GRAPHICAL MODELS

CS5011- MACHINE LEARNING

Bishop sections 8.1 and 8.3

BAYESIAN NETWORKS

- In order to motivate the use of directed graphs to describe probability distributions, consider first an arbitrary joint distribution $p(a, b, c)$ over three variables a, b, and c.
- By application of the product rule of probability, we can write the joint distribution in the form

$$p(a, b, c) = p(c|a, b)p(a, b). \tag{8.1}$$

- A second application of the product rule, this time to the second term on the right hand side of (8.1), gives

$$p(a, b, c) = p(c|a, b)p(b|a)p(a). \tag{8.2}$$

- Note that this decomposition holds for any choice of the joint distribution.

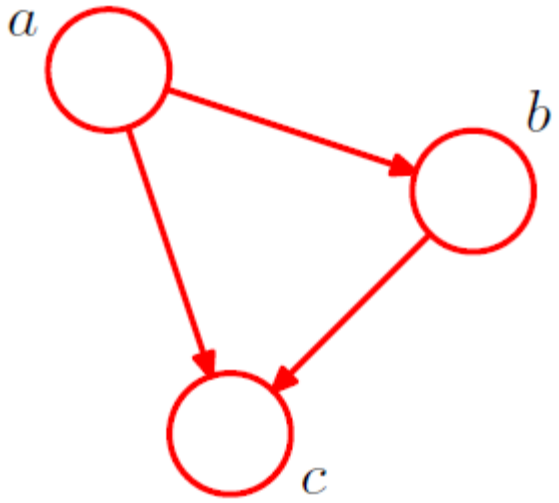
BAYESIAN NETWORKS

- We now represent the right-hand side of (8.2) in terms of a simple graphical model as follows.

$$p(a, b, c) = p(c|a, b)p(b|a)p(a). \qquad (8.2)$$

- First we introduce a node for each of the random variables a, b, and c and associate each node with the corresponding conditional distribution on the right-hand side of (8.2).
- Then, for each conditional distribution we add directed links (arrows) to the graph from the nodes corresponding to the variables on which the distribution is conditioned.
- Thus for the factor $p(c|a, b)$, there will be links from nodes a and b to node c, whereas for the factor $p(a)$ there will be no incoming links.
- If there is a link going from a node a to a node b, then we say that node a is the *parent* of node b, and we say that node b is the *child* of node a.

BAYESIAN NETWORKS



A directed graphical model representing the joint probability distribution over three variables a, b, and c, corresponding to the decomposition on the right-hand side of

$$p(a, b, c) = p(c|a, b)p(b|a)p(a). \tag{8.2}$$

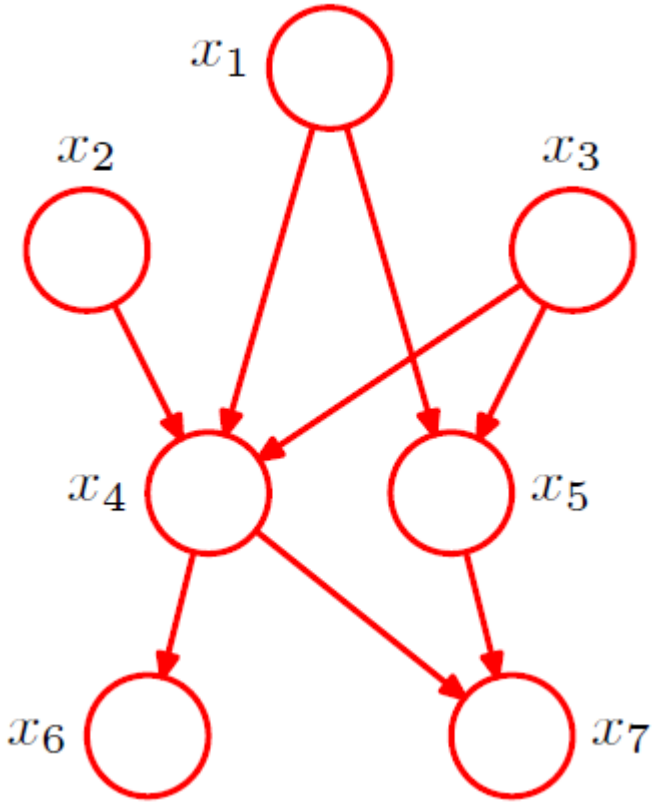
BAYESIAN NETWORKS

- Consider the joint distribution over K variables given by $p(x_1, \ldots, x_K)$.
- By repeated application of the product rule of probability, this joint distribution can be written as a product of conditional distributions, one for each of the variables

$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1). \tag{8.3}$$

- For a given choice of K, we can again represent this as a directed graph having K nodes, one for each conditional distribution on the right-hand side of (8.3), with each node having incoming links from all lower numbered nodes.
- We say that this graph is *fully connected* because there is a link between every pair of nodes.

BAYESIAN NETWORKS



- Example of a directed acyclic graph describing the joint distribution over variables $x_1, \ldots, x_7$.
- This is not a fully connected graph because, for instance, there is no link from x_1 to x_2 or from x_3 to x_7.

BAYESIAN NETWORKS

- The joint distribution of all 7 variables is given by

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5). \tag{8.4}$$

- We can now state in general terms the relationship between a given directed graph and the corresponding distribution over the variables.
- The joint distribution defined by a graph is given by the product, over all of the nodes of the graph, of a conditional distribution for each node conditioned on the variables corresponding to the parents of that node in the graph.
- Thus, for a graph with K nodes, the joint distribution is given by

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \tag{8.5}$$

BAYESIAN NETWORKS

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \qquad (8.5)$$

where pa_k denotes the set of parents of x_k, and $\mathrm{x} = \{x_1, \ldots, x_K\}$.

- This key equation expresses the *factorization* properties of the joint distribution for a directed graphical model.
- The directed graphs that we are considering are subject to an important restriction namely that there must be no *directed cycles.*
- Such graphs are also called *directed acyclic graphs*, or *DAGs*.
- This is equivalent to the statement that there exists an ordering of the nodes such that there are no links that go from any node to any lower numbered node.

