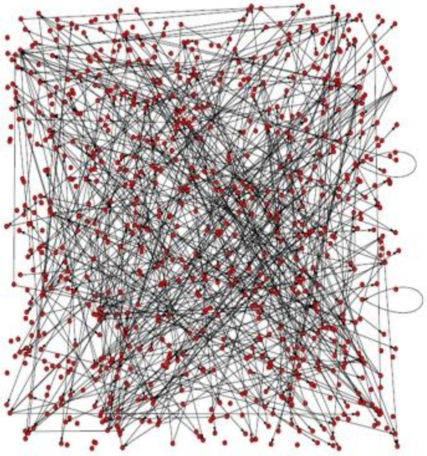
L7. Bayesian Network (Modeling)



Probability + Statistics + Graph Theory

Degree of Belief and Probability

How to compare the plausibility of different statements?

G : "we can be a billionaire if we go to graduate school"

vs

S : "we can be a billionaire if we go to Samsung"

- If you believe G more than S, you can write $G \succ S$
- If you believe S more than G, you can write $G \prec S$
- If you have the same belief, you can write $G \sim S$

Assumptions about relationships of $\succ$ and $\sim$

- *Universal comparability* : either $G \succ S$, $G \prec S$ or $G \sim S$
- *Transitivity* : if $G \succ S$ and $S \succ V$, then $G \succ V$

Due to the two assumptions, the degree of belief can be represented by a real-valued function:

- $P(G) > P(S)$ if and only if $G \succ V$
- $P(G) = P(S)$ if and only if $G \sim V$

Properties of probabilities for Bayesian Networks

We are going to use very simple probability theories to construct Probabilistic Graphical Model

- conditional probability :

$$P(A|B) = \frac{P(B|A)}{P(B)}$$

- Law of total probability :

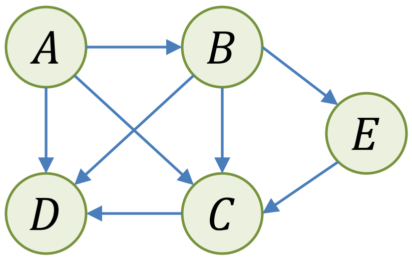
$$P(A) = \sum_{B \in \mathcal{B}} P(A|B)\, P(B)$$

- Bayes' rule:

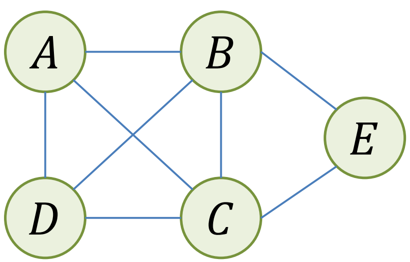
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Graph

- A graph G consists of nodes (also called vertices) and edges (also called links) between the nodes.



A directed graph G consists of directed edges between nodes

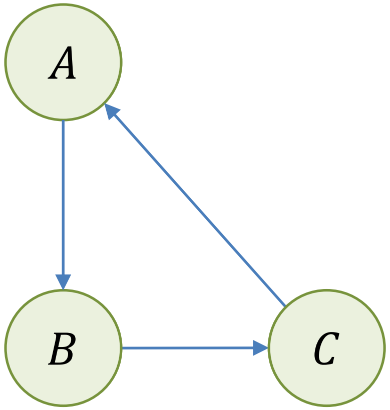


An undirected graph G consists of undirected edges between nodes

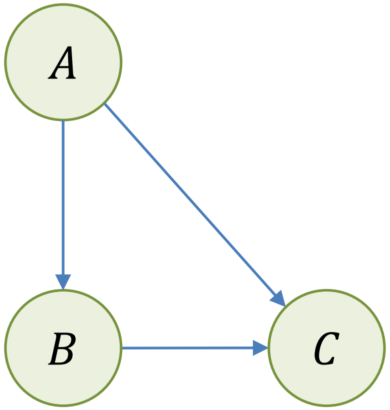
Directed Acyclic Graph (DAG)

- A DAG is a graph G with directed edges (arrows on each link) between the nodes such that by following a path of nodes from one node to another along the direction of each edge no path will revisit a node.

Cyclic Graph



Acyclic Graph



- DAG will play a central role in modeling environments with many variables
  → will be used for the belief networks
  → can encode the direction dependence between the parent nodes and child nodes.

Path

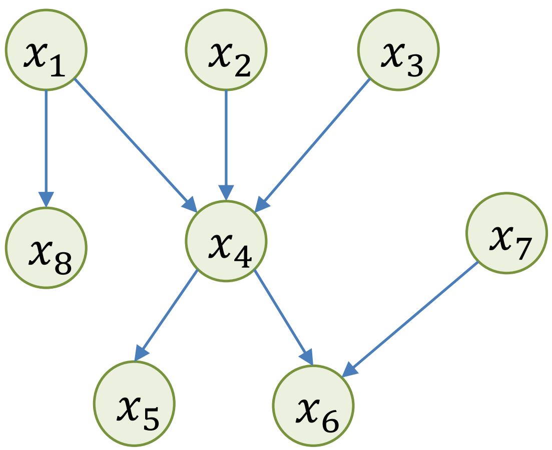
- A path $A \rightarrow B$ from node A to node B is a sequence of nodes that connects A to B

Ancestors

- In directed graph, the nodes A such that $A \rightarrow B$ and $B \nrightarrow A$ are the ancestors of B

Descendants

- In directed graph, the nodes B such that $A \rightarrow B$ and $B \nrightarrow A$ are the descendants of A



Representations

- Edge list

$$L = \{(x_1, x_4), (x_2, x_4), (x_3, x_4), (x_1, x_8), (x_4, x_5), (x_4, x_6), (x_7, x_6)\}$$

- Adjacency matrix

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- ✓ A path $x_1 \rightarrow x_6$ is $x_1 \rightarrow x_4 \rightarrow x_6$
- ✓ The ancestors of x_6 are $\mathrm{ac}(x_4) = \{x_1, x_2, x_3, x_4\}$
- ✓ The descendants of x_2 are $\mathrm{dc}(x_2) = \{x_4, x_5, x_6\}$
- ✓ The parents of x_4 are $\mathrm{pa}(x_4) = \{x_1, x_2, x_3\}$
- ✓ The children of x_4 are $\mathrm{ch}(x_4) = \{x_5, x_6\}$

