CS540 Machine learning
Directed graphical models

Outline

- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling

Conditional independence

- Recall the naïve Bayes assumption

$$X_j \perp X_k | Y$$

- This lets us factorize the class conditional density

$$p(\mathbf{x}|y) = \prod_{j=1}^{n_x} p(x_j|y)$$

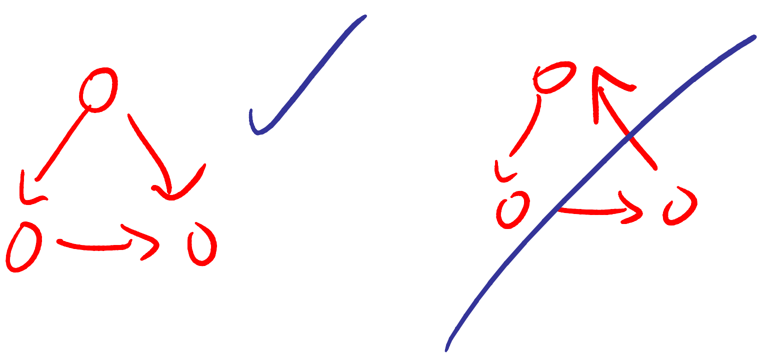
- Hence the joint distribution is

$$p(\mathbf{x}, y) = p(y) \prod_{j=1}^{n_x} p(x_j|y)$$

- Graphical models are ways to represent CI statements pictorially. This provides a compact way to define joint probability distributions.

Kinds of graphical models

- Undirected graphical models – aka Markov Random fields – see later in class.
- Directed graphical models – aka Bayesian (belief) networks.
  - BNs require that the graph is a DAG (directed acyclic graphs).
  - No directed cycles allowed.

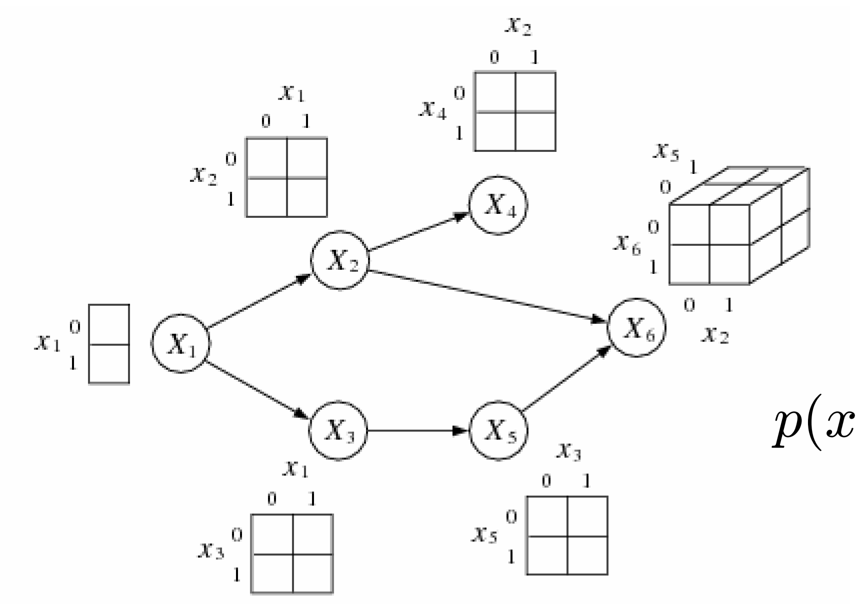


Directed graphical models

- A prob distribution factorizes according to a DAG if it can be written as

$$p(\mathbf{x}) = \prod_{j=1}^{d} p(x_j|\mathbf{x}_{\pi_j})$$

where π_j are the parents of j , and the nodes are ordered topologically (parents before children).



Each row of the conditional probability table (CPT) defines the distribution over the child's values given its parents values. The model is locally normalized.

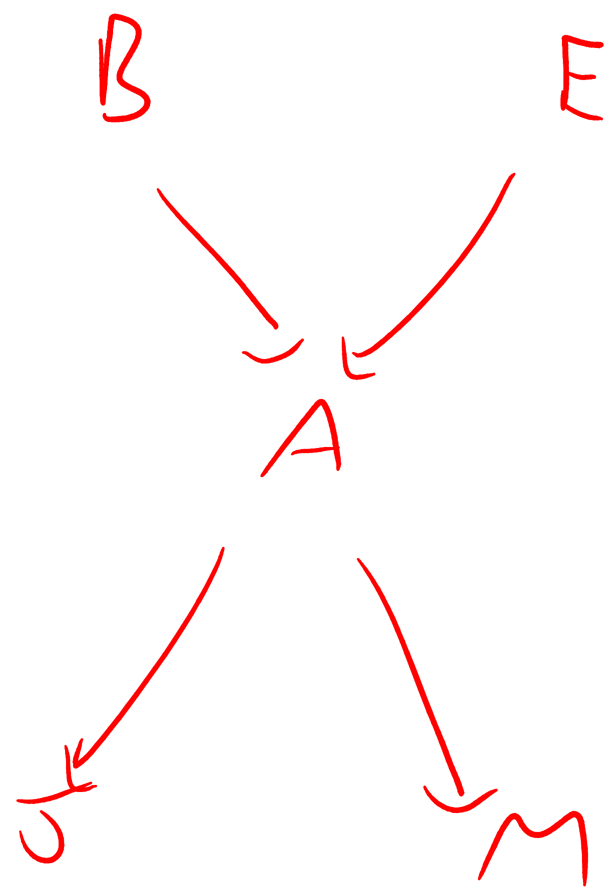
$$\begin{aligned} p(x_{1:6}) &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3) \\ &\quad p(x_5|x_2,x_3)p(x_6|x_2,x_5) \end{aligned}$$

Example model

$$p(B,E,A,J,M) = p(B)p(E)p(A|B,E)p(J|A)p(M|A)$$

0	1
0.001	0.999

0	1
0.002	0.998



B	E	p(A=0)	p(A=1)
0	0	0.999	0.001
0	1	0.71	0.29
1	0	0.06	0.94
1	1	0.05	0.95

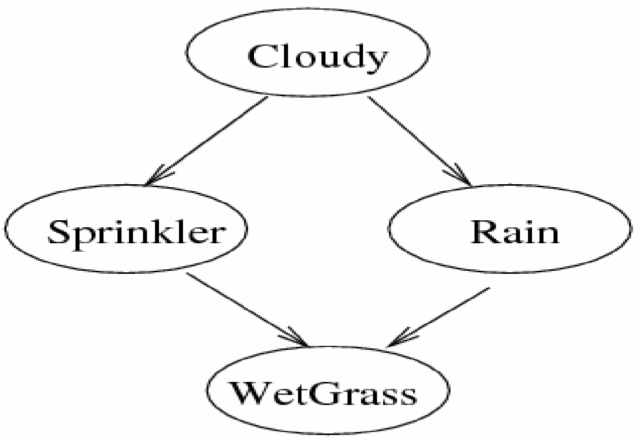
A	p(J=0)	p(J=1)
0	0.10	0.90
1	0.95	0.05

A	p(M=0)	p(M=1)
0	0.99	0.01
1	0.30	0.70

Source: Russell & Norvig

Example model

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

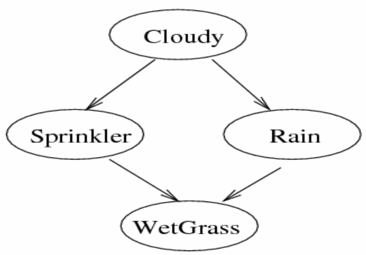
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

$$p(C, S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R)$$

Joint distribution

$$p(C,S,R,W) = p(C)p(S|C)p(R|C)p(W|S,R)$$

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

c	s	r	w	prob
0	0	0	0	0.200
0	0	0	1	0.000
0	0	1	0	0.005
0	0	1	1	0.045
0	1	0	0	0.020
0	1	0	1	0.180
0	1	1	0	0.001
0	1	1	1	0.050
1	0	0	0	0.090
1	0	0	1	0.000
1	0	1	0	0.036
1	0	1	1	0.324
1	1	0	0	0.001
1	1	0	1	0.009
1	1	1	0	0.000
1	1	1	1	0.040

Inference

- Prior that sprinkler is on

$$p(S=1)=\sum_{c=0}^{1}\sum_{r=0}^{1}\sum_{w=0}^{1}p(C=c,S=1,R=r,W=w)=0.3$$

- Posterior that sprinkler is on given that grass is wet

$$p(S=1|W=1)=\frac{p(S=1,W=1)}{p(W=1)}=0.43$$

- Posterior that sprinkler is on given that grass is wet and it is raining

$$p(S=1|W=1,R=1)=\frac{p(S=1,W=1,R=1)}{p(W=1,R=1)}=0.19$$

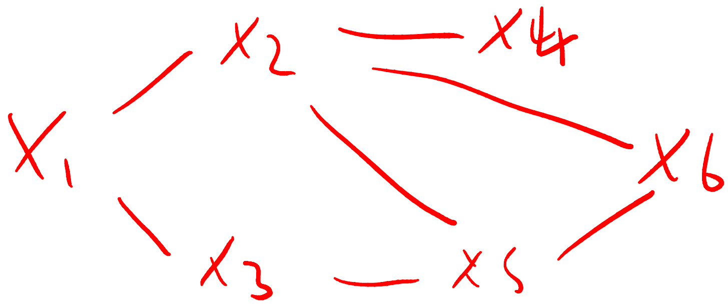
Explaining away!