GENERATIVE MODELS

- Consider an object recognition task in which each observed data point corresponds to an image (comprising a vector of pixel intensities) of one of the objects.
- In this case, the latent variables might have an interpretation as the position and orientation of the object.
- Given a particular observed image, our goal is to find the posterior distribution over objects, in which we integrate over all possible positions and orientations.

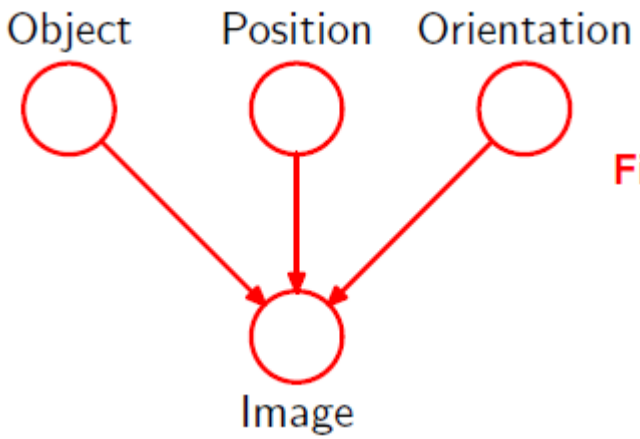


Figure 8.8 A graphical model representing the process by which images of objects are created, in which the identity of an object (a discrete variable) and the position and orientation of that object (continuous variables) have independent prior probabilities. The image (a vector of pixel intensities) has a probability distribution that is dependent on the identity of the object as well as on its position and orientation.

DISCRETE VARIABLES

- Here, the parent and child node each correspond to discrete variables.
- The probability distribution $p(\mathrm{x}|\mu)$ for a single discrete variable x having K possible states (using the 1-of-K representation) is given by:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

and is governed by the parameters $\mu = (\mu_1, \ldots, \mu_K)^T$.

- Due to the constraint $\sum_k \mu_k = 1$, only $K - 1$ values for μ_k need to be specified in order to define the distribution.

DISCRETE VARIABLES

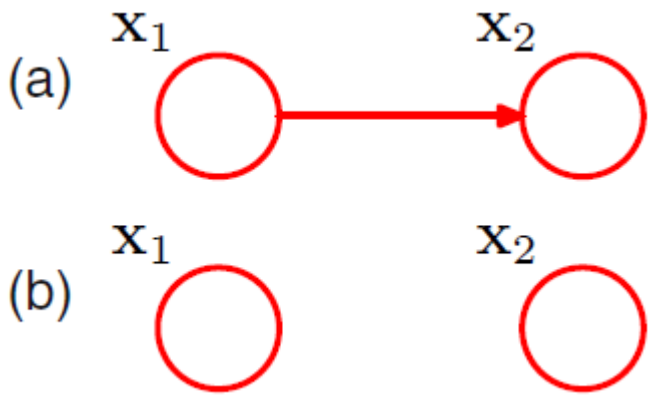
- Now suppose that we have two discrete variables, x_1 and x_2, each of which has K states, and we wish to model their joint distribution.
- We denote the probability of observing both $x_{1k} = 1$ and $x_{2l} = 1$ by the parameter μ_{kl}, where x_{1k} denotes the k^{th} component of x_1, and similarly for x_{2l}.
- The joint distribution can be written:

$$p(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}.$$

- Because the parameters μ_{kl} are subject to the constraint $\sum_k \sum_l \mu_{kl} = 1$, this distribution is governed by $K^2 - 1$ parameters.

DISCRETE VARIABLES

- It is easily seen that the total number of parameters that must be specified for an arbitrary joint distribution over M variables is $K^M - 1$ and therefore grows exponentially with the number M of variables.
- Using the product rule, we can factor the joint distribution $p(\mathrm{x}_1, \mathrm{x}_2)$ in the form $p(\mathrm{x}_2|\mathrm{x}_1)p(\mathrm{x}_1)$, which corresponds to a two-node graph with a link going from the x_1 node to the x_2 node as shown below:



(a) This fully-connected graph describes a general distribution over two K-state discrete variables having a total of $K^2 - 1$ parameters. (b) By dropping the link between the nodes, the number of parameters is reduced to $2(K - 1)$.

DISCRETE VARIABLES

- The marginal distribution $p(\mathrm{x}_1)$ is governed by $K - 1$ parameters.
- Similarly, the conditional distribution $p(\mathrm{x}_2|\mathrm{x}_1)$ requires the specification of $K - 1$ parameters for each of the K possible values of x_1.
- The total number of parameters that must be specified in the joint distribution is therefore $(K - 1) + K(K - 1) = K^2 - 1$.
- Now suppose that the variables x_1 and x_2 were independent.
- Each variable is then described by a separate multinomial distribution, and the total number of parameters would be $2(K - 1)$.

DISCRETE VARIABLES

- For a distribution over M independent discrete variables, each having K states, the total number of parameters would be $M(K-1)$, which therefore grows linearly with the number of variables.
- From a graphical perspective, we have reduced the number of parameters by dropping links in the graph, at the expense of having a restricted class of distributions.
- If we have M discrete variables $\mathrm{x}_1, \ldots, \mathrm{x}_M$ and if the graph is fully connected then we have a completely general distribution having $K^M - 1$ parameters, whereas if there are no links in the graph the joint distribution factorizes into the product of the marginals, and the total number of parameters is $M(K-1)$.

DISCRETE VARIABLES

- The marginal distribution $p(\mathrm{x}_1)$ requires $K - 1$ parameters, whereas each of the $M - 1$ conditional distributions $p(\mathrm{x}_i|\mathrm{x}_{i-1})$, for $i = 2, \ldots, M$, requires $K(K - 1)$ parameters.
- This gives a total parameter count of $K - 1 + (M - 1)K(K - 1)$, which is quadratic in K and which grows linearly (rather than exponentially) with the length M of the chain.
- An alternative way to reduce the number of independent parameters in a model is by *sharing* parameters (also known as *tying* of parameters). For instance, in the chain example given below, we can arrange that all of the conditional distributions $p(\mathrm{x}_i|\mathrm{x}_{i-1})$, for $i = 2, \ldots, M$, are governed by the same set of $K(K - 1)$ parameters.
- Together with the $K - 1$ parameters governing the distribution of $\mathbf{x}_1$, this gives a total of $K^2 - 1$ parameters that must be specified in order to define the joint distribution.