Motivation of Bayesian Network

Full Joint Distribution

Example distribution

A	B	C	$P(A,B,C)$
0	0	0	0.08
0	0	1	0.15
0	1	0	0.05
0	1	1	0.10
1	0	0	0.14
1	0	1	0.18
1	1	0	0.19
1	1	1	0.11

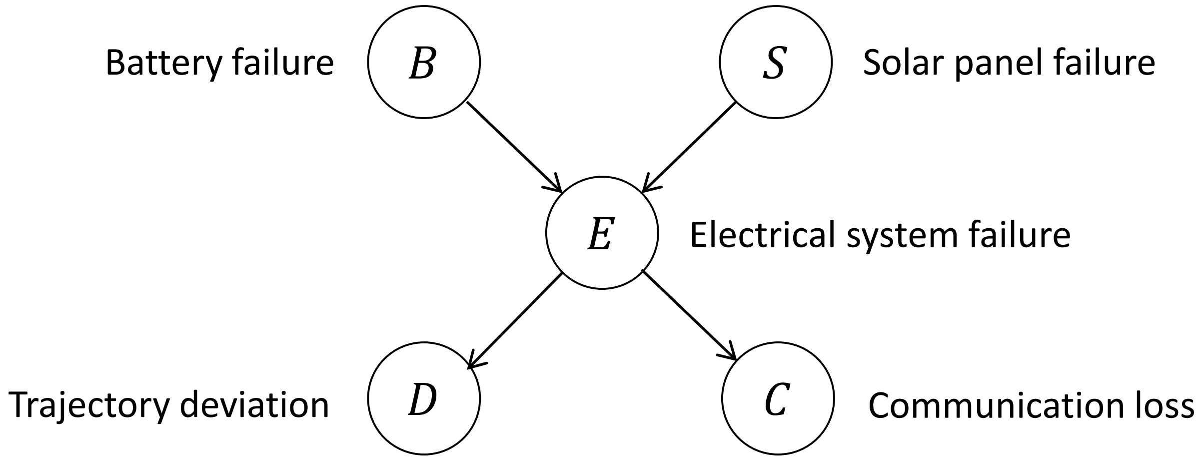
- Binary variables: A, B, C (e.g., $A = 1\ or\ 0$)
- 2^3 entities are required to construct the table
- $2^3 - 1$ independent parameters are required to fully specify the joint probability distribution
- $2^N - 1$ parameters are required for N binary variables
- If each variable has M different choices, $M^N - (M-1)$ parameters are required

The number of parameters grows exponentially

→ Difficult to represent Probability distribution and learn the parameters from data

Motivation of Bayesian Network

Full Joint Probability Distribution



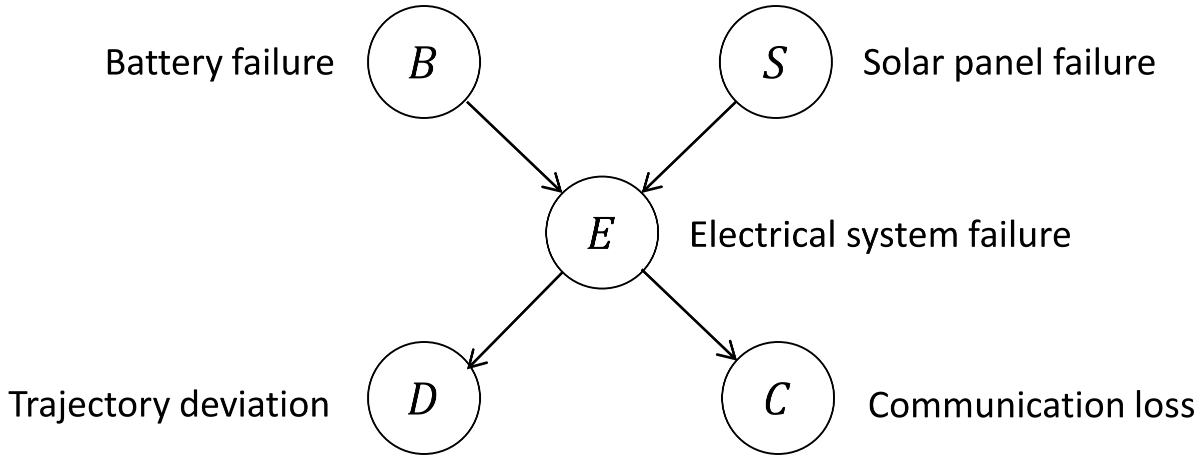
- Binary variables: B, S, E, D, C (e.g., $B = 1\ or\ 0$)
- 2^5 entities are required to construct the table
- $2^5 - 1$ independent parameters are required to fully specify the joint probability distribution
- $2^N - 1$ parameters are required for N binary variables
- If each variable has M different choices,
  $M^N - (M - 1)$ parameters are required

The number of parameters grows exponentially

→ Difficult to represent Probability distribution and learn the parameters from data

Motivation of Bayesian Network

A Bayesian network is a compact representation of a joint distribution

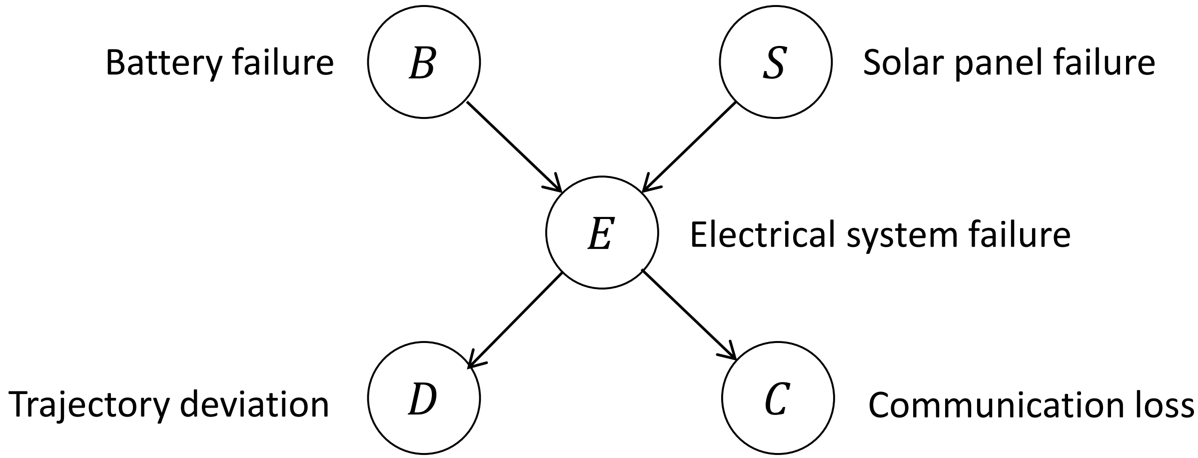


Probability + Statistics + Graph Theory

- A Bayesian Network introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables. For inferencing, it use statistics
- Provide a good representation to model the probabilistic structures between random variables.
    - Nodes represent random variables
    - Edges represent probabilistic dependency, namely conditional probability among variables
- Conditional independence described by the graph, greatly reduce the computational effort to learn the model and inferencing random variables.

Motivation of Bayesian Network

A Bayesian network is a compact representation of a joint distribution



- Each node corresponds to a random variable
- Directed edges connect pairs of nodes, indicating direct probabilistic relationships
- $P(x_i|\mathrm{pa}_{x_i})$ represents the probability distribution of x_i conditional on the parent nodes pa_{x_i} of X_i e.g., $P(E|B,S)$: B and S are the parent nodes of E

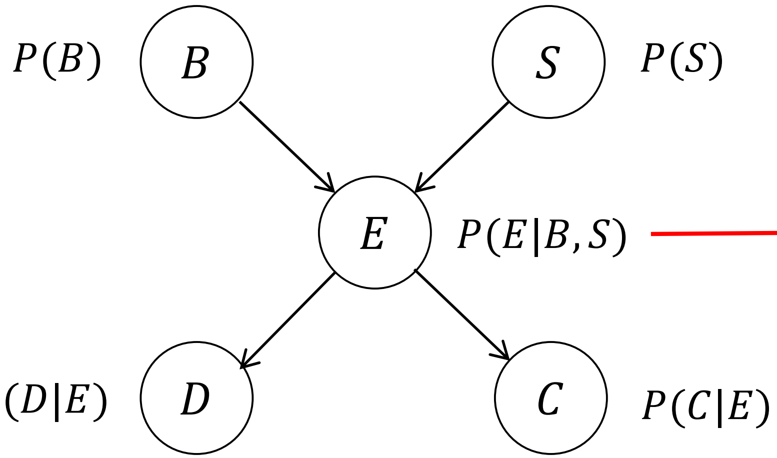
The chain rule for Bayesian networks specifies how to construct a joint distribution from the local conditional probability distribution

