

10707

Deep Learning

Russ Salakhutdinov

Machine Learning Department

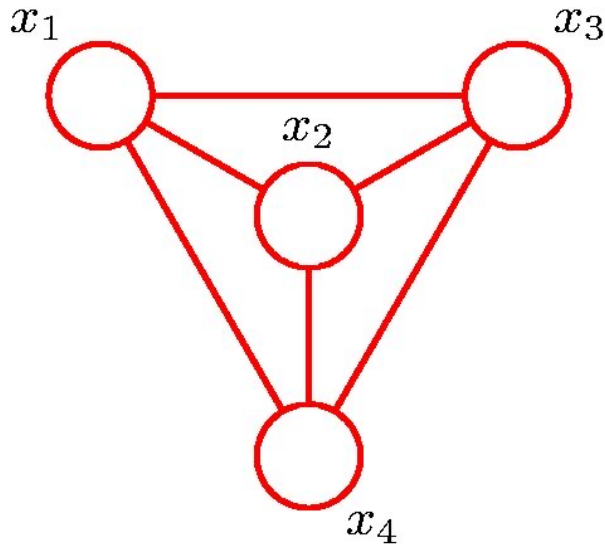
Graphical Models

Graphical Models

- Probabilistic graphical models provide a powerful framework for representing **dependency structure between random variables**.
- Graphical models offer several useful properties:
 - They provide **a simple way to visualize the structure of a probabilistic model** and can be used to motivate new models.
 - They provide **various insights into the properties of the model**, including conditional independence.
 - Complex computations (e.g. inference and learning in sophisticated models) can be expressed in terms of **graphical manipulations**.

Graphical Models

- A graph contains a set of nodes (vertices) connected by links (edges or arcs)



- In a probabilistic graphical model, each **node** represents a random variable, and **links** represent probabilistic dependencies between random variables.
- The graph specifies the way in which the joint distribution over all random variables decomposes into a **product of factors**, where each factor depends on a subset of the variables.

- Two types of graphical models:

- **Bayesian networks**, also known as Directed Graphical Models (the links have a particular directionality indicated by the arrows)
- **Markov Random Fields**, also known as Undirected Graphical Models (the links do not carry arrows and have no directional significance).

- **Hybrid graphical models** that combine directed and undirected graphical models, such as Deep Belief Networks.

Bayesian Networks

- Directed Graphs are useful for expressing **causal relationships** between random variables.
- Let us consider an arbitrary joint distribution $p(a, b, c)$ over three random variables a, b , and c .
- Note that at this point, we do not need to specify anything else about these variables (e.g. whether they are discrete or continuous).
- By application of the **product rule of probability** (twice), we get

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

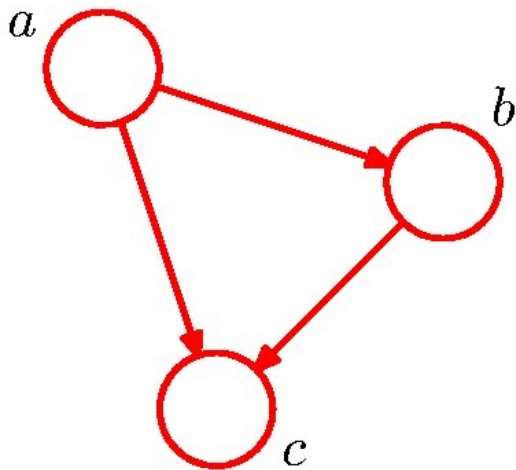
- This decomposition holds for any choice of the joint distribution.

Bayesian Networks

- By application of the product rule of probability (twice), we get

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

- Represent the joint distribution in terms of a simple graphical model:



- Introduce a node for each of the random variables.
- Associate each node with the corresponding conditional distribution in above equation.
- For each conditional distribution we add directed links to the graph from the nodes corresponding to the variables on which the distribution is conditioned.

- Hence for the factor $p(c|a, b)$, there will be links from nodes a and b to node c.
- For the factor $p(a)$, there will be no incoming links.

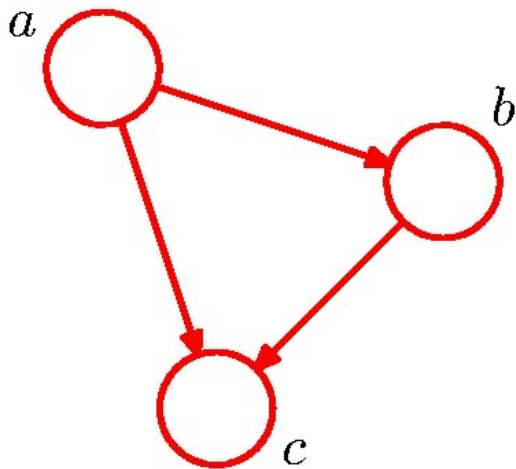
Bayesian Networks

- By application of the product rule of probability (twice), we get

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

- If there is a link going from node a to node b, then we say that:

- node a is a **parent** of node b.
- node b is a **child** of node a.



- For the decomposition, we choose **a specific ordering** of the random variables: a,b,c.
- If we chose a **different ordering**, we would get a **different graphical representation** (we will come back to that point later).