LINEAR-GAUSSIAN MODELS

- Consider an arbitrary directed acyclic graph over D variables in which node i represents a single continuous random variable x_i having a Gaussian distribution.
- The mean of this distribution is taken to be a linear combination of the states of its parent nodes pa_i of node i

$$p(x_i|\mathrm{pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j\in\mathrm{pa}_i} w_{ij}x_j + b_i, v_i\right) \tag{8.11}$$

where w_{ij} and b_i are parameters governing the mean, and v_i is the variance of the conditional distribution for x_i.

LINEAR- GAUSSIAN MODELS

- The log of the joint distribution is then the log of the product of these conditionals over all nodes in the graph and hence takes the form:

$$\ln p(\mathbf{x}) = \sum_{i=1}^{D} \ln p(x_i|\mathrm{pa}_i) \tag{8.12}$$

$$= -\sum_{i=1}^{D} \frac{1}{2v_i} \left(x_i - \sum_{j \in \mathrm{pa}_i} w_{ij} x_j - b_i \right)^2 + \text{const} \tag{8.13}$$

where $\mathrm{x} = (x_1, \ldots, x_D)^T$ and 'const' denotes terms independent of x.

- We see that this is a quadratic function of the components of $\mathbf{x}$, and hence the joint distribution $p(\mathrm{x})$ is a multivariate Gaussian.

LINEAR- GAUSSIAN MODELS

- We can determine the mean and covariance of the joint distribution recursively as follows.
- Each variable x_i has (conditional on the states of its parents) a Gaussian distribution of the form (8.11) and so

$$x_i = \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i + \sqrt{v_i}\epsilon_i \qquad (8.14)$$

where ϵ_i is a zero mean, unit variance Gaussian random variable satisfying $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = I_{ij}$, where I_{ij} is the i, j element of the identity matrix.

- Taking the expectation, we have

$$\mathbb{E}[x_i] = \sum_{j \in \mathrm{pa}_i} w_{ij} \mathbb{E}[x_j] + b_i. \qquad (8.15)$$

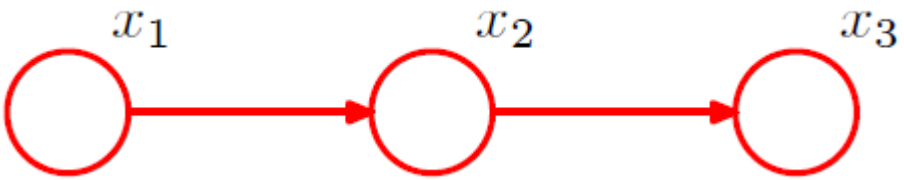
LINEAR- GAUSSIAN MODELS

- We can use (8.14) and (8.15) to obtain the i, j element of the covariance matrix for $p(\mathrm{x})$ in the form of a recursion relation

$$\begin{aligned}
\mathrm{cov}[x_i, x_j] &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])\right] \\
&= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\left\{\sum_{k \in \mathrm{pa}_j} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\right\}\right] \\
&= \sum_{k \in \mathrm{pa}_j} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij}v_j \qquad (8.16)
\end{aligned}$$

and so the covariance can be evaluated recursively starting from the lowest numbered node.

- Consider the example



which has a link missing between variables x_1 and x_3.

- Using the recursion relations (8.15) and (8.16), we see that the mean and covariance of the joint distribution are given by

$$\boldsymbol{\mu} = (b_1, b_2 + w_{21}b_1, b_3 + w_{32}b_2 + w_{32}w_{21}b_1)^{\mathrm{T}} \tag{8.17}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 \\ w_{21}v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2(v_2 + w_{21}^2 v_1) \end{pmatrix}. \tag{8.18}$$

Bayesian Networks: Definition

Definition (Bayesian Network)

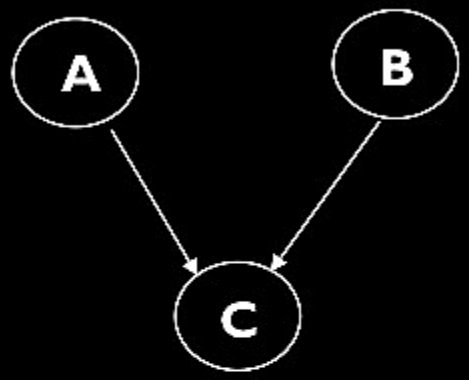
A Bayesian network consists of

- A directed acyclic graph (V, E) whose nodes are labeled with random variables
- A domain for each random variable
- A conditional probability distribution for each variable V
  - Specifies $P(V|Parents(V))$
  - $Parents(V)$ is the set of variables V' with $(\mathrm{V}', \mathrm{V}) \in E$
    - For nodes V without predecessors, $Parents(V) = \{\}$

- The parents of variable V are those V directly depends on
- A Bayesian network is a compact representation of the JPD: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i))$

Example of a simple Bayesian network

$p(A,B,C) = p(A)p(B)\, p(C|A,B)$ ⟷



- Probability model has simple factored form
- Directed edges => direct dependence
- Absence of an edge => conditional independence
- Also known as belief networks, graphical models, causal networks
- Other formulations, e.g., undirected graphical models

Conditional Independence

Consider three variables a, b, and c, and
suppose that the conditional distribution of a, given b and c, is such that it does not depend on the value of b, so that:

$$p(a|b,c) = p(a|c)$$

We say that a is conditionally independent of b given c

This can be expressed in a slightly different way if we consider the joint distribution of a and b conditioned on c, which we can write in the form

p(a, b|c) = p(a|b,c) p(b|c) = p(a|c) p(b|c).