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} \underbrace{P(x_i|\mathrm{pa}_{x_i})}_{\text{local conditional probability distribution}}$$

Motivation of Bayesian Network

A Bayesian network is a compact representation of a joint distribution

B	$P(B)$
0	
1	



E	B	S	$P(E|B,S)$
0	0	0	
1	0	0	
0	0	1	
1	0	1	
0	1	0	
1	1	0	
0	1	1	
1	1	1	

D	E	$P(D|E)$
0	0	
1	0	
0	1	
1	1	

- Chain rule: $P(B,S,E,D,C) = P(B)P(S)P(E|B,S)P(D|E)P(C|E)$
- Required independent parameters to fully specify the joint PDF

  $P(B)$: 1, $P(S)$: 1, $P(E|B,S)$: 4, $P(D|E)$: 2, $P(C|E)$: 2 (total 10 compared to 2^5-1 = 31)

Bayesian network can greatly reduce the number of parameters

Formal Definition Bayesian Network

- A Bayesian network (BN) is a distribution of the form

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_{x_i})$$

  - ✓ pa_{x_i} represents the parental variables of variable x_i
  - ✓ BN is represented as a directed acyclic graph with an arrow pointing from a parent variable to child variable

- Every probability distribution can be written as a BN:

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_n|x_1, \ldots, x_{n-1}) p(x_1, \ldots, x_{n-1}) \\ &= p(x_n|x_1, \ldots, x_{n-1}) p(x_{n-1}|x_1, \ldots, x_{n-2}) p(x_1, \ldots, x_{n-2}) \\ &= p(x_1) \prod_{i=2}^{n} p(x_i|\mathrm{pa}_{x_{i-1}}) \end{aligned}$$

- The particular role of BN is that the structure of the DAG corresponds to a set of conditional independence assumptions, namely which ancestral parental variables are sufficient to specify each conditional probability table

Conditional Independence

Definition : Independence

$$X \perp Y$$

$$p(X,Y) = p(X)p(Y) \text{ for all states of } X, Y$$

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

Definition : Conditional Independence

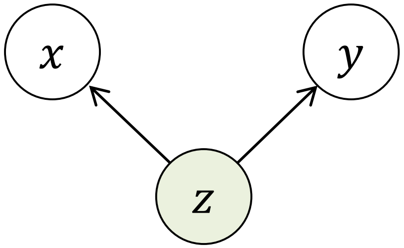
$$X \perp Y|Z$$

$$p(X,Y|Z) = p(X|Z)p(Y|Z) \text{ for all states of } X, Y, Z$$

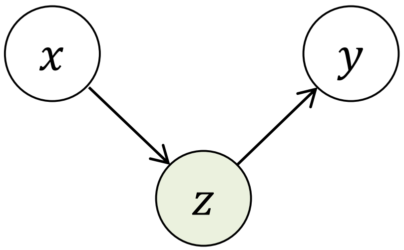
or equivalently $P(X|Y,Z) = P(X|Z)$

- ✓ The two sets of variables X and Y are independent of each other provided we know the state of the set of variables Z
- ✓ The information of Y does not give further information on X

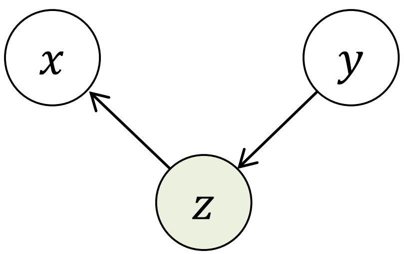
V-structure (or collider)



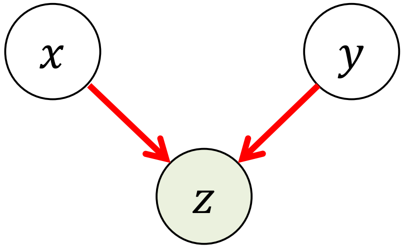
$$p(x,y|z)=\frac{p(x,y,z)}{p(z)}=\frac{p(z)p(x|z)p(y|z)}{p(z)}=p(x|z)p(y|z)$$



$$p(x,y|z)=\frac{p(x,y,z)}{p(z)}=\frac{p(x)p(z|x)p(y|z)}{p(z)}$$
$$=\frac{p(x,z)p(y|z)}{p(z)}=p(x|z)p(y|z)$$



$$p(x,y|z)=\frac{p(x,y,z)}{p(z)}=\frac{p(y)p(z|y)p(x|z)}{p(z)}$$
$$=\frac{p(y,z)p(x|z)}{p(z)}=p(y|z)p(x|z)$$



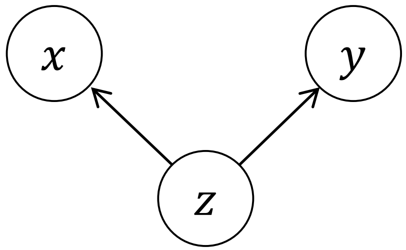
$$p(x,y|z)=\frac{p(x,y,z)}{p(z)}=\frac{p(x)p(y)p(z|x,y)}{p(z)}\neq p(y|z)p(x|z)$$

BN with $x\rightarrow z\leftarrow y$

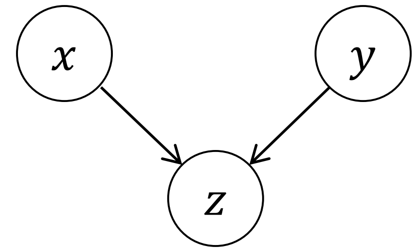
- ✓ x and y are unconditionally independent
- ✓ x and y are dependent conditional on z

V-structure (or collider)