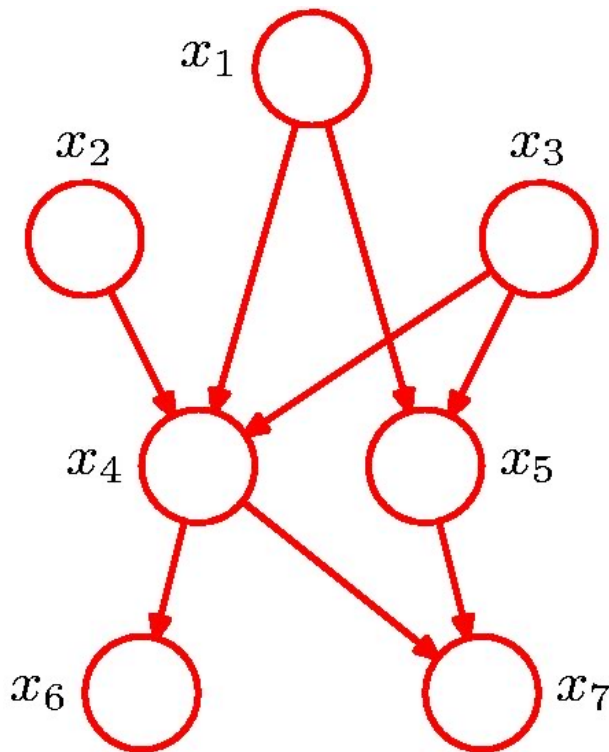
- The joint distribution over K variables factorizes:

$$p(x_1, \dots, x_K) = p(x_K|x_1, \dots, x_{K-1}) \dots p(x_2|x_1)p(x_1)$$

- If each node has incoming links from all lower numbered nodes, then the graph is **fully connected**; there is a link between all pairs of nodes.

Bayesian Networks

- **Absence of links** conveys certain information about the properties of the class of distributions that the graph conveys.



- Note that this graph is not fully connected (e.g. there is no link from x_1 to x_2).
- The joint distribution over x_1, \dots, x_7 can be written as **a product of a set of conditional distributions**.

$$p(x_1, \dots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

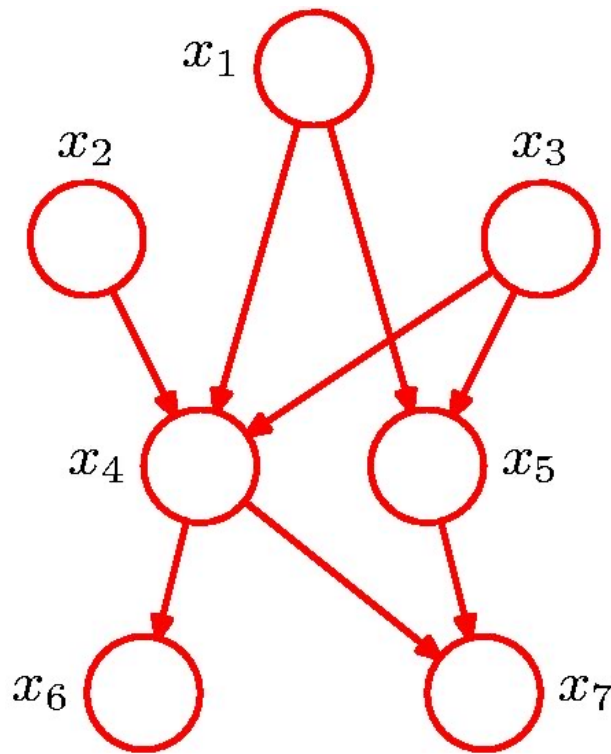
- Note that according to the graph, x_5 will be conditioned only on x_1 and x_3 .

Factorization Property

- The joint distribution defined by the graph is given by **the product of a conditional distribution** for each node conditioned on its parents:

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

where pa_k denotes a set of parents for the node x_k .



- This equation expresses a **key factorization property of the joint distribution** for a directed graphical model.

- Important restriction: There must be **no directed cycles!**

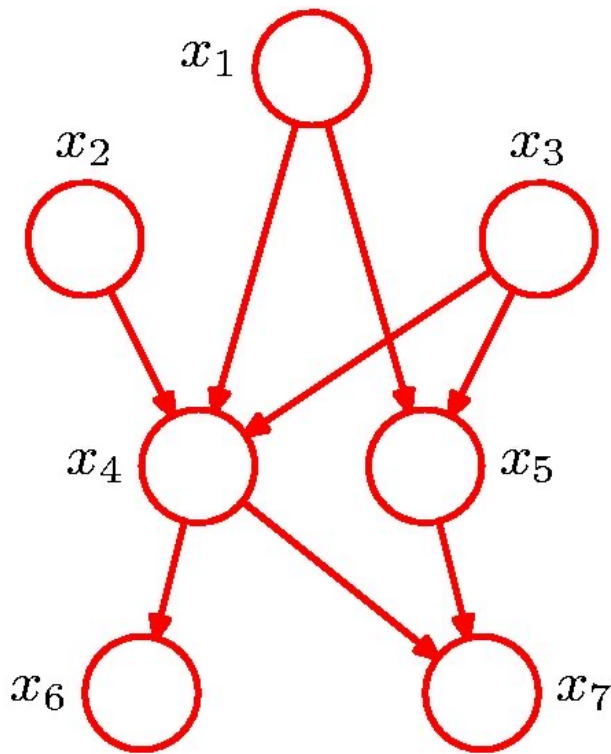
- Such graphs are also called **directed acyclic graphs (DAGs)**.