$$p(a, b|c) = p(a|c)\, p(b|c).$$

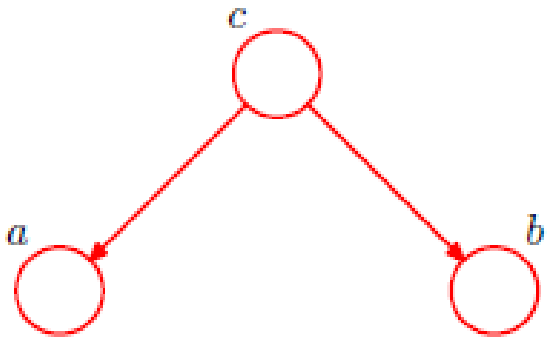
We see that, conditioned on *c*, the joint distribution of *a* and *b* factorizes into the product of the marginal distribution of *a* and the marginal distribution of *b* (again both conditioned on *c*). This says that the variables *a* and *b* are statistically independent, given *c*.

$$a \perp\!\!\!\perp b \mid c$$

This notation denotes that a is conditionally independent of b given c and is equivalent to

$$p(a|b, c) = p(a|c)$$

$p(a, b, c) = p(a|c)\, p(b|c)\, p(c)$.



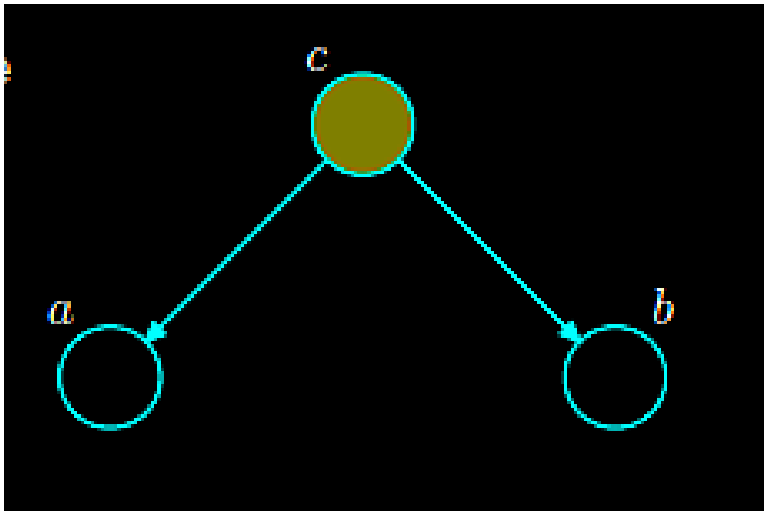
by marginalizing both sides of above with respect to *c* gives

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c).$$

$$a \not\perp b \mid \emptyset$$

where $\emptyset$ denotes the empty set, and the symbol $\not\perp$ means that the conditional independence property does not hold, in general.

From
$p(a, b, c) = p(a|c)\, p(b|c)\, p(c)$, we also have:

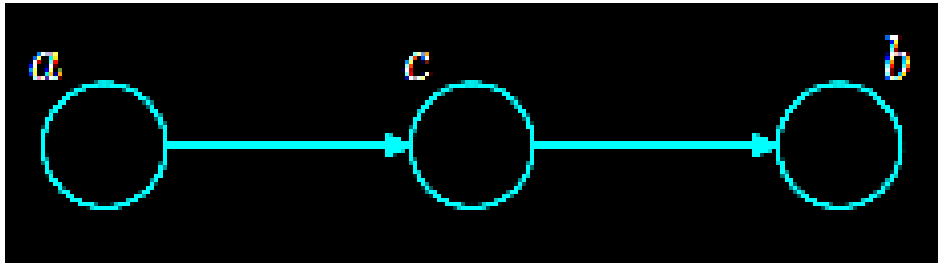


$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)$$

$$a \perp\!\!\!\perp b \mid c$$

The node *C* is said to be *tail-to-tail* with respect to this path because the node is connected to the tails of the two arrows, and the presence of such a path connecting nodes *a* and *b* causes these nodes to be dependent.

However, when we condition on node *c*, as in Figure, the conditioned node 'blocks' the path from *a* to *b* and causes *a* and *b* to become (conditionally) independent.

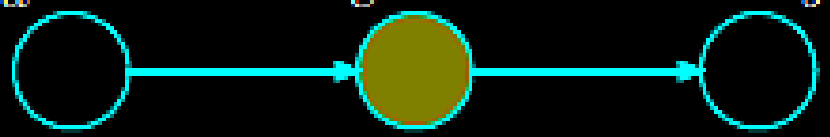


$$p(a, b, c) = p(a)p(c|a)p(b|c).$$

if a and b are independent by marginalizing over c to give

$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

conditioning on node c.



Using Bayes' Theorem, and

$$p(a, b, c) = p(a)p(c|a)p(b|c).$$

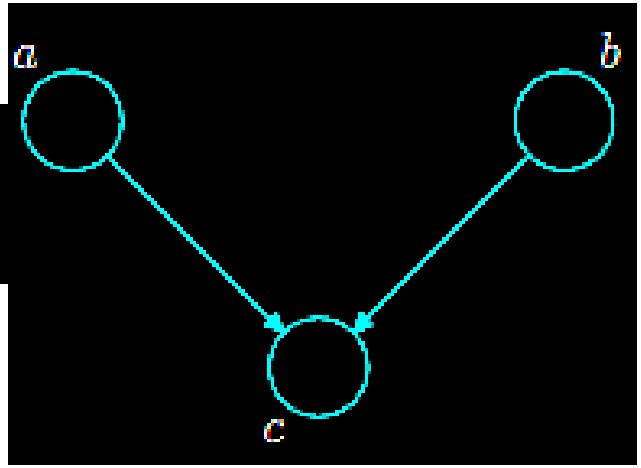
$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

and so again we obtain the conditional independence property

$$a \perp\!\!\!\perp b \mid c.$$

The node *c* is said to be *head-to-tail* with respect to the path from node *a* to node *b*. Such a path connects nodes *a* and *b* and renders them dependent. If we now observe *c*, as in Figure, then this observation 'blocks' the path from *a* to *b* and so we obtain the conditional independence property.

