COMP 790-125: Goals for today

- Conditional independence
- Representation of probabilistic models (Bayes nets)
- D-separation, Bayes ball, explaining away
- Undirected graphical models
- Examples of simple graphical models

Conditional independence

Marginal independence you are familiar with

$$\begin{aligned} p(X|Y,Z) &= p(X) \\ p(X,Y) &= p(X)p(Y) \end{aligned}$$

Conditional independence

$$\begin{aligned} p(X|Y,Z) &= p(X|Z) \\ p(X,Y|Z) &= p(X|Z)p(Y|Z) \end{aligned}$$

For some random variables X, Y and Z with joint $p(X, Y, Z)$ we say that X and Y are independent given Z if

$$p(X = x|Z = z) = p(X = x|Y = y, Z = z)$$

for all $x \in \mathbf{dom}\, X, y \in \mathbf{dom}\, Y, z \in \mathbf{dom}\, Z$ such that $p(z) > 0$.

And shorthand for this relationship is

$$X \perp Y|Z$$

Note that the relationship is symmetric.

Conditional independence between sets of variables

This generalizes to sets of variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$

$$\mathcal{X} \perp \mathcal{Y} | \mathcal{Z} \quad \Leftrightarrow \quad \forall X \in \mathcal{X}, Y \in \mathcal{Y} \quad X \perp Y | \mathcal{Z}$$

We can then also say that the distribution factors

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

We will also use notation X_A where A is a set of indices to denote a set of random variables, e.g. $A = \{1, 2, 5\}$ then $X_A = \{X_1, X_2, X_5\}$.

Examples of conditional independence

Shoe size $\perp$ Gray hair | Age

$\text{System}_t \perp \text{System}_{t-2} | \text{System}_{t-1}$

Temperature inside $\perp$ Temperature outside | Air conditioning is working

Pepsi or Coke $\perp$ Sugar content in your drink | Restaurant chain

Federal funds rate $\perp$ State of economy | Federal Reserve meeting notes

Dice roll outcome $\perp$ Previous dice rolls | Dice is not loaded

Benefits of capturing conditional independence

The obvious benefit is in compact representation.

Suppose we have a probability distribution $p(X, Y, Z)$ and random variables X, Y, Z can each assume 5 states.

To represent the distribution we need to store 5^3 probabilities.[1]

If we know that

$$X \perp Z|Y$$

then we can write

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

and instead of storing one large table we store 3 significantly smaller tables.[2]

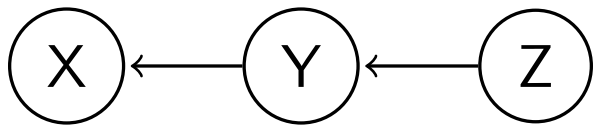
---

[1] ok $5^3 - 1$ because they need to sum to 1

[2] 124 vs 44 entries, alternatively 125 vs. 55

Benefits of capturing conditional independence

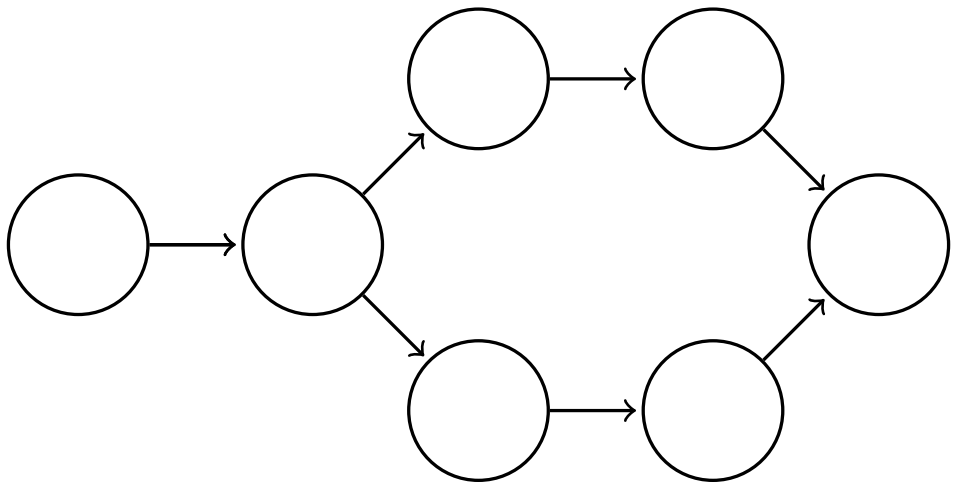
We can also capture such independencies in a graphical form.



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

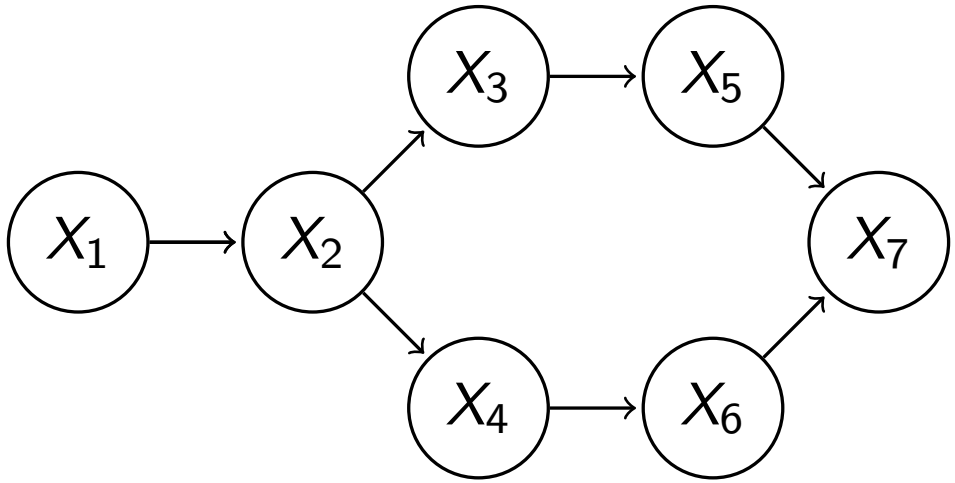
Specifying a graphical model

The starting point is a directed acyclic graph (DAG) over n nodes.



Specifying a graphical model

Each of these nodes corresponds to a random variable.



