Belief networks

Artificial intelligence (CK0031/CK0248)

Francesco Corona

Department of Computer Science
Federal University of Ceará, Fortaleza

Belief networks

We make a first connection between probability and graph theory

Belief networks (**BNs**) introduce structure into a probabilistic model

- Graphs are used to represent **independence assumptions**
- Details about the model can be 'read' from the graph

Probability operations (marginalisation/conditioning) as graph operations

- A benefit in terms of computational efficiency

Belief networks (cont.)

Belief networks cannot capture all possible relations among variables

- They are a natural choice for representing ‘causal’ relations

They belong to the family of **probabilistic graphical models**

Benefits of structure
Belief networks

Benefits of structure

The many possible ways random variables can interact is extremely large

- Without assumptions, we are unlikely to make a useful model

Example

Consider a model with N random variables x_i, with $i = 1, \dots, n$

Independently specify all entries of a table $p(x_1, \dots, x_N)$

Consider the case of binary variables x_i

$\rightsquigarrow$ It takes $\mathcal{O}(2^N)$ space

It might be impractical for more than a handful of variables

Benefits of structure (cont.)

We deal with distributions on potentially hundreds to millions of variables

This grow is infeasible in many application areas

Structure is crucial for tractability of inference

Benefits of structure (cont.)

Example

Consider a model with N binary random variables, $p(x_1, \dots, x_N)$

Computing a marginal $p(x_i)$ requires summing over 2^{N-1} states

Even on the most optimistically fast computer, this would take too long

- Even for a $N = 100$ variable system

Benefits of structure (cont.)

We need a way to render specification/inference in such systems tractable

- We must constrain the nature of variable interactions
- This is only way with such distributions

The idea is to specify which variables are independent of others

⇝ A **structured factorisation** of the **joint probability distribution**

Benefits of structure (cont.)

Belief networks are a framework for representing independence assumptions

- They play a (quasi) natural role as '*causal*' models

Example

Consider a distribution on a chain

$$p(x_1, \ldots, x_{100}) = \sum_{i=1}^{99} \phi(x_i, x_{i+1})$$

Computing a marginal $p(x_1)$ is fast

Benefits of structure (cont.)

Belief networks (BN, or Bayes' networks or Bayesian belief networks)

- A way to depict independence assumptions in a distribution

The application domain of the general framework is widespread

- ⇝ Expert reasoning under uncertainty
- ⇝ Machine learning
- ⇝ ...

Modelling independences

Benefits of structure

Modelling independencies

Example

One morning Tracey leaves her house and realises that her grass is wet

- Is it due to overnight rain or did she forget the sprinkler on?

Next, she notices that the grass of her neighbour (Jack) is also wet

- This explains away to some extent that her sprinkler was left on

She concludes that it has probably been raining

We can model the situation by defining the variables we wish to include

$$R \in \{0, 1\}: \quad R = 1 \text{ It has been raining} \quad (R = 0, \text{ otherwise})$$
$$S \in \{0, 1\}: \quad S = 1 \text{ Tracey' sprinkler was on} \quad (S = 0, \text{ otherwise})$$
$$J \in \{0, 1\}: \quad J = 1 \text{ Jack's grass is wet} \quad (J = 0, \text{ otherwise})$$
$$T \in \{0, 1\}: \quad T = 1 \text{ Tracey's grass is wet} \quad (T = 0, \text{ otherwise})$$

Modelling independencies (cont.)

A model of Tracey's world

$$p(T, J, R, S)$$

A distribution on the joint set of variables of interest (order is irrelevant)

- Each of the variables can take one of two states

We have to specify the values for each of the $2^4 = 16$ states

⇝ $p(T = 1, J = 0, R = 0, S = 1) = 0.7$

⇝ ...

(This is not truly true, there are normalisation conditions for probabilities)

Modelling independencies (cont.)

How many states need to be specified?

Consider the following decomposition

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= P(T|J, R, S)p(J|R, S)p(R, S) \\ &= P(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned} \qquad (1)$$

The joint distribution is factorised as a product of conditional distributions

Modelling independencies (cont.)

$$p(T, J, R, S) = P(T|J, R, S)p(J|R, S)p(R|S)p(S)$$

The first term $p(T|J, R, S)$ requires us to specify $2^3 = 8$ values

- $p(T = 1|J, R, S)$ for the 8 joint states of (J, R, S)
- $p(T = 0|J, R, S) = 1 - p(T = 1|J, R, S)$, by normalisation

- $p(J = 1|R, S)$ for the 4 joint states of (R, S)
- $p(J = 0|R, S) = 1 - p(J = 1|R, S)$, by normalisation

- $p(R = 1|S)$ for the 2 states of (S)
- $p(R = 0|S) = 1 - p(R = 1|S)$, by normalisation

- $p(S = 1)$
- $p(S = 0) = 1 - p(S = 0)$, by normalisation

A total of 15 values

Modelling independencies (cont.)

Remark

In general, consider a distribution on n binary variables

We need to specify $2^n - 1$ values in the range $[0, 1]$

■

The number of values that need to be specified scales exponentially

- with the number of variables in the model
- (in general)

This is impractical, in general, and motivates simplifications

Modelling independencies - Conditional independence

The modeller often knows some constraints on the system

- We may assume that ...

Modelling independencies - Conditional independence (cont.)

Example

Whether Tracey's grass (T) is wet only depends directly on whether or not it has been raining (R) and whether or not her sprinkler (S) was on

⇝ That is, we make a conditional independence assumption

$$\rightsquigarrow \quad p(T|J,R,S) = p(T|\not{J},R,S) \tag{2}$$

Whether Jack's grass (J) is wet is influenced only directly by whether or not it has been raining (R)

$$\rightsquigarrow \quad p(J|R,S) = p(J|R,\not{S}) \tag{3}$$

The rain (R) is not directly influenced by the sprinkler (S)

$$\rightsquigarrow \quad p(R|S) = p(R|\not{S}) \tag{4}$$

Modelling independencies - Conditional independence (cont.)

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S) \tag{5}$$

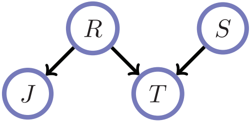
This reduces to $4 + 2 + 1 + 1 = 8$ the number of values to be specified

- A saving over the 15 values in the case where no conditional independencies had been assumed

Modelling independencies - Conditional independence (cont.)

We can represent these conditional independencies graphically

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$



Each node in the graph represents a variable in the joint distribution

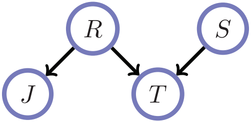
Variables which feed in (parents) to another variable (children) represent which variables are to the right of the conditioning bar

To complete the model, we need to specify the aforementioned 8 values

- The conditional probability tables (CPTs)

Modelling independencies - Conditional independence (cont.)