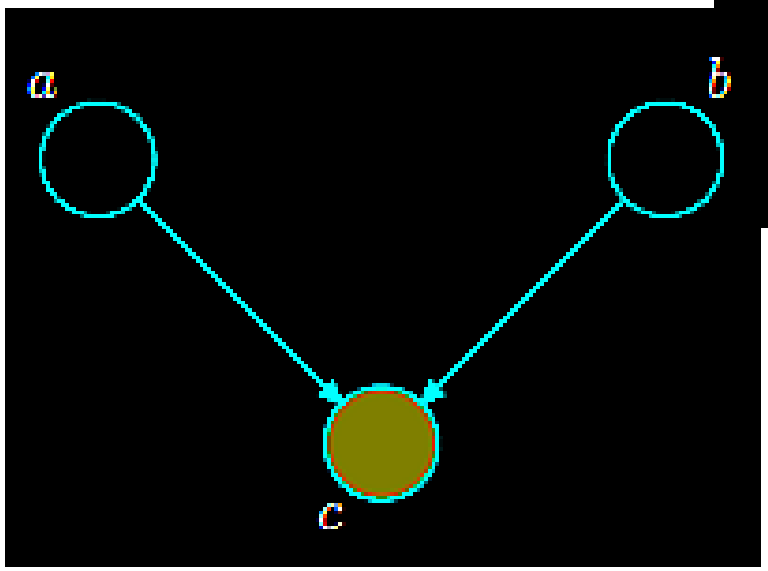
$$p(a, b, c) = p(a)p(b)p(c|a, b).$$



$$p(a, b) = p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset.$$

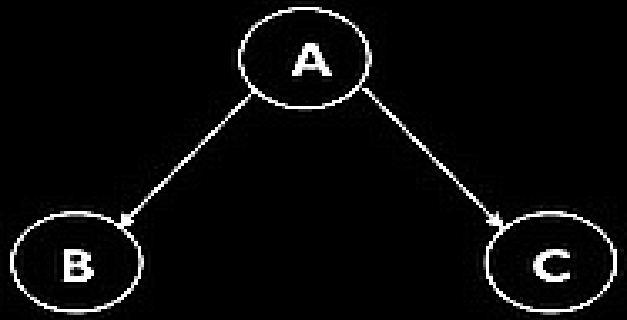
$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a, b)}{p(c)} \end{aligned}$$



$$a \not\perp\!\!\!\perp b \mid c.$$

Marginal Independence:
$p(A,B,C) = p(A)\,p(B)\,p(C)$

Conditionally independent effects:
$p(A,B,C) = p(B|A)p(C|A)p(A)$



B and C are conditionally independent
Given A

e.g., A is a disease, and we model
B and C as conditionally independent
symptoms given A

Independent Causes:
$p(A,B,C) = p(C|A,B)p(A)p(B)$

Markov dependence:
$p(A,B,C) = p(C|B)\,p(B|A)p(A)$

A and B are (marginally) independent
but become dependent once C is known

MARKOV RANDOM FIELDS

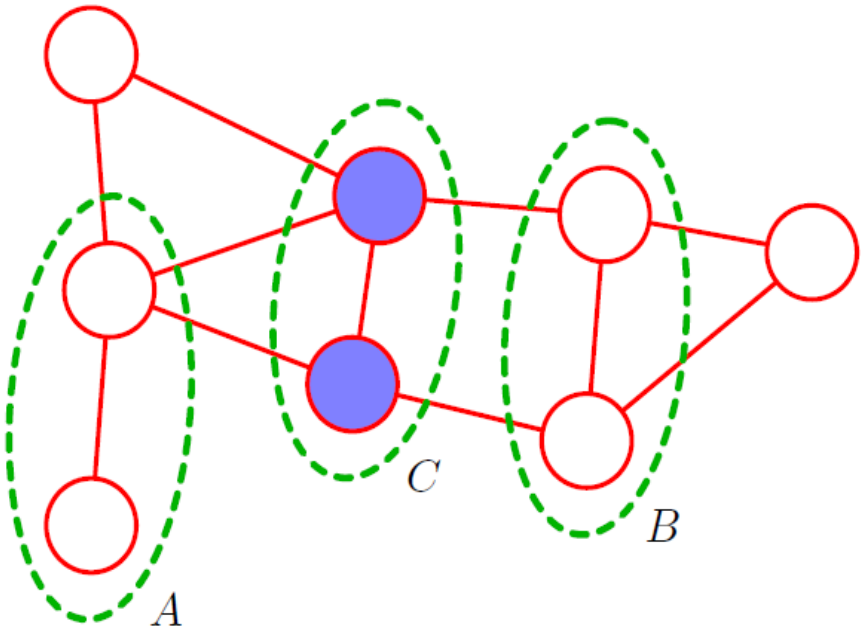
- A *Markov random field*, also known as a *Markov network* or an *undirected graphical model*, has a set of nodes each of which corresponds to a variable or group of variables, as well as a set of links each of which connects a pair of nodes. The links are undirected, that is they do not carry arrows.
- It is possible to define graphical semantics for probability distributions such that conditional independence is determined by simple graph separation.
- By removing the directionality from the links of the graph, the asymmetry between parent and child nodes is removed.

MARKOV RANDOM FIELDS

- Suppose that in an undirected graph we identify three sets of nodes, denoted A, B, and C, and that we consider the conditional independence property

$$A \perp\!\!\!\perp B \mid C. \tag{8.37}$$

i.e., A is conditionally independent of B given C.