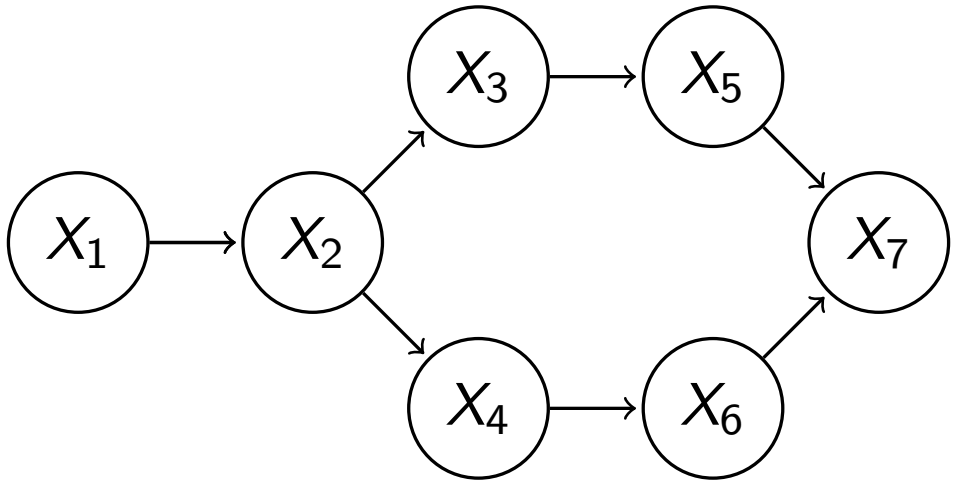
and each node has a conditional probability associated with it

$$p(X_i|X_{\mathbf{pa}(i)})$$

where $\mathbf{pa}(i)$ is a list of parent nodes of node i, e.g.

$$p(X_7|X_5, X_6)$$

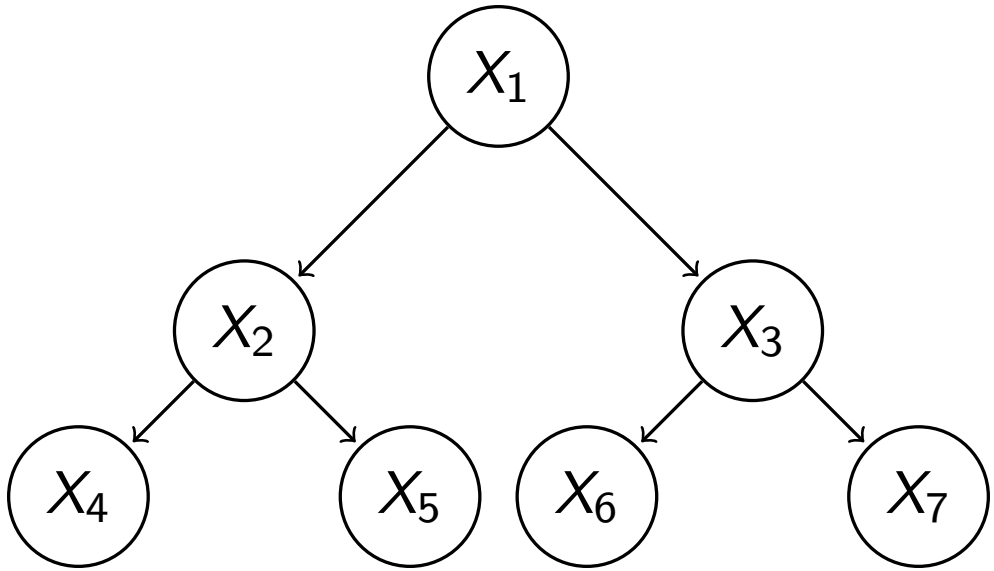
Specifying a graphical model



This graphical model specifies a joint distribution

$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) &= \prod_i p(X_i | X_{\mathbf{pa}(i)}) \\
&= p(X_1)p(X_2|X_1)p(X_3|X_2)p(X_4|X_2) \\
&\quad p(X_5|X_3)p(X_6|X_4)p(X_7|X_5, X_6)
\end{aligned}
$$

Another example of a graphical model



$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) &= \prod_i p(X_i | X_{\mathbf{pa}(i)}) \\
&= p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2) \\
&\quad p(X_5|X_2)p(X_6|X_3)p(X_7|X_3)
\end{aligned}
$$

Determining conditional independencies from graphs

In the topological order of nodes of a DAG, parent nodes precede child nodes. There can be many topological orders.

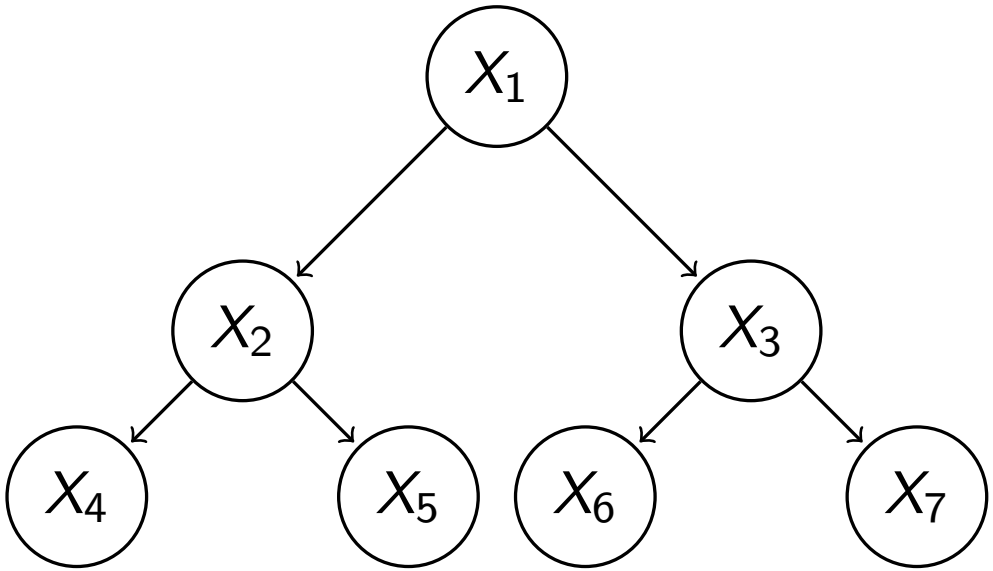
Given an order O, let $\mathbf{pnp}_O(i)$ denote a set of nodes that precede node i in a topological order but are not its parents.

We can show that

$$X_i \perp X_{\mathbf{pnp}_O(i)} | X_{\mathbf{pa}(i)}$$

These are basic conditional independence relationships.

Obtaining basic conditional independencies



A topological order:
$X_1, X_2, X_3, X_4, X_5, X_6, X_7$

$$
\begin{aligned}
X_1 \perp \emptyset \mid \emptyset \\
X_2 \perp \emptyset \mid X_1 \\
X_3 \perp X_2 \mid X_1 \\
X_4 \perp \{X_1, X_3\} \mid X_2 \\
X_5 \perp \{X_1, X_3, X_4\} \mid X_2 \\
X_6 \perp \{X_1, X_2, X_4, X_5\} \mid X_3 \\
X_7 \perp \{X_1, X_2, X_4, X_5, X_6\} \mid X_3
\end{aligned}
$$

And we are going to verify one of these because it highlights the distributive property that is crucial for message passing algorithm derivations.