Ancestral Sampling

- Consider a joint distribution over K random variables $p(x_1, x_2, \dots, x_K)$ that factorizes as:

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

- Our goal is draw a **sample from this distribution**.
- Start at the top and sample in order.



$$\hat{x}_1 \sim p(x_1)$$

$$\hat{x}_2 \sim p(x_2)$$

$$\hat{x}_3 \sim p(x_3)$$

$$\hat{x}_4 \sim p(x_4 | \hat{x}_1, \hat{x}_2, \hat{x}_3)$$

$$\hat{x}_5 \sim p(x_5 | \hat{x}_1, \hat{x}_3)$$

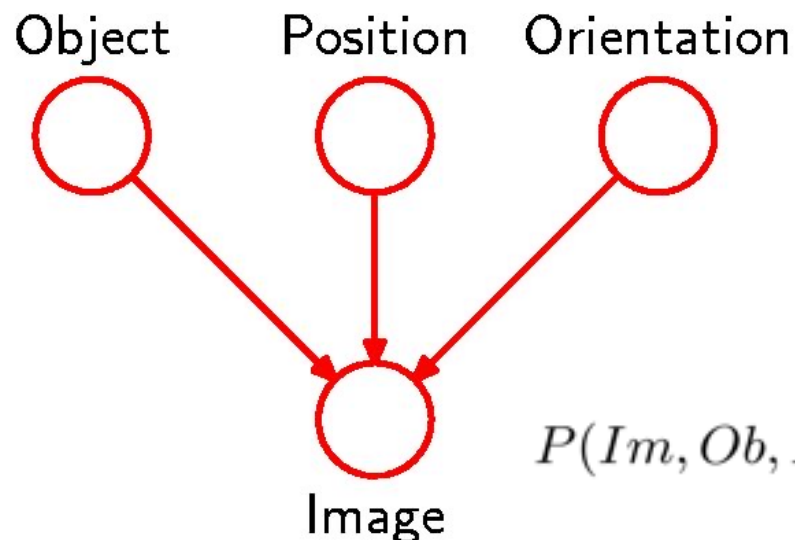
The parent variables are set to their sampled values

- To obtain a sample from **the marginal distribution**, e.g. $p(x_2, x_5)$, we sample from the full joint distribution, retain \hat{x}_2, \hat{x}_5 , and discard the remaining values.

Generative Models

- Higher-level nodes will typically represent **latent (hidden) random variables**.
- The primary role of the latent variables is to allow a complicated distribution over observed variables to be constructed from simpler (**typically exponential family**) conditional distributions.

Generative Model of an Image



- Object identity, position, and orientation have independent **prior probabilities**.
- The image has a probability distribution that depends on the object identity, position, and orientation (**likelihood function**).

$$P(Im, Ob, Po, Or) = \underbrace{P(Im|Ob, Po, Or)}_{\text{Likelihood}} \underbrace{P(Ob)P(Po)P(Or)}_{\text{Prior}}$$

- The graphical model captures the **causal process**, by which the observed data was generated (hence the name **generative models**).

Discrete Variables

- We now examine the discrete random variables.
- Assume that we have two discrete random variables x_1 and x_2 , each of which has K states.



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

- Using 1-of- K encoding, we denote the probability of **observing both** $x_{1k}=1$, $x_{2l}=1$ by the parameter μ_{kl} , where x_{1k} denotes the k^{th} component of x_1 (similarly for x_2).
- This distribution is governed by $K^2 - 1$ parameters.
- The total number of parameters that must be specified for an arbitrary joint distribution over M random variables is $K^M - 1$ (corresponds to a **fully connected graph**).
- **Grows exponentially** in the number of variables M !

Discrete Variables

- General joint distribution: K^2-1 parameters.



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

- Independent joint distribution: $2(K-1)$ parameters.



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

- We dropped the link between the nodes, so each variables is described by a separate multinomial distribution.

Discrete Variables

- In general:
 - Fully connected graphs have completely general distributions and have exponential K^M-1 number of parameters (**too complex**).
 - If there are no links, the joint distribution fully factorizes into the product of the marginals, and has $M(K-1)$ parameters (**too simple**).
 - Graphs that have an **intermediate level of connectivity** allow for more general distributions compared to the fully factorized one, while requiring fewer parameters than the general joint distribution.
- Let us look at the example of the chain graph.

Chain Graph

- Consider an M-node Markov chain:



- The marginal distribution $p(\mathbf{x}_1)$ requires $K-1$ parameters.
- The remaining conditional distributions $p(\mathbf{x}_i|\mathbf{x}_{i-1}), i = 2, \dots, M$ require $K(K-1)$ parameters.
- Total number of parameters: $K-1 + (M-1)(K-1)K$, which is quadratic in K and linear in the length M of the chain.
- This graphical model forms the basis of a simple **Hidden Markov Model**.

