

Bayesian Network

Jin Young Choi

Seoul National University

Outline

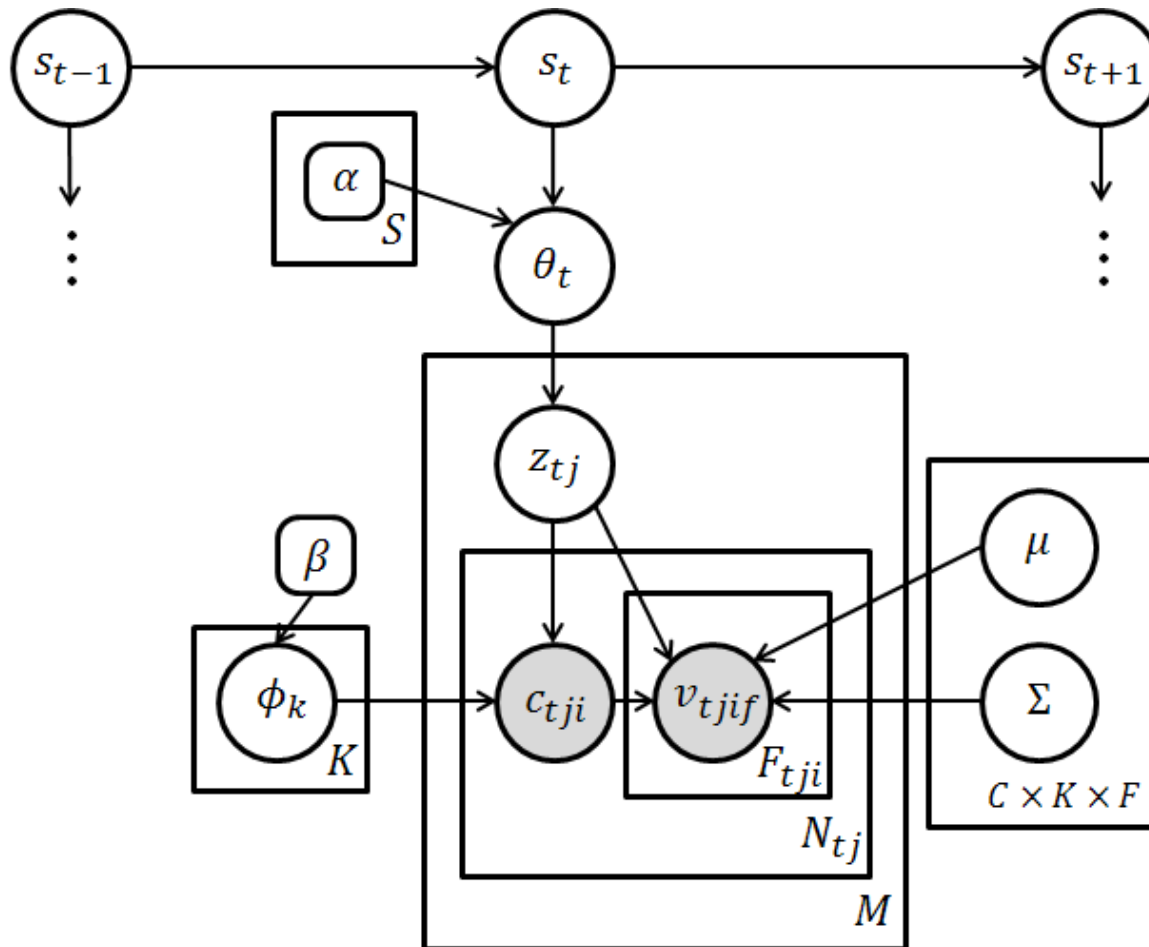
- Bayesian Networks
 - Applications:
 - Traffic Pattern Analysis
 - Topic Model (Document Analysis)
 - Directed Acyclic Graph
 - Conditional Independence
 - D-separation
 - Bayesian Parameters
 - Parameterized Conditional Distributions
 - Multinomial, Dirichlet Distribution, Conjugate Prior
 - Markov Blanket
-

Application: Traffic Pattern Analysis

- Surveillance in crowded scenes

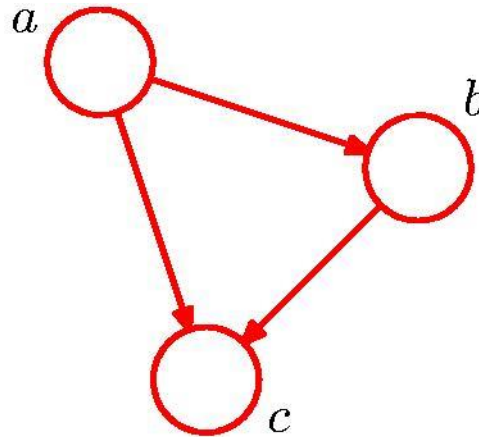


Graphical Inference Model



Bayesian Networks

- Directed Acyclic Graph (DAG)