Verifying a conditional independence

We want to show

$$p(X_3|X_2,X_1) = \frac{p(X_3,X_2,X_1)}{p(X_2,X_1)} = p(X_3|X_1)$$

Marginal $p(X_3,X_2,X_1)$

$$p(X_1,X_2,X_3) =$$

$$\sum_{X_4}\sum_{X_5}\sum_{X_6}\sum_{X_7} p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2)p(X_5|X_2)p(X_6|X_3)p(X_7|X_3)$$

$$\sum_{X_4}\sum_{X_5}\sum_{X_6} p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2)p(X_5|X_2)p(X_6|X_3)\sum_{X_7} p(X_7|X_3)$$

$$\sum_{X_4}\sum_{X_5}\sum_{X_6} p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2)p(X_5|X_2)p(X_6|X_3) =$$

$$p(X_1)p(X_2|X_1)p(X_3|X_1)\sum_{X_4} p(X_4|X_2)\sum_{X_5} p(X_5|X_2)\sum_{X_6} p(X_6|X_3) =$$

$$p(X_1)p(X_2|X_1)p(X_3|X_1)\sum_{X_4} p(X_4|X_2)\sum_{X_5} p(X_5|X_2) =$$

$$p(X_1)p(X_2|X_1)p(X_3|X_1)\sum_{X_4} p(X_4|X_2) = p(X_1)p(X_2|X_1)p(X_3|X_1)$$

Verifying a conditional independence

We want to show

$$p(X_3|X_2, X_1) = \frac{p(X_3, X_2, X_1)}{p(X_2, X_1)} = p(X_3|X_1)$$

Marginals are

$$\begin{aligned} p(X_3, X_2, X_1) &= p(X_1)p(X_2|X_1)p(X_3|X_1) \\ p(X_2, X_1) &= p(X_1)p(X_2|X_1) \end{aligned}$$

plugging them in we get

$$p(X_3|X_2, X_1) = \frac{p(X_3, X_2, X_1)}{p(X_2, X_1)} = \frac{p(X_1)p(X_2|X_1)p(X_3|X_1)}{p(X_1)p(X_2|X_1)} = p(X_3|X_1)$$

and this confirms $X_3 \perp X_2|X_1$

Directed separation

We say $\mathcal{Z}$ blocks an undirected path if there is a node B on the path such that

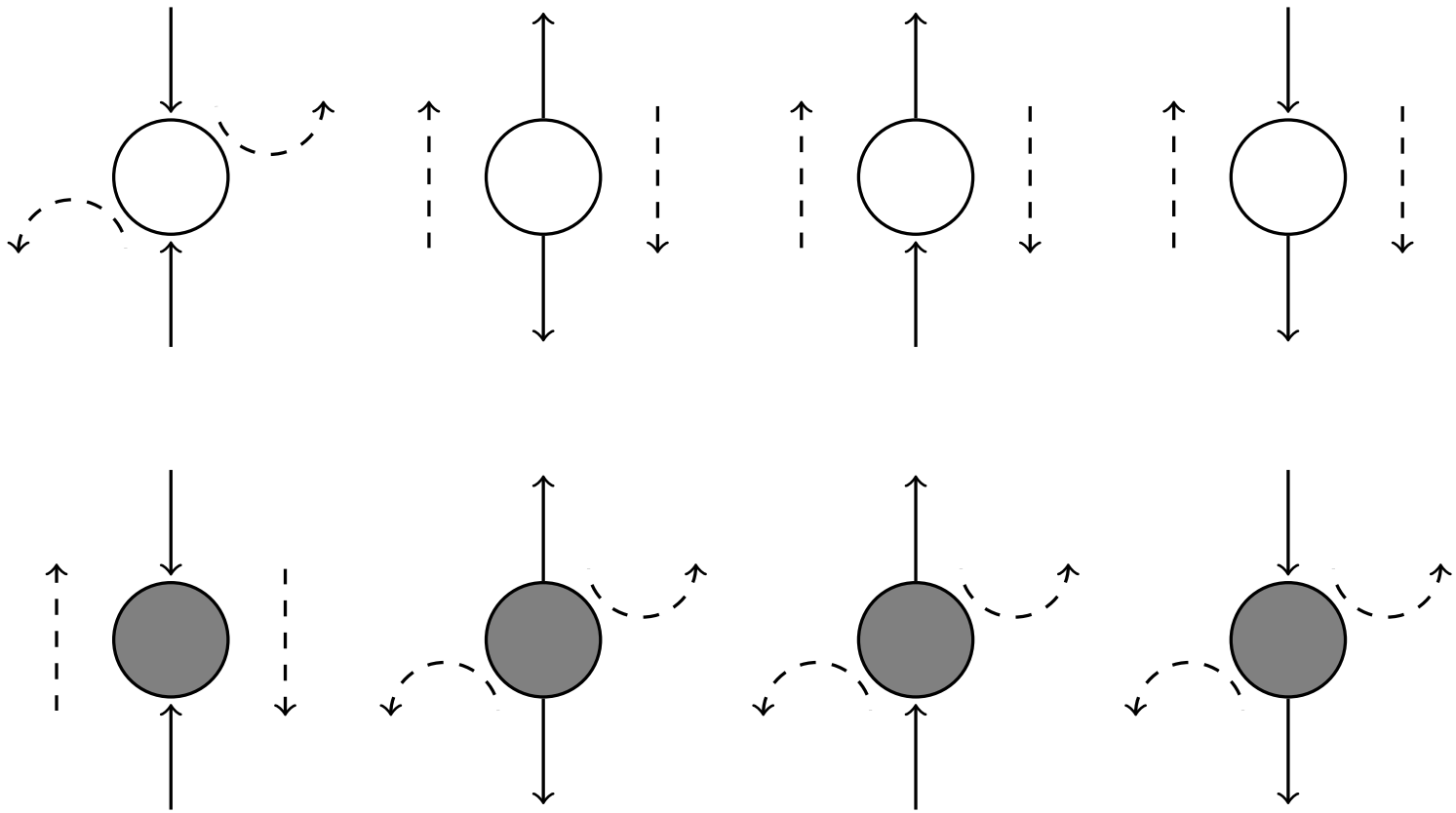
- either passes through a V-structure on B ($\rightarrow B \leftarrow$) and neither B nor its descendants are in $\mathcal{Z}$
- or it does *not* pass through V-structure on B ($\rightarrow B \rightarrow$ or $\leftarrow B \rightarrow$) but $B \in \mathcal{Z}$

We say that $\mathcal{Z}$ d-separates X from Y if every undirected path between X and Y is blocked by $\mathcal{Z}$.

If every variable in $\mathcal{X}$ is d-separated from every variable in $\mathcal{Y}$ by some variables in $\mathcal{Z}$ then $\mathcal{X} \perp \mathcal{Y} | \mathcal{Z}$.