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$



Prior probabilities for R and S

- $p(R = 1) = 0.2$
- $p(S = 1) = 0.1$

We can set the remaining probabilities

- $p(J = 1|R = 1) = 1.0$
- $p(J = 1|R = 0) = 0.2$ ⊗

- $p(T = 1|R = 1, S = 0) = 1.0$
- $p(T = 1|R = 1, S = 1) = 1.0$
- $p(T = 1|R = 0, S = 1) = 0.9$ ⊙
- $p(T = 1|R = 0, S = 0) = 0.0$

⊗ Jack's grass is wet due to unknown effects, other than rain

⊙ There is a small chance that even though the sprinkler was left on, it did not wet the grass noticeably

Modelling independencies - Inference

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$

We made a full model of the environment (as it was described)

- We can start performing some **inference**

Modelling independencies - Inference

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$

Let us calculate the probability that the sprinkler was on overnight

- Given that Tracey's grass is wet

$$p(S = 1|T = 1)$$

We use conditional probability

$$\begin{aligned} p(S = 1|T = 1) = \frac{p(S = 1, T = 1)}{p(T = 1)} &= \frac{\sum_{J,R} p(T = 1, J, R, S = 1)}{\sum_{J,R,S} p(T = 1, J, R, S)} \\ &= \frac{\sum_{J,R} p(J|R)p(T = 1|R, S = 1)p(R)p(S = 1)}{\sum_{J,R,S} p(J|R)p(T = 1|R, S)p(R)p(S)} \\ &= \frac{\sum_{R} p(T = 1|R, S = 1)p(R)p(S = 1)}{\sum_{R,S} p(T = 1|R, S)p(R)p(S)} \end{aligned} \tag{6}$$

Modelling independencies - Inference (cont.)

$$p(S = 1 | T = 1) = \frac{(0.9 \cdot 0.8 \cdot 0.1) + (1 \cdot 0.2 \cdot 0.1)}{0.9 \cdot 0.8 \cdot 0.1 + 1 \cdot 0.2 \cdot 0.1 + 0 \cdot 0.8 \cdot 0.9 + 1 \cdot 0.2 \cdot 0.9}$$
$$= 0.3382$$

The (posterior) belief that the sprinkler is on

- It increases above the prior probability $p(S = 1) = 0.1$
- This is due to the evidence that the grass is wet

Modelling independencies - Inference (cont.)

Remark

In Equation (6), the summation over J in the numerator is unity

For any function $f(R)$, a sum of the form $\sum_J p(J|R)f(R)$ equals $f(R)$

- From the definition that a distribution $p(J|R)$ must sum to one
- $f(R)$ does not depend on J

A similar effect occurs for the summation over J in the denominator

Modelling independencies - Inference (cont.)

$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$

Let us calculate the probability that Tracey's sprinkler was on overnight

- Given that her and Jack's grass are wet

$$p(S = 1|T = 1, J = 1)$$

We use conditional probability

$$\begin{aligned} p(S = 1|T = 1, J = 1) &= \frac{p(S = 1, T = 1, J = 1)}{p(T = 1, J = 1)} \\ &= \frac{\sum_R p(T = 1, J = 1, R, S = 1)}{\sum_{R,S} p(T = 1, J = 1, R, S)} \\ &= \frac{\sum_R p(J = 1|R)p(T = 1|R, S = 1)p(R)p(S)}{\sum_{R,S} p(J = 1|R)p(T = 1)p(R)p(S)} \end{aligned} \tag{7}$$

Modelling independencies - Inference (cont.)

$$p(S = 1 | T = 1, J = 1) = \frac{0.0344}{0.2144} = 0.1604$$

The (posterior) probability that the sprinkler is on

- It is lower than it is given only that Tracey's grass is wet (0.34)
- This is due to the extra evidence (Jack's wet grass)

This occurs since the fact that Jack's grass is also wet increases the chance that the rain has played a role in making Tracey's grass wet

Modelling independencies (cont.)

Example

Sally comes home to find that the burglar alarm is sounding ($A = 1$)

- Has she been burgled ($B = 1$)
- Or, was it an earthquake ($E = 1$)?

Soon, she finds that the radio broadcasts an earthquake alert ($R = 1$)

We can write

$$p(B, E, A, R) = p(A|B, E, R)p(R|B, E)p(E|B)p(B) \tag{8}$$

However, the alarm is surely not directly influenced by radio reports

$$P(A|B, E, R) = p(A|B, E, \not{R})$$

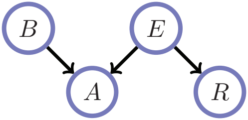
We can make other conditional independence assumptions

$$p(B, E, A, R) = p(A|B, E)p(R|\not{B}, E)P(E|\not{B})p(B) \tag{9}$$

Modelling independencies (cont.)

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

Graphical representation of the factorised joint and CPT specification



$A = 1$ (Alarm is on)	B (Burglar)	E (Earthquake)
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

$$p(B = 1) = 0.01$$
$$p(E = 1) = 0.000001$$

$R = 1$ (Earthquake alert)	E (Earthquake)
1	1
0	0

The tables and graphical structure fully specify the distribution

Modelling independencies (cont.)

What happens when we observe evidence?

Initial evidence

⇝ The alarm is sounding

$$\begin{aligned} p(B=1|A=1) &= \frac{\sum_{E,R} p(B=1,E,A=1,R)}{\sum_{B,E,R} p(B,E,A=1,R)} \\ &= \frac{\sum_{E,R} p(A=1|B=1,E)p(B=1)p(E)p(R|E)}{\sum_{B,E,R} p(A=1|B,E)p(B)p(E)p(R|E)} \\ &\simeq 0.99 \end{aligned} \quad (10)$$

Additional evidence

⇝ The earthquake alarm is broadcasted

$$p(B=1|A=1,R=1) \simeq 0.01$$

■

Modelling independencies (cont.)

Remark

Causal intuitions

Belief networks, as defined, express independence statements

- In expressing these independencies it can be useful, though potentially misleading, to think of '*what causes what*'

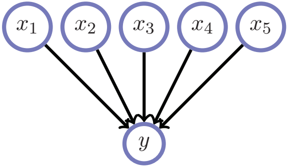
The ordering of variables is used to reflect our intuition on root causes

Reducing specifications

Benefits of structure

Reducing specifications

Consider a discrete variable y with discrete parental variables $x_1, \dots, x_n$



Formally, the structure of the graph implies nothing about the form of the parameterisation of the table

$$p(y|x_1, \dots, x_5)$$

If all variables are binary, then $2^5 = 32$ states to specify

$$p(y|x_1, \dots, x_5)$$

Reducing specifications (cont.)

Remark

Suppose that each parent x_i variable has $\dim(x_i)$ states

Suppose that there are no constraint on the table

Then, $p(y|x_1, \dots, x_n)$ contains $\big[\dim(y) - 1\big] \prod_i \dim(x_i)$ entries

If stored explicitly for each state, this is a potentially huge storage

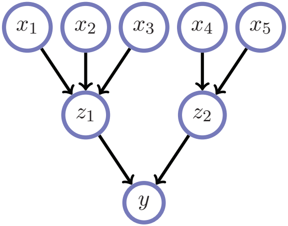
- An alternative is to constrain the table
- Use a simpler parametric form

Reducing specifications (cont.)