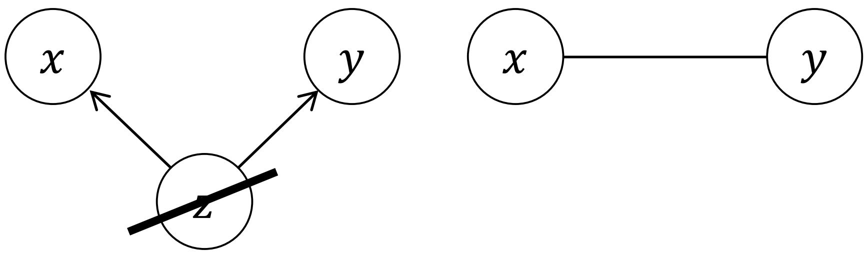
$$p(x, y, z) = p(x|z)p(y|z)$$



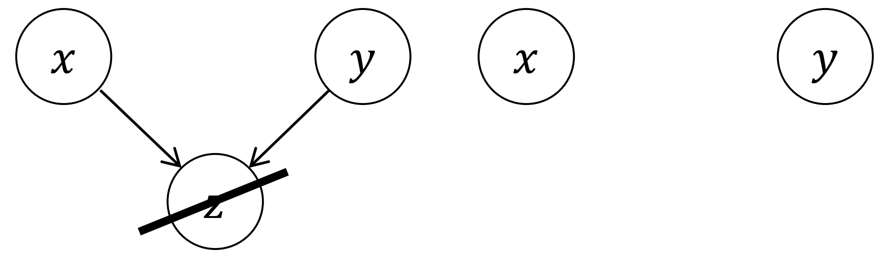
$$p(x, y, z) = p(z|x, y)p(x)p(y)$$



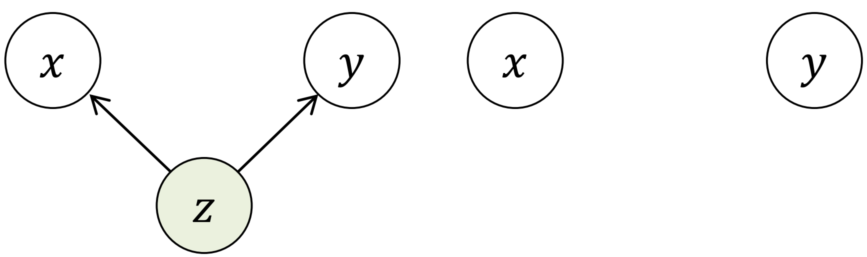
Marginalization over z



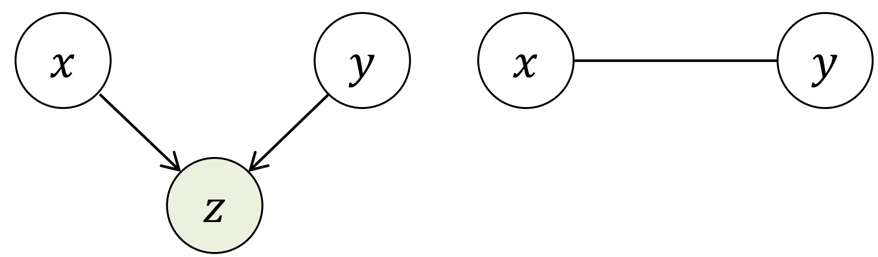
Marginalization over z



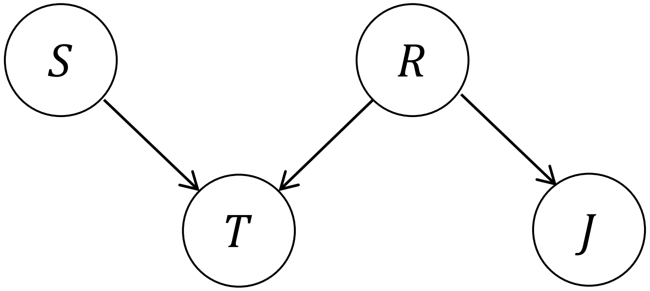
Conditionalization on z



Conditionalization on z



Example : Wet Grass



$R \in \{0,1\}$: $R = 1$ means that it has been raining
$S \in \{0,1\}$: $S = 1$ Sprinkler is turned on
$J \in \{0,1\}$: $J = 1$ Jack's grass is wet
$T \in \{0,1\}$: $T = 1$ Tracey's grass is wet

Joint distribution based on chain rule

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

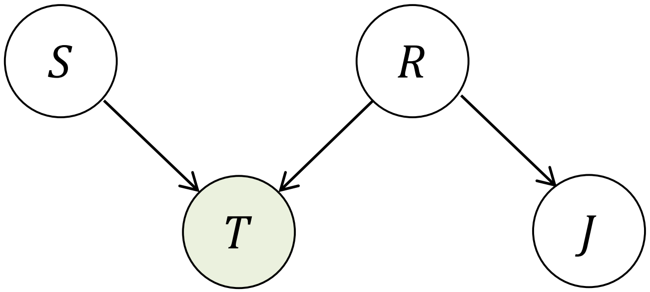
$8 + \quad 4 + \quad 2 + 1 \; = \; 2^4 - 1 = 15$ parameters are required

Joint distribution conditional independence

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

Example : Wet Grass

Modeling



$R \in \{0,1\} : R = 1$ means that it has been raining
$S \in \{0,1\} : S = 1$ Sprinkler is turned on
$J \in \{0,1\} : J = 1$ Jack's grass is wet
$T \in \{0,1\} : T = 1$ Tracey's grass is wet

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

$p(T|S,E)$

Tracey's Grass wet=1	Rain	Sprinkler
1	1	1
1	1	0
0.9	0	1
0	0	0

$p(J|R)$

Jack's Grass wet=1	Rain
1	1
0.2	0

$p(S = 1) = 0.1$

$p(R = 1) = 0.2$

The tables and graphical structure fully specify the distribution

Example : Wet Grass

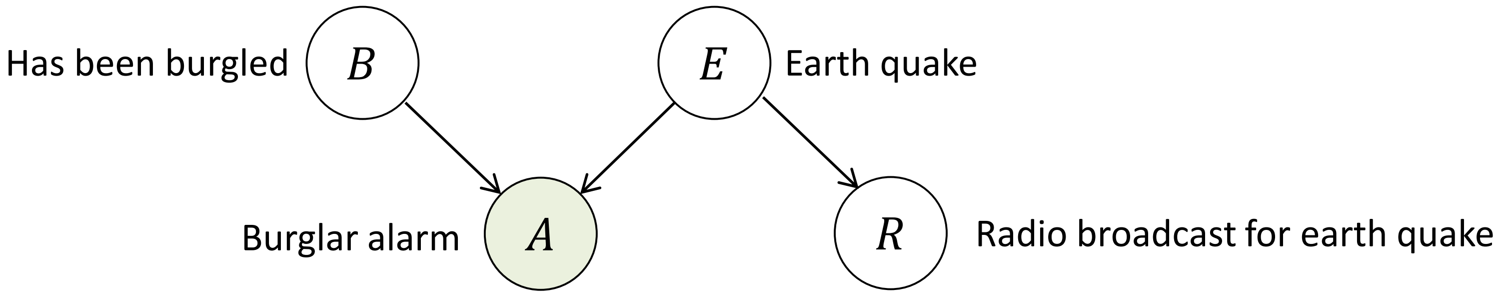
Inference

$$
\begin{aligned}
p(S=1|T=1) &= \frac{p(S=1,T=1)}{p(T=1)} = \frac{\sum_{J,R} p(T=1,J,R,S=1)}{\sum_{J,R,S} p(T=1,J,R,S)} \\
&= \frac{\sum_{J,R} p(J|R) p(T=1|R,S=1)p(R)p(S=1)}{\sum_{J,R,S} p(J|R) p(T=1|R,S)p(R)p(S)} \\
&= \frac{\sum_{R} p(T=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(T=1|R,S)p(R)p(S)} \qquad \because \sum_J p(J|R) = 1 \\
&= \frac{0.9\times 0.8\times 0.1 + 1\times 0.2\times 0.1}{0.9\times 0.8\times 0.1 + 1\times 0.2\times 0.1 + 0\times 0.8\times 0.9 + 1\times 0.2\times 0.9} = 0.3382
\end{aligned}
$$

$$
\begin{aligned}
p(S=1|T=1,J=1) &= \frac{p(S=1,T=1,J=1)}{p(T=1,J=1)} \\
&= \frac{\sum_{R} p(T=1,J=1,R,S=1)}{\sum_{R,S} p(T=1,J=1,R,S)} \\
&= \frac{\sum_{R} p(J=1|R)p(T=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(J=1|R)p(T=1|R,S)p(R)p(S)} \\
&= \frac{0.0344}{0.2144} = 0.1604
\end{aligned}
$$

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

Example : Burglar Alarm



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

$p(A|B, E)$

Alarm=1	Burglar	Earthquake
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

$p(R|E)$

Radio=1	Earthquake
1	1
0	0

$p(E = 1) = 0.01$

$p(E = 1) = 0.000001$

$$p(B = 1|A = 1) = \frac{p(B, A = 1)}{p(A = 1)} = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)}$$

$$= \frac{\sum_{E,R} p(A = 1|B = 1, E)p(R|E)p(E)p(B = 1)}{\sum_{B,E,R} p(A = 1|B, E)p(R|E)p(E)p(B)} \approx 0.99$$

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

Conditional Independence

Where causes the number of parameters to be reduced?

