?

x	y	$p(x, y)$
0	0	0.08
0	1	0.12
1	0	0.32
1	1	0.48

Factorization theorem

G is an I-map of p if and only if p factorizes according to G:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}), \forall x_i \in \Omega_{X_i}$$

Proof: $\Longrightarrow$
Assume that G is an I-map for p. By the chain rule, $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$. Without loss of generality, we can order the variables x_i according to G. From this assumption, $X_{\pi_i} \subseteq \{X_1, \ldots, X_{i-1}\}$. This means that $\{X_1, \ldots, X_{i-1}\} = X_{\pi_i} \cup Z$, where $Z \subseteq$ Nondescendents(X_i). Since G is an I-map, we have $X_i \perp\!\!\!\perp$Nondescendents$(X_i) | X_{\pi_i}$, so:

$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | z, x_{\pi_i}) = p(x_i | x_{\pi_i})$$

and the conclusion follows.

Factorization theorem (2)

Proof: $\Longleftarrow$

Assume the p factorizes over G. Let X_{D_i} denote the descendents of X_i and X_{N_i} denote nondescendents. Note that $\{X_1 \ldots X_n\} = \{X_i\} \cup X_{\pi_i} \cup X_{D_i} \cup X_{N_i}$. We have:

$$p(x_i|x_{\pi_i}, x_{N_i}) = \frac{p(x_i, x_{\pi_i}, x_{N_i})}{\sum_{x_i \in \Omega_{X_i}} p(x_i, x_{\pi_i}, x_{N_i})}$$

We compute the numerator:

$$\begin{aligned} p(x_i, x_{\pi_i}, x_{N_i}) &= \sum_{x_{D_i}} p(x_i, x_{\pi_i}, x_{N_i}, x_{D_i}) = \sum_{x_{D_i}} \prod_{j=1}^{n} p(x_j|x_{\pi_j}) \\ &= p(x_i|x_{\pi_i}) \prod_{x_j \in x_{N_i}} p(x_j|x_{\pi_j}) \prod_{x_k \in x_{\pi_i}} p(x_k|x_{\pi_k}) \sum_{x_{D_i}} \prod_{l \in X_{D_i}} p(x_l|x_{\pi_l}) \end{aligned}$$

Factorization theorem (3)

The last factor above is 1

The denominator of the fraction is:

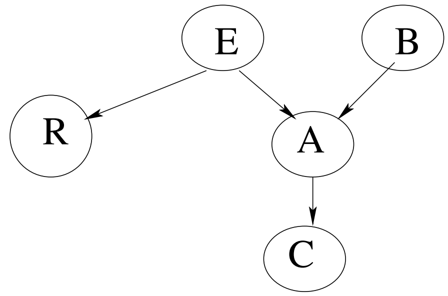
$$\begin{aligned} \sum_{x_i \in \Omega_{X_i}} p(x_i, x_{\pi_i}, x_{N_i}) &= \sum_{x_i \in \Omega_{X_i}} p(x_i|x_{\pi_i}) \prod_{x_j \in x_{N_i}} p(x_j|x_{\pi_j}) \prod_{x_k \in x_{\pi_i}} p(x_k|x_{\pi_k}) \\ &= \prod_{x_j \in x_{N_i}} p(x_j|x_{\pi_j}) \prod_{x_k \in x_{\pi_i}} p(x_k|x_{\pi_k}) \end{aligned}$$

Putting these back together in the fraction, we get:

$$p(x_i|x_{\pi_i}, x_{N_i}) = p(x_i|x_{\pi_i}) \Longrightarrow X_i \perp\!\!\!\perp X_{N_i} | X_{\pi_i}$$

which means that G is an I-map of p.

Factorization example



The factorization theorem allows us to represent $p(c, a, r, e, b)$ as:

$$p(c, a, r, e, b) = p(b)p(e)p(a|b, e)p(c|a)p(r|e)$$

instead of:

$$p(c, a, r, e, b) = p(b)p(e|b)p(a|e, b)p(c|a, e, b)p(r|a, e, c, b)$$

Complexity of factorized representations

- If k is the maximum number of ancestors for any node in the graph, and we have binary variables, then every conditional probability distribution will require $\leq 2^k$ numbers to specify
- The whole joint distribution can then be specified with $\leq n \cdot 2^k$ numbers, instead of 2^n
- The savings are big if the graph is sparse ($k \ll n$).

Minimal I-maps

- The fact that a DAG G is an I-map for a joint distribution p might not be very useful.
  E.g. Complete DAGs (where all arcs that do not create a cycle are present) are I-maps for *any distribution* (because they do not imply any independencies).
- A DAG G is **minimal I-map** of p if:
  1. G is an I-map of p
  2. If $G' \subseteq G$ then G' is not an I-map for p

Constructing minimal I-maps

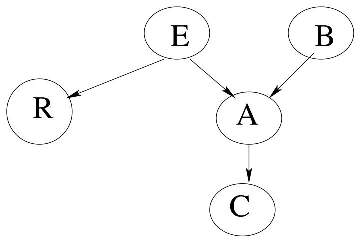
The factorization theorem suggests an algorithm:

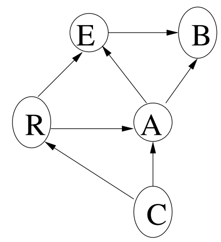
1. Fix an ordering of the variables: $X_1, \dots, X_n$
2. For each X_i, select its parents X_{π_i} to be the minimal subset of $\{X_1, \dots, X_{i-1}\}$ such that
   $X_i \perp\!\!\!\perp (\{X_1, \dots, X_{i-1}\} - X_{\pi_i}) \,|X_{\pi_i}.$

This will yield a minimal I-map

Non-uniqueness of the minimal I-map

- Unfortunately, a distribution can have *many minimal I-maps*, depending on the variable ordering we choose!
- The initial choice of variable ordering can have a big impact on the complexity of the minimal I-map:
  Example:



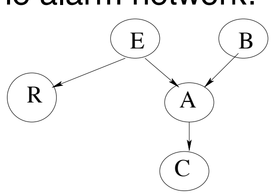


Ordering: E, B, A, R, C Ordering: C, R, A, E, B

- A good heuristic is to use causality in order to generate an ordering.

DAGs and independencies

- Given a graph G, what sort of independence assumptions does it imply? E.g. Consider the alarm network:

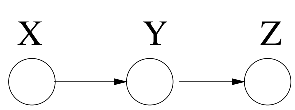


- In general the *lack of an edge* corresponds to lack of a variable in the conditional probability distribution, so it must imply some independencies

Implied independency

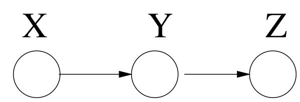
- The fact that a Bayes net is an I-map for a distribution implies a set of conditional independencies that always hold, and allows us to compute join probabilities (and hence make inference) a lot faster in practice
- In practice, we also have *evidence* about the values of certain variables.
- Is there a way to say what are *all* the independence relations implied by a Bayes net?

A simple case: Indirect connection



- Think of X as the past, Y as the present and Z as the future
- This is a simple **Markov chain**
- We interpret the lack of an edge between X and Z as a conditional independence, $X \perp\!\!\!\perp Z | Y$. Is this justified?

Indirect connection (continued)



- We interpret the lack of an edge between X and Z as a conditional independence, $X \perp\!\!\!\perp Z | Y$. Is this justified?
- Based on the graph structure, we have:

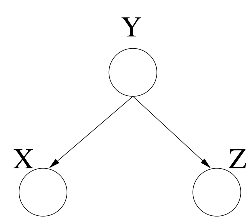
$$p(X, Y, Z) = p(X)p(Y|X)p(Z|Y)$$

- Hence, we have:

$$p(Z|X, Y) = \frac{p(X, Y, Z)}{p(X, Y)} = \frac{p(X)p(Y|X)p(Z|Y)}{p(X)p(Y|X)} = p(Z|Y)$$

- Note that the edges that are *present* do *not* imply dependence. But the edges that are *missing* do imply independence.

A more interesting case: Common cause

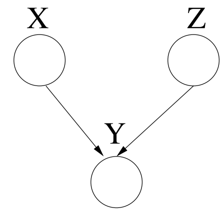


- Again, we interpret the lack of edge between X and Z as $X \perp\!\!\!\perp Z | Y$. Why is this true?

$$p(X|Y, Z) = \frac{p(X, Y, Z)}{p(Y, Z)} = \frac{p(Y)p(X|Y)p(Z|Y)}{p(Y)p(Z|Y)} = p(X|Y)$$

- This is a "hidden variable" scenario: if Y is unknown, then X and Z could appear to be dependent on each other

The most interesting case: V-structure



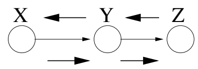
- In this case, the lacking edge between X and Z is a statement of *marginal independence*: $X \perp\!\!\!\perp Z$.
- In this case, once we know the value of Y, X and Z might depend on each other.
- E.g., suppose X and Z are independent coin flips, and Y is true if and only if both X and Z come up heads.
- Note that in this case, X is *not* independent of Z given Y!
- This is the case of "explaining away".

Bayes ball algorithm

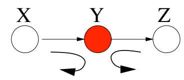
- Suppose we want to decide whether $X \perp\!\!\!\perp Z | Y$ for a general Bayes net with corresponding graph G.
- We shade all nodes in the evidence set, Y
- We put balls in all the nodes in X, and we let them bounce around the graph according to rules inspired by these three base cases
- Note that the balls can go in any direction along an edge!
- If any ball reaches any node in Z, then the conditional independence assertion is *not* true.

Base rules

- *Head-to-tail*

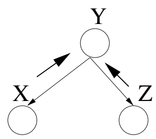


Y unknown, path unblocked

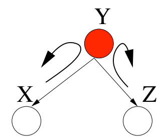


Y known, path blocked

- *Tail-to-tail*

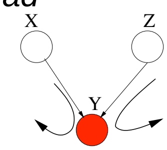


Y unknown, path unblocked

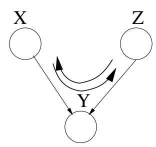


Y known, path blocked

- *Head-to-head*



Y unknown, path BLOCKED



Y known, path UNBLOCKED