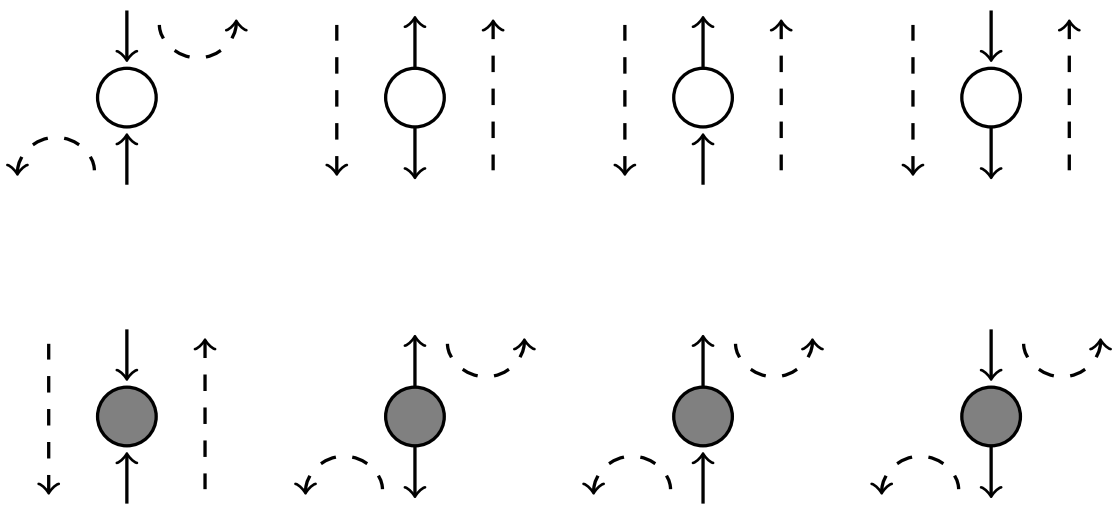
Bayes ball (http://uai.sis.pitt.edu/papers/98/p480-shachter.pdf)

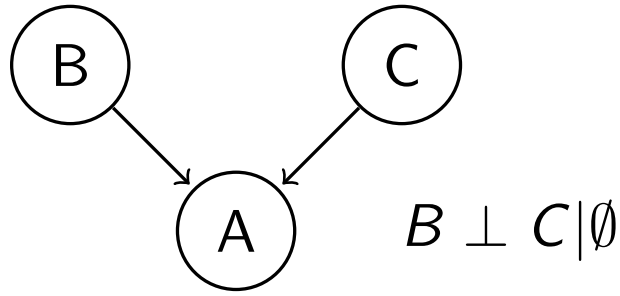
You want to check if $X \perp Y | \mathcal{Z}$. Imagine passing a "ball" from a node to a node, if the ball cannot make it from X to Y then you can assert conditional independence. Nodes in $\mathcal{Z}$ are gray and rules for passing are:



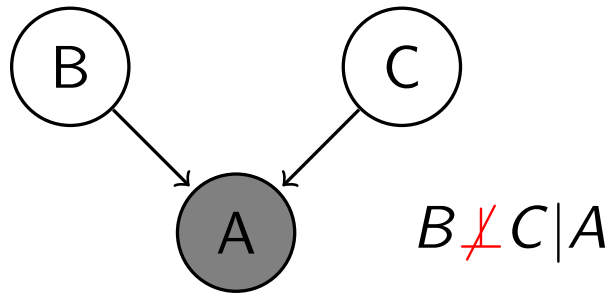
V-structure and Bayes ball

$X \perp Y | \mathcal{Z}$ is equivalent to no path from X to Y using these rules (nodes in $\mathcal{Z}$ are gray)

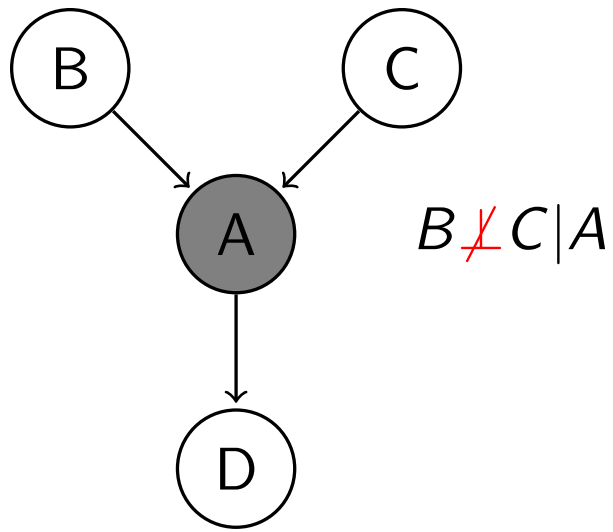




$B \perp C | \emptyset$

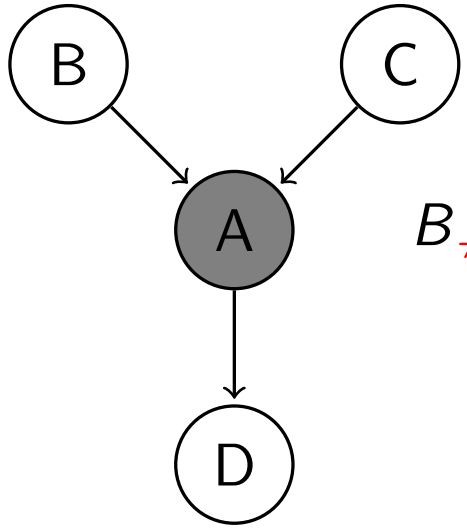


$B \not\perp C | A$



$B \not\perp C | A$

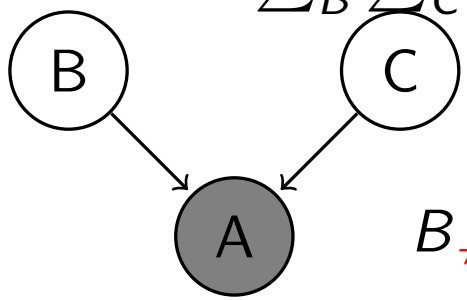
V-structure with a hanging unobserved variable



$B \not\perp C|A$

$$p(B,C|A) = \frac{p(A,B,C)}{p(A)} = \frac{\sum_D p(B)p(C)p(A|B,C)p(D|A)}{\sum_B \sum_C \sum_D p(B)p(C)p(A|B,C)p(D|A)}$$

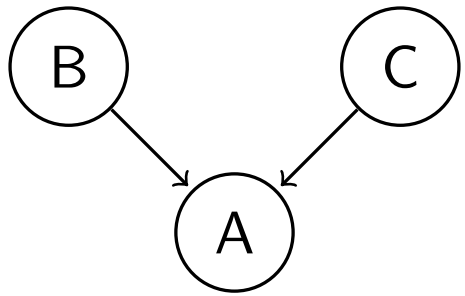
$$= \frac{p(B)p(C)p(A|B,C)\sum_D p(D|A)}{\sum_B \sum_C p(B)p(C)p(A|B,C)\sum_D p(D|A)}$$



$B \not\perp C|A$

V-structures and explaining away

Suppose we know of two competing explanations of an outcome.