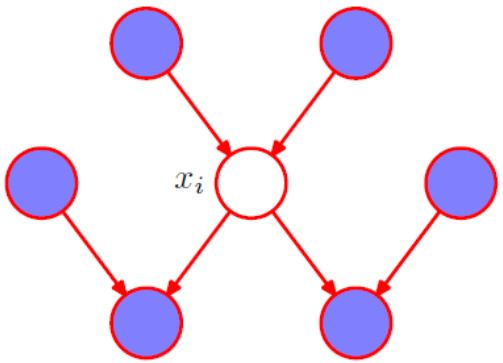
An example of an undirected graph in which every path from any node in set A to any node in set B passes through at least one node in set C. Consequently the conditional independence property $A \perp\!\!\!\perp B \mid C$ holds for any probability distribution described by this graph.

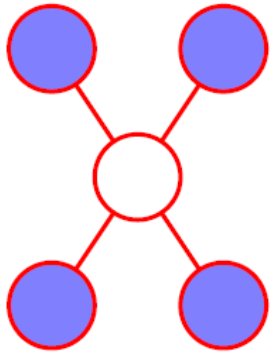
MARKOV RANDOM FIELDS

- To test whether this property is satisfied by a probability distribution defined by a graph we consider all possible paths that connect nodes in set A to nodes in set B.
- If all such paths pass through one or more nodes in set C, then all such paths are ‘blocked’ and so the conditional independence property holds.
- However, if there is at least one such path that is not blocked, then the property does not necessarily hold, or more precisely there will exist at least some distributions corresponding to the graph that do not satisfy this conditional independence relation.

MARKOV RANDOM FIELDS

- An alternative way to view the conditional independence test is to imagine removing all nodes in set C from the graph together with any links that connect to those nodes.
- We then ask if there exists a path that connects any node in A to any node in B. If there are no such paths, then the conditional independence property must hold.
- The **Markov blanket** for an undirected graph takes a particularly simple form, because a node will be conditionally independent of all other nodes conditioned only on the neighboring nodes.
- Consider a node x_i. The set of nodes comprising the parents, the children and the co-parents of x_i is called the Markov blanket.





MARKOV RANDOM FIELDS

- If we consider two nodes x_i and x_j that are not connected by a link, then these variables must be conditionally independent given all other nodes in the graph.
- This follows from the fact that there is no direct path between the two nodes, and all other paths pass through nodes that are observed, and hence those paths are blocked.
- This conditional independence property can be expressed as

$$p(x_i, x_j|\mathbf{x}_{\setminus\{i,j\}}) = p(x_i|\mathbf{x}_{\setminus\{i,j\}})p(x_j|\mathbf{x}_{\setminus\{i,j\}}) \qquad (8.38)$$

MARKOV RANDOM FIELDS

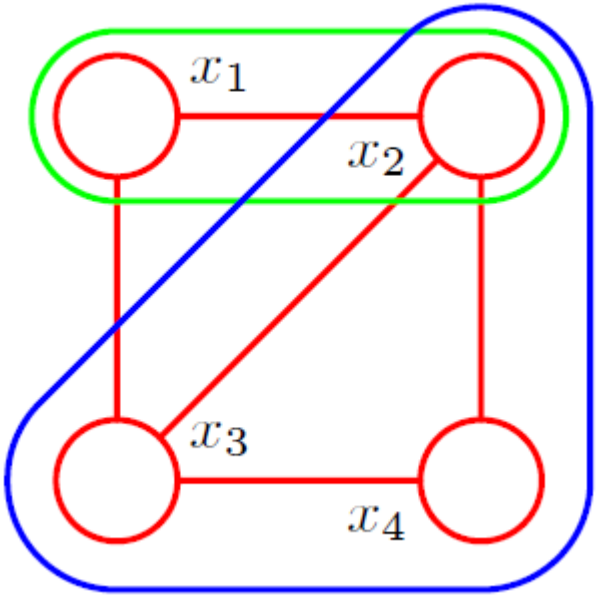
$$p(x_i, x_j|\mathbf{x}_{\setminus\{i,j\}}) = p(x_i|\mathbf{x}_{\setminus\{i,j\}})p(x_j|\mathbf{x}_{\setminus\{i,j\}}) \quad (8.38)$$

where $\mathrm{x}_{\setminus\{i,j\}}$ denotes the set x of all variables with x_i and x_j removed.

- The factorization of the joint distribution must therefore be such that x_i and x_j do not appear in the same factor in order for the conditional independence property to hold for all possible distributions belonging to the graph.
- This leads us to consider a graphical concept called a *clique*, which is defined as a subset of the nodes in a graph such that there exists a link between all pairs of nodes in the subset.
- A *maximal clique* is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.

MARKOV RANDOM FIELDS

A four-node undirected graph showing a clique (outlined in green) and a maximal clique (outlined in blue).



- This graph has five cliques of two nodes given by $\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\}, \{x_4, x_2\}$, and $\{x_1, x_3\}$, as well as two maximal cliques given by $\{x_1, x_2, x_3\}$ and $\{x_2, x_3, x_4\}$. The set $\{x_1, x_2, x_3, x_4\}$ is not a clique because of the missing link from x_1 to x_4.

MARKOV RANDOM FIELDS

- Let us denote a clique by C and the set of variables in that clique by x_C. Then the joint distribution is written as a product of *potential functions* $\psi_C(\mathrm{x}_C)$ over the maximal cliques of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C). \tag{8.39}$$

- Here the quantity Z, sometimes called the *partition function*, is a normalization constant and is given by

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C) \tag{8.40}$$

which ensures that the distribution $p(\mathrm{x})$ given by (8.39) is correctly normalized.

- To establish the connection between conditional independence and factorization for undirected graphs, we need to restrict attention to potential functions $\psi_C(\mathrm{x}_C)$ that are strictly positive.
- Given this restriction, we can make a precise relationship between factorization and conditional independence.
- Because we are restricted to potential functions which are strictly positive it is convenient to express them as exponentials

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\} \qquad (8.41)$$

where $E(\mathrm{x}_C)$ is called an *energy function*, and the exponential representation is called the *Boltzmann distribution*.

- The joint distribution is defined as the product of potentials, and so the total energy is obtained by adding the energies of each of the maximal cliques.