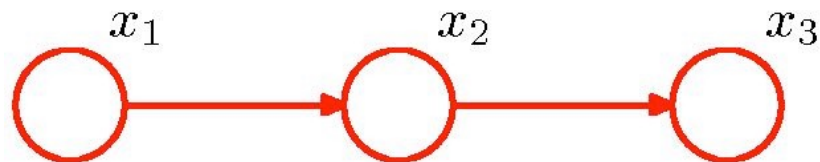
Linear Gaussian Models

- The log of the joint distribution takes form:

$$\ln p(\mathbf{x}) = \sum_{i=1}^D \ln p(x_i | \text{pa}_i) = - \sum_{i=1}^D \frac{1}{2v_i} \left(x_i - \sum_{j \in \text{pa}_i} w_{ij} x_j - b_i \right)^2 + \text{const},$$

where 'const' denotes terms independent of \mathbf{x} .

- This is a quadratic function of \mathbf{x} , and hence the joint distribution $p(\mathbf{x})$ is a **multivariate Gaussian**.
- For example, consider a directed graph over three Gaussian variables with one missing link:



Computing the Mean

- We can determine the mean and covariance of the joint distribution.

Remember:

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

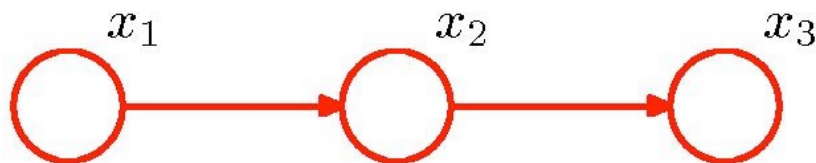
hence

$$x_i = \sum_{j \in \text{pa}_i} w_{ij} x_j + b_i + \sqrt{v_i} \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1),$$

so its expected value:

$$\mathbb{E}[x_i] = \sum_{j \in \text{pa}_i} w_{ij} \mathbb{E}[x_j] + b_i.$$

- Hence we can find components: $\mathbb{E}[\mathbf{x}] = [\mathbb{E}[x_1], \dots, \mathbb{E}[x_D]]$ by doing **ancestral pass**: start at the top and proceed in order (see example):



Computing the Covariance

- We can obtain the i, j element of the covariance matrix in the form of a recursion relation:

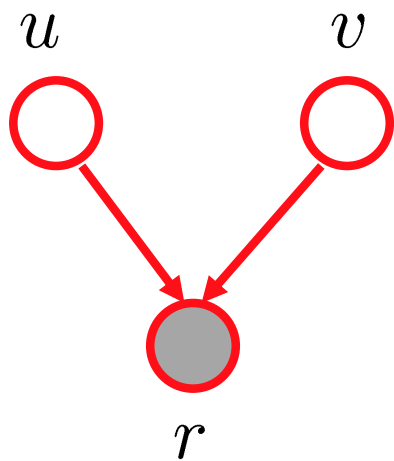
$$\begin{aligned}\text{cov}[x_i, x_j] &= \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] \\ &= \mathbb{E} \left[(x_i - \mathbb{E}[x_i]) \left(\sum_{k \in \text{pa}_j} w_{jk} (x_k - \mathbb{E}[x_k]) + \sqrt{v_j} \epsilon_j \right) \right] \\ &= \sum_{k \in \text{pa}_j} w_{jk} \text{cov}[x_i,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

- Consider two cases:
 - There are no links in the graph (**graph is fully factorized**), so that w_{ij} 's are zero. In this case: $\mathbb{E}[\mathbf{x}] = [b_1, \dots, b_D]^T$, and the covariance is diagonal $\text{diag}(v_1, \dots, v_D)$. The joint distribution represents D independent univariate Gaussian distributions.
 - The graph is **fully connected**. The total number of parameters is $D + D(D-1)/2$. The covariance corresponds to a general symmetric covariance matrix.

Bilinear Gaussian Model

- Consider the following model:

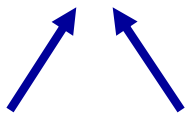
	🎵	🎵	🎵	🎵	🎵
👤	★★☆	?	?	★★☆	★★☆
👤	?	★★☆	★★★★	?	★★★★
👤	★★★★	?	★★☆	★★★★	?



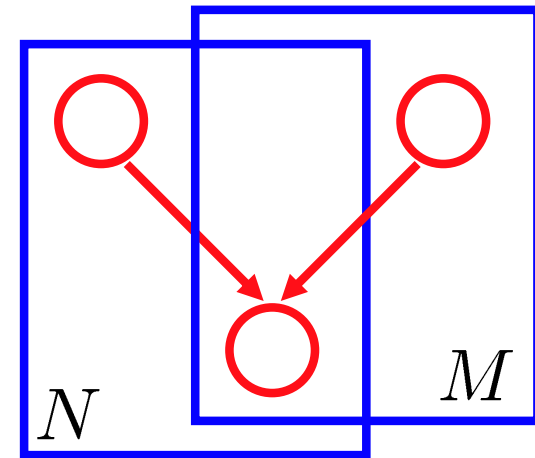
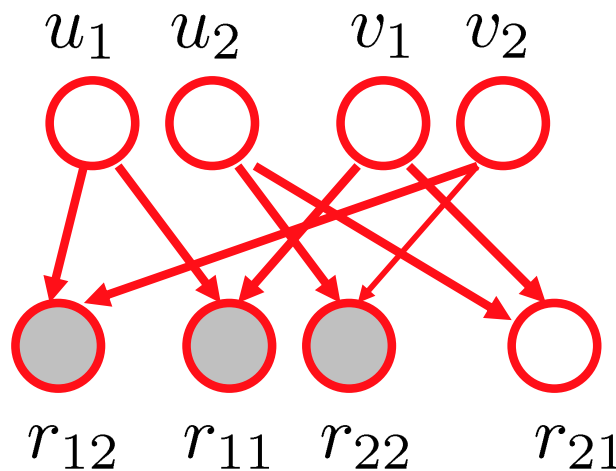
$$u \sim \mathcal{N}(0, 1),$$