Divorcing parents

A decomposition with only a limited number of parental interactions

Assume all variables are binary



Constrained case

- States that require specification

$$2^3 + 2^2 + 2^2 = 16$$

Unconstrained case

- States that require specification

$$2^5 = 32$$

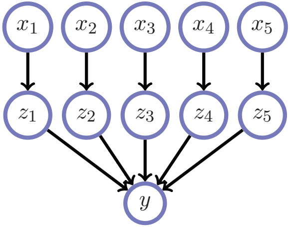
$$p(y|x_1, \ldots, x_5) = \sum_{z_1, z_2} p(y|z_1, z_2) p(z_1|x_1, x_2, x_3) p(z_2|x_4, x_5) \qquad (11)$$

Reducing specifications (cont.)

Logical gates

Simple classes of conditional tables

Use a logical OR gate on binary z_i



$$p(y|z_1,\dots,z_5) = \begin{cases} 1 & \text{if at least one } z_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

We can make table $p(y|x_1,\dots,x_5)$

- By including terms $p(z_i = 1|x_i)$

Consider the case in which each x_i is binary

There are $2+2+2+2+2 = 10$ quantities required for specifying $p(y|x)$

Reducing specifications (cont.)

Remark

The graph can be used to represent any **noisy logical state**

- The **noisy OR** or **noisy AND**

The number of parameters needed to specify the noisy gate is linear

- In the number of parents

The noisy-OR is particularly common in disease-symptom networks

- Many diseases x can give rise to the same symptom y

The probability that the symptom will be present is high

- Provided that at least one of the diseases is present

Belief networks

Belief networks

Belief networks

Definition

Belief networks

A belief network is a distribution of form

$$p(x_1, \dots, x_D) = \prod_{i=1}^{D} p\big[x_i | pa(x_i)\big] \quad (13)$$

$pa(x_i)$ *denotes the* ***parental variables*** *of variable* x_i

As a directed graph, a BN corresponds to a Directed Acyclic Graph (DAG)[1]

- The i-th node in the graph corresponds to factor $p\big[x_i \big| \text{pa}(x_i)\big]$

[1] DAG: A graph with directed edges such that by following a path from one node to another along the direction of the edges no path will revisit a node.

Belief networks (cont.)

Remark

Graphs and distributions

A subtle point is whether a BN corresponds to an instance of a distribution

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p\big[x_i | \text{pa}(x_i)\big]$$

- Requiring specification of the CPTs

Or, whether it refers to any distribution consistent with the graph structure

Belief networks (cont.)

In the case of a graph-consistent distribution, one can distinguish two cases

- A BN distribution (with numerical specification)
- A BN graph (without numerical specification)

Important to clarify the scope of independence/dependence statements

Remark

Consider the grass and burglar cases

- WE chose how to recursively factorise

■

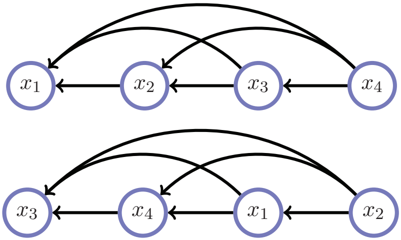
Belief networks (cont.)

Consider the general four-variable case

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1|x_2, x_3, x_4)p(x_2|x_3, x_4)p(x_3|x_4)p(x_4) \\ &= p(x_3|x_4, x_1, x_2)p(x_4|x_1, x_2)p(x_1|x_2)p(x_2) \end{aligned} \quad (14)$$

These two choices of factorisation are equivalently valid

The two associated graphs represent the same independence assumptions



Both graphs represent the same joint distribution $p(x_1, \ldots, x_4)$

- They say nothing about the content of the CPTs
- They represent the same (lack of) assumptions

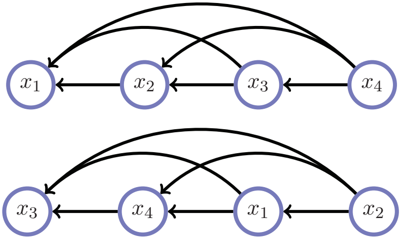
Belief networks (cont.)

In general, different graphs may represent equal independence assumptions

- ⇝ To make independence assumptions, the factorisation is crucial

Belief networks (cont.)

We observe that any distribution may be written in the **cascade form**



This cascade can be extended to many variables

- ⇝ The result is always a DAG

Belief networks (cont.)

This suggests an algorithm for constructing a BN on variables $x_1, \ldots, x_n$

1. Write the n-node cascade graph
2. Label the nodes with the variables in any order
3. Independence statement corresponds to deleting some of the edges

Belief networks (cont.)

Definition

More formally, this corresponds to an ordering of the variables

Without loss of generality, we may write as $x_1, \dots, x_n$*, from Bayes' rule*

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_1 | x_2, \dots, x_n) p(x_2, \dots, x_n) \\ &= p(x_1 | x_2, \dots, x_n) p(x_2 | x_3, \dots, x_n) p(x_3, \dots, x_n) \\ &= \cdots \\ &= p(x_n) \prod_{i=1}^{n-1} p(x_i | x_{i+1}, \dots, x_n) \end{aligned} \tag{15}$$

The representation of any BN is thus a ***Direct Acyclic Graph*** *(**DAG**)*

Belief networks (cont.)

Remark

Every probability distribution can be written as a Belief network

Though it may correspond to a fully connected 'cascade' DAG

■

The role of a BN is that the structure of the DAG corresponds to a set of conditional independence assumptions of variables on their ancestors

- Which ancestral parental variables are sufficient to specify each CPT

Belief networks (cont.)

This does not mean that non-parental variables have no influence

Example

Consider the distribution

$$p(x_1|x_2)p(x_2|x_3)p(x_3)$$

The DAG

$$x_1 \leftarrow x_2 \leftarrow x_3$$

This does not imply $p(x_2|x_1, x_3) = p(x_2|x_3)$

The DAG specifies conditional independence statements

- CI statements of variables on their ancestors
- (which ancestors are direct 'causes' for the variable)

The 'effects' will generally be dependent on the variable

- (given by the descendants of the variable)

Belief networks (cont.)

Remark

Dependencies and **Markov blanket**

Consider a distribution on a set of variables $\mathcal{X}$ and consider a variable $x_i \in \mathcal{X}$

Let the corresponding Belief network be represented by a DAG $\mathcal{G}$

- Let $\text{MB}(x_i)$ be the variables in the Markov blanket[2] of x_i

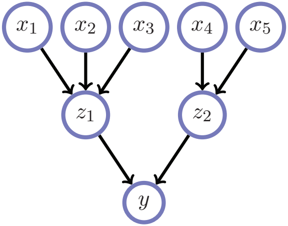
For any other variable y not in the Markov blanket of x_i, $x_i \perp\!\!\!\perp y | \text{MB}(x_i)$

- $y \in \mathcal{X} \backslash \{ x_i \cup \text{MB}(x_i) \}$

[2] Markov blanket of a node: Parents and children, and the parents of its children.

Belief networks (cont.)