Outline

- Undirected graphical models
- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling

Graph separation

- We say S *separates* A and B in G if, when we remove edges connected to S, all paths from A to B are blocked



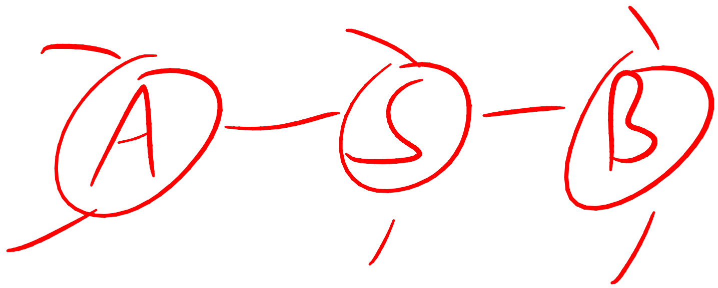
eg {2,5} separates 1 and 4

- Hammersley-Clifford Theorem: if p(x)>0 for all x, and p factorizes over G, then graph separation iff conditional independence

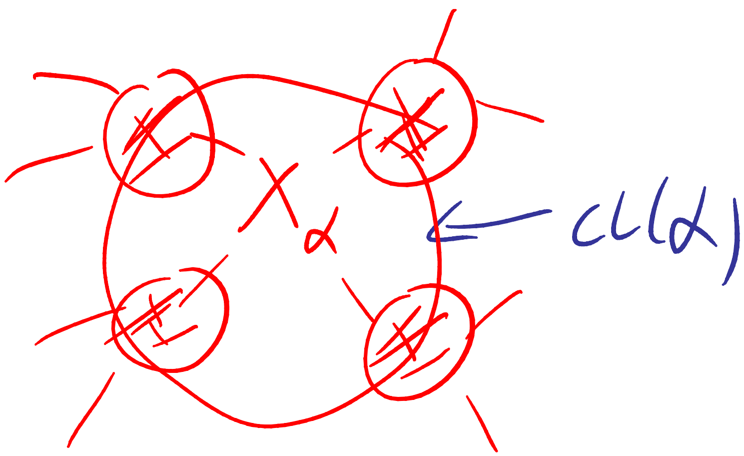
$$A \perp_G B|S \Leftrightarrow A \perp_p B|S$$

Markov properties of UGMs

- Global $A \perp B | S$



- Local $\alpha \perp V \setminus cl(\alpha) | bd(\alpha)$



bd = boundary,
cl = closure = boundary + node

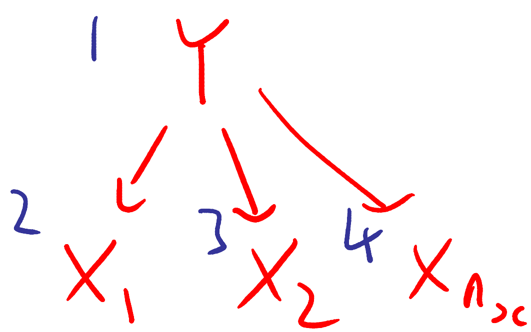
A node is independent of the rest given its Markov blanket

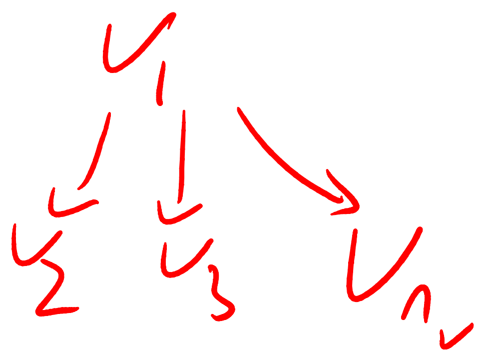
Conditional independence properties of DAGs

- For UGMs, independence $\equiv$ separation.
- For DGMs, independence $\equiv$ d-separation.
- Alternatively, we can convert a DGM to a UGM and use simple separation.

DAGs

- DAGs admit a total ordering (parents before children).
- Local Markov property: A node is independent of its predecssors given its parents.

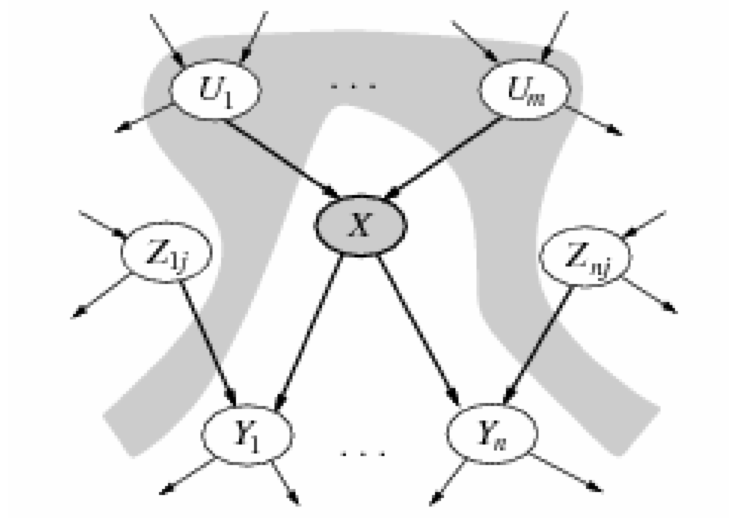




$$X_j \perp X_{1:j} \mid Y$$

Local directed Markov property

- A node is independent of its non-descendants given its parents



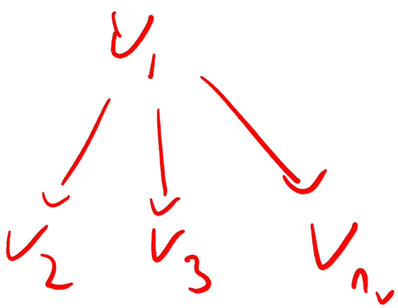
Chain rule

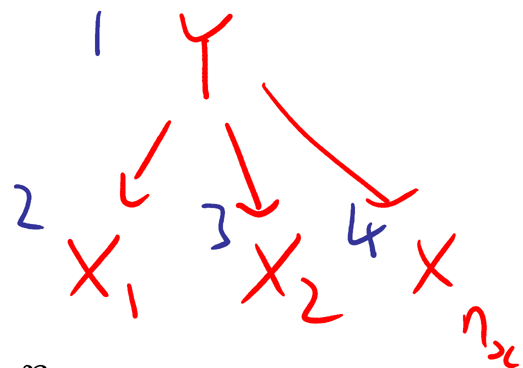
- By the chain rule

$$p(v_{1:n_v}) = p(v_1)p(v_2|v_1)p(v_3|v_1,v_2)\ldots p(v_{n_v}|v_{1:n_v-1}$$

- By the local Markov property

$$p(v_{1:n}) = p(v_1)p(v_2|v_{\pi_2})p(v_3|v_{\pi_3})\ldots p(v_n|\overset{v}{\cancel{z}}_{\pi_n})$$



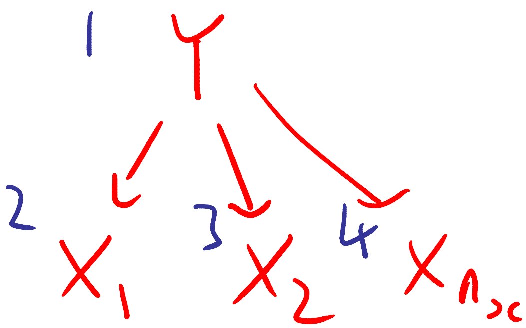


$$p(y, x_{1:n_x}) = p(y)\prod_{j=1}^{n_x} p(x_j|y)$$

Local Markov property is not enough

- NB property is $X_j \perp X_k \mid Y$ for all k, including k > j
- But local Markov property only tells us $X_j \perp X_k \mid Y$ for k < j
- Want to be able to answer the following for any sets of variables a,b,c: $Z_a \perp Z_b \mid Z_c$?

$V_a \perp V_b \mid V_c$



$X_j \perp X_{1:j} \mid Y$

Global Markov property

- By chaining together local independencies, one can infer global independencies.
- The general definition/ algorithm is complex, so we will break it into pieces.

Chains

- Consider the chain

$X \to Y \to Z$

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

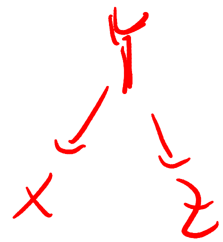
- If we condition and y, x and z are independent

$$\begin{aligned} p(x, z|y) &= \frac{p(x)p(y|x)p(z|y)}{p(y)} \\ &= \frac{p(x, y)p(z|y)}{p(y)} \\ &= p(x|y)p(z|y) \end{aligned}$$

$X \to Y \to Z$

Tents

- Consider the "tent"

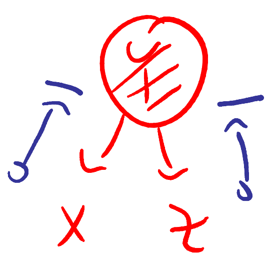


$$p(x, y, z) = p(y)p(x|y)p(z|y)$$

- Conditioning on Y makes X and Z independent

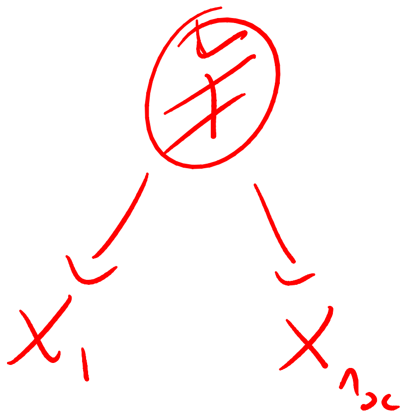
$$p(x, z|y) = \frac{p(x, y, z)}{p(y)}$$

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



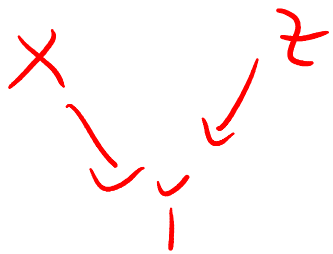
Naïve Bayes assumption

- Conditional on class, features are independent



V-structure

- Consider the v-structure



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

- X and Z are unconditionally independent

$$p(x, z) = \int p(x, y, z)dy = \int p(x)p(z)p(y|x, z)dy = p(x)p(z)$$

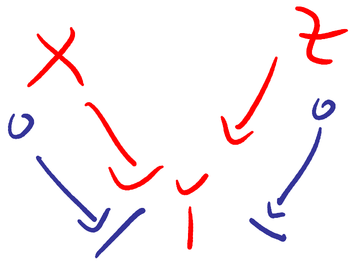
but are conditionally dependent

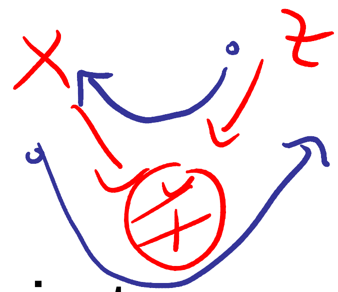
$$p(x, z|y) = \frac{p(x)p(z)p(y|x, z)}{p(y)} \neq f(x)g(z)$$