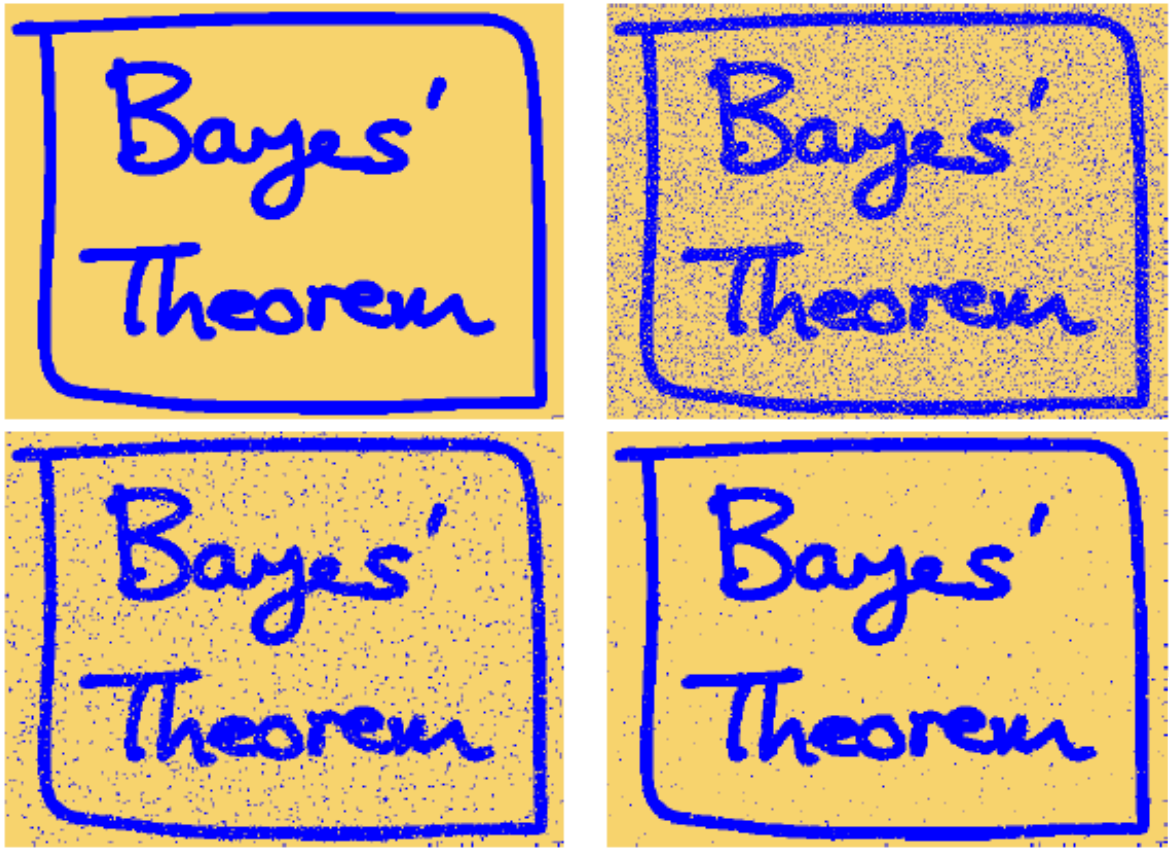


Figure 8.30 Illustration of image de-noising using a Markov random field. The top row shows the original binary image on the left and the corrupted image after randomly changing 10% of the pixels on the right. The bottom row shows the restored images obtained using iterated conditional models (ICM) on the left and using the graph-cut algorithm on the right. ICM produces an image where 96% of the pixels agree with the original image, whereas the corresponding number for graph-cut is 99%.

Figure 8.31 An undirected graphical model representing a Markov random field for image de-noising, in which x_i is a binary variable denoting the state of pixel i in the unknown noise-free image, and y_i denotes the corresponding value of pixel i in the observed noisy image.

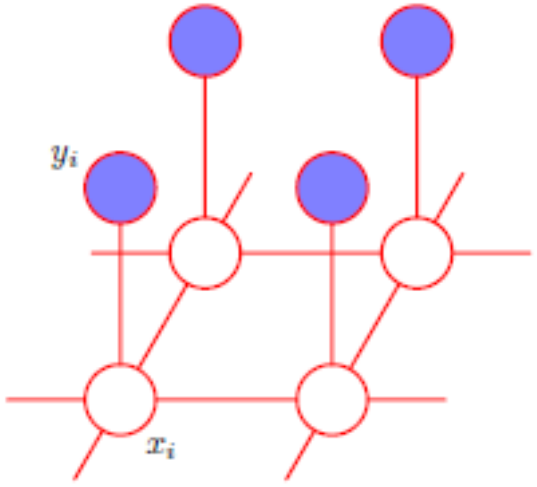
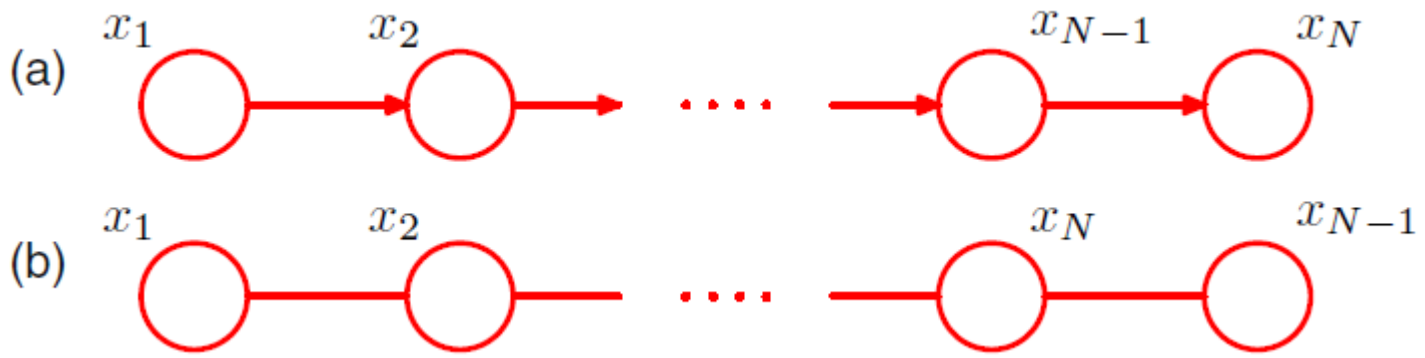


Illustration: Image de-noising

RELATION TO DIRECTED GRAPHS

- We have introduced two graphical frameworks for representing probability distributions, corresponding to directed and undirected graphs.
- Consider first the problem of taking a model that is specified using a directed graph and trying to convert it to an undirected graph.