The Markov blanket of x_i carries all information about x_i



$$\text{MB}(z_1) = \{x_1, x_2, x_3, y, z_2\}$$
$$z_1 \perp\!\!\!\perp x_4 | \text{MB}(z_1)$$

Belief networks (cont.)

The DAG corresponds to a statement of conditional independencies

- We need to define all elements of the CPTs $p\big[x_i\big|\text{pa}(x_i)\big]$
- This complete the specification of the BN

Once the structure is defined, then the entries of the CPTs can be expressed

A value for each state of x_i (except one, normalisation) needs to be specified

- For every possible state of the parental variables $\text{pa}(x_i)$

Belief networks (cont.)

For a large number of parents, this kind of specification is intractable

- Tables can parameterised in a low-dimensional manner
- (Belief networks in machine learning)

Conditional independence

Belief networks

Conditional independence

A BN corresponds to sets of conditional independence assumptions

Is a set of variables conditionally independent of a set of other variables?

$$p(\mathcal{X}, \mathcal{Y} | \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z}) p(\mathcal{Y} | \mathcal{Z}), \quad \text{or } \mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

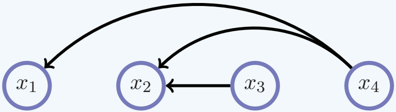
This is not always immediately clear from the DAG whether

Conditional independence (cont.)

Example

Consider the four-variable case

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



Are x_1 and x_2 independent, given the state of x_4?

Conditional independence (cont.)

$$\begin{aligned} p(x_1, x_2|x_4) &= \frac{1}{p(x_4)} \sum_{x_3} p(x_1, x_2, x_3, x_4) \\ &= \frac{1}{p(x_4)} \sum_{x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4) \\ &= p(x_1|x_4) \sum_{x_3} p(x_2|x_3, x_4)p(x_3) \end{aligned} \quad (16)$$

$$\begin{aligned} p(x_2|x_4) &= \frac{1}{p(x_4)} \sum_{x_1, x_3} p(x_1, x_2, x_3, x_4) \\ &= \frac{1}{p(x_4)} \sum_{x_1, x_3} p(x_1|x_4)p(x_2|x_3, x_4)p(x_3)p(x_4) \\ &= \sum_{x_3} p(x_2|x_3, x_4)p(x_3) \end{aligned} \quad (17)$$

Combining the two results, we have $P(x_1, x_2|x_4) = p(x_1|x_4)p(x_2|x_4)$

- Hence, variable x_1 and x_2 are independent conditioned on x_4

Conditional independence (cont.)

We would like to avoid doing such tedious manipulations

We would like to have some sort of algorithm for that

- ⇝ Read the results directly from a graph

We can develop intuition towards building such algorithm

Conditional independence (cont.)

Example

Consider a three-variable joint distribution

$$p(x_1, x_2, x_3)$$

We can write the distribution in a total of six ways

$$p(x_1, x_2, x_3) = p(x_{i_1} | x_{i_2}, x_{i_3}) p(x_{i_2} | x_{i_3}) p(x_{i_3}) \tag{18}$$

(i_1, i_2, i_3) is any of the six permutations of $(1, 2, 3)$

Conditional independence (cont.)

Each of the resulting factorisations produces a different DAG

- All of the DAGs represent the very same distribution
- None of the DAGs makes independence statement

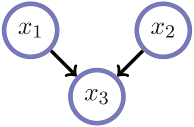
If DAGs are cascades, no independence assumptions were made

Conditional independence (cont.)

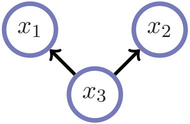
Minimal independence assumptions correspond to dropping any link

Say, we cut the link between x_1 and x_2

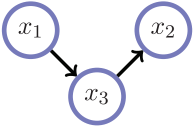
- ⇝ This gives rise to four graphs



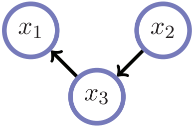
(a) $x_1 \rightarrow x_2 / x_2 \rightarrow x_1$



(b) $x_1 \rightarrow x_2 / x_2 \rightarrow x_1$

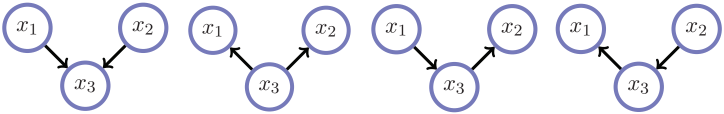


(c) $x_1 \rightarrow x_2$



(d) $x_2 \rightarrow x_1$

Conditional independence (cont.)



(a) (b) (c) (d)

Are theses graphs equivalent in representing some distribution?

$$\underbrace{p(x_2|x_3)p(x_3|x_1)p(x_1)}_{\text{graph (c)}} = \frac{p(x_2, x_3)p(x_3, x_1)}{p(x_3)} = p(x_1|x_3)p(x_2, x_3)$$
$$= \underbrace{p(x_1|x_3)p(x_3|x_2)p(x_2)}_{\text{graph (d)}} = \underbrace{p(x_1|x_3)p(x_2|x_3)p(x_3)}_{\text{graph (b)}} \quad (19)$$

(b), (c) and (d) represent the same conditional independence assumptions

- (given x_3, x_1 and x_2 are independent $x_1 \perp\!\!\!\perp x_2 | x_3$)

DAG (a) is fundamentally different, $p(x_1, x_2) = p(x_1)p(x_2)$

- There is no way to transform $p(x_3|x_1, x_2)p(x_1)p(x_2)$ into the others

Conditional independence (cont.)

Remark

Graphical dependence

Belief networks (graphs) are good for encoding conditional independence

- They are not appropriate for encoding dependence

Graph $a \rightarrow b$ may seem to encode a relation that a and b are dependent

- However, a specific numerical instance of a BN distribution could be such that $p(b|a) = p(b)$ for which we have $a \perp\!\!\!\perp b$

When a graph appears to show 'graphical' dependence, there can be instances of the distributions for which dependence does not follow

The impact of collisions

Belief networks

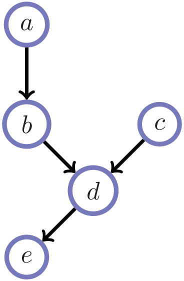
Impact of collisions

Definition

Collider

*Given a path $\mathcal{P}$, a **collider** is a node c on $\mathcal{P}$ with neighbours a and b on $\mathcal{P}$ such that $a \rightarrow c \leftarrow b$*

Impact of collisions (cont.)

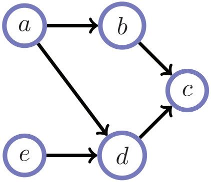


Variable d is a collider along path

$$a - b - d - c$$

but not along path

$$a - b - d - e$$



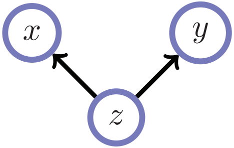
Variable d is a collider along path

$$a - d - e$$

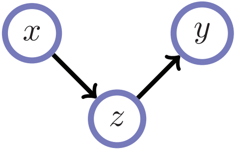
but not along path

$$a - b - c - d$$

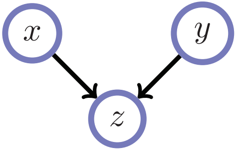
Impact of collisions (cont.)