Explaining away

- Consider the v-structure

$X \perp Z$





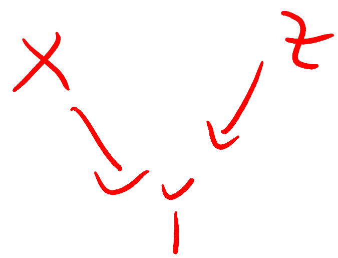
$X \not\perp Z \mid Y$

- Let X, Z $\in$ {0,1} be iid coin tosses.
- Let Y = X + Z.
- If we observe Y, X and Z are coupled.

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	2

Explaining away

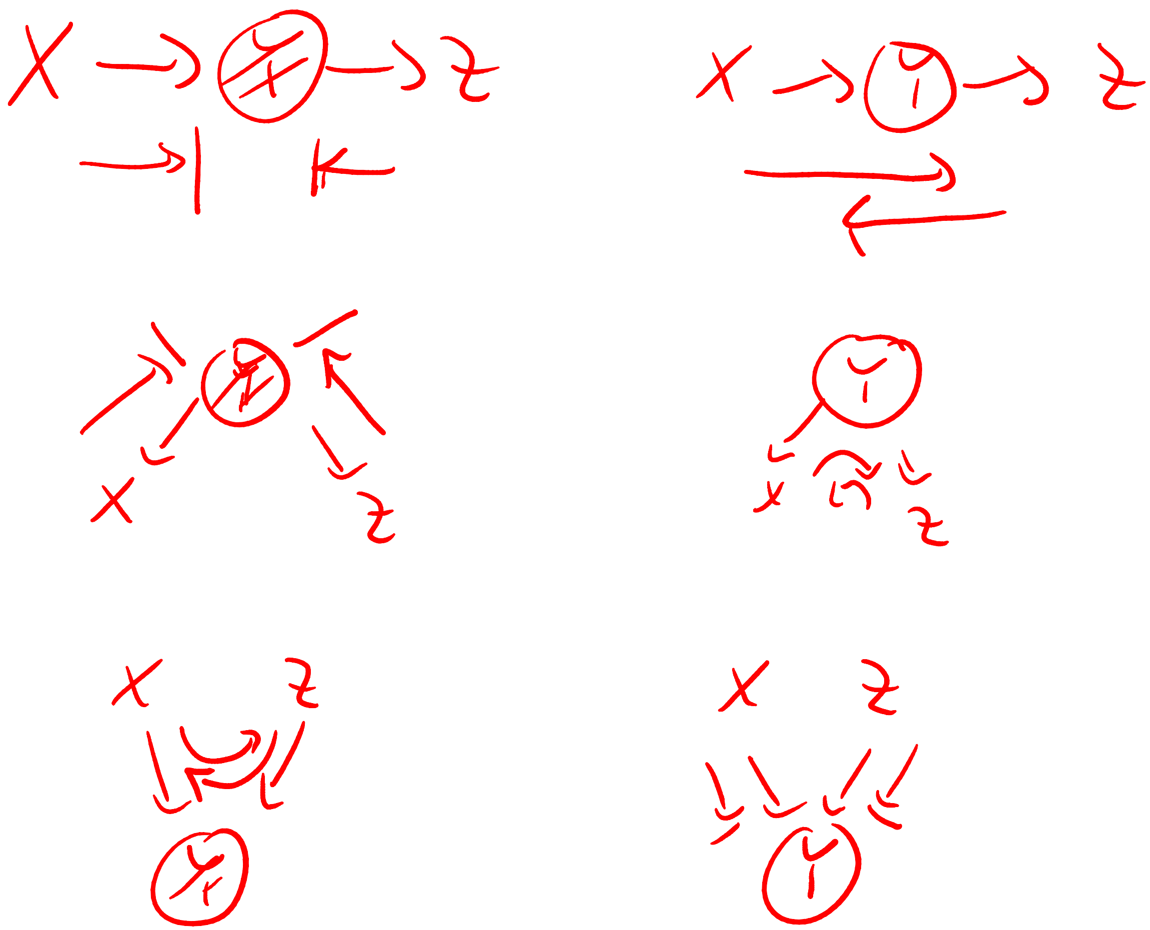
- Let Y = 1 iff burglar alarm goes off,
- X=1 iff burglar breaks in
- Z=1 iff earthquake occurred



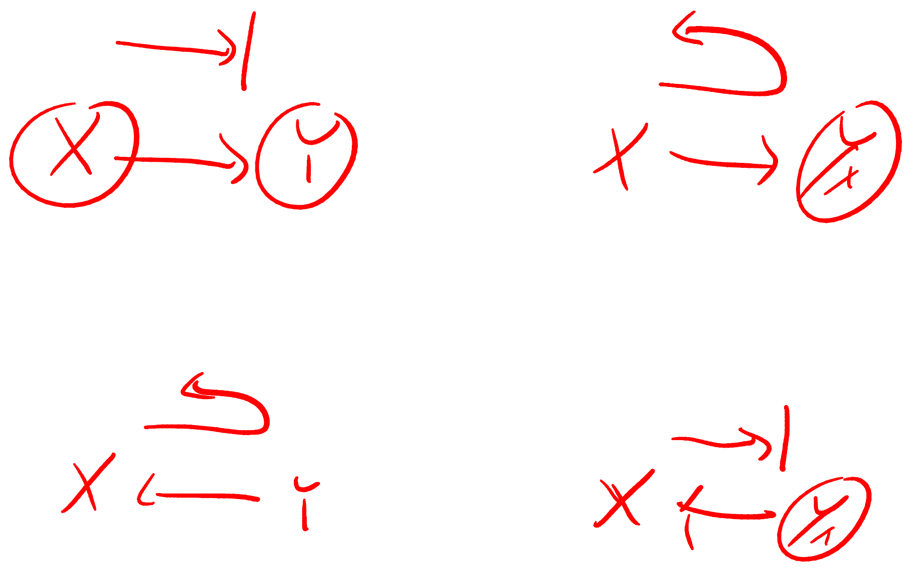
- X and Z compete to explain Y, and hence become dependent
- Intuitively, $p(X=1|Y=1) > p(X=1|Y=1,Z=1)$

Bayes Ball Algorithm

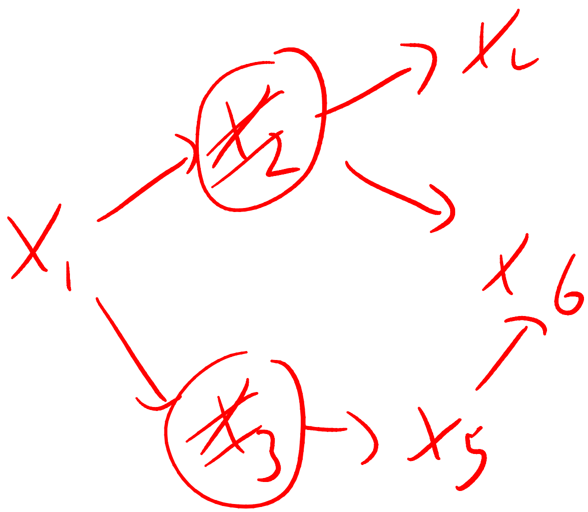
- $Z_A \perp Z_B \mid Z_C$ if every variable in A is d-separated from every variable in B when we shade the variables in C



Boundary conditions

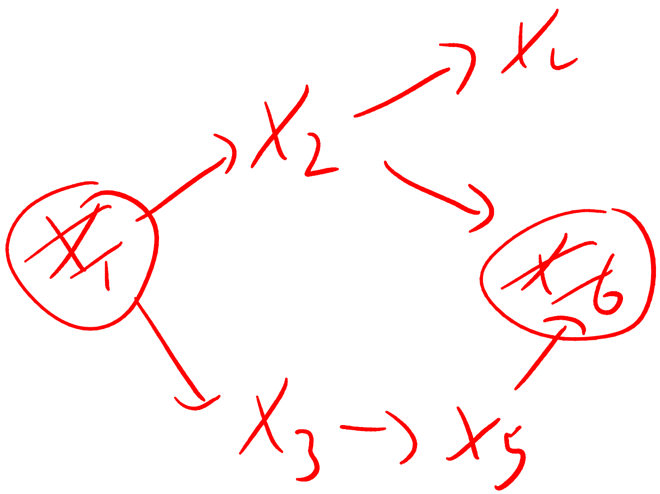


Example



$X_1 \perp X_6 \mid X_2, X_3$?

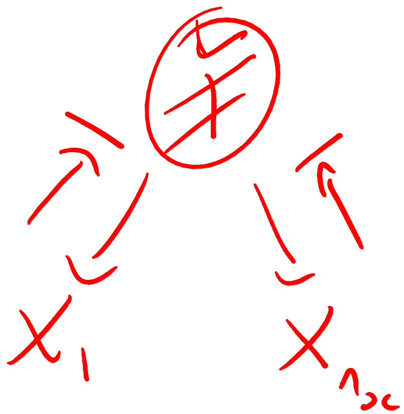
Example



$X_2 \perp X_3 \mid X_1, X_6$?