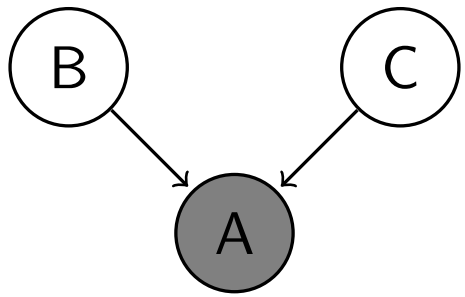


variables B and C are independent

$$\begin{aligned} p(B,C) &= \sum_A p(A|B,C)p(B)p(C) = p(B)p(C)\sum_A p(A|B,C) \\ &= p(B)p(C) \end{aligned}$$

V-structures and explaining away

But as soon as we observe A the variables B and C become dependent



$$
\begin{aligned}
p(B, C|A = a) &= \frac{p(A = a|B, C)p(B)p(C)}{\sum_B \sum_C p(A = a|B, C)p(B)p(C)} \\
&\neq p(B)p(C)
\end{aligned}
$$

Examples of explaining away

Outcome	Explanation 1	Explanation 2
wet grass	sprinkler	rain
student admitted	student brainy	student athletic
house jumps	truck hits house	earthquake

Also called Berkson's paradox.

Worked out example of explaining away

E - earthquake, T - truck hits the house, H - house moves

$$p(E=1)=0.01 \quad p(T=1)=0.01 \quad p(H=1|E,T)=\begin{cases}1\\0, \text{otherwise.}\end{cases}$$

We will compare $p(T=1|H=1)$ and $p(T=1|H=1,E=1)$ and see that if house moved, the probability that a truck hit it drops if there was an earthquake[3].

$$\begin{aligned} p(T=1|H=1) &= 0.5025 \\ p(T=1|H=1,E=1) &= 0.01 \end{aligned}$$

We will do in a pedestrian way in class.

[3] So, perfect storms are unlikely; no, it does not pour when it rains, and blues songs are a bit overdramatic ... in case you wondered

Undirected models

So far we looked at the graphical models specified in terms of a DAG (directed acyclic graph) and conditional probabilities. These were Bayes(ian belief) net(works).

We will come back to them, but we also want to familiarize ourselves with other common representations.
Two types of *undirected* graphical models we will consider today

- Markov random fields (markov nets)
- Factor graphs

Markov Random Fields

The building blocks for Bayes Nets were conditional probabilities.

For MRFs the building blocks are potentials, e.g.

$$\phi_i(X_{C_i}).$$

where potential operates on a subset of variables whose indices are in set C_i.

Sets C_i correspond to cliques of the graph.

The set of random variables X_{C_i} is sometimes called scope of potential ϕ_i (paralleling function scope).

MRFs

The full joint probability is then given as

$$p(X) = \frac{1}{Z} \prod_i \phi_i(X_{C_i}).$$

Note that the subsets C_i for different i are not required to be disjoint.

The constant Z is called the partition function or normalization constant.

Just to tie things together a little bit suppose we had d potentials

$$\phi_i(X_i) = \exp\left\{-\frac{1}{2\sigma^2}(X_i^2 - 2X_i\mu + \mu^2)\right\}$$

then $Z = (2pi)^{\frac{d}{2}} \prod_i \sigma_i$ and this is just a Gaussian distribution written in an MRF form.

Constructing a graph for an MRF

The graph is constructed by connecting each pair of nodes corresponding to random variables X_i,X_j if $\exists k, i \in C_k, j \in C_j$.