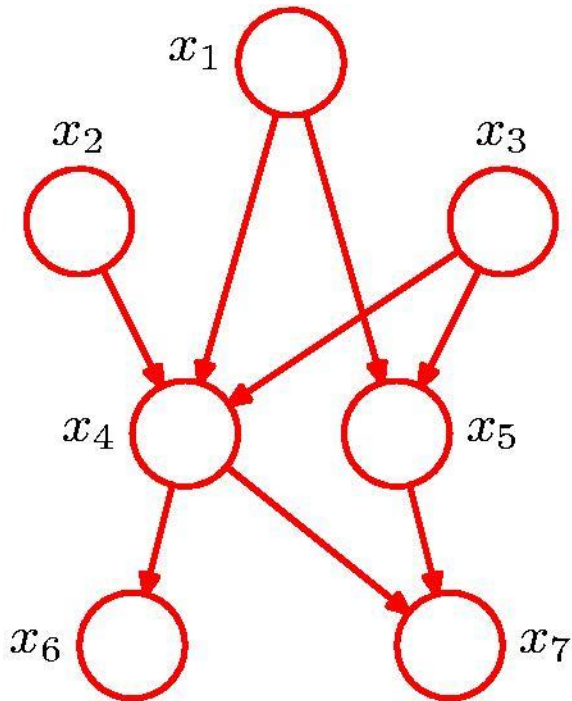


$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

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

Bayesian Networks

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



General Factorization

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

LDA Model (Topic Modelling)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

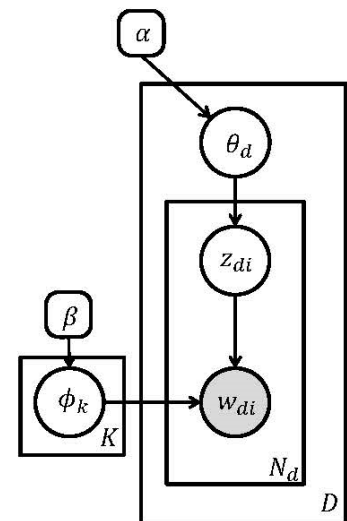
COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions** "are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a biologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

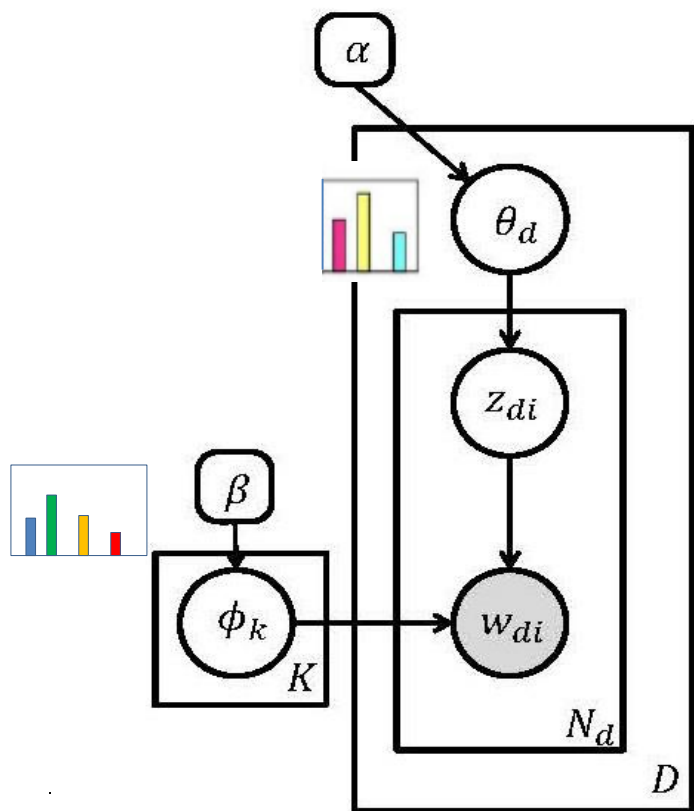


LDA Model

Likelihood: $p(w|z, \theta, \phi, \alpha, \beta)$

Posteriori: $p(z, \theta, \phi|w, \alpha, \beta)$

- $\text{Dir}(K, \alpha): p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \mu_1^{\alpha_1-1} \dots \mu_K^{\alpha_K-1}$
- $\text{Mul}(K, \mu): p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$



Notations

D : the number of documents.

N_d : the number of words in d -th document.

K : the number of topics.

α : Dirichlet prior on the per-document topic distributions.

β : Dirichlet prior on the per-topic word distribution.

θ_d : topic distribution for d -th document.

ϕ_k : word distribution for topic k .

z_{di} : the topic for the i -th word in d -th document.

w_{di} : the specific word.

$$\{w_{d1}, w_{d2}, \dots, w_{dN_d}\}$$

Mathematical description

Choose $\theta_d \sim \text{Dir}(\alpha)$.

Choose $\phi_k \sim \text{Dir}(\beta)$.