Naïve Bayes assumption

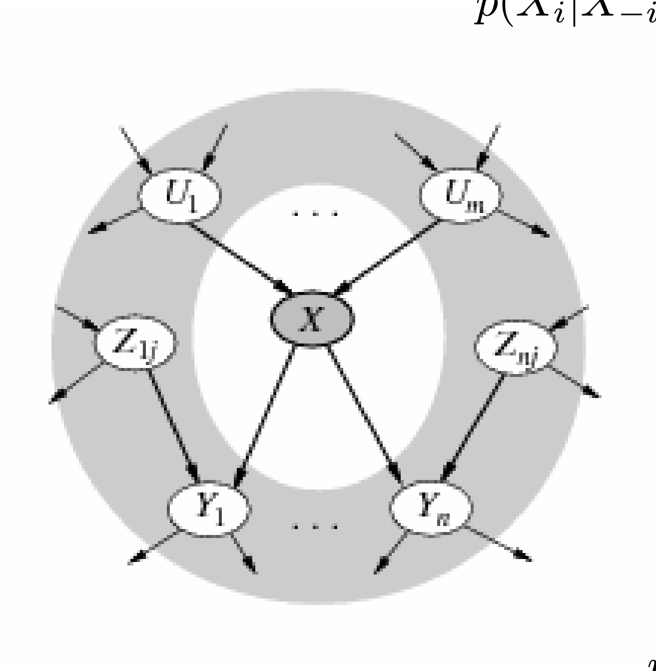
- Conditional on class, features are independent



$$X_j \perp X_k \mid Y$$

Markov blankets for DAGs

- The Markov blanket of a node is the set that renders it independent of the rest of the graph.
- This is the parents, children and co-parents.



$$
\begin{aligned}
p(X_i|X_{-i}) &= \frac{p(X_i, X_{-i})}{\sum_x p(X_i, X_{-i})} \\
&= \frac{p(X_i, U_{1:n}, Y_{1:m}, Z_{1:m}, R)}{\sum_x p(x, U_{1:n}, Y_{1:m}, Z_{1:m}, R)} \\
&= \frac{p(X_i|U_{1:n})[\prod_j p(Y_j|X_i, Z_j)]P(U_{1:n}, Z_{1:m}, R)}{\sum_x p(X_i = x|U_{1:n})[\prod_j p(Y_j|X_i = x, Z_j)]P(U_{1:n}, Z_{1:m}, R)} \\
&= \frac{p(X_i|U_{1:n})[\prod_j p(Y_j|X_i, Z_j)]}{\sum_x p(X_i = x|U_{1:n})[\prod_j p(Y_j|X_i = x, Z_j)]}
\end{aligned}
$$

$$p(X_i|X_{-i}) \propto p(X_i|Pa(X_i)) \prod_{Y_j \in ch(X_i)} p(Y_j|Pa(Y_j))$$

Useful for Gibbs sampling

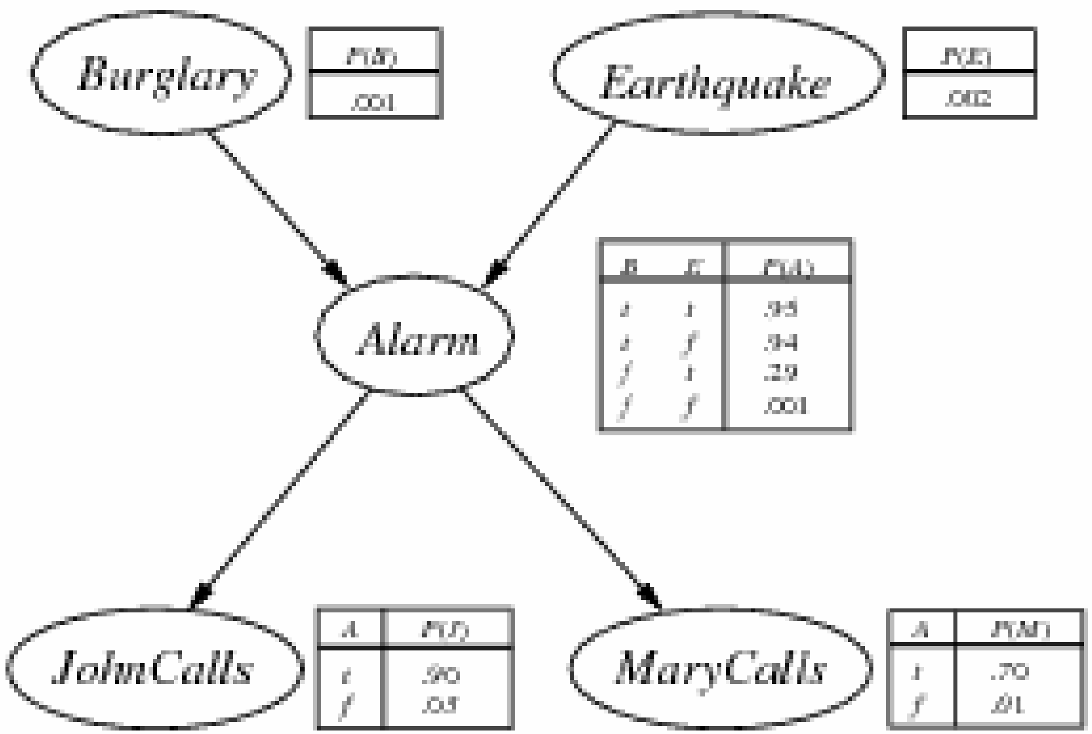
Outline

- Undirected graphical models
- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling

Example model

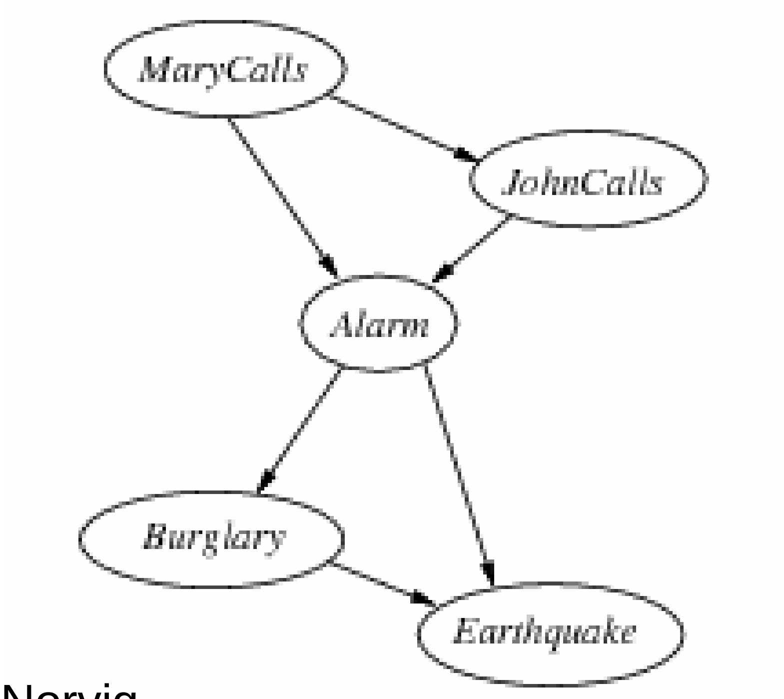
- Suppose the true distribution is

$$p(B, E, A, J, M) = p(B)p(E)p(A|B, E)p(J|A)p(M|A)$$



Choosing the "wrong" ordering

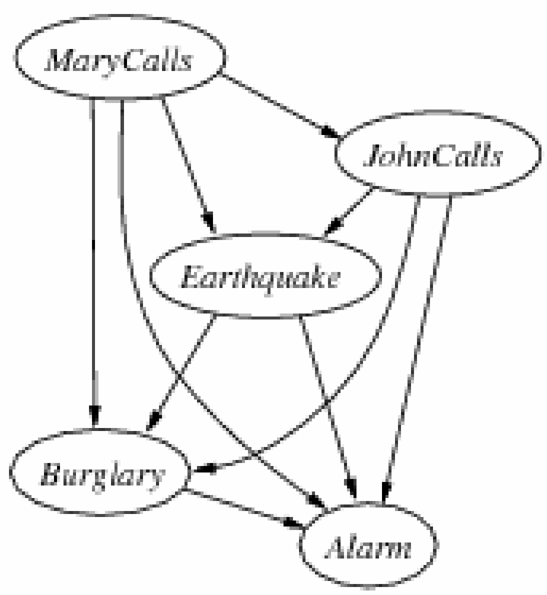
- If we choose the order MJABE, we get a more densely connected network, otherwise this will make independence statements that are not true.
- Eg in original model we have $E \perp M|A, \;\; E \perp J|A, \;\; E \not\perp B|A$ so we must connect E to B,A but not M,J



Source: Russell & Norvig

A worse ordering

- If we pick the order MJEBA, the graph becomes fully connected, and thus makes no independence statements (and therefore includes the true distribution).