(a) z is a not collider



(b) z is a not collider

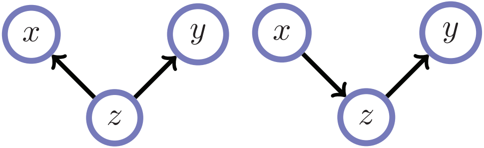


(c) z is a collider

Impact of collisions (cont.)

In a general BN, how can we check if $x \perp\!\!\!\perp y | z$?

Impact of collisions (cont.)



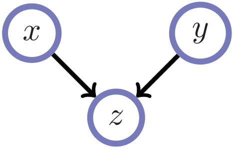
(a) $x \perp\!\!\!\perp y | z$ (b) $x \perp\!\!\!\perp y | z$

In these DAGs, x and y are independent, given z

(a) Since $p(x, y|z) = p(x|z)p(y|z)$

(b) Since $p(x, y|z) \propto \underbrace{p(z|x)p(x)}_{f(x)}\underbrace{p(y|z)}_{g(y)}$

Impact of collisions (cont.)

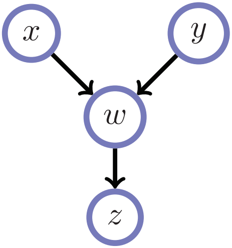


(a) $x \top\!\!\top y | z$

In this DAG, x and y are graphically dependent, given z

(c) Since $p(x, y|z) \propto p(z|x, y)p(x)p(y)$

Impact of collisions (cont.)



(a) $x \perp\!\!\!\perp y | z$

When we condition on z, x and y will be graphically dependent

$$p(x, y, | z) = \frac{p(x, y, z)}{p(z)} = \frac{1}{p(z)} \sum_{w} p(z|w) p(w|x, y) p(x) p(y)$$
$$\neq p(x|z) p(y|z)$$

Impact of collisions (cont.)

$$p(x, y, |z) = \frac{1}{p(z)} \sum_{w} p(z|w) p(w|x, y) p(x) p(y) \neq p(x|z) p(y|z)$$

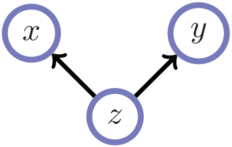
The inequality holds due to the term $p(w|x, y)$

In special cases such as $p(w|x, y) = const$ would x and y be independent

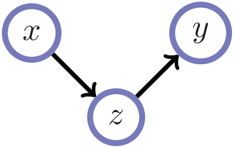
w becomes dependent on the value of z

- x and y are conditionally dependent on w
- They are conditionally dependent on z

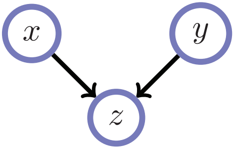
Impact of collisions (cont.)



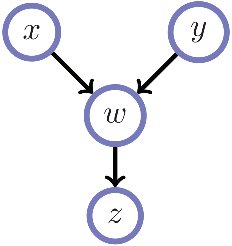
(a) $x \perp\!\!\!\perp y | z$



(b) $x \perp\!\!\!\perp y | z$



(c) $x \top\!\!\!\top y | z$

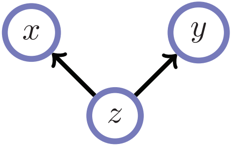


(d) $x \top\!\!\!\top y | z$

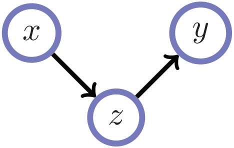
Suppose there is a non-collider z, conditioned on the path between x and y

- This path does not induce dependence between x and y

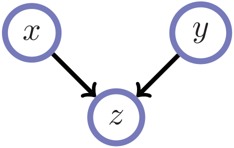
Impact of collisions (cont.)



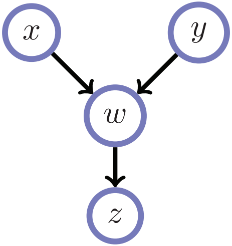
(a) $x \perp\!\!\!\perp y|z$



(b) $x \perp\!\!\!\perp y|z$



(c) $x \top\!\!\!\top y|z$



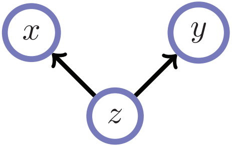
(d) $x \top\!\!\!\top y|z$

Suppose there is a path between x and y which contains a collider

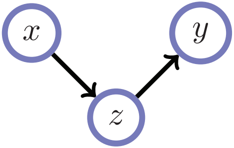
Suppose this collider is not in the conditioned set, neither are its descendants

- This path does not make x and y dependent

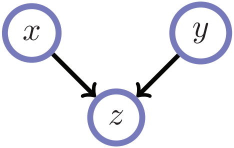
Impact of collisions (cont.)



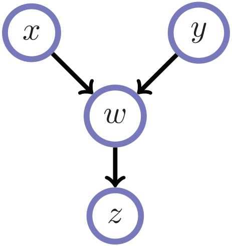
(a) $x \perp\!\!\!\perp y | z$



(b) $x \perp\!\!\!\perp y | z$



(c) $x \top\!\!\!\top y | z$



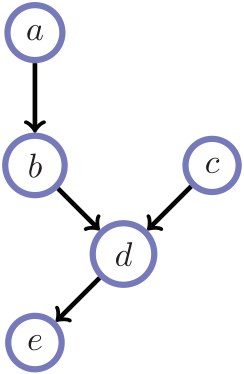
(d) $x \top\!\!\!\top y | z$

Suppose there is path between x and y which contains no colliders

Suppose that no conditioning variables are along the path

- This path '**d-connects**' x and y

Impact of collisions (cont.)



Variable d is a collider along the path

$$a - b - d - c$$

but not along the path

$$a - b - d - e$$

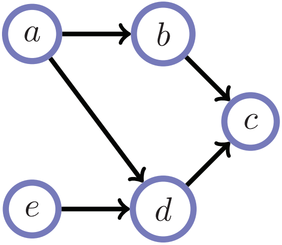
- Is $a \perp\!\!\!\perp e|b$?

a and e are not d-connected (no colliders on the path between them)

Moreover, there is a non-collider b which is in the conditioning set

- Hence, a and e are d-separated by b and $a \perp\!\!\!\perp e|b$

Impact of collisions (cont.)



Variable d is a collider along the path

$$a - d - e$$

but not along the path

$$a - b - c - d - e$$

- Is $a \perp\!\!\!\perp e | c$?

There are two paths between a and c

- ($a - b - c - d - e$ and $a - d - e$)

Path $a - d - e$ is not blocked

Although d is a collider on this path and d is not in the conditioning set

A descendant of the collider d is in the conditioning set (namely, node c)

Impact of collisions (cont.)

Some properties of belief networks

It is useful to understand what effect conditioning or marginalising a variable has on a belief network

- We state how these operations effect the remaining variables in the graph
- We use this intuition to develop a more complete description

Impact of collisions (cont.)