Choose a topic $z_{ji} | \theta_d \sim \text{Multi}(\theta_d)$.

Choose a word $w_{ji} | \phi_k, z_{di} \sim \text{Multi}(\phi_{z_{di}})$.

Inference of LDA Model

Likelihood: $p(w|z, \theta, \phi, \alpha, \beta)$

Posteriori: $p(z, \theta, \phi|w, \alpha, \beta)$

- Maximum A posteriori Probability (MAP) given observation w, α, β

$$\hat{\phi}, \hat{\theta}, \hat{z} = \arg \max_{\phi, \theta, z} p(\phi, \theta, z|w, \alpha, \beta),$$

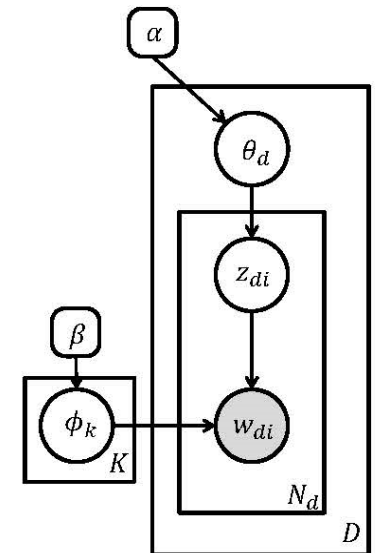
Not Convex

Closed-form solution is not available

- Bayesian Inference (Learning)

$$\begin{aligned} p(\phi, \theta, z|w, \alpha, \beta) &= \frac{p(\phi, \theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}, \\ &= \frac{p(\phi, \theta, z, w|\alpha, \beta)}{\int_{\phi} \int_{\theta} \sum_z p(\phi, \theta, z, w|\alpha, \beta) d\theta d\phi}. \end{aligned}$$

Likelihoods



$$p(\phi, \theta, z, w|\alpha, \beta) = \left(\prod_{k=1}^K p(\phi_k|\beta) \right) \prod_{d=1}^D p(\theta_d|\alpha) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) p(w_{di}|z_{di}, \phi).$$

Conditional Independence

- a is independent of b given c

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

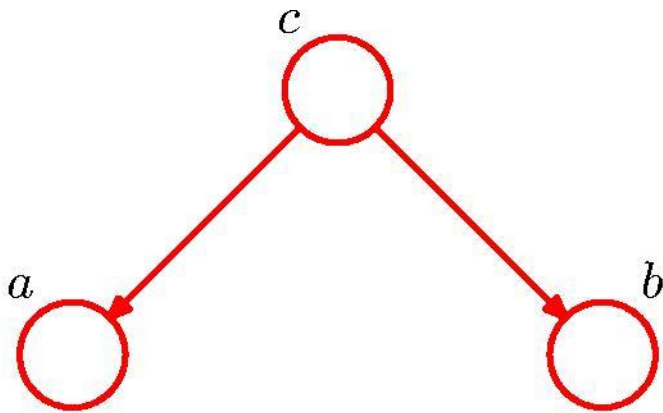
- Equivalently

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

- Notation

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

Conditional Independence: Example 1



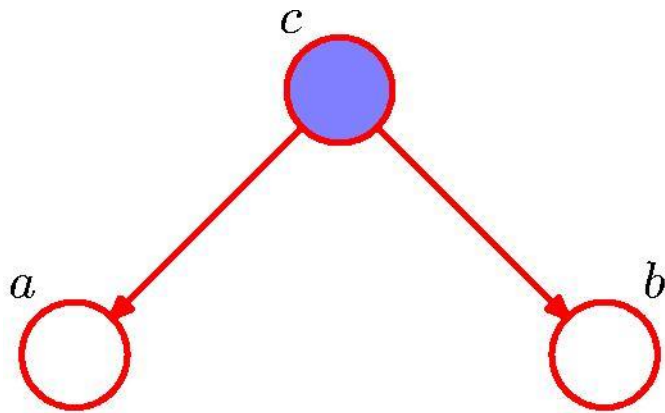
$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

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

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

U, V , and c are independent. $a = U + c, b = V + c$; a, b independent?

Conditional Independence: Example 1



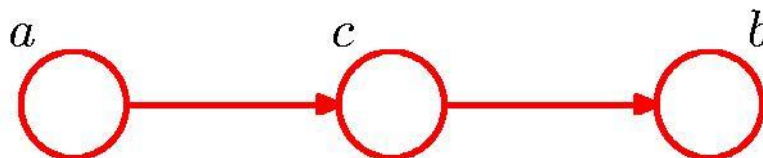
$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

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

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

U, V , and c are independent. $a = U + c, b = V + c, c = 1$; a, b independent?