(b)

Outline

- Undirected graphical models
- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling

Markov equivalence

- The following 3 graphs all assert the same set of conditional independencies, namely X indep Y | Z; hence they are equivalent

$X \rightarrow Y \rightarrow Z$ $\quad$ $X \leftarrow Y \leftarrow Z$ $\quad$ $X \leftarrow Y \rightarrow Z$

$X \not\perp Z$

$X \perp Z \mid Y$

This v-structure is not equivalent

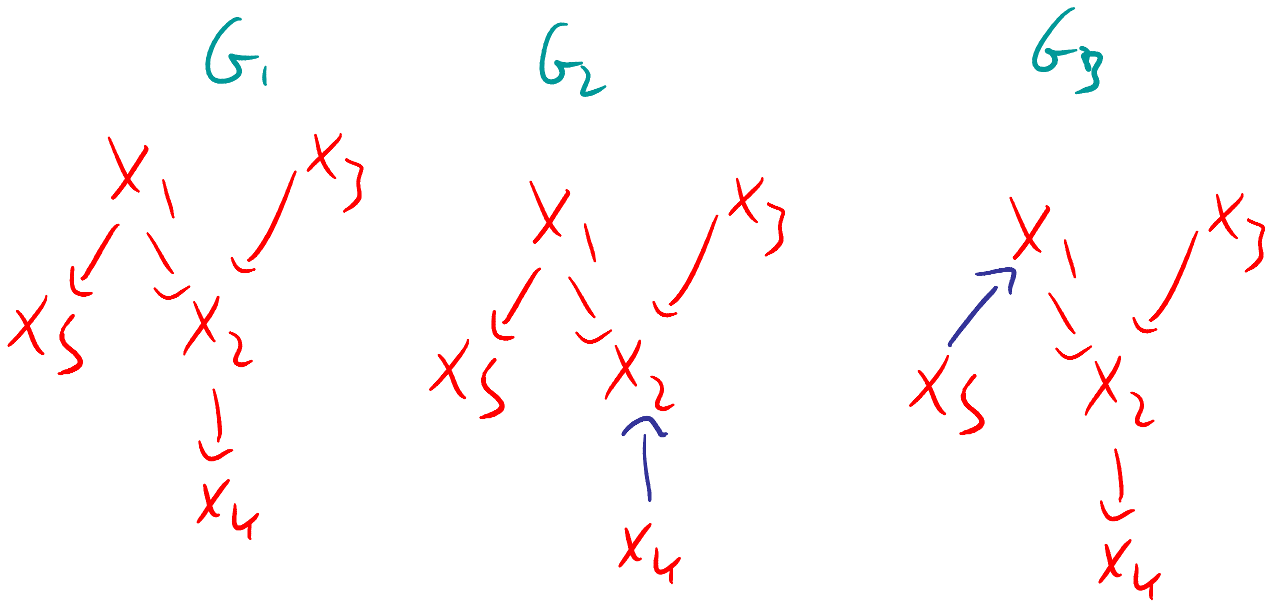
$X \rightarrow Y \leftarrow Z$

$X \perp Z$

$X \not\perp Z \mid Y$

Markov equivalence

- Thm: 2 DAGs are Markov equivalent iff they have the same undirected skeleton and the same set of v-structures

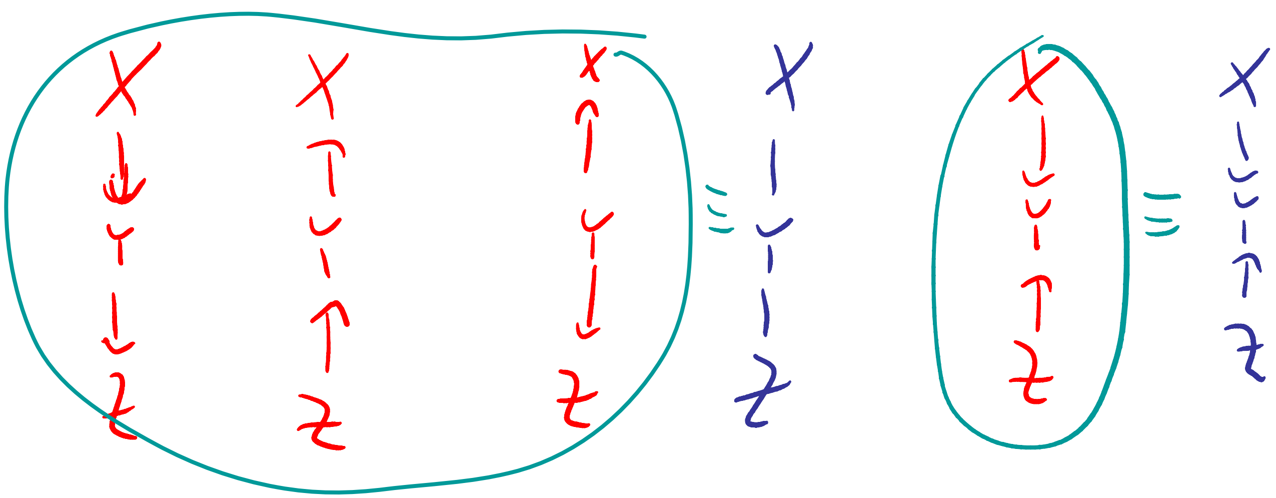


$G_1 \equiv G_2$?

$G_1 \equiv G_3$?

PDAGs

- We can uniquely represent each equivalence class using a partially directed acyclic graph (aka essential graph).
- This uses undirected edges if they are reversible, and directed edges if they are compelled.

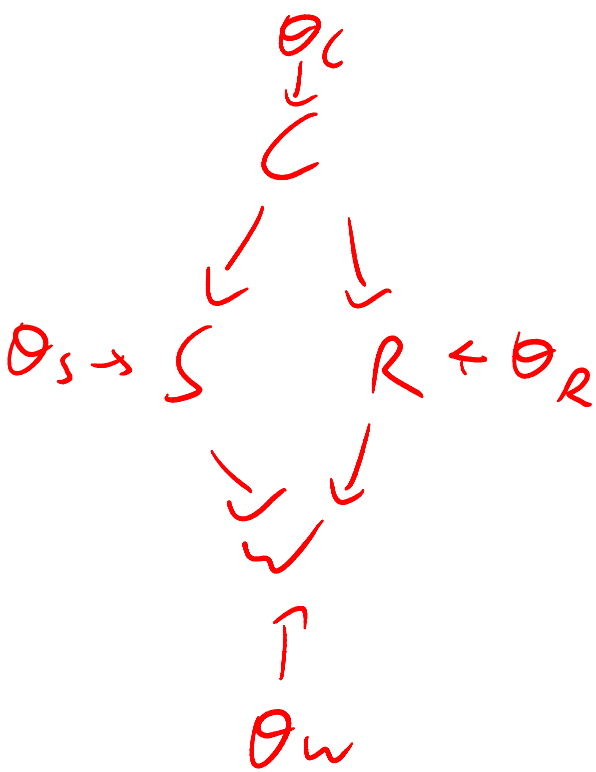


Outline

- Undirected graphical models
- Directed graphical models
- Conditional independence
- Effects of node ordering
- Markov equivalence
- Bayesian modeling

Parameter nodes

- If we treat the parameters as random variables, we can add them as nodes to the graph.
- Here we assume global parameter independence.



Repetitive structure

- If we have iid samples, the variables get replicated but the parameters are tied / shared