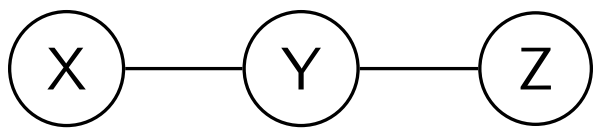
For example an MRF

$$p(X, Y, Z) = \frac{1}{Z}\phi(X, Y)\psi(Y, Z)$$

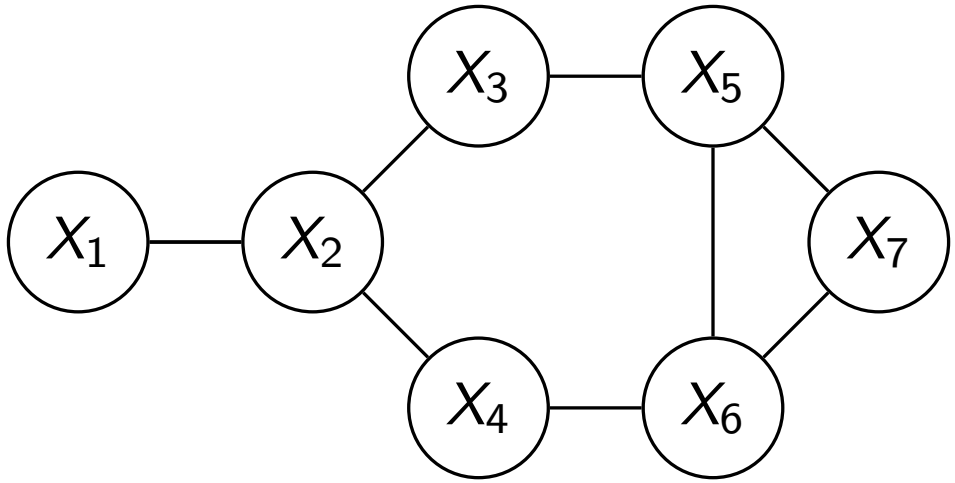
and

$$Z = \sum_{X,Y,Z} \phi(X, Y)\psi(Y, Z)$$

corresponds to an undirected graph



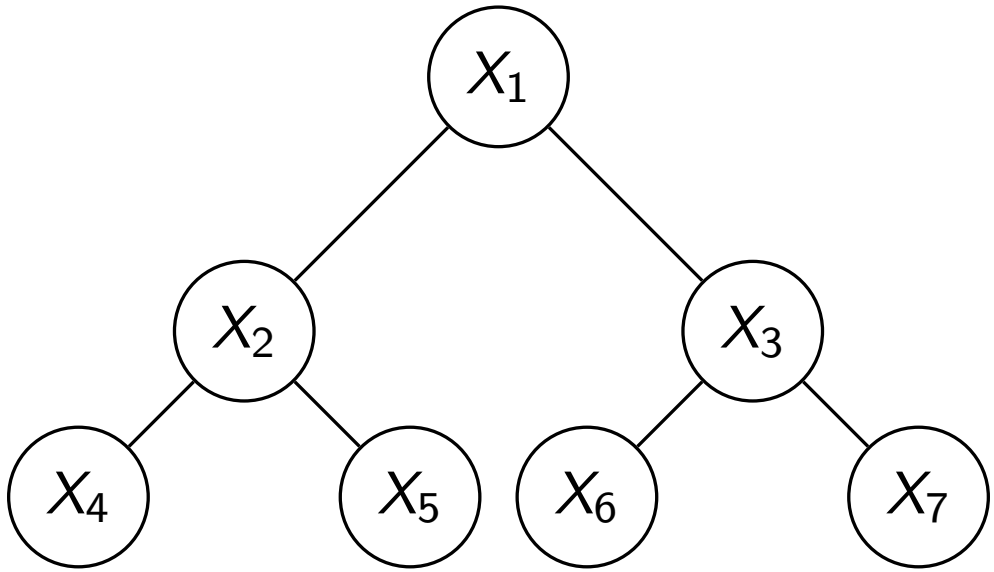
More examples of MRFs



This graphical model specifies a joint distribution

$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) &= \frac{1}{Z} \prod_i \phi(X_{C_i}) \\
&= \frac{1}{Z} \phi_1(X_1)\phi_2(X_2, X_1)\phi_3(X_3, X_2)\phi_4(X_4, X_2) \\
&\quad \phi_5(X_5, X_3)\phi_6(X_6, X_4)\phi_7(X_7, X_5, X_6)
\end{aligned}
$$

Another example of an MRF



$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) &= \frac{1}{Z} \prod_i \phi(X_{C_i}) \\
&= \frac{1}{Z} \phi_1(X_1)\phi_2(X_2, X_1)\phi_3(X_3, X_1)\phi_4(X_4, X_2) \\
&\quad \phi_5(X_5, X_2)\phi_6(X_6, X_3)\phi_7(X_7, X_3)
\end{aligned}
$$

Differences from Bayes Nets

No worrying about orderings, conditional probabilities, acyclicity.

Conditional independence is much easier: $X \perp Y | \mathcal{Z}$ if every path between X and Y contains a node corresponding to a variable in $\mathcal{Z}$

And immediately from this if we let **neighbors**(X) denote the set of neighbors of X then

$$X \perp Y | \textbf{neighbors}(X)$$

for $Y \notin X \cup \textbf{neighbors}(X)$. The set of neighbors is also called **Markov blanket**.

MRF applications

Very popular in computer vision; a bit less in comp bio.

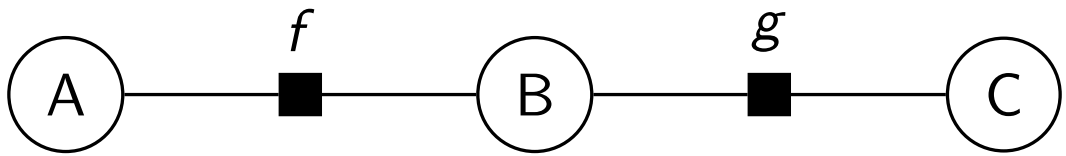
Can be seen as "energy" models where the energy of configuration is given by

$$E(X) = -\sum \log \phi_i(X_{C_i})$$

Gaussian MRFs have several nice properties that enable learning the model structure (the graph). We will come back to this.

Factor graphs: Another undirected representation

These are the easiest to get off the ground.



$$p(A, B, C) = \frac{1}{Z} f(A, B) g(B, C)$$

Two types of nodes: variable (one for each random variable) and factor nodes (one for each potential).

Factor graphs

Joint probability is given in the same form as in MRF

$$p(X) = \prod \phi_i(X_{C_i})$$

but the scope and the different factors are easier to read off of the graph.

Conditional independence for factor graphs

Neighbors as in MRF, if two variables appear in scope of the same potential.

Reading this off of the graph is much easier, since scope is obvious based on connections to factor node.

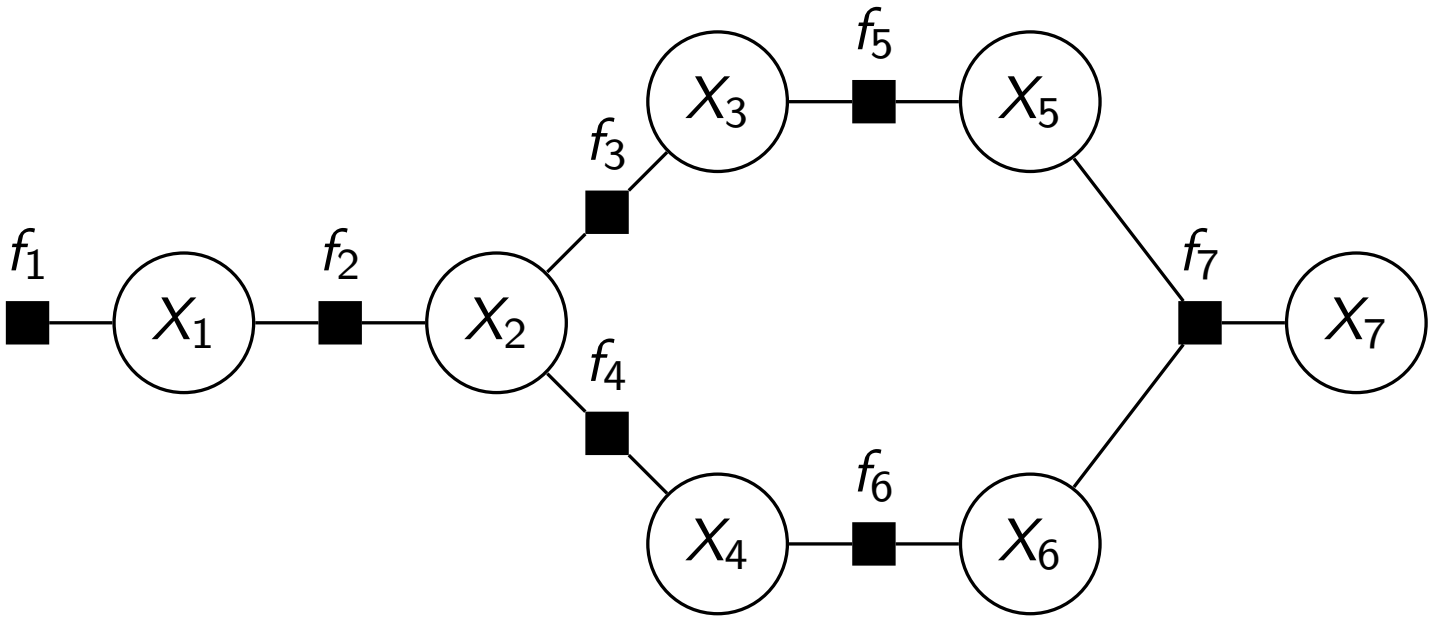
Conditional independence is easy again: $X \perp Y | \mathcal{Z}$ if every path between X and Y contains a node corresponding to variable $\mathcal{Z}$.

And in general

$$X \perp Y | \textbf{neighbors}(X)$$

for $Y \notin X \cup \textbf{neighbors}(X)$.