Conditional Independence: Example 2



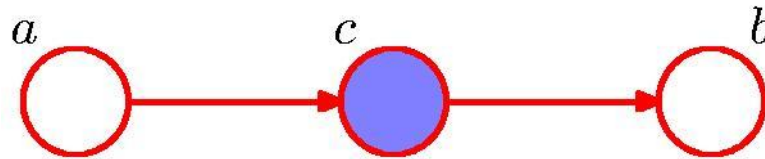
$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

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

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

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

Conditional Independence: Example 2

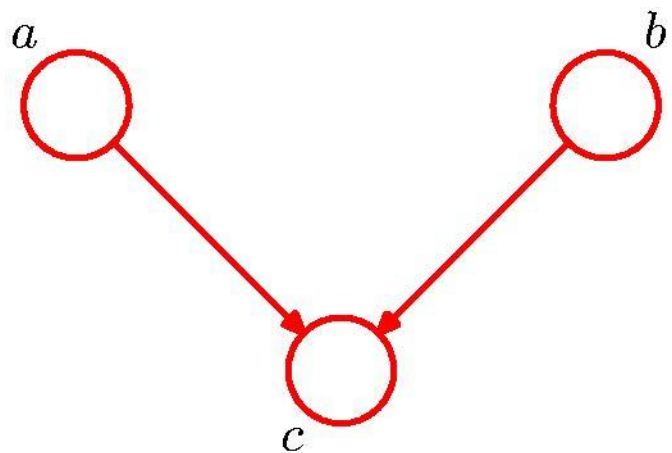


$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

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

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

Conditional Independence: Example 3



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

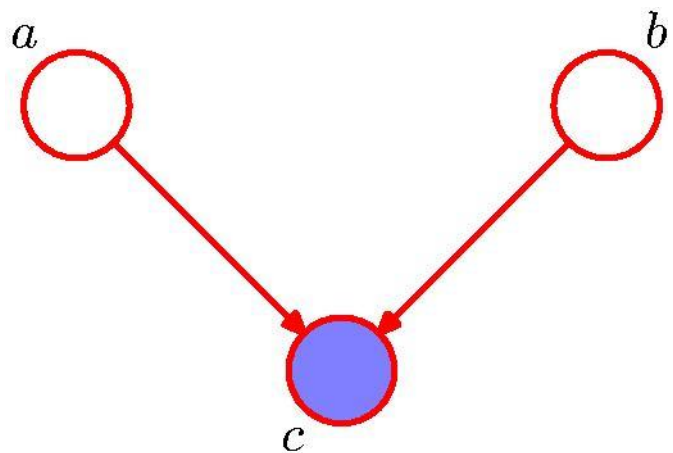
$$p(a, b) = p(a)p(b)$$

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

Note: this is the opposite of Example 1, with c unobserved.

a and b are independent Bernoulli rvs. $c = a + b$

Conditional Independence: Example 3



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

$$a \not\perp b | c$$

Note: this is the opposite of Example 1, with C observed.

a and b are independent Bernoulli rvs. $c = a + b$

“Am I out of fuel?”

$$p(G = 1 | B = 1, F = 1) = 0.8$$

$$p(G = 1 | B = 1, F = 0) = 0.2$$

$$p(G = 1 | B = 0, F = 1) = 0.2$$

$$p(G = 1 | B = 0, F = 0) = 0.1$$

➤ G is dependent to B and F

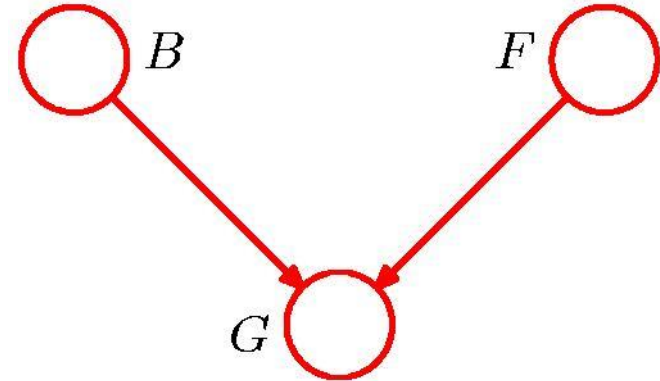
$$p(B = 1) = 0.9$$

$$p(F = 1) = 0.9$$

and hence

$$p(F = 0) = 0.1$$

➤ F is independent to B



B = Battery (0=flat, 1=fully charged)

F = Fuel Tank (0=empty, 1=full)

G = Fuel Gauge Reading
(0=empty, 1=full)

“Am I out of fuel?”