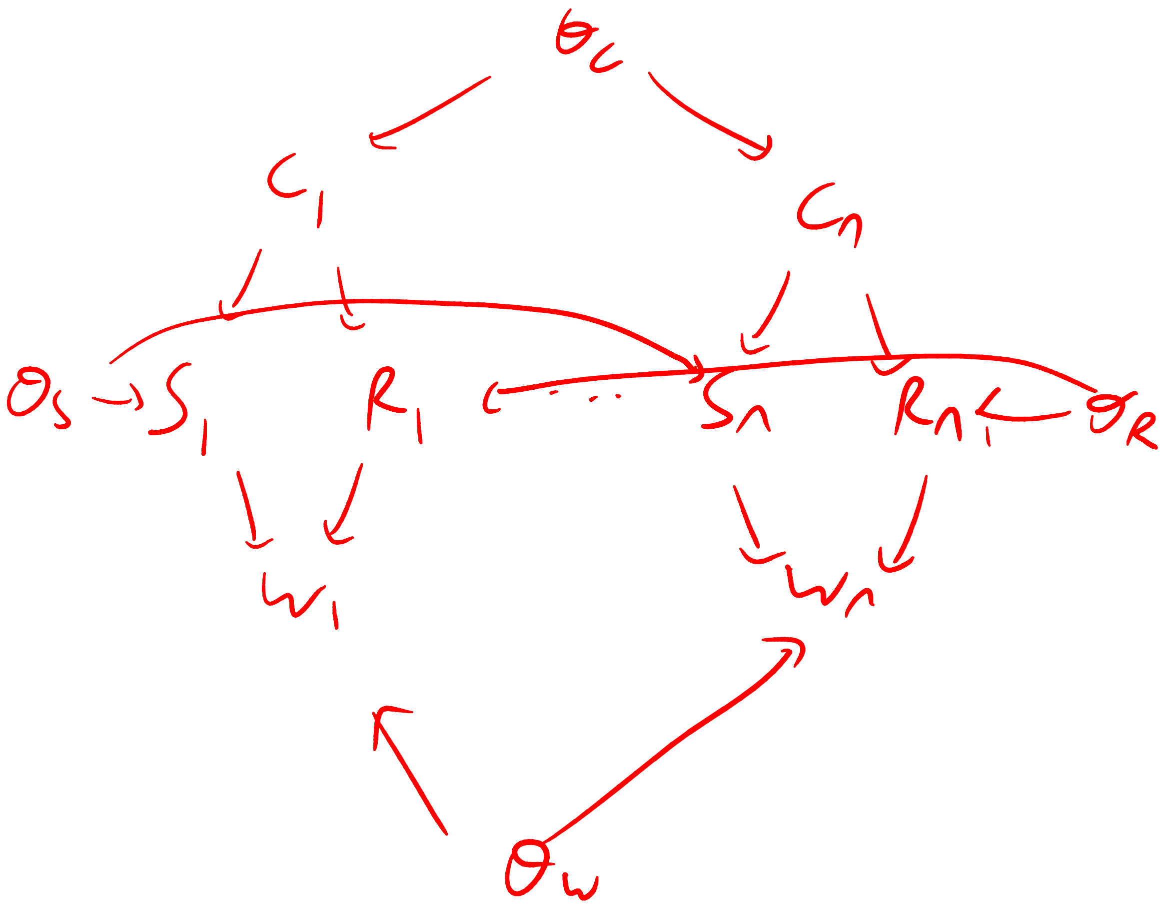
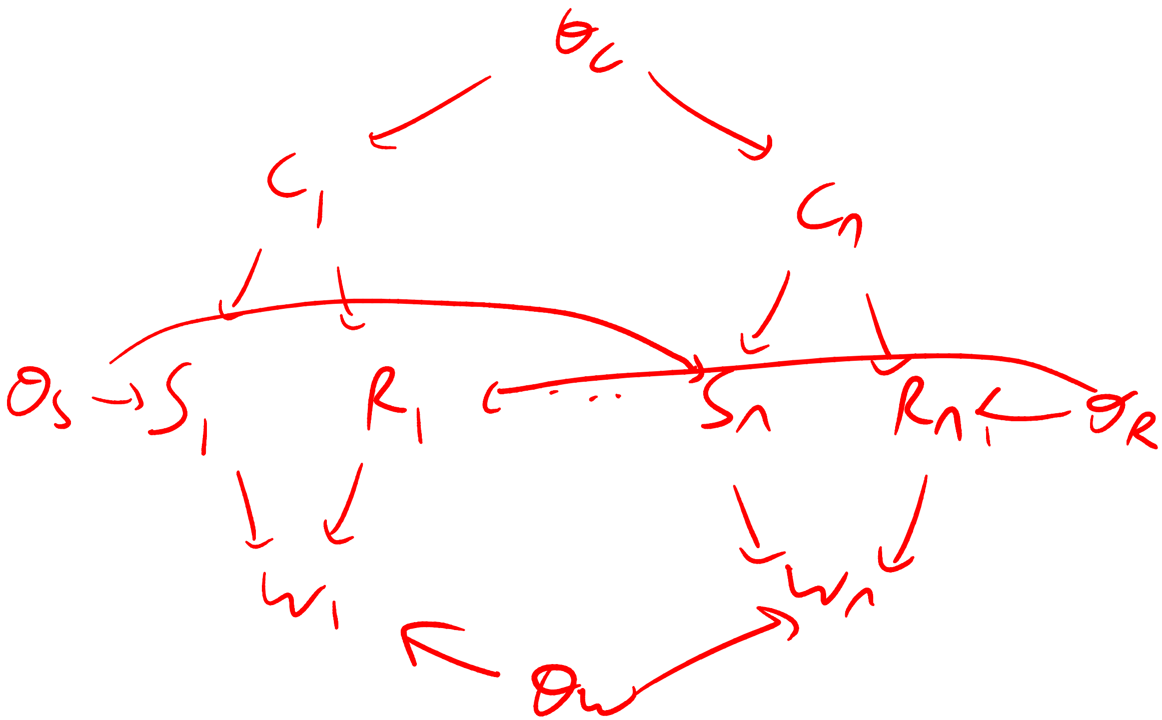
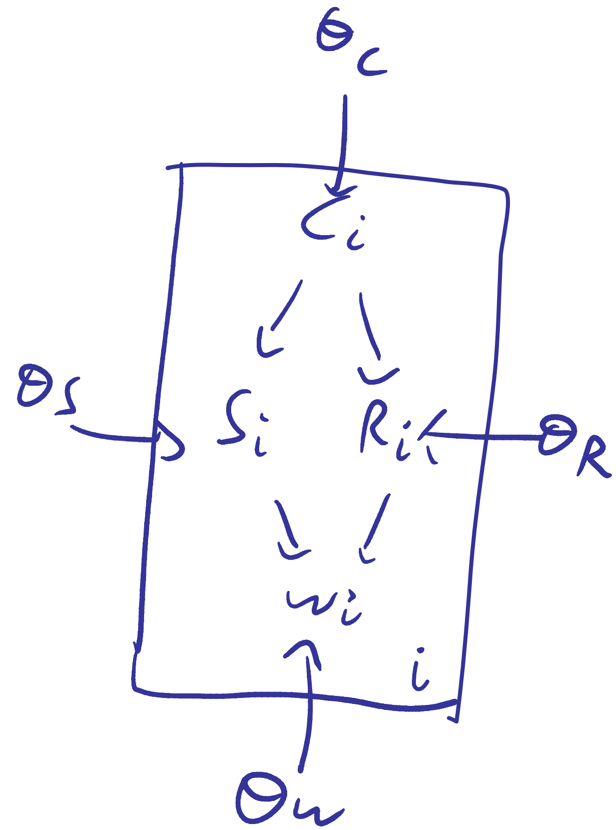


Plate notation

- For shorthand, we use plates





$$
\begin{aligned}
p(D,\theta) &= p(\theta_c)p(\theta_s)p(\theta_r)p(\theta_w) \\
&\times \prod_{i=1}^{n} p(c_i|\theta_c)p(s_i|c_i,\theta_s)p(r_i|c_i,\theta_r)p(w_i|s_i,r_i,\theta_w
\end{aligned}
$$

Factored prior, likelihood, posterior

- Since the parameters are independent in the prior, and the likelihood is factorized, they are also independent in the posterior

$$
\begin{aligned}
p(\theta|D) &\propto p(\theta)p(D|\theta) \\
&= p(\theta_c)\prod_i p(c_i|\theta_c) \\
&\times p(\theta_s)\prod_i p(s_i|c_i,\theta_s) \\
&\times p(\theta_r)\prod_i p(r_i|c_i,\theta_r) \\
&\times p(\theta_w)\prod_i p(w_i|s_i,r_i,\theta_s)
\end{aligned}
$$

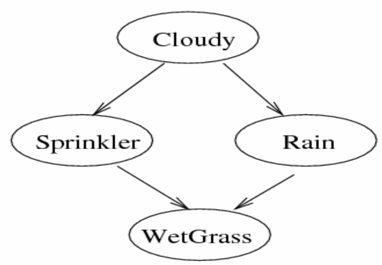
Local parameter independence

- Each row of CPT is a different multinomial distribution. We typically assume these are independent.

$$p(\boldsymbol{\theta}_R) = \prod_{k=0}^{1} p(\boldsymbol{\theta}_{R|C=k}) = \prod_k Dir(\boldsymbol{\theta}_{R|C=k}|\boldsymbol{\alpha}_{R|C=k})$$

P(C=F)	P(C=T)
0.5	0.5

$\leftarrow \theta_C = (0.5, 0.5)$



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

$\leftarrow \theta_{R,C=0} = (0.8, 0.2)$

$\leftarrow \theta_{R,C=1} = (0.2, 0.8)$

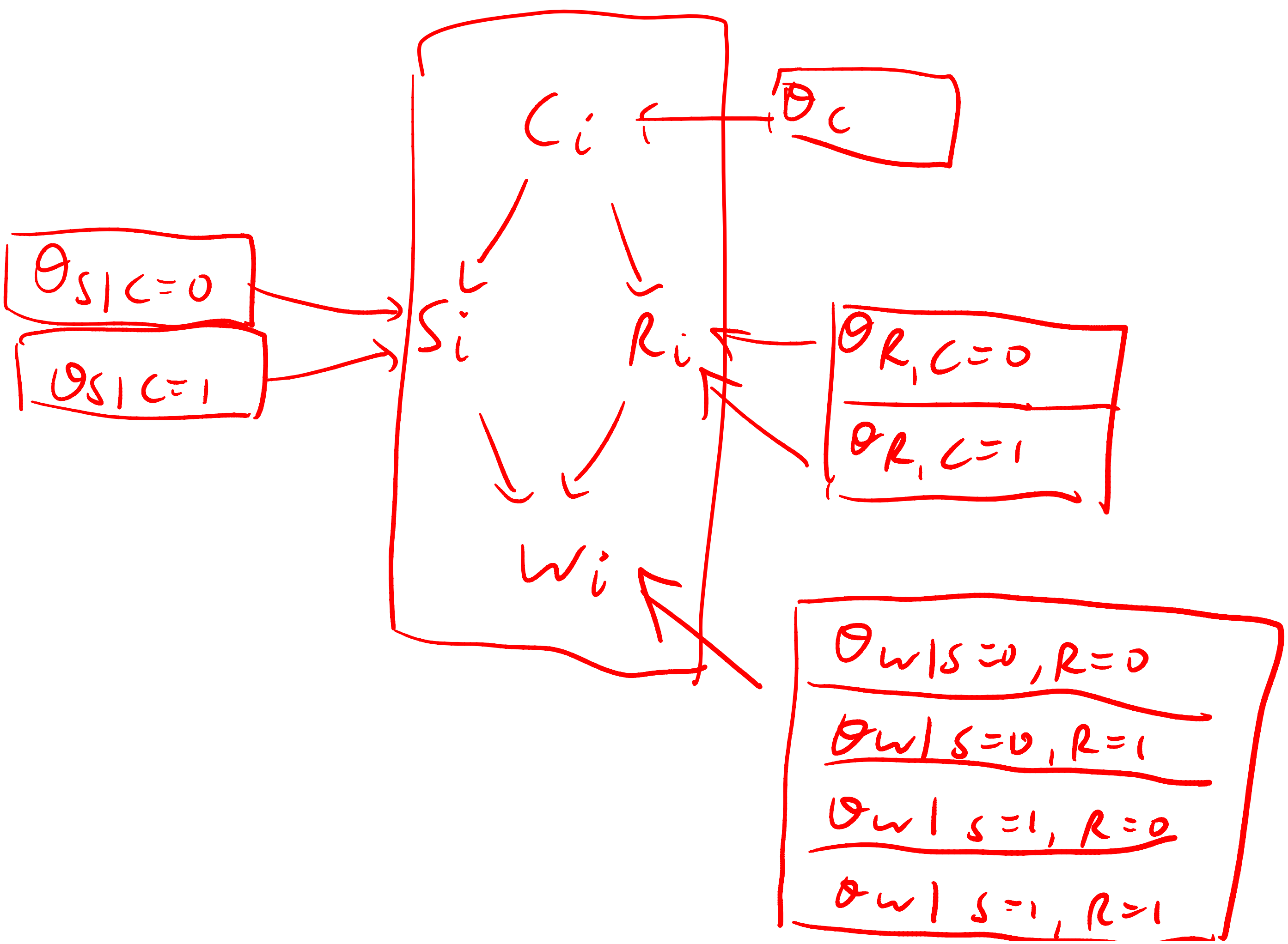
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