Consider $A \rightarrow B \leftarrow C$ with A and C (unconditionally) independent

$$p(A, B, C) = p(C|A, B)p(A)p(B)$$

Conditioning of B makes them 'graphically' dependent

Impact of collisions (cont.)

From a 'causal' perspective, this models the 'causes' A and B as a priori independent

⇝ Both determining effect C

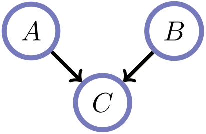
Remark

Intuitively

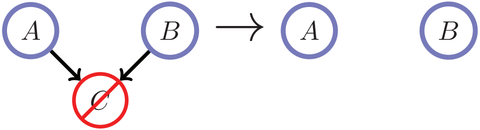
Whilst we believe the root causes are independent given the value of the observation, this tells us something about the state of both the causes coupling them and making them (generally) dependent

Impact of collisions (cont.)

Conditioning/marginalisation effects on the graph of the remaining variables



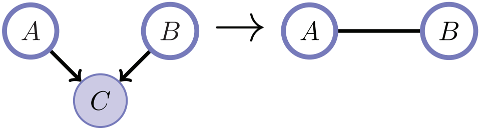
Impact of collisions (cont.)



Marginalising over C makes A and B independent

- A and B are conditionally independent $p(A, B) = p(A)p(B)$
- In the absence of any info about effect C, we retain this belief

Impact of collisions (cont.)



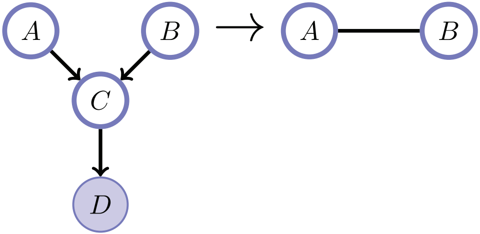
Conditioning on C makes A and B (graphically) dependent

- In general, $p(A, B|C) \neq p(A|C)p(B|C)$

Remark

Although the causes are a priori independent, knowing the effect, in general, tells us something about how the causes colluded to bring about the effect observed

Impact of collisions (cont.)



Conditioning on D makes A and B (graphically) dependent

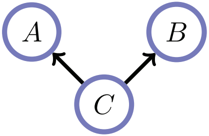
- In general, $p(A, B|D) \neq p(A|D)p(B|D)$

D is a descendent of collider C

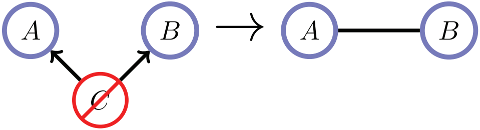
Impact of collisions (cont.)

A case in which there is a 'cause' C and independent 'effects' A and B

$$P(A|, B, C) = p(A|C)p(B|C)p(C)$$



Impact of collisions (cont.)



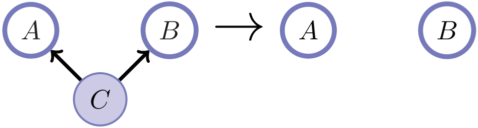
Marginalising over C makes A and B (graphically) dependent

In general, $p(A, B) \neq p(A)P(B)$

Remark

Though we do not know the ‘cause’, the ‘effects’ will be dependent

Impact of collisions (cont.)



Conditioning on C makes A and B independent

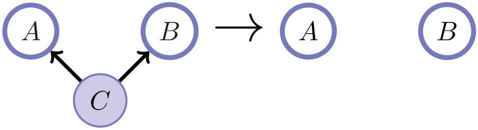
$$p(A, B|C) = p(A|C)p(B|C)$$

Remark

If you know 'cause' C, you know everything about how each effect occurs

- independent of the other effect

Impact of collisions (cont.)

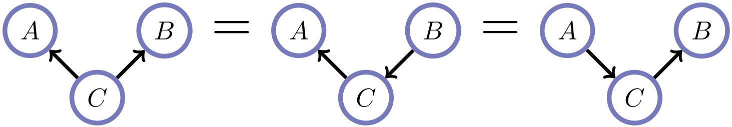


This is also true from reversing the arrow from A to C

- A would ‘cause’ C and then C would ‘cause’ B

Conditioning on C blocks the ability of A to influence B

Impact of collisions (cont.)



These graphs express the same conditional independence assumptions

Path manipulations

Belief networks

Path manipulations for independence

We now understand when x is independent of y, conditioned on z $(x \perp\!\!\!\perp y|z)$

- ⇝ We need to look at each path between x and y

Colouring x as red, y as green and the conditioning node z as yellow

- ⇝ We need to examine each path between x and y
- ⇝ We adjust the edges, following some intuitive results

Path manipulations for independence (cont.)

Remark

$$x \perp\!\!\!\perp y | z$$

After the manipulations, if there is no undirected path between x and y

- ⇝ Then, x and y are independent, conditioned on z

■

Path manipulations for independence (cont.)

The graphical rules we define here differ from those provided earlier

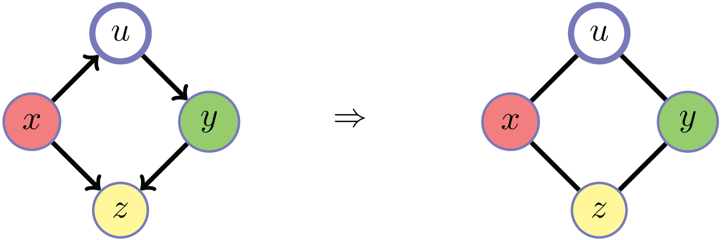
We considered the effect on the graph having eliminated a variable

- (via conditioning or marginalisation)

Rules for determining independence, from graphical representation

- The variables remain in the graph

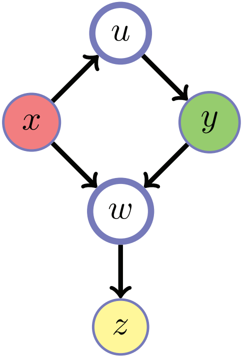
Path manipulations for independence (cont.)



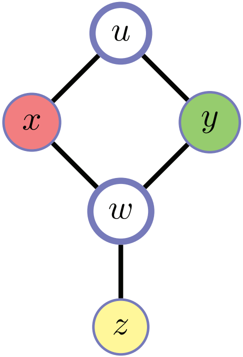
Suppose z is a collider (bottom path)

- We keep undirected links between the neighbours of the collider

Path manipulations for independence (cont.)



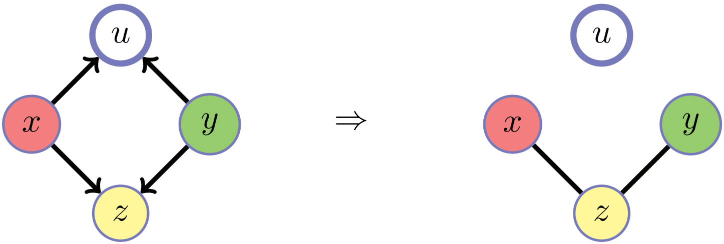
$\Rightarrow$



Suppose z is a descendant of a collider (this could induce dependence)

- We retain the links, making them undirected

Path manipulations for independence (cont.)