→ The conditional independence assumptions encoded by the structure of a Bayesian network

- X and Y are independent if and only if

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

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

$$\because P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y)}{P(Y)} = P(X)$$

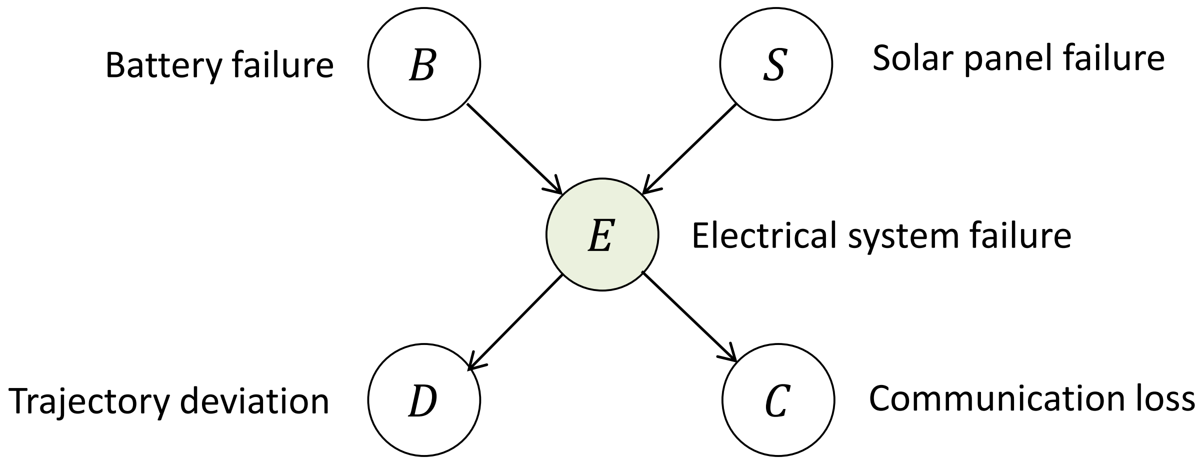
- X and Y are conditionally independent given Z if and only if

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

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

Independence assumptions reduce the number of parameters used to represent a joint pdf

Conditional Independence examples



- C is independent of B given E : $(C \perp B|E)$

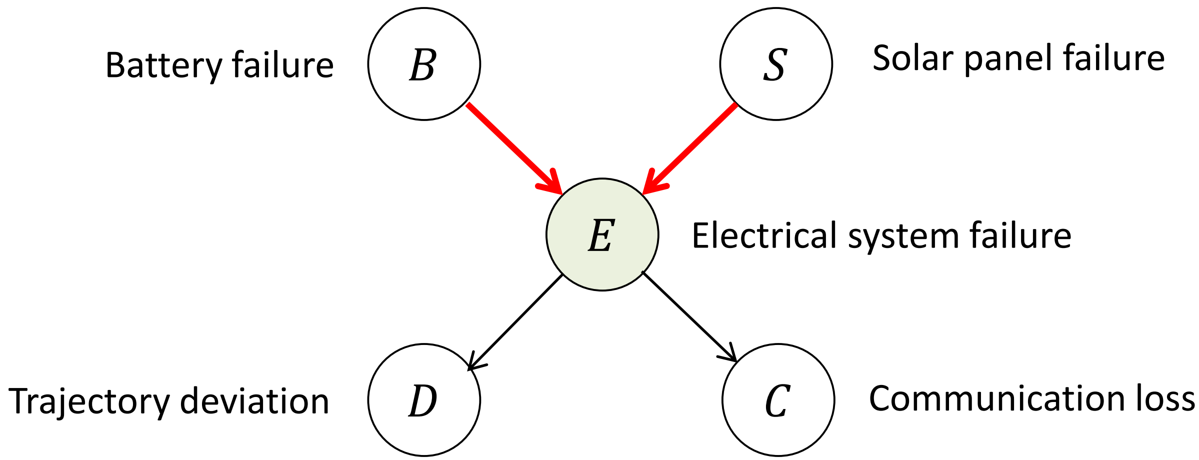
  → Information about Battery failure does not affect my belief on communication loss if I already know (observed) the status of electrical system failure

- D is independent of S given E : $(D \perp S|E)$

  → Information about Solar failure does not affect my belief on a trajectory deviation if I already know (observed) the status of electrical system failure

Conditional Independence examples

V-structure



- B is independent S (E is not observed)

  → Knowing there is a battery failure does not affect my belief regarding solar panel failure

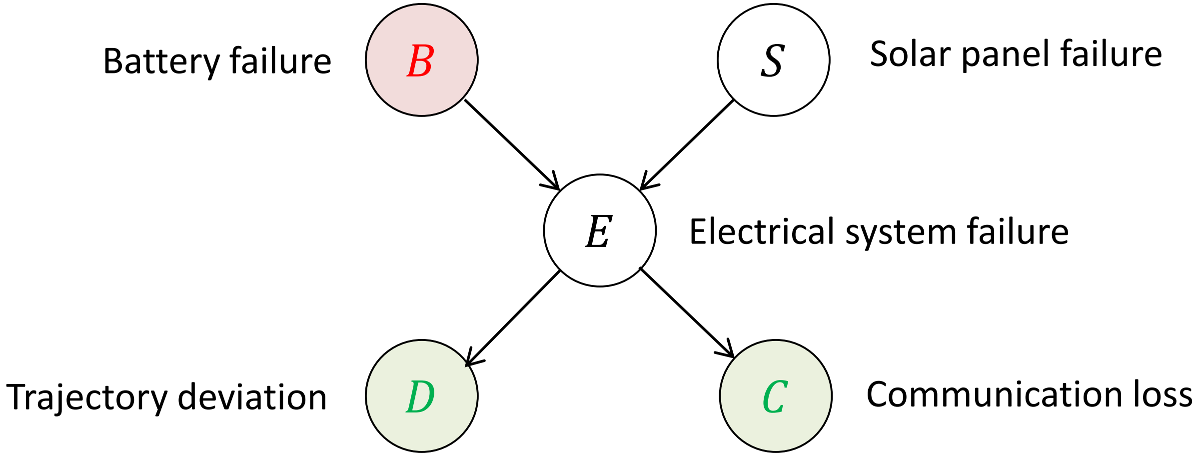
- B is dependent S given E

  → If there was an electrical system failure (observed) and there was no battery failure, there it is likely that a solar panel fails

- Influence flows only through $B \rightarrow E \leftarrow S$ when E is known

Inference

Once a joint probability distribution is constructed, inference can be performed to determine the distribution over on or more unobserved variables given the values associated with a set of observed variables

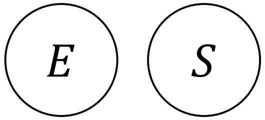


$P(B|d^1, c^1)$ Probability distribution of Battery failure

Query variable

given the trajectory deviation and the communication loss

Evidence variable



: Hidden variables

Inference

How to compute $P(B|d^1, c^1)$?