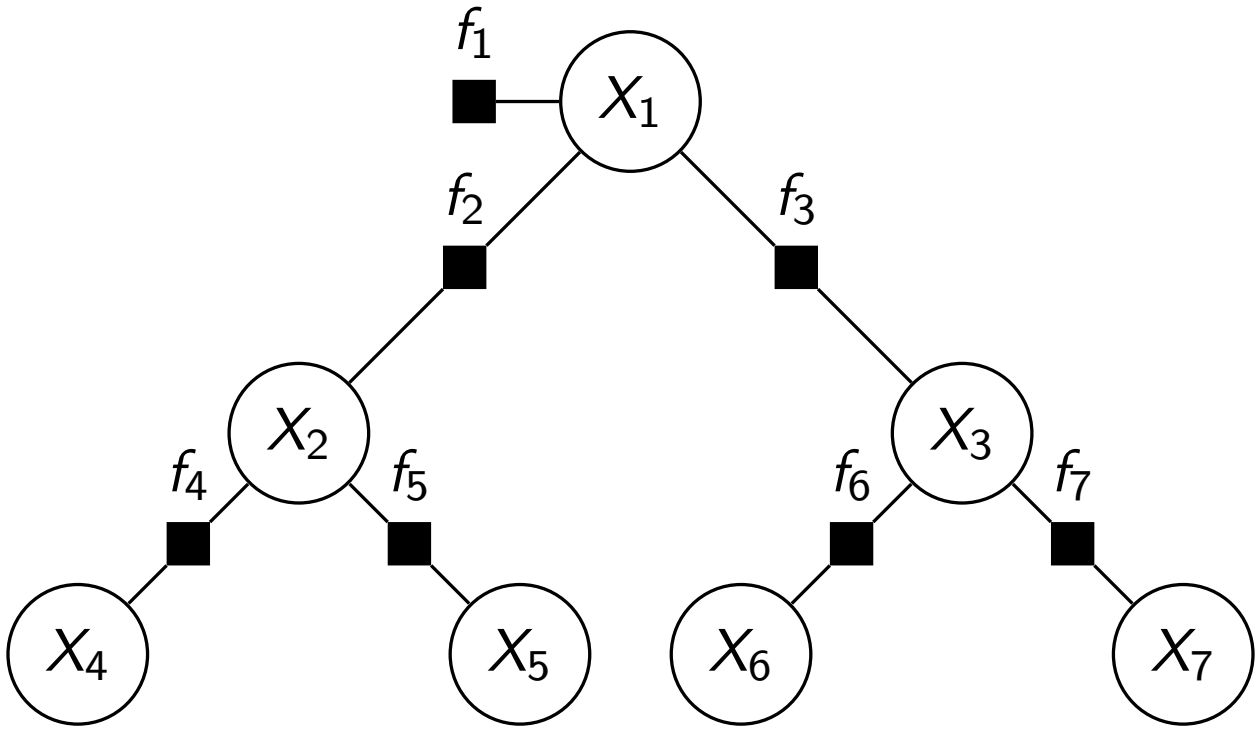
More examples of Factor Graphs



This graphical model specifies a joint distribution

$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) &= \frac{1}{Z} \prod_i \phi(X_{C_i}) \\
&= \frac{1}{Z} f_1(X_1) f_2(X_2, X_1) f_3(X_3, X_2) f_4(X_4, X_2) \\
&\quad f_5(X_5, X_3) f_6(X_6, X_4) f_7(X_7, X_5, X_6)
\end{aligned}
$$

Another example of a Factor Graph



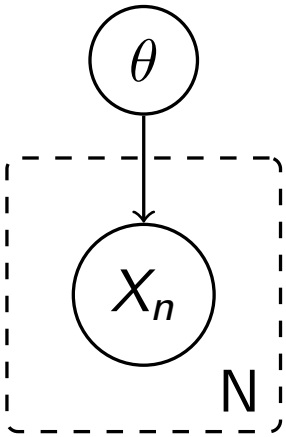
$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5, X_6, X_7) &= \frac{1}{Z} \prod_i \phi(X_{C_i}) \\
&= \frac{1}{Z} f_1(X_1) f_2(X_2, X_1) f_3(X_3, X_1) f_4(X_4, X_2) \\
&\quad f_5(X_5, X_2) f_6(X_6, X_3) f_7(X_7, X_3)
\end{aligned}
$$

Plate notation

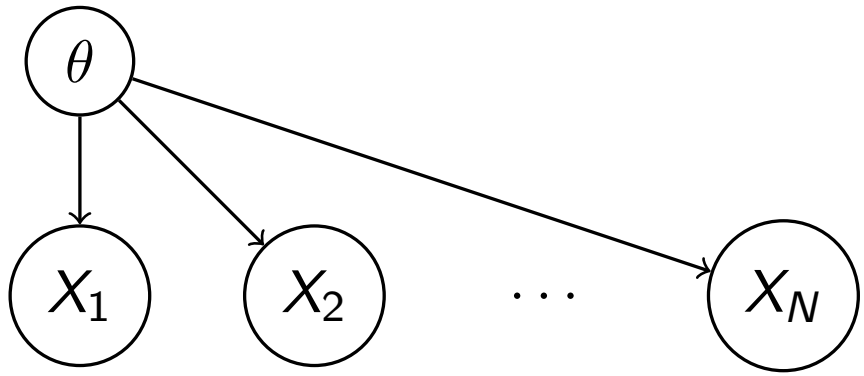
In some cases we may have a particular model structure that is repeatedly reused.

- Same parameter reused across data instances
- Same structure reused across instances
- Any other repetitive regularity