$\leftarrow \theta_{W,S=1,R=0} = (0.1, 0.9)$

$\leftarrow \theta_{W,S=1,R=1} = (0.01, 0.99)$

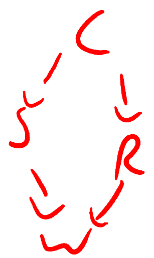
Local parameter independence

- In the case of CPTs, we assume each row of the table is an independent multinomial



Posterior over parameters factorizes

$$p(\boldsymbol{\theta}_R|D) = \prod_{k=0}^{1} p(\boldsymbol{\theta}_{R|C=k}) \prod_{i=1}^{n} I(c_i = k) p(r_i|\theta_{R|C=k})$$



$$= \prod_{k} Dir(\boldsymbol{\theta}_{R|C=k}|\boldsymbol{\alpha}_{R|C=k}) Mu(\mathbf{n}_{R,C=k}|\boldsymbol{\theta}_{R|C=k}, n)$$

i	C	S	R	W
1	0	0	0	0
2	0	0	1	1
3	1	1	1	1

$p(\theta_C)$

1	1
2	1
3	1
3	2

$p(\theta_{R|C=0})$

1	1
2	1
2	2
2	2

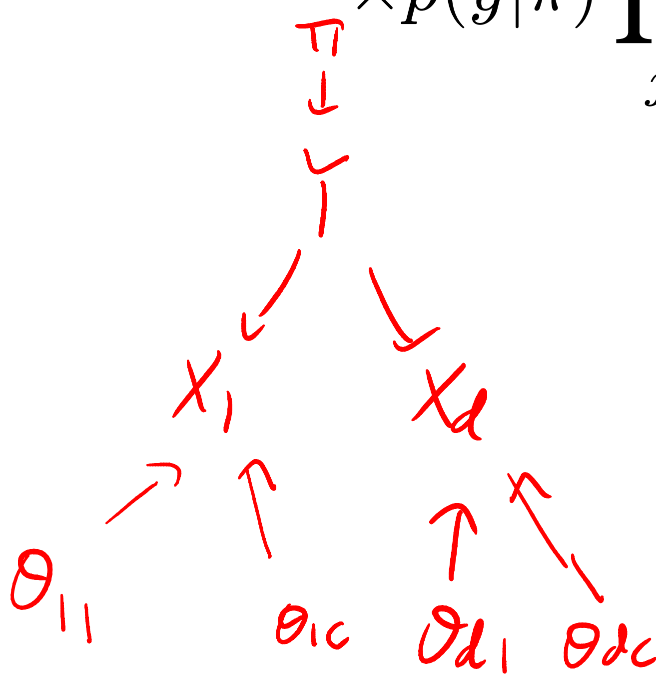
$p(\theta_{R|C=1})$

1	1
1	1
1	1
1	2

$$p(\boldsymbol{\theta}|D) = \prod_{j=1}^{d} \prod_{k \in Pa(j)} Dir(\boldsymbol{\theta}_{jk}|\boldsymbol{\alpha}_{jk} + \mathbf{n}_{jk})$$

Parameters are rv's, too!

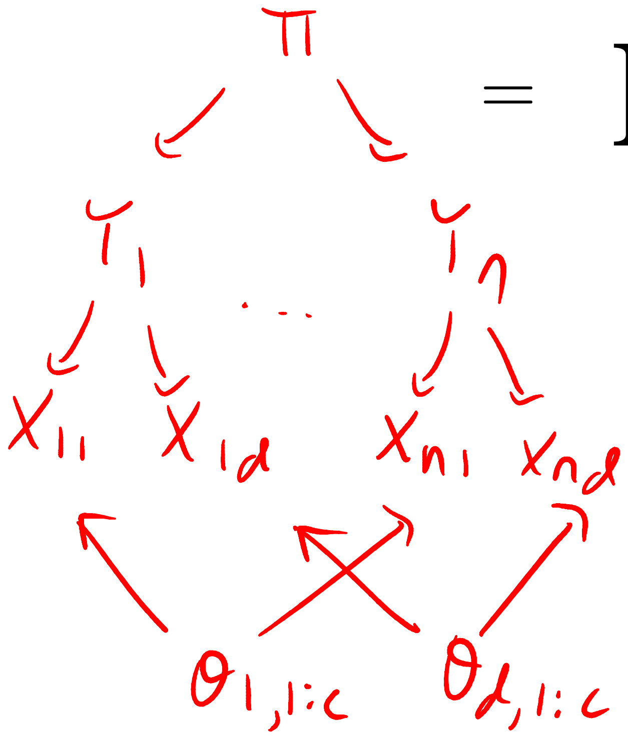
$$
\begin{aligned}
p(\mathbf{x}, y, \pi, \boldsymbol{\theta}) &= p(\pi) p(y|\pi) \prod_{j=1}^{d} \left[p(x_j | y, \boldsymbol{\theta}_j) \prod_{c=1}^{C} p(\theta_{jc}) \right] \\
&= p(\pi) \prod_{j} \prod_{c} p(\theta_{jc}) \\
&\quad \times p(y|\pi) \prod_{j} p(x_j | y, \boldsymbol{\theta}_j)
\end{aligned}
$$



Repetitive structure

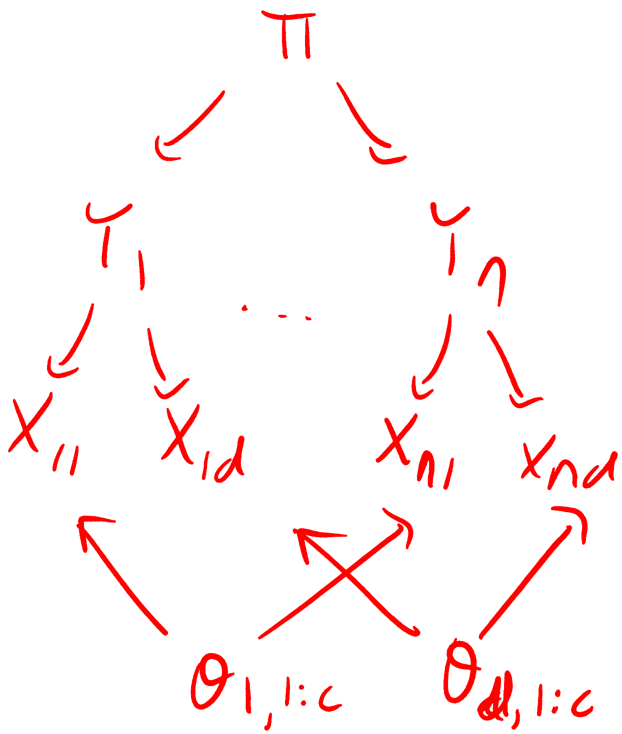
- When we have multiple samples, we replicate the variables, but the params are fixed

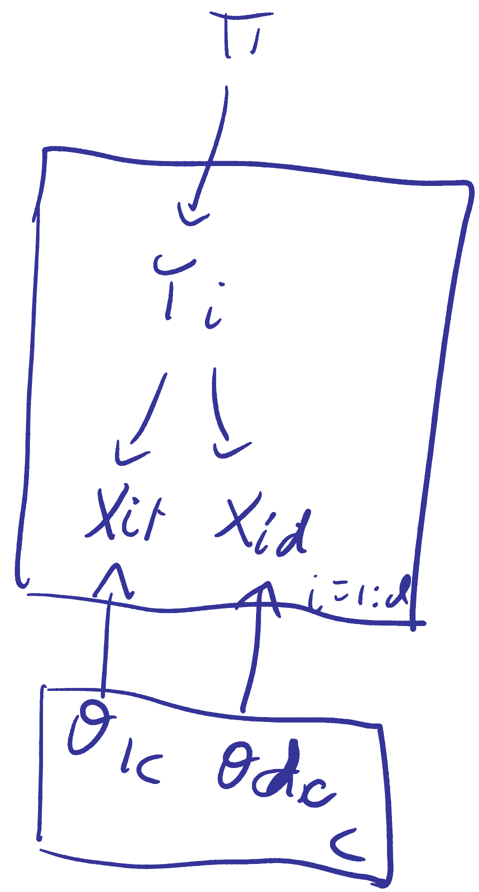
$$
\begin{aligned}
p(D, \pi, \boldsymbol{\theta}) &= p(\pi, \boldsymbol{\theta}) p(D|\pi, \boldsymbol{\theta}) \\
p(D|\pi, \boldsymbol{\theta}) &= \prod_i p(y_i|\pi) \prod_j p(x_{ij}|y_i, \boldsymbol{\theta}_j) \\
&= \prod_c \prod_{i:y_i=c} \pi_c p(x_{ij}|\theta_{jc})
\end{aligned}
$$