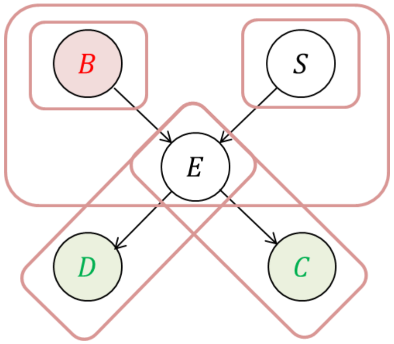
Exact inference

$$\begin{aligned} P(b^1|d^1, c^1) &\propto \sum_s \sum_e P(b^1, s, e, d^1, c^1) \\ &= \sum_s \sum_e P(b^1)P(s)P(e|b^1, s)P(d^1|e)p(c^1|e) \quad \text{By conditional independence} \\ &= P(b^1) \sum_e P(d^1|e)p(c^1|e) \sum_s P(s)P(e|b^1, s) \end{aligned}$$

The number of terms to be added together can grow exponentially with the number of hidden variables

How to compute $P(B|d^1, c^1)$?

Variable Elimination



Conditional distributions are represented by the following tables

$$T_1(B)T_2(S)T_3(E,B,S)T_4(d^1,E)T_5(c^1,E)$$

$$T_1(B)T_2(S)T_3(E,B,S)T_6(E)T_7(E)$$

Observe evidence (d^1 and c^1)

$$T_1(B)T_2(S)T_8(B,S)$$

$$T_8(B,S) = \sum_e T_3(e,B,S)T_6(e)T_7(e)$$

$$T_1(B)T_9(B)$$

$$T_9(B) = \sum_s T_2(s)T_8(B,s)$$

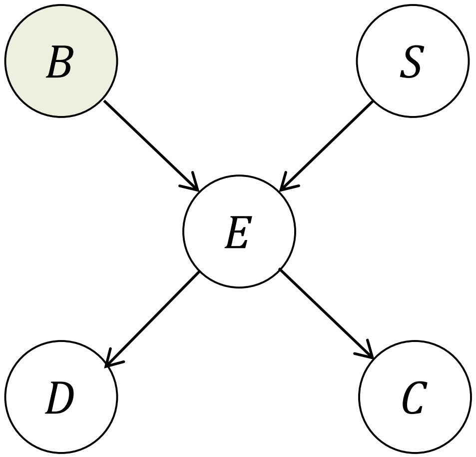
Normalizing the product of the two factors ($T_1(B)$ and $T_9(B)$) results in $P(B|d^1, c^1)$

Variable elimination algorithm relies on heuristic ordering of variables to eliminate in sequence
→ Often linear but sometimes exponential