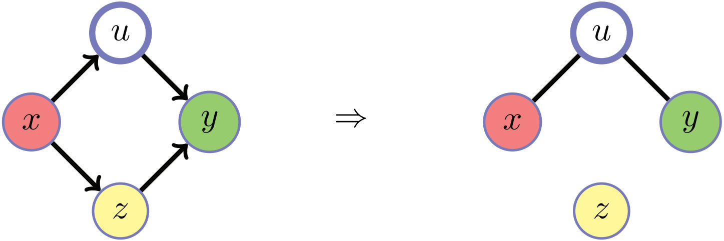


Suppose there is a collider not in the conditioning set (upper path)

- We cut the links to the collider variables

Here, the upper path between x and y is blocked

Path manipulations for independence (cont.)

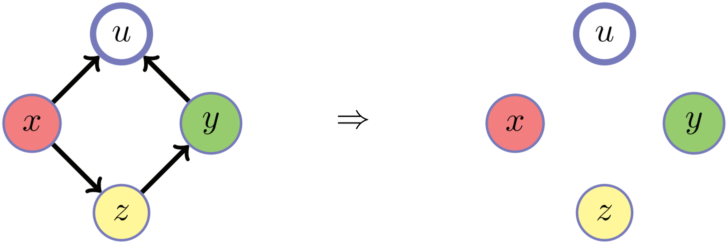


Suppose there is a non-collider from the conditioning set (bottom path)

- We cut the link between the neighbours of this non-collider
- Those that cannot induce dependence between x and y

Here, the bottom path is blocked

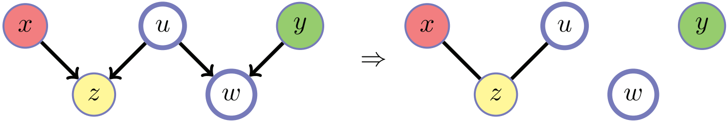
Path manipulations for independence (cont.)



Neither path contributes to dependence, hence $x \perp\!\!\!\perp y | z$

- Both paths are blocked

Path manipulations for independence (cont.)



Suppose w is a collider that is not in the conditioning set

Suppose z is a collider in the conditioning set

This means that there is no path between x and y

- Hence, x and y are independent, given z

d-Separation

Belief networks

d-separation

We need a formal treatment that is amenable to implementation

- The graphical description is intuitive

This is straightforward to get from intuitions

We define the DAG concepts of the **d-separation** and **d-connection**

- They are central to determining conditional independence
- (in any BN with structure given by the DAG)

d-separation (cont.)

Definition

d-connection *and* ***d-separation***

Let $\mathcal{G}$ be a directed graph in which $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are disjoint sets of vertices

Then, $\mathcal{X}$ and $\mathcal{Y}$ are ***d-connected*** *by $\mathcal{Z}$ in $\mathcal{G}$ if and only if there exists an undirected path U between some vertex in $\mathcal{X}$ and some vertex in $\mathcal{Y}$ such that for every collider c on U, either c or a descendant of c is in $\mathcal{Z}$ and no non-collider on U is in $\mathcal{Z}$*

$\mathcal{X}$ and $\mathcal{Y}$ are ***d-separated*** *by $\mathcal{Z}$ in $\mathcal{G}$ if and only if they not d-connected by $\mathcal{Z}$ in $\mathcal{G}$*

■

d-separation (cont.)

One may also phrase this differently as follows

'*For every variable* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$, *check every path* U *between* x *and* y, *a path* U *is said to be* ***blocked*** *if there is a node* w *on* U *such that either*:

- w *is a collider and neither* w *nor any of its descendants is in* $\mathcal{Z}$
- w *is not a collider on* U *and* w *is in* $\mathcal{Z}$

If all such paths are blocked, then $\mathcal{X}$ *and* $\mathcal{Y}$ *are d-separated by* $\mathcal{Z}$

If variables sets $\mathcal{X}$ *and* $\mathcal{Y}$ *are d-separated by* $\mathcal{Z}$, *then they are independent conditional on* $\mathcal{Z}$ *in all probability distributions such a graph can represent*'

d-separation (cont.)

Remark

Bayes ball

The Bayes ball is a linear time complexity algorithm

Given a set of nodes $\mathcal{X}$ and $\mathcal{Z}$ the Bayes ball determines the set of nodes $\mathcal{Y}$ such that $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$

- $\mathcal{Y}$ is called the set of irrelevant nodes for $\mathcal{X}$ given $\mathcal{Z}$

Graphical and distributional in/dependence

Belief networks

Graphical and distributional in/dependence

We have that $\mathcal{X}$ and $\mathcal{Y}$ d-separated by $\mathcal{Z}$ leads to $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$

- In all distributions consistent with the BN structure

Consider any instance of distro P factorising according to the BN structure

Write down a list $\mathcal{L}_p$ of all CI statements that can be obtained from P

1. If $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, list $\mathcal{L}_p$ must contain the statement

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

2. List $\mathcal{L}_p$ could contain more statements than those from the graph

Graphical and distributional in/dependence (cont.)

Example

Consider the network graph $p(a, b, c) = p(c|a, b)p(a)p(b)$

- This is representable by the DAG $a \rightarrow c \leftarrow b$

Then, $a \perp\!\!\!\perp b$ is the only graphical independence statement we can make

Consider a distribution consistent with $p(a, b, c) = p(c|a, b)p(a)p(b)$

For example, on binary variables $\text{dom}(a) = \text{dom}(b) = \text{dom}(c) = \{0, 1\}$

$$p_{[1]}(c = 1|a, b) = (a - b)^2$$
$$p_{[1]}(a = 1) = 0.3$$
$$p_{[1]}(b = 1) = 0.4$$

Numerically, we must have $a \perp\!\!\!\perp b$ for this distribution $p_{[1]}$

- $\mathcal{L}_{[1]}$ contains only the statement $a \perp\!\!\!\perp b$

Graphical and distributional in/dependence (cont.)

We can also consider the distribution

$$p_{[2]}(c = 1 | a, b) = 0.5$$
$$p_{[2]}(a = 1) = 0.3$$
$$p_{[2]}(b = 1) = 0.4$$

Here, $\mathcal{L}_{[2]} = \{a \perp\!\!\!\perp b, a \perp\!\!\!\perp c, b \perp\!\!\!\perp c\}$

Graphical and distributional in/dependence (cont.)

A question is whether or not d-connection similarly implies dependence

- *Do all distributions P, consistent with the BN possess the dependencies implied by the graph?*

Graphical and distributional in/dependence (cont.)

Example

Consider the BN equation $p(a, b, c) = p(c|a, b)p(a)p(b)$

- a and b are d-connected by c
- So, a and b are dependent, conditioned on c, graphically

Consider instance, $p_{[1]}$

- Numerically, $a \perp\!\!\!\perp b | c$
- The list of dependence statements for $p_{[1]}$ contains the graphical dependence statement

Consider For instance $p_{[2]}$

- The list of dependence statements for $p_{[2]}$ is empty

Graphical and distributional in/dependence (cont.)