Plate notation: IID



same as

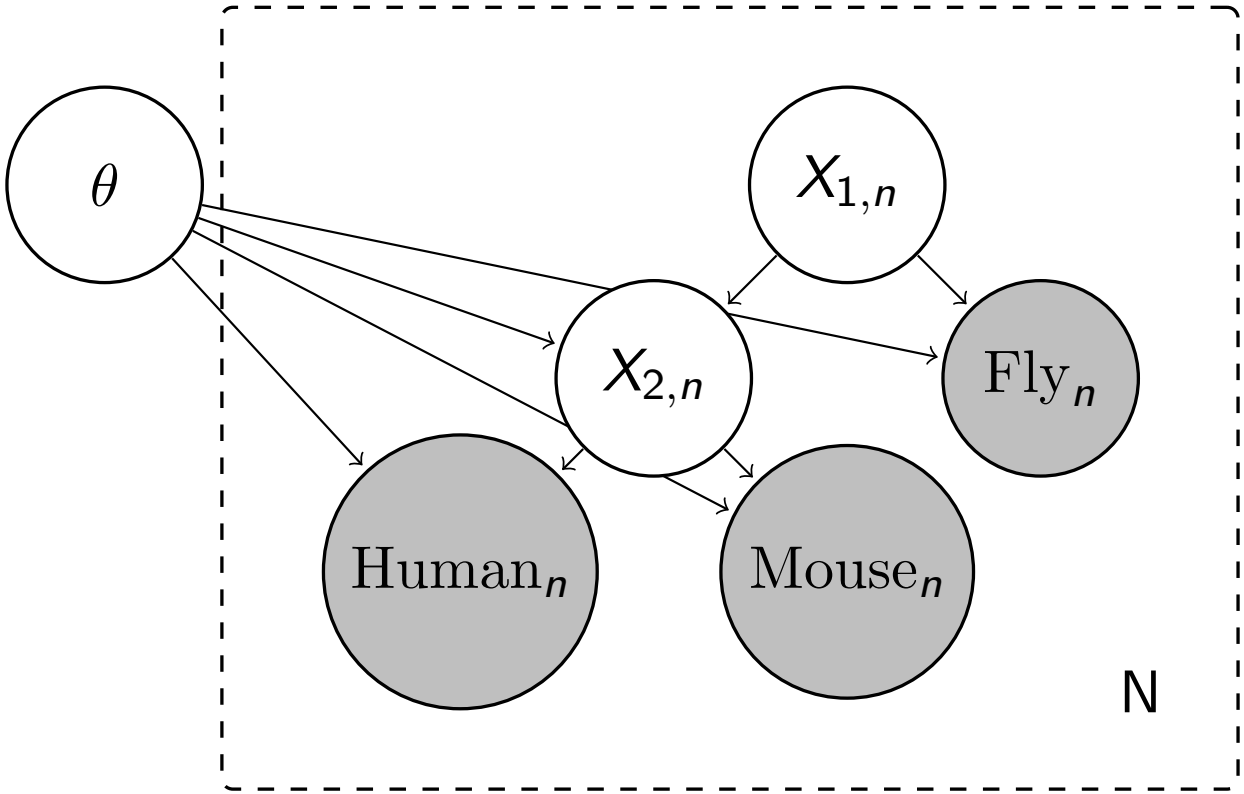


$$p(X_1, X_2, \ldots, X_N|\theta) = p(X_1|\theta)p(X_2|\theta)\ldots p(X_N|\theta) = \prod_i p(X_i|\theta)$$

Random variables X_i are independently and identically distributed.

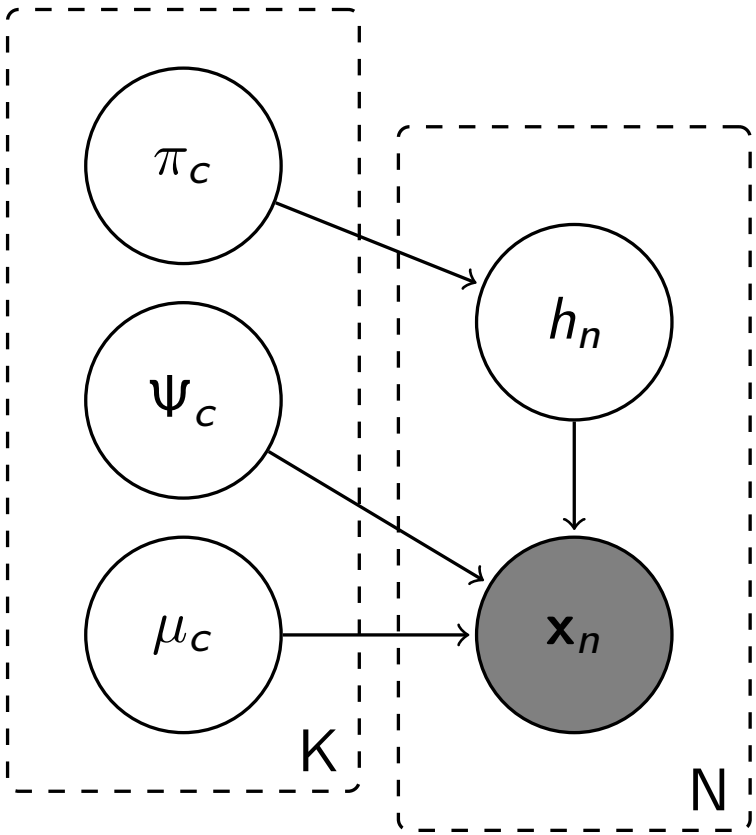
Phylogenies

An early example of a comp bio graphical model is a phylogeny



Each of the random variables is a nucleotide. The conditional probabilities $p(X_{i,n}|X_{j,n})$ are specified by a single 4×4 mutation matrix θ shared across N positions in aligned genome sequences.

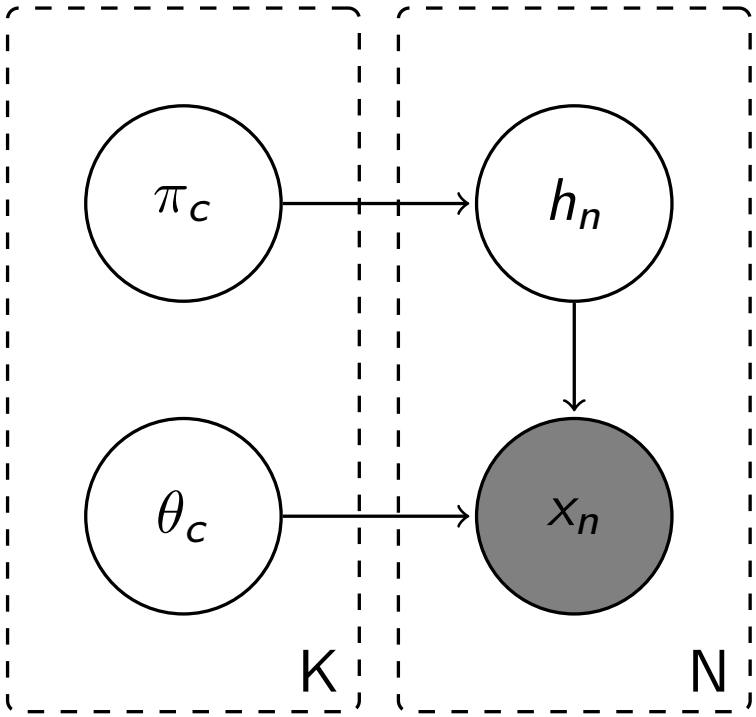
Graphical model for MoG



$$p(h_n = k) = \pi_k$$
$$p(\mathbf{x}_n | h_n = k) = \mathcal{N}(\mathbf{x}_n | \mu_k, \Psi_k)$$

Mixture of Gaussians model for N data points and K classes (mixture components)

Graphical model for MoPWM



$$p(h_n = k) = \pi_k$$
$$p(\mathbf{x}_n | h_n = k) = \prod_j \prod_v \theta_{k,j,v}^{[\mathbf{x}_{n,j}=v]}$$

Mixture of PWMs model for N data points and K classes (mixture components)

We did ...

- Conditional independence
- Graphical models (Bayes Nets,MRFs,Factor graphs)