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



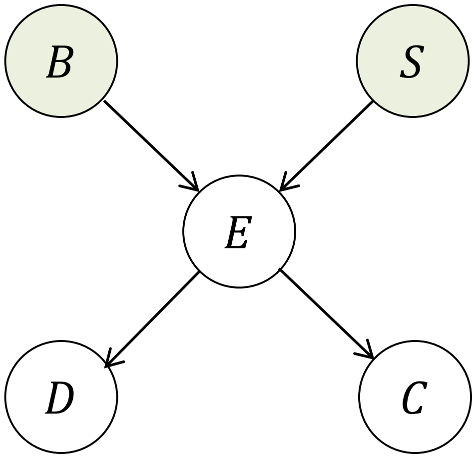
Sample from $P(B)$

B	S	E	D	C
1				

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



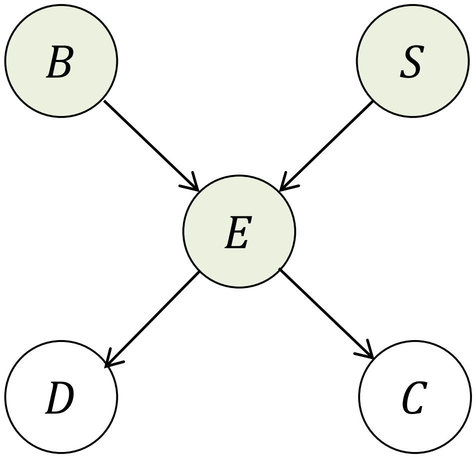
B	S	E	D	C
1	1			

Sample from $P(S)$

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



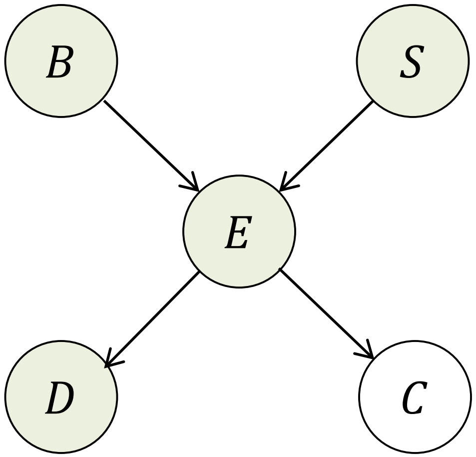
B	S	E	D	C
1	1	1		

Sample from $P(E|B=1, S=1)$

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



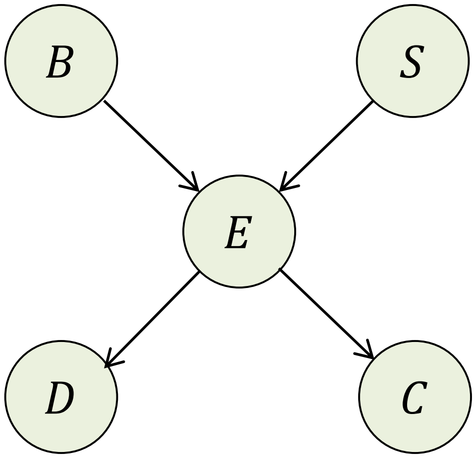
B	S	E	D	C
1	1	1	0	

Sample from $P(D|E = 1)$

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



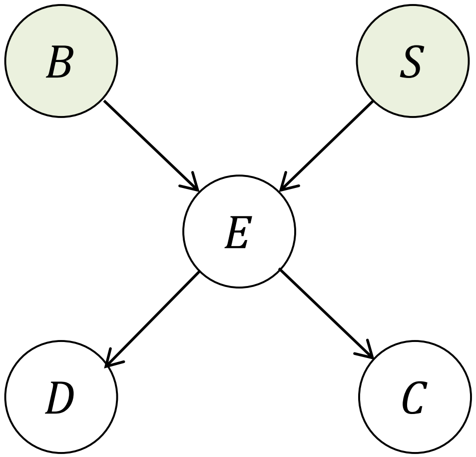
B	S	E	D	C
1	1	1	0	0

Sample from $P(C|E = 1)$

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



$$P(b^1|d^1, c^1) = 1/3$$
$$P(b^0|d^1, c^1) = 2/3$$

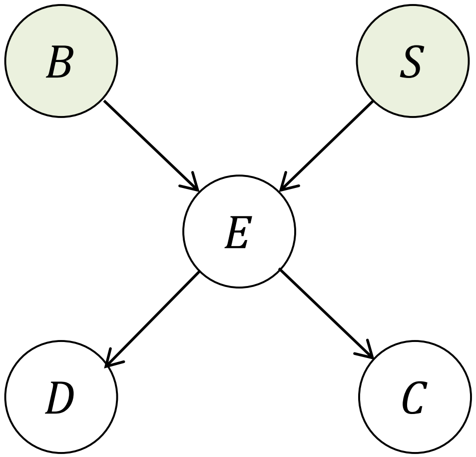
B	S	E	D	C
1	1	1	0	0
0	1	0	1	0
1	0	1	1	1
0	1	0	0	1
0	1	1	1	1
0	1	0	0	1
0	0	0	1	0
0	1	1	1	0
0	1	0	1	1

Three cases coincide observations d^1, c^1

Inference

How to compute $P(B|d^1, c^1)$?

Approximate inference (Sampling based methods)



$P(b^1|d^1, c^1) = 1/3$
$P(b^0|d^1, c^1) = 2/3$

B	S	E	D	C
1	1	1	0	0
0	1	0	1	0
1	0	1	1	1
0	1	0	0	1
0	1	1	1	1
0	1	0	0	1
0	0	0	1	0
0	1	1	1	0
0	1	0	1	1

Three cases coincide observations d^1, c^1

If likelihood of evidence is small, then many samples are required!!

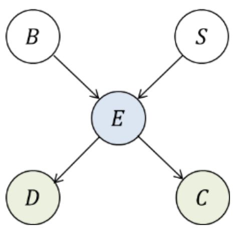
Inference

How to compute $P(B|d^1, c^1)$?

Likelihood sampling

Algorithm 2.5 Likelihood-weighted sampling from a Bayesian network

```
function LikelihoodWeightedSample(B, o_{1:n})
    X_{1:n} ← a topological sort of nodes in B
    w ← 1
    for i ← 1 to n
        if o_i = NIL
            x_i ← a random sample from P(X_i | pa_{x_i})
        else
            x_i ← o_i
            w ← w × P(x_i | pa_{x_i})
    return (x_{1:n}, w)
```



B	S	E	D	C
1	1	1	0	0
0	1	0	1	0
1	0	1	1	1
0	1	0	0	1
0	1	1	1	1
0	1	0	0	1
0	0	0	1	0
0	1	1	1	0
0	1	0	1	1

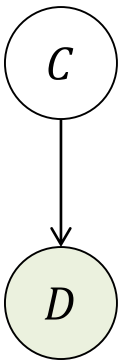
B	S	E	D	C	weight
1	0	1	1	1	$P(d^1|e^1)\ P(c^1|e^1)$
0	1	1	1	1	$P(d^1|e^1)\ P(c^1|e^1)$
0	1	0	1	1	$P(d^1|e^0)\ P(c^1|e^0)$

$$P(b^1|d^1, c^1) = \frac{P(d^1|e^1)\ P(c^1|e^1)}{P(d^1|e^1)\ P(c^1|e^1) + P(d^1|e^1)\ P(c^1|e^1) + P(d^1|e^0)\ P(c^1|e^0)}$$

Inference

How to compute $P(B|d^1, c^1)$?

Likelihood sampling has a still problem!



$P(c^1) = 0.001$

$P(d^1|c^0) = 0.001$

$P(d^1|c^1) = 0.999$

Bayesian approach :

$$P(c^1|d^1) = \frac{P(d^1|c^1)P(c^1)}{P(d^1|c^1)P(c^1) + P(d^1|c^0)P(c^0)}$$

$$= \frac{0.999 \times 0.001}{0.999 \times 0.001 + 0.001 \times 0.999}$$

$$= 0.5$$

To use likelihood weighting sampling approach:

$$c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, c^0, \dots, c^1$$

$P(d^1|c^1) = 0$ because c^1 is not sampled due to the low prior

Inference

How to compute $P(B|d^1, c^1)$?

Gibbs sampling, a kind of Markov chain Monte Carlo technique

- The sequence of samples forms a Markov chain
- In the limit, samples are drawn exactly from the joint distribution over the unobserved variables given the observations
- Simulate samples by sweeping through all the posterior conditionals, one random variables at a time

---

Algorithm : Gibbs sampler

Initialize $X^{(0)} \sim q(x)$

for iteration $i = 1, \dots do$

$$x_1^{(i)} \sim P\left(X_1 = x_1 \middle| X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)}\right)$$
$$x_2^{(i)} \sim P\left(X_2 = x_2 \middle| X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \dots, x_D = x_D^{(i-1)}\right)$$
$$x_3^{(i)} \sim P\left(X_3 = x_2 \middle| X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \dots, x_D = x_D^{(i-1)}\right)$$
$$\vdots$$
$$x_D^{(i)} \sim P\left(X_D = x_D \middle| X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \dots, X_{D-1} = x_{D-1}^{(i)}\right)$$

end for

Because samples from the early iterations are not from the target posterior, it is common to discard these samples "burn-in" period"

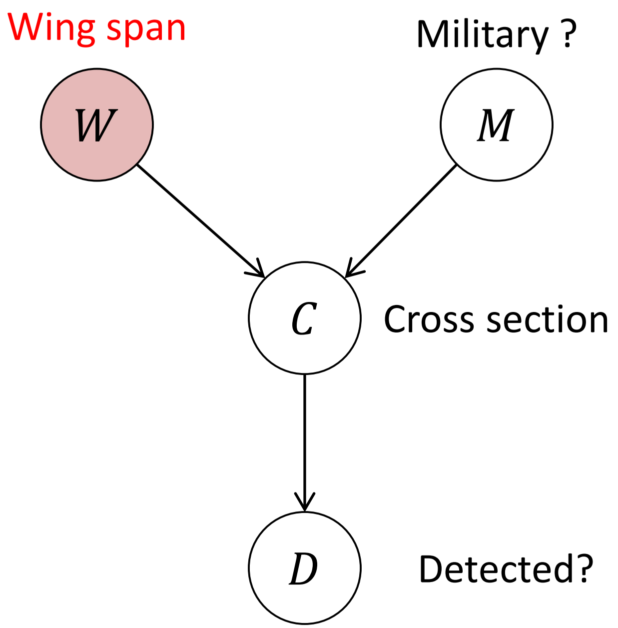
Inference

Sampling method comparisons

Jupyter Demo Simulation
Wet grass (PyMC)

Hybrid Bayesian Networks

Bayesian networks can contain a mixture of both discrete and continuous variables