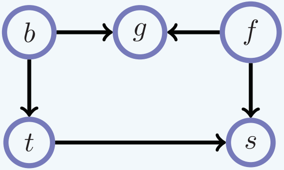
Graphical dependence statements are not necessarily found in all distributions consistent with the belief network

$\mathcal{X}$ and $\mathcal{Y}$ d-connected by $\mathcal{Z}$ does NOT lead to $\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z}$ in all distributions consistent with the belief network

Graphical and distributional in/dependence (cont.)

Example

Variables t and f are d-connected by variable g

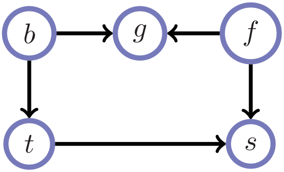


Are the variables t and f unconditionally independent ($t \perp\!\!\!\perp f | \emptyset$)?

There are two colliders, g and s, they are not in the conditioning set (empty)

- Hence, t and f are d-separated
- Therefore, they are unconditionally independent

Graphical and distributional in/dependence (cont.)



What about $t \perp\!\!\!\perp f | g$?

There is a path between t and f

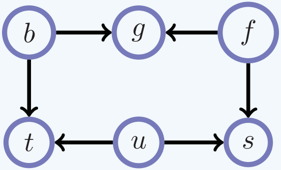
- For this path all colliders are in the conditioning set
- Hence, t and f are d-connected by g

Thus, t and f are graphically dependent conditioned on g

Graphical and distributional in/dependence (cont.)

Example

Variables b and f are d-separated by variable u



Is $\{b, f\} \perp\!\!\!\perp u | \emptyset$?

The conditioning set is empty

Every path from either b or f to u contains a collider

b and f are unconditionally independent of u

Markov equivalence in BNs

Belief networks

Markov equivalence in BNs

We studied how to read conditional independence relations from a DAG

We determine whether two DAGs represent the same set of CI statements

- A relatively simple rule

It works even when we do not know what they are!

Definition

Markov equivalence

Two graphs are Markov equivalent if they both represent the same set of conditional independence statements

This definition holds for both directed and undirected graphs

Markov equivalence in BNs (cont.)

Example

Consider the belief network with edges $A \rightarrow C \leftarrow B$

- The set of conditional independence statements is $A \perp\!\!\!\perp B | \emptyset$

For the belief network with edges $A \rightarrow C \leftarrow B$ and $A \rightarrow B$

- The set of conditional independence statements is empty

The two belief networks are not Markov equivalent

■

Markov equivalence in BNs (cont.)

Pseudo-code

Determine Markov equivalence

*Define an **immorality** in a DAG*

- *A configuration of three nodes A, B and C*
- *C is child of both A and B, with A and B not directly connected*

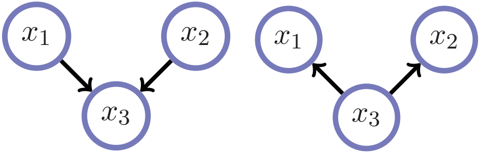
*Define the **skeleton** of a graph*

- *Remove the directions of the arrows*

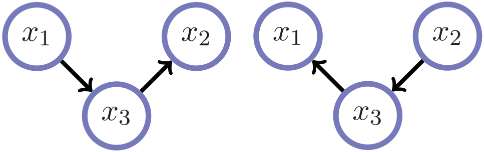
Two DAGS represent the same set of independence assumption if and only if they share the same skeleton and the same immoralities

- ***Markov equivalence***

Markov equivalence in BNs (cont.)



(a) (b)



(c) (d)

(b), (c) and (d) are equivalent

- They share the same skeleton with no immoralities

(a) has an immorality

- It is not equivalent to the others

Expressibility of BNs

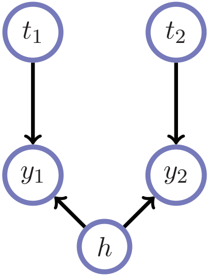
Belief networks

Expressibility of BNs

Belief networks fit with our notion of modelling 'causal' independencies

- They cannot necessarily represent all the independence properties
- (graphically)

Consider the DAF used to represent two successive experiments



t_1 and t_2 are two treatments

y_1 and y_2 are two outcomes of interest

- h: Underlying health status of the patient

The first treatment has no effect on the second outcome

⇝ Hence, there is no edge from y_1 and y_2

Expressibility of BNs (cont.)

Now consider the implied independencies in the marginal distribution

$$p(t_1, t_2, y_1, y_2)$$

They are obtained by marginalising the full distribution over h

There is no DAG containing only the vertices t_1, y_1, t_2, y_2

- No DAG represents the independence relations

It does not imply some other independence relation not implied in the figure

Expressibility of BNs (cont.)

Consequently, any DAG on vertices t_1, y_1, t_2 and y_2 alone will either fail to represent an independence relation of $p(t_1, y_2, t_2, y_2)$, or will impose some additional independence restriction that is not implied by the DAG

In general, consider $p(t_1, y_1, t_2, y_2) = p(t_1)p(t_2) \sum_h p(y_1|t_1, h)p(y_2|t_2, h)p(h)$

- Cannot be expressed as product of functions on a limited set of variables

CI conditions $t_1 \perp\!\!\!\perp (t_2, y_2)$ and $t_2 \perp\!\!\!\perp (t_1, y_1)$ hold in $p(t_1, t_2, y_1, y_2)$

- They are there encoded in the form of the CPTs

We cannot see this independence

- Not in the structure of the marginalised graph
- Though it can be inferred in a larger graph

$$p(t_1, t_2, y_1, y_2, h)$$

Expressibility of BNs (cont.)

Consider the BN with link from y_2 to y_1

We have,

$$t_1 \perp\!\!\!\perp t_2 | y_2$$

For $p(t_1, y_1, t_2, y_2) = p(t_1)p(t_2) \sum_h p(y_1|t_1, h)p(y_2|t_2, h)p(h)$

Similarly, consider the BN with $y_1 \rightarrow y_2$

The implied statement $t_1 \perp\!\!\!\perp t_2 | y_1$ is also not true for that distribution

Expressibility of BNs (cont.)

BNs cannot express all CI statements from that set of variables

- The set of conditional independence statements can be increased
- (by considering additional variables however)

This situation is rather general

Graphical models have limited expressibility of independence statements

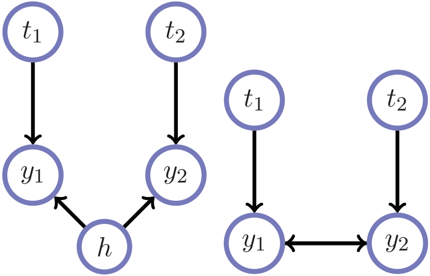
Expressibility of BNs (cont.)

BNs may not always be the most appropriate framework

- Not to express one's independence assumptions

A natural consideration

- Use a bi-directional arrow when a variable is marginalised



One could depict the marginal distribution using a bi-directional edge