$$v \sim \mathcal{N}(0, 1),$$

$$r \sim \mathcal{N}(uv, 1).$$



Gaussian terms



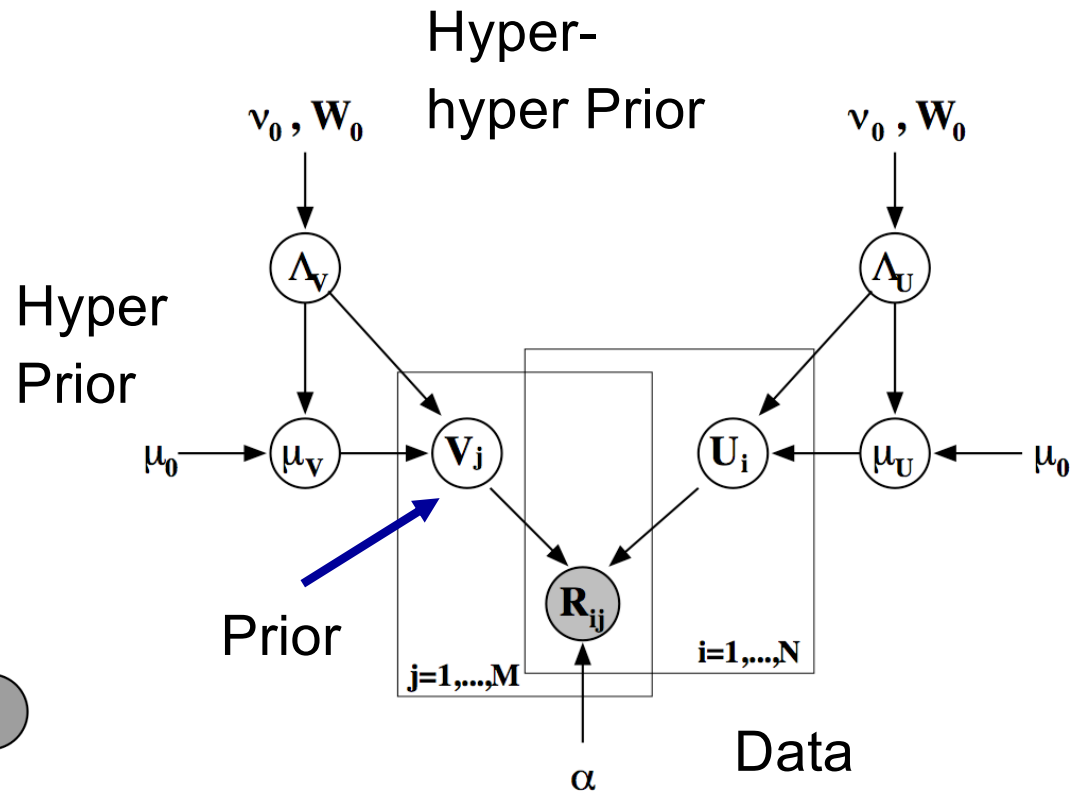
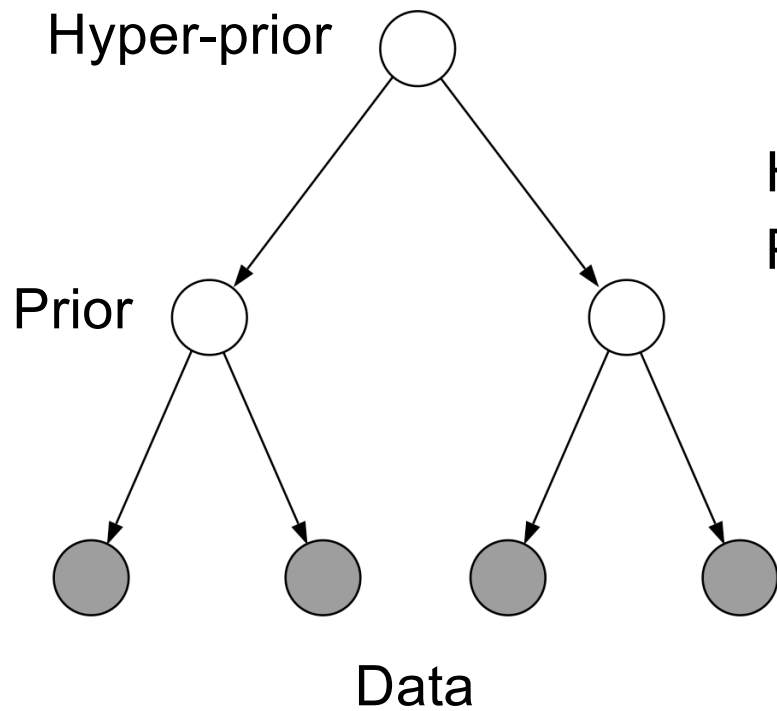
$$u_i \sim \mathcal{N}(0, 1), \quad i = 1, \dots, N$$

$$v_j \sim \mathcal{N}(0, 1), \quad j = 1, \dots, M$$

$$r_{ij} \sim \mathcal{N}(u_i v_j, 1).$$

- The mean is given by the product of two Gaussians.