$$p(G = 1|B = 1, F = 1) = 0.8$$

$$p(G = 1|B = 1, F = 0) = 0.2$$

$$p(G = 1|B = 0, F = 1) = 0.2$$

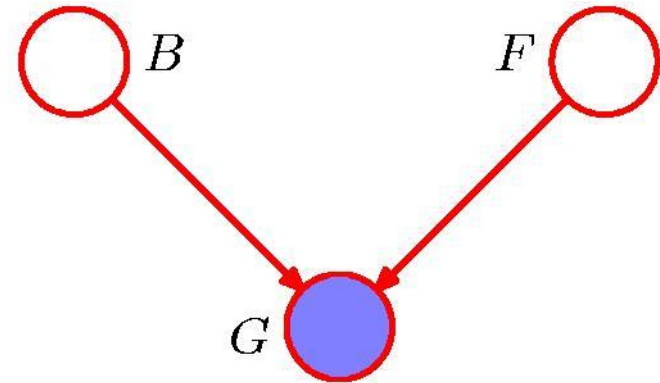
$$p(G = 1|B = 0, F = 0) = 0.1$$

$$p(B = 1) = 0.9$$

$$p(F = 1) = 0.9$$

and hence

$$p(F = 0) = 0.1$$



$$\begin{aligned} p(F = 0|G = 0) &= \frac{p(G = 0|F = 0)p(F = 0)}{p(G = 0)} \\ &\simeq 0.257 \end{aligned}$$

Probability of an empty tank increased by observing $G = 0$.

“Am I out of fuel?”

$$p(G = 1|B = 1, F = 1) = 0.8$$

$$p(G = 1|B = 1, F = 0) = 0.2$$

$$p(G = 1|B = 0, F = 1) = 0.2$$

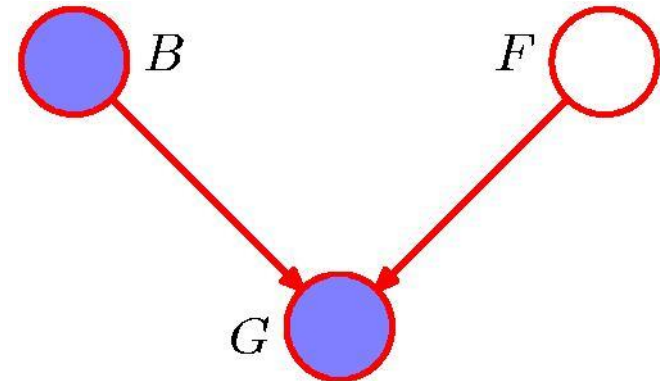
$$p(G = 1|B = 0, F = 0) = 0.1$$

$$p(B = 1) = 0.9$$

$$p(F = 1) = 0.9$$

and hence

$$p(F = 0) = 0.1$$



$$\begin{aligned} p(F = 0|G = 0, B = 0) &= \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0|B = 0, F)p(F)} \\ &\simeq 0.111 \end{aligned}$$

Probability of an empty tank reduced by observing $B = 0$.

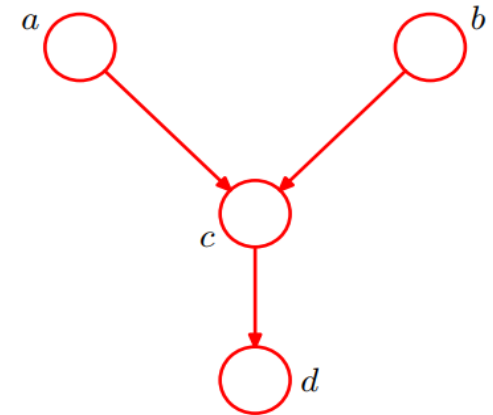
This referred to as “explaining away”.

➤ F is dependent to B given G

Exercise

Answer the following questions for the right-hand Bayesian network.

- 1) When any random variables are not observed, show that a and b are independent to each other.
- 2) When d is observed, show that a and b are dependent to each other.



1) Nothing observed

$$\begin{aligned}
 p(a, b, c, d) &= [\quad \quad \quad] \\
 p(a, b) &= \Sigma_c \Sigma_d p(a, b, c, d) = p(a)p(b) \Sigma_c \Sigma_d [\quad \quad] \\
 &= p(a)p(b) \Sigma_c [\quad \quad] \Sigma_d [\quad \quad] \\
 &= [\quad \quad]
 \end{aligned}$$

2) d is observed

$$\begin{aligned}
 p(a, b|d) &= \frac{p(a, b, d)}{p(d)} = \Sigma_c \left\{ \quad \quad \quad \right\} \\
 &= \Sigma_c \left\{ \quad \quad \quad \right\} = \frac{p(a)p(b)}{p(d)} \Sigma_c \left\{ \quad \quad \quad \right\} \\
 &= \frac{[\quad \quad \quad]}{p(d)} \neq p(a|d)p(b|d)
 \end{aligned}$$

Exercise

Answer the following questions for the right-hand Bayesian network.

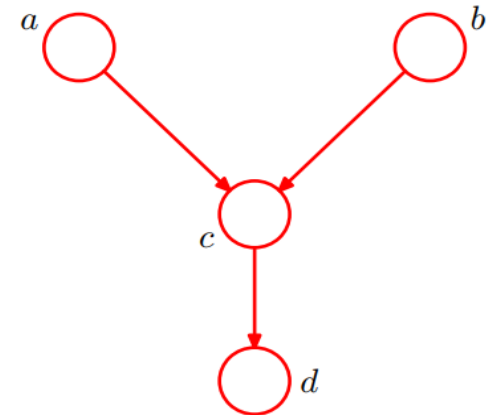
- 1) When any random variables are not observed, show that a and b are independent to each other.
- 2) When d is observed, show that a and b are dependent to each other.