(a) Example of a directed graph. (b) The equivalent undirected graph.

RELATION TO DIRECTED GRAPHS

- Here the joint distribution for the directed graph is given as a product of conditionals in the form

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1}). \tag{8.44}$$

- Now let us convert this to an undirected graph representation.
- In the undirected graph, the maximal cliques are simply the pairs of neighboring nodes, and so we get the joint distribution in the form

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1}, x_N). \tag{8.45}$$

RELATION TO DIRECTED GRAPHS

- This is easily done by identifying

$$\begin{aligned}
\psi_{1,2}(x_1, x_2) &= p(x_1)p(x_2|x_1) \\
\psi_{2,3}(x_2, x_3) &= p(x_3|x_2) \\
&\vdots \\
\psi_{N-1,N}(x_{N-1}, x_N) &= p(x_N|x_{N-1})
\end{aligned}$$

where we have absorbed the marginal $p(x_1)$ for the first node into the first potential function. Note that in this case, the partition function $Z = 1$.

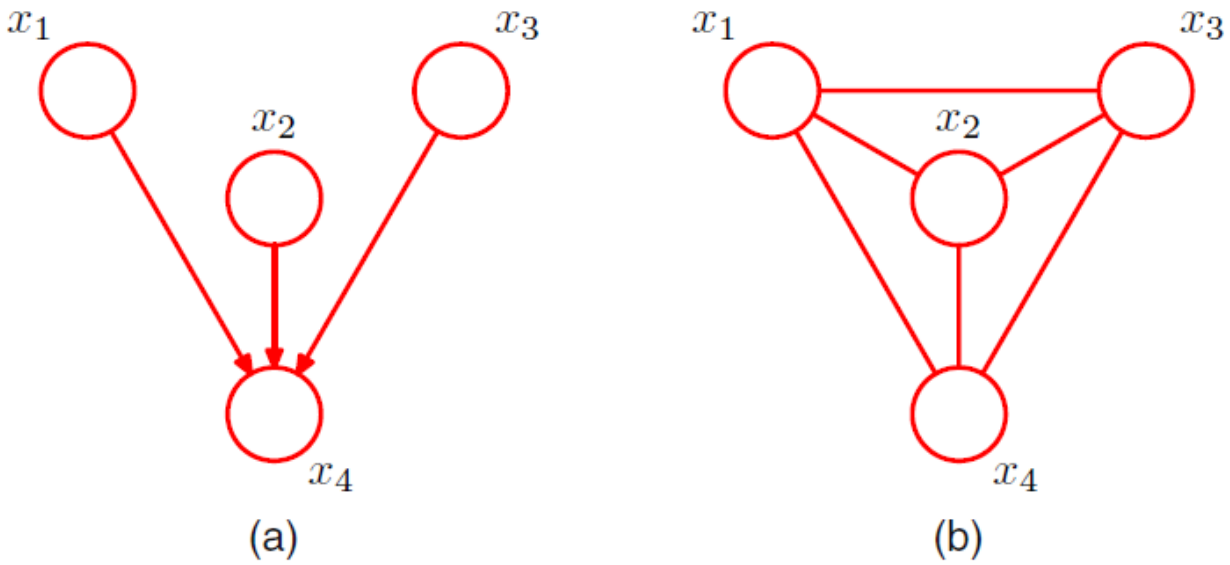
- Let us consider how to generalize this construction, so that we can convert any distribution specified by a factorization over a directed graph into one specified by a factorization over an undirected graph.

RELATION TO DIRECTED GRAPHS

- This can be achieved if the clique potentials of the undirected graph are given by the conditional distributions of the directed graph.
- In order for this to be valid, we must ensure that the set of variables that appears in each of the conditional distributions is a member of at least one clique of the undirected graph.
- For nodes on the directed graph having just one parent, this is achieved simply by replacing the directed link with an undirected link.
- However, for nodes in the directed graph having more than one parent, this is not sufficient.

RELATION TO DIRECTED GRAPHS

- Consider a simple directed graph over 4 nodes



Example of a simple directed graph (a) and the corresponding moral graph (b).

RELATION TO DIRECTED GRAPHS

- The joint distribution for the directed graph takes the form

$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3). \qquad (8.46)$$

- We see that the factor $p(x_4|x_1, x_2, x_3)$ involves the four variables x_1, x_2, x_3, and x_4, and so these must all belong to a single clique if this conditional distribution is to be absorbed into a clique potential.
- To ensure this, we add extra links between all pairs of parents of the node x_4.
- This process of 'marrying the parents' has become known as *moralization*, and the resulting undirected graph, after dropping the arrows, is called the *moral graph*.

RELATION TO DIRECTED GRAPHS

- It is important to observe that the moral graph in this example is fully connected and so exhibits no conditional independence properties, in contrast to the original directed graph.
- Thus in general to convert a directed graph into an undirected graph, we first add additional undirected links between all pairs of parents for each node in the graph and then drop the arrows on the original links to give the moral graph.
- Then we initialize all of the clique potentials of the moral graph to 1.
- We then take each conditional distribution factor in the original directed graph and multiply it into one of the clique potentials.

RELATION TO DIRECTED GRAPHS

- There will always exist at least one maximal clique that contains all of the variables in the factor as a result of the moralization step. Note that in all cases the partition function is given by $Z = 1$.
- The process of converting a directed graph into an undirected graph plays an important role in exact inference techniques.
- Converting from an undirected to a directed representation is much less common and in general presents problems due to the normalization constraints.