Hierarchical Models



Conditional Independence

- We now look at the concept of conditional independence.
- **a** is independent of **b** given **c**:

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

- Equivalently:

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

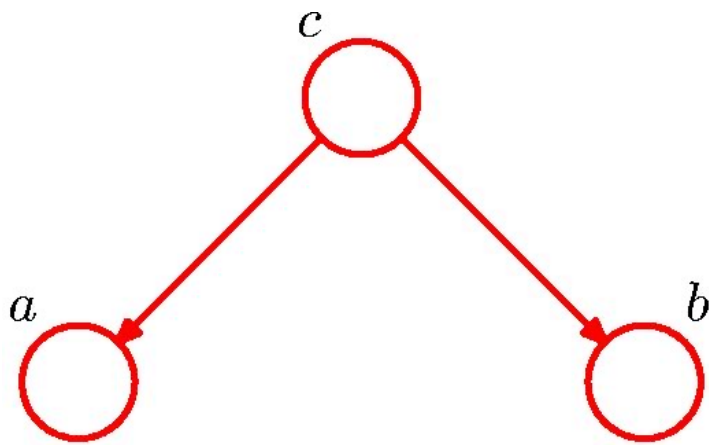
- We will use the notation:

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

- An important feature of graphical models is that **conditional independence properties** of the joint distribution can be read directly from the graph without performing any analytical manipulations
- The general framework for achieving this is called **d-separation**, where d stands for 'directed' (Pearl 1988).

Example 1: Tail-to-Tail Node

- The joint distribution over three variables can be written:



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

- If none of the variables are observed, we can examine whether **a** and **b** are independent:

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

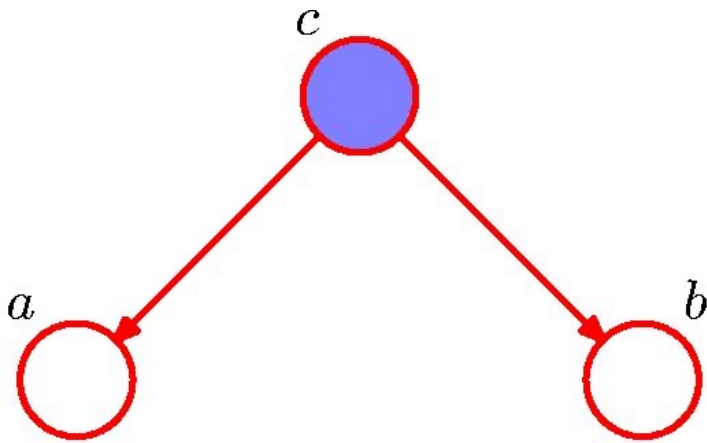
- In general, this does not factorize into the product $p(a, b) = p(a)p(b)$.

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

- **a** and **b** have a **common cause**.
- The node **c** is said to be **tail-to-tail node** with respect to this path (the node is connected to the tails of the two arrows).

Example 1: Tail-to-Tail Node

- Suppose we condition on the variable c :



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

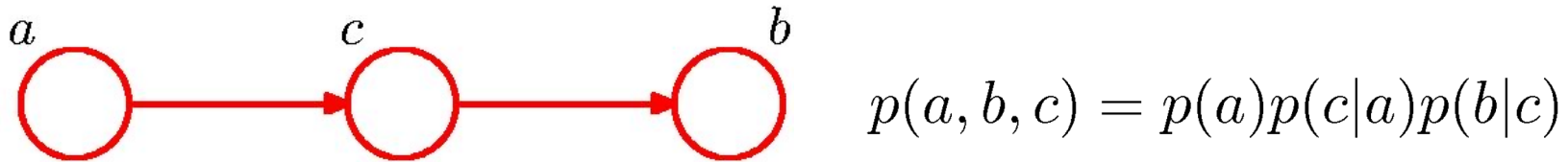
- We obtain **conditional independence property**:

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

- Once c has been **observed**, a and b can no longer have any effect on each other. They become independent.

Example 2: Head-to-Tail Node

- The joint distribution over three variables can be written:



- If none of the variables are observed, we can examine whether **a** and **b** are independent:

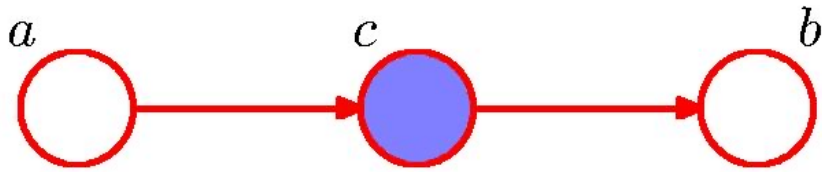
$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

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

- If **c** is not observed, **a** can influence **c**, and **c** can influence **b**.
- The node **c** is said to be **head-to-tail node** with respect to the path from node **a** to node **b**.