1) Nothing observed

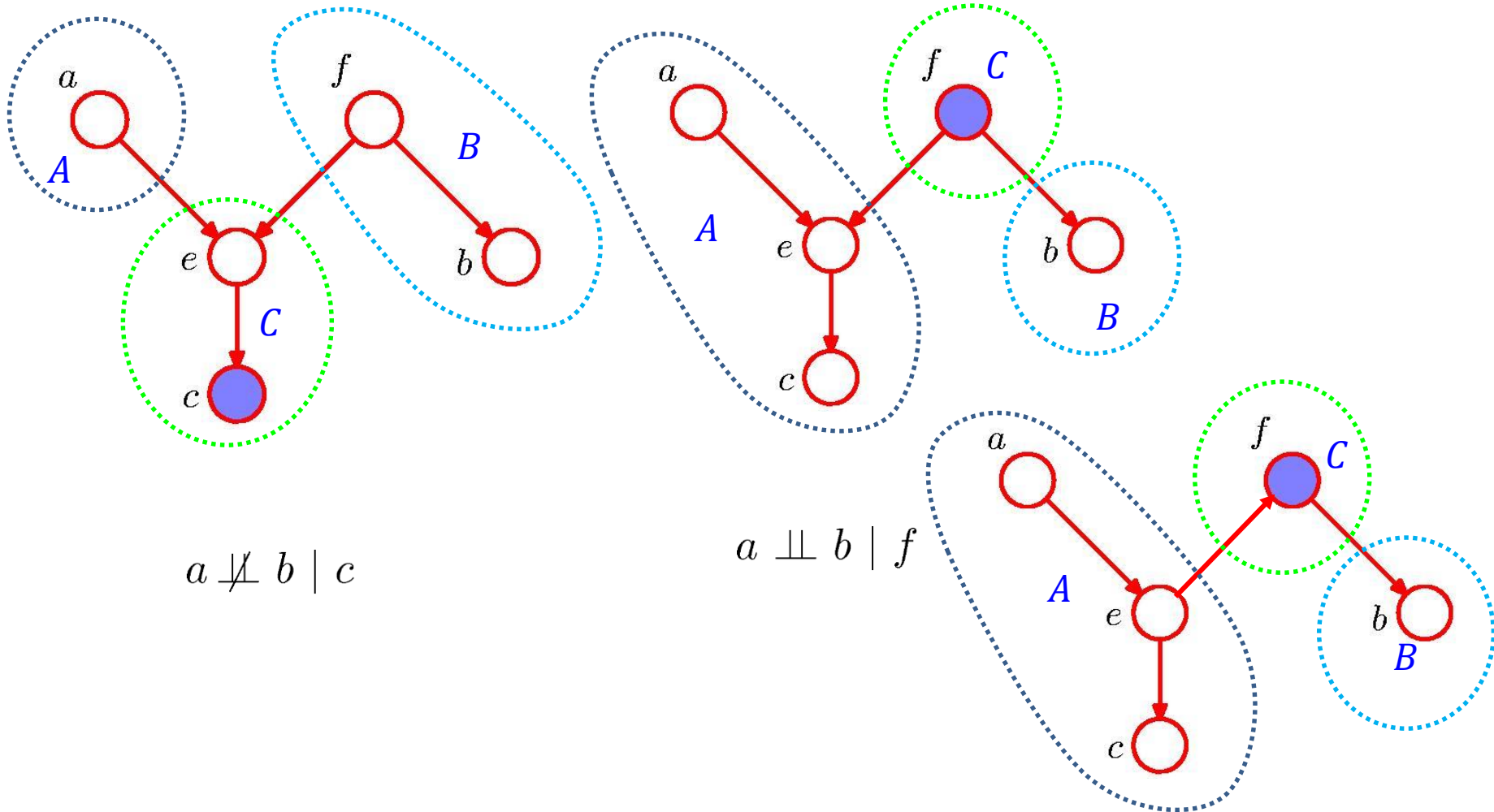
$$\begin{aligned} p(a, b, c, d) &= [p(a)p(b)p(c|a, b)p(d|c)] \\ p(a, b) &= \Sigma_c \Sigma_d p(a, b, c, d) = p(a)p(b) \Sigma_c \Sigma_d [p(c|a, b)p(d|c)] \\ &= p(a)p(b) \Sigma_c [p(c|a, b)] \Sigma_d [p(d|c)] \\ &= [p(a)p(b)] \end{aligned}$$

