Lecture 4: Independence maps. Factorization

- Independence maps
- A more formal definition of Bayes nets
- Factorization theorem

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:

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

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

X	Y	$P(X,Y)$
x=0	y=0	0.08
x=0	y=1	0.32
x=1	y=0	0.32
x=1	y=1	0.28

Factorization

Let G be a DAG over variables $X_1, \ldots, X_n$. We say that a distribution **P factorizes according to G** if P can be expressed as a product:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i))$$

The individual factors $P(X_i|Parents(X_i))$ are called **local probabilistic models** or **conditional probability distributions(CPD)**.

Bayesian network definition

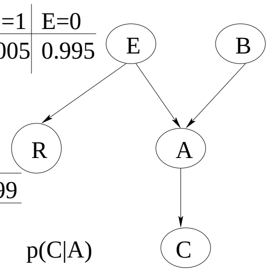
A Bayesian network is a DAG G over variables $X_1, \ldots, X_n$, together with a distribution P that factorizes over G. P is specified as the set of conditional probability distributions associated with G's nodes.

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

.

Factorization theorem

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

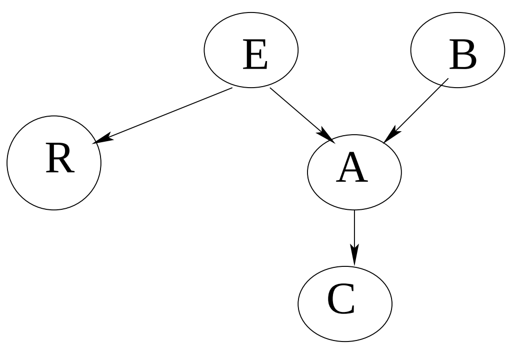
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i))$$

Proof: One direction: 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, $\text{Parents}(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$. This means that $\{X_1, \ldots, X_{i-1}\} = \text{Parents}(X_i) \cup Z$, where $Z \subseteq \text{Nondescendents}(X_i)$. Since G is an I-map, we have $X_i \perp\!\!\!\perp \text{Nondescendents}(X_i) | \text{Parents}(X_i)$, so:

$$P(X_i | X_1, \ldots, X_{i-1}) = P(X_i | Z, \text{Parents}(X_i)) = P(X_i | \text{Parents}(X_i))$$

and the conclusion follows.

Factorization example



The factorization theorem allows us to represent $P(C, A, R, E, B)$ as:

$$P(C, A, R, E, B) = P(B)P(E)P(R|E)P(A|E, B)P(C|A)$$

instead of:

$$P(C, A, R, E, B) = P(B)P(E|B)P(R|E, B)P(A|E, B, R)P(C|A, E, B, R)$$

Complexity of factorized representations

- If $|\text{Parents}(X_i)| \leq k, \forall i$, 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 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 G:
  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:

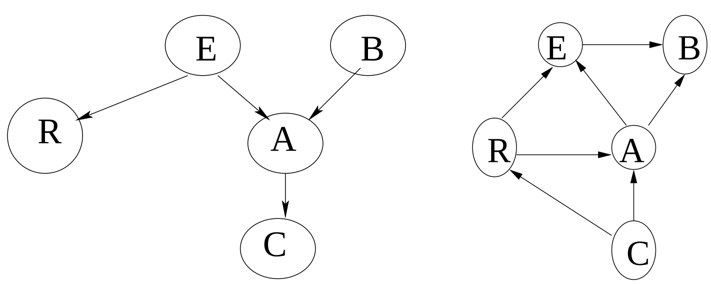
1. Fix an ordering of the variables: $X_1, \ldots, X_n$
2. For each X_i, select Parents(X_i) to be the minimal subset of $\{X_1, \ldots, X_{i-1}\}$ such that $X_i \perp\!\!\!\perp (\{X_1, \ldots, X_{i-1}\} -$ Parents$(X_i)) \,|$Parents(X_i).

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.