Example 2: Head-to-Tail Node

- Suppose we condition on the variable c :



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

- We obtain **conditional independence property**:

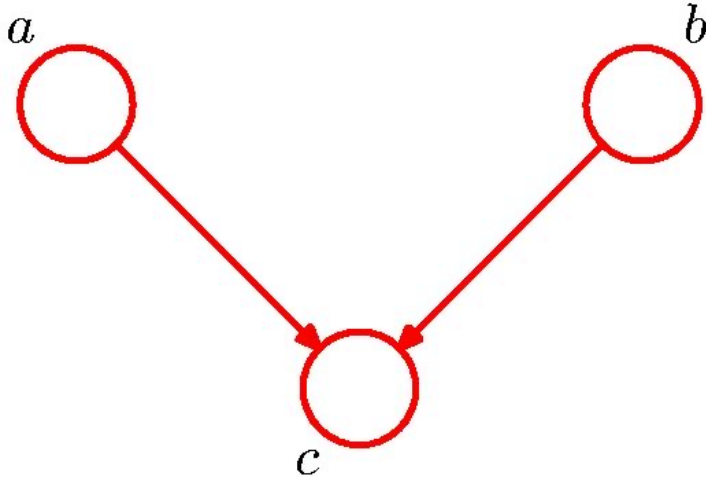
$$a \perp\!\!\!\perp b \mid c$$

- If c is observed, the value of a can no longer influence b .

Example 3: Head-to-Head Node

- The joint distribution over three variables can be written:

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



- If none of the variables are observed, we can examine whether **a** and **b** are independent:

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

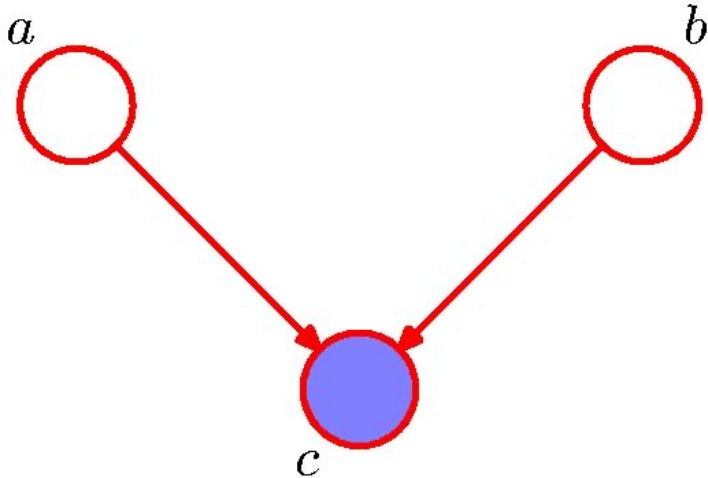
$$a \perp\!\!\!\perp b \mid \emptyset$$

- Opposite to Example 1.

- An unobserved descendant has no effect.
- The node **c** is said to be **head-to-head** node with respect to the path from **a** to **b** (because it connects to the heads of two arrows).

Example 3: Head-to-Head Node

- Suppose we condition on the variable c :



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

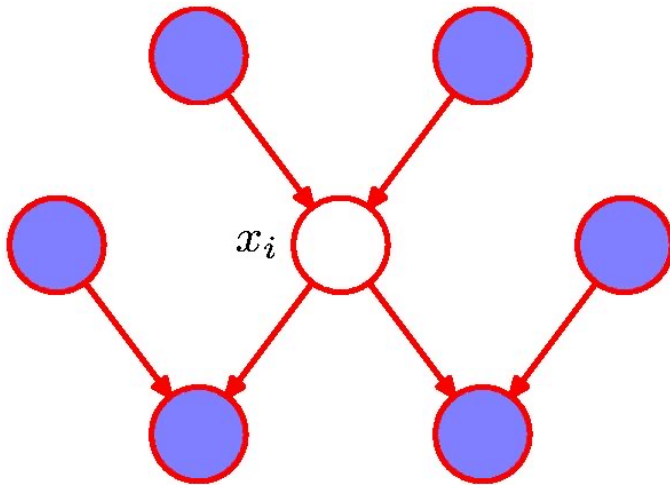
- In general, this does not factorize into the product.

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

- Opposite to Example 1.
- If the descendant (or any of its descendants) is observed, its value has implications for both a and b ,

Markov Blanket in Directed Models

- The **Markov blanket** of a node is the minimal set of nodes that must be observed to make this node independent of all other nodes
- In a directed model, the Markov blanket includes **parents, children** and **co-parents** (i.e. all the parents of the node's children) due to explaining away.

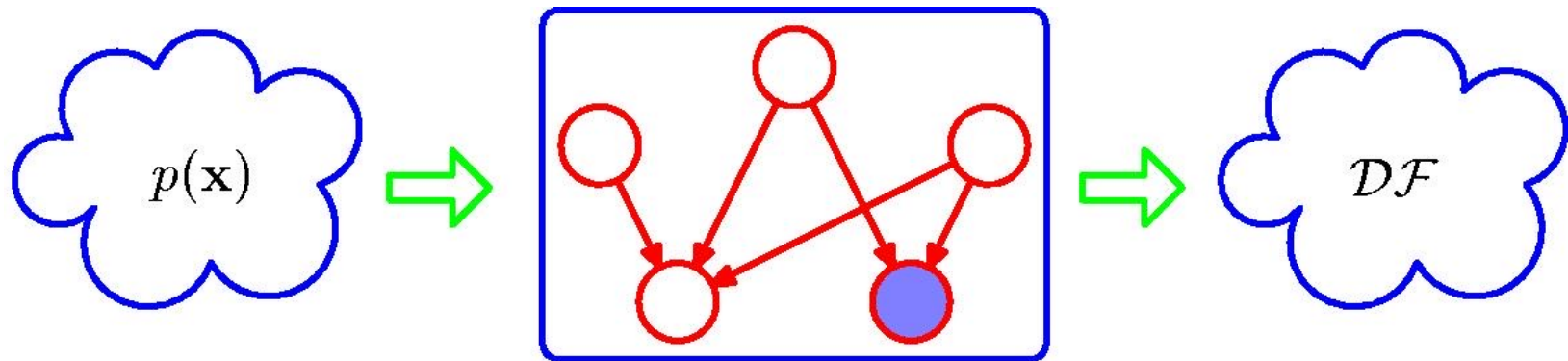


$$\begin{aligned} p(\mathbf{x}_i | \mathbf{x}_{\{j \neq i\}}) &= \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_M)}{\int p(\mathbf{x}_1, \dots, \mathbf{x}_M) d\mathbf{x}_i} \\ &= \frac{\prod_k p(\mathbf{x}_k | \text{pa}_k)}{\int \prod_k p(\mathbf{x}_k | \text{pa}_k) d\mathbf{x}_i} \end{aligned}$$

Factors independent of x_i cancel between numerator and denominator.

Directed Graphs as Distribution Filters

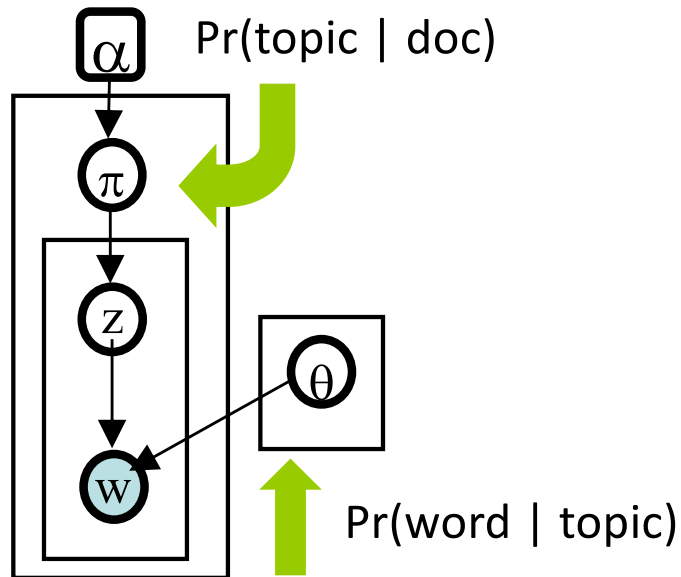
- We can view the graphical model as a filter.



- The joint probability distribution $p(\mathbf{x})$ is allowed through the filter if and only if it satisfies the factorization property.
- Note: The fully connected graph exhibits **no conditional independence properties** at all.
- The fully disconnected graph (no links) corresponds to a joint distribution that factorizes into the **product of marginal distributions**.

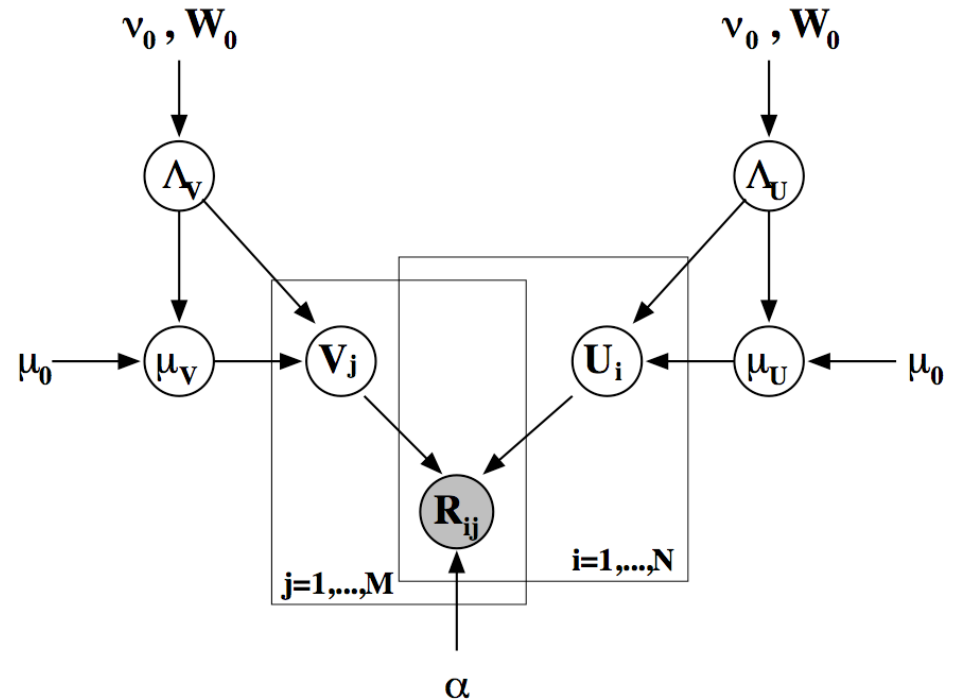
Popular Models

Latent Dirichlet Allocation



- One of the popular models for modeling word count vectors. We will see this model later.

Bayesian Probabilistic Matrix Factorization



- One of the popular models for collaborative filtering applications.