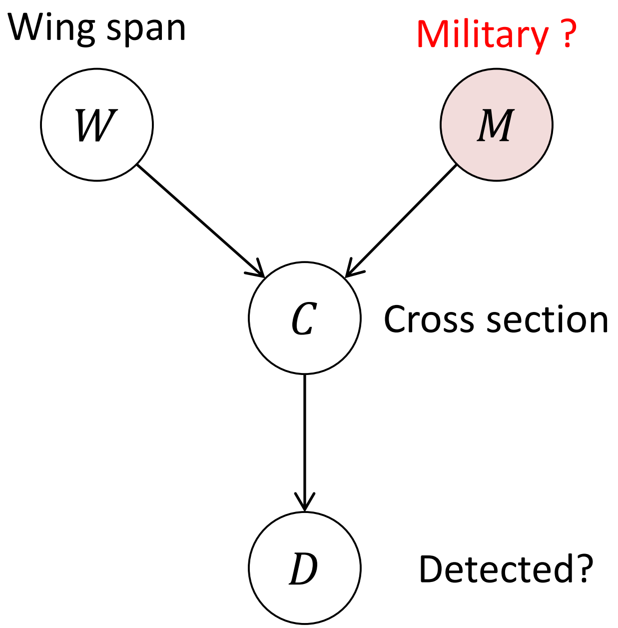


Wing span is a continuous variable and modeled as a Gaussian distribution

$$P(w) = N(w|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2}$$

Hybrid Bayesian Networks

Bayesian networks can contain a mixture of both discrete and continuous variables

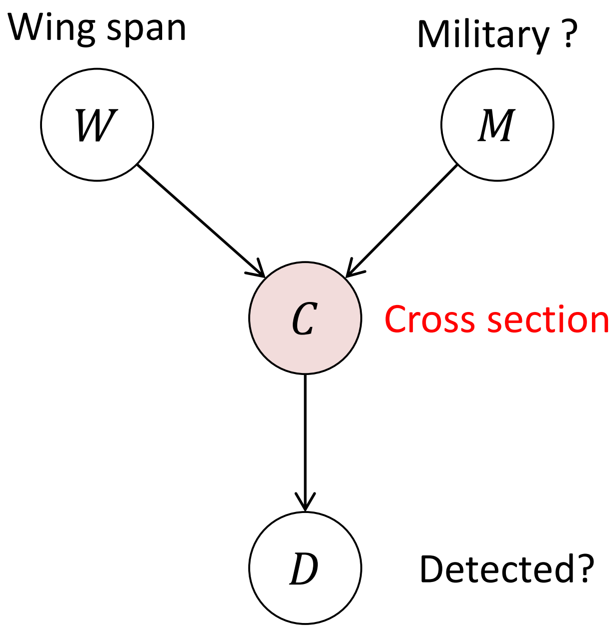


Whether a target is a military vehicle can be modeled with a single parameter θ

$$P(m^1) = \theta$$
$$P(m^0) = 1 - \theta$$

Hybrid Bayesian Networks

Bayesian networks can contain a mixture of both discrete and continuous variables

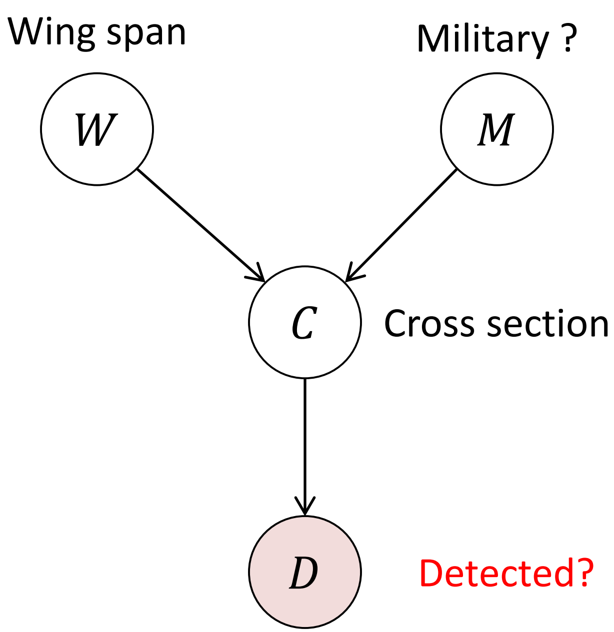


Radar cross section can be modeled as a conditional Gaussian

$$P(c|w,m) = \begin{cases} N(c|a_0 w + b_0, \sigma_0^2) & \text{if } m = m^0 \\ N(c|a_1 w + b_1, \sigma_1^2) & \text{if } m = m^1 \end{cases}$$

(Conditional linear Gaussian)

Bayesian networks can contain a mixture of both discrete and continuous variables

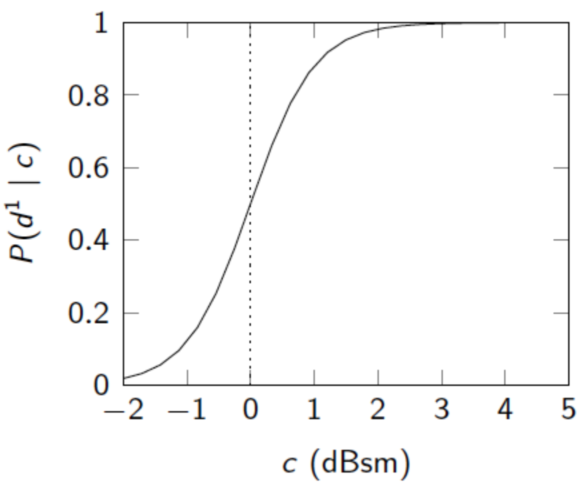


- Logit model:

$$P(d^1|c) = \frac{1}{1 + \exp\left(-2\frac{c-\alpha}{\beta}\right)}$$

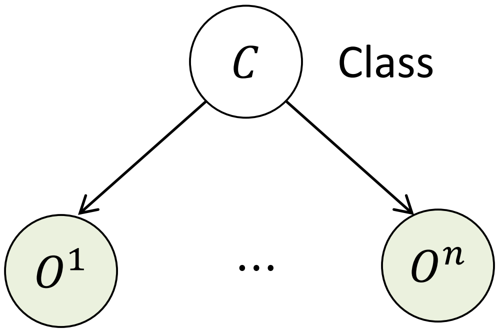
- Probit model:

$$P(d^1|c) = \Phi\left(\frac{c-\alpha}{\beta}\right)$$



Bayesian Network for Classification

Naïve Bayes Model



Prior: $\mathrm{P}(C)$

Class conditional distribution $P(O^i|C)$

$$P(C|O^{1:n}) = \frac{P(C, O^{1:n})}{P(O^{1:n})} = \frac{P(C)\prod_{i=1}^{n} P(O^i|C)}{P(O^{1:n})}$$

$$P(O^{1:n}) = \sum_{c} P(C, O^{1:n})$$

$$P(C|O^{1:n}) \propto P(C)\prod_{i=1}^{n} P(O^i|C)$$

We already know how to estimate the parameters for probability distributions

MLE or Bayesian approach

Structure learning

- Bayesian Score $P(G|D)$ for a certain graph G given data D is defined as

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$
$$= \frac{P(G) \int_\theta P(D|\theta, G)P(\theta|G)d\theta}{P(D)}$$

- A Bayesian approach to structure learning involves finding the graph G that maximizes the Bayesian Score $P(G|D)$ as

$$G^* = \underset{G}{\operatorname{argmax}}\, P(G|D)$$

- Not feasible to enumerate every possible structure, so use local search for graph with largest Bayesian score

Reference