2) d is observed

$$\begin{aligned} p(a, b|d) &= \frac{p(a, b, d)}{p(d)} = \Sigma_c \left\{ \frac{p(a, b, c, d)}{p(d)} \right\} \\ &= \Sigma_c \left\{ \frac{p(a)p(b)p(c|a, b)p(d|c)}{p(d)} \right\} = \frac{p(a)p(b)}{p(d)} \Sigma_c \{ p(c|a, b)p(d|c) \} \\ &= \frac{[p(a)p(b)p(d|a, b)]}{p(d)} \neq p(a|d)p(b|d) \end{aligned}$$



D-separation: Example

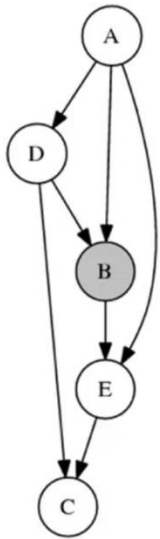


D-separation

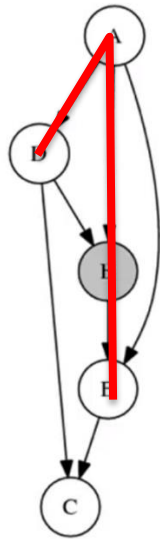
- A , B , and C are non-intersecting subsets of nodes in a directed graph.
 - A path from A to B is blocked if it contains a node such that either
 - the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C , or
 - the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set C .
 - If all paths from A to B are blocked, A is said to be d -separated from B by C .
 - If A is d -separated from B by C , the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$
-

Exercise

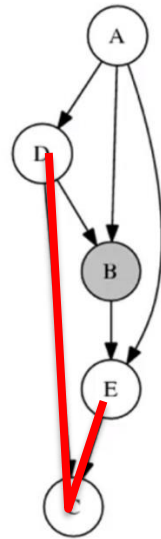
When B is observed in the following Bayesian network, decide whether every path from D to E is blocked (d -separated) or not and determine the dependency between D and E .



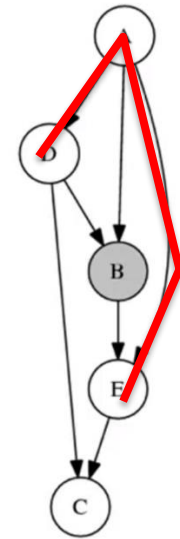
(a)



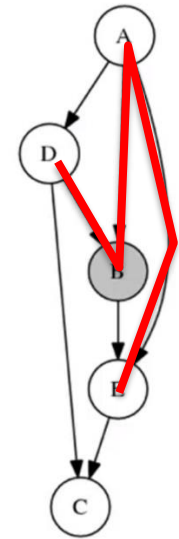
(b)



(c)



(d)



(e)

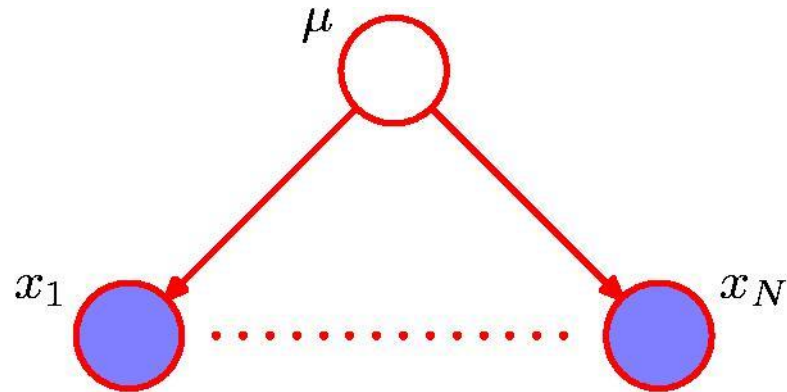
Exercise

- a. path 1 ($D \leftarrow A \rightarrow B \rightarrow E$) or ($D \rightarrow B \rightarrow E$) : Since the connection in B is head to tail and B is observed, the path 1 becomes [].
 - b. path 2 ($D \rightarrow C \leftarrow E$) : Since the connection in C is head to head and C is not observed, the path 2 becomes [].
 - c. path 3 ($D \leftarrow A \rightarrow E$) : Since the connection in A is tail to tail and A is not observed, the path 3 becomes [].
 - d. path 4 ($D \rightarrow B \leftarrow A \rightarrow E$) : Since the connection in B is head to head and B is observed, the path $D \rightarrow B \leftarrow A$ becomes [] by B. And since the connection in A is tail to tail and A is not observed, the path $B \leftarrow A \rightarrow E$ becomes []. Hence path 4 becomes [].
 - e. Among the above 4 paths, there [exists or does not exist] at least one non-blocking path and thus D and E are [dependent or independent] to each other.
-

Exercise

- a. path 1 ($D \leftarrow A \rightarrow B \rightarrow E$) or ($D \rightarrow B \rightarrow E$) : Since the connection in B is head to tail and B is observed, the path 1 becomes [blocking].
 - b. path 2 ($D \rightarrow C \leftarrow E$) : Since the connection in C is head to head and C is not observed, the path 2 becomes [blocking].
 - c. path 3 ($D \leftarrow A \rightarrow E$) : Since the connection in A is tail to tail and A is not observed, the path 3 becomes [non-blocking].
 - d. path 4 ($D \rightarrow B \leftarrow A \rightarrow E$) : Since the connection in B is head to head and B is observed, the path $D \rightarrow B \leftarrow A$ becomes [non-blocking] by B. And since the connection in A is tail to tail and A is not observed, the path $B \leftarrow A \rightarrow E$ becomes [non-blocking]. Hence path 4 becomes [non-blocking] .
 - e. Among the above 4 paths, there [exists or does not exist] at least one non-blocking path and thus D and E are [dependent or independent] to each other.
-

D-separation: I.I.D. Data



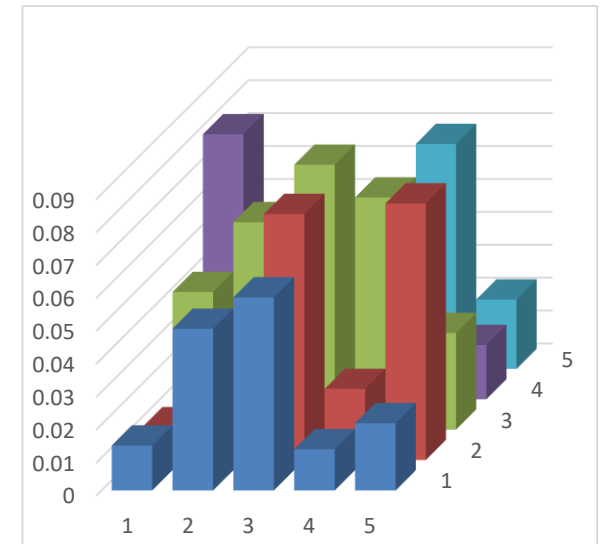
$$p(\mathcal{D}|\mu) = \prod_{n=1}^N p(x_n|\mu)$$

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu) d\mu \neq \prod_{n=1}^N p(x_n)$$

Discrete Variables, Multinomial

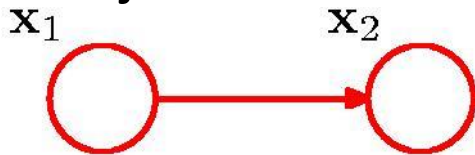
$$p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$$

$$p(x_{11}, \dots, x_{1K}, x_{21}, \dots, x_{2K} | \mu_{11}, \dots, \mu_{KK}) = \frac{n!}{x_{11}! \dots x_{1K}!} \frac{n!}{x_{21}! \dots x_{2K}!} \mu_{11}^{x_{11} x_{21}} \dots \mu_{KK}^{x_{1K} x_{2K}}$$



Discrete Variables (1), Multinomial

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

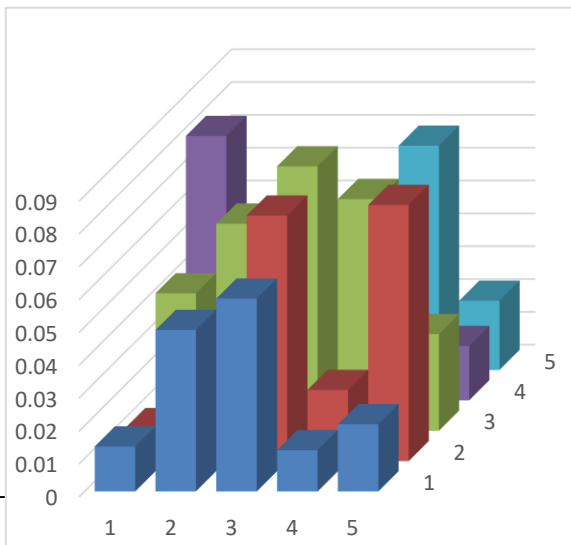


$$p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^K \prod_{l=1}^K \mu_{kl}^{x_{1k} x_{2l}}$$

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



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



$$p(x_1, x_2) = p(x_1 | x_2) p(x_2)$$

$$K - 1 + K(K - 1)$$

$$p(x_1, x_2) = p(x_1) p(x_2)$$

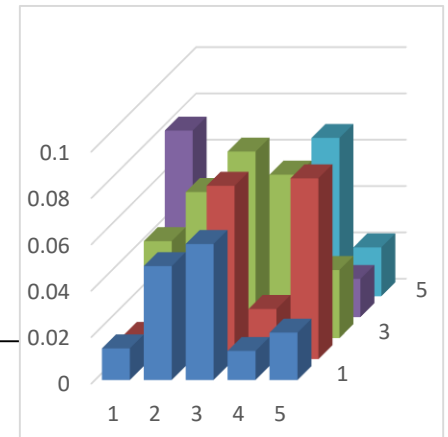
$$K - 1 + K - 1$$

Discrete Variables, Dirichlet

- The posterior distributions are in the