Plates

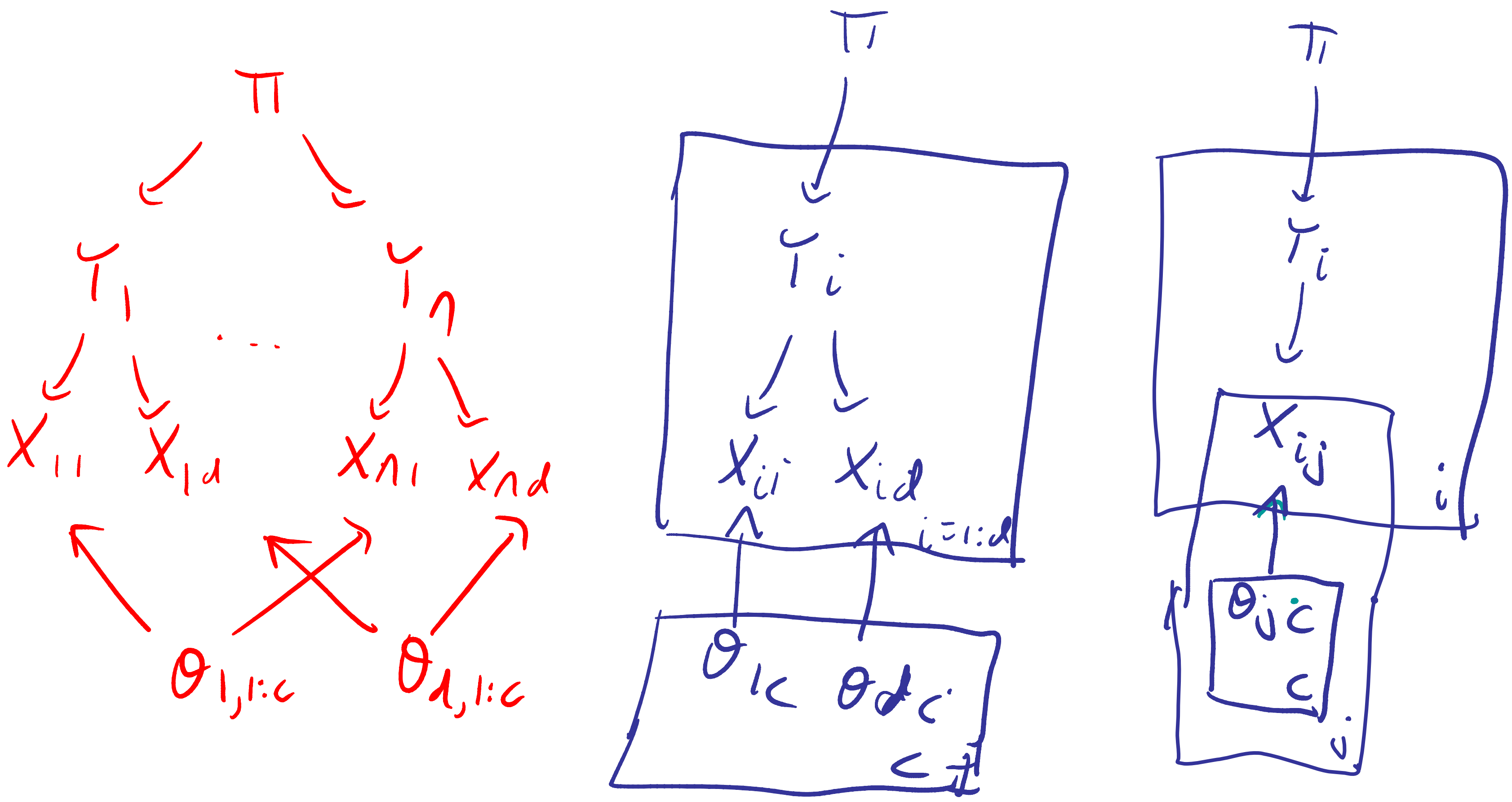
- We introduce a shorthand for repetitive structure





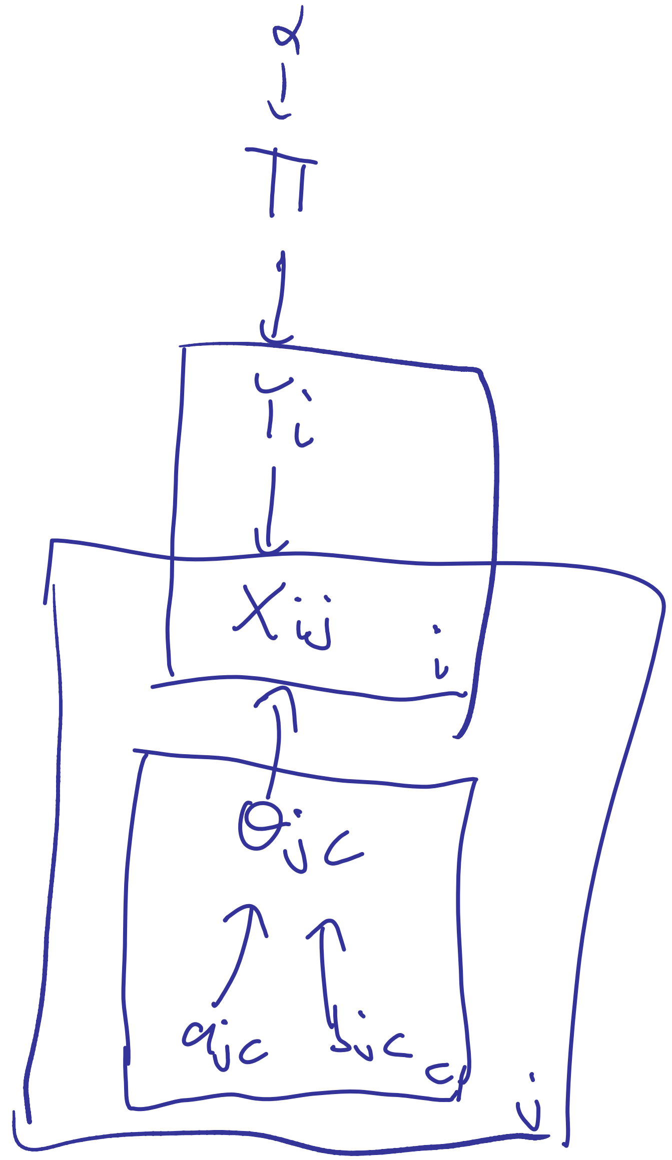
Nested plates

- Doubly indexed nodes



Hyper-parameters

- If the hyper-parameters are fixed, they will be root nodes in the graph.

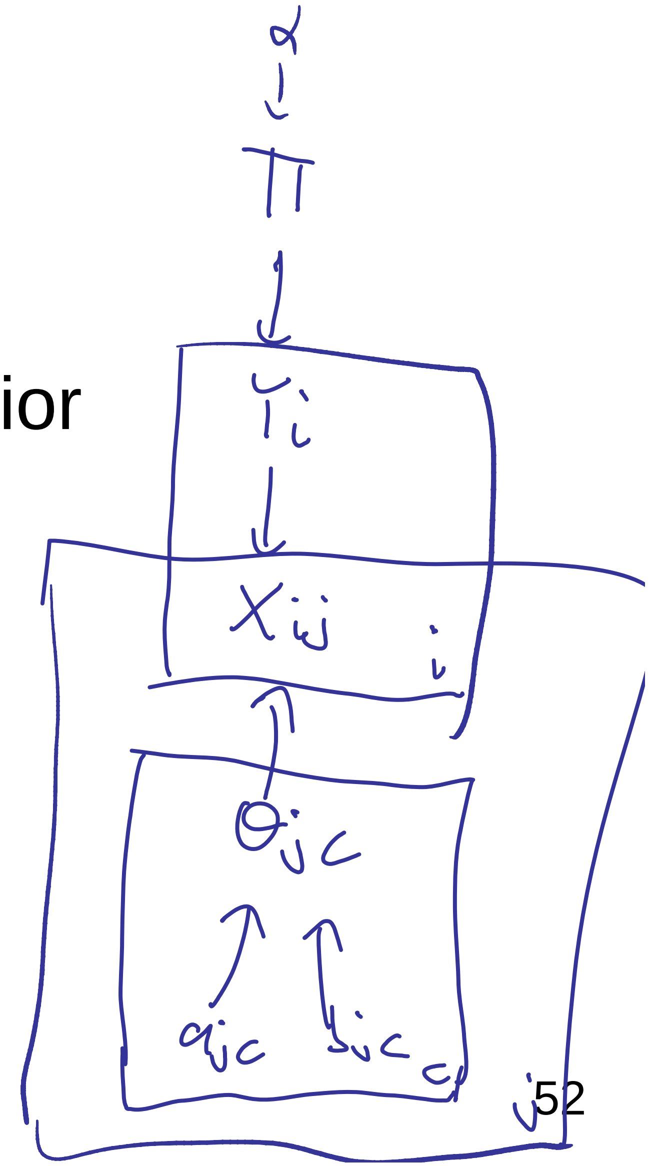


Factored prior/ likelihood/ posterior

- Since the prior and likelihood are factorized over parameters, so is the posterior

eg $$\theta_{jc} \perp \theta_{j'c'} | D$$

Hence we can compute the posterior (or MLE/MAP) of each parameter separately



Example: Binary features

$$
\begin{aligned}
&p(D, \boldsymbol{\pi}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \mathbf{a}, \mathbf{b}) \\
&= \quad p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_i p(y_i|\boldsymbol{\pi}) \prod_c \left[\prod_j \prod_{i:y_i=c} p(x_{ij}|\theta_{jc}) \right] p(\theta_{jc}) \\
&= \quad Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) Mu(\mathbf{n}|\boldsymbol{\pi}) \prod_c \prod_j Bin(n_{jc1}|\theta_{jc}, n_{jc}) Beta(\theta_{jc}|a_{jc}, b_{jc}) \\
&= \quad Dir(\boldsymbol{\pi}|\boldsymbol{\alpha} + \mathbf{n}) \prod_c \prod_j Beta(\theta_{jc}|a_{jc} + n_{jc1}, b_{jc} + n_{jc0})
\end{aligned}
$$

$$
\begin{aligned}
n_{jc1} &= \sum_i I(y_i = c) I(x_{ij} = 1) \\
n_{jc0} &= \sum_i I(y_i = c) I(x_{ij} = 0) \\
n_{jc} &= n_c = \sum_i I(y_i = c) \\
\mathbf{n} &= (n_1, \ldots, n_C)
\end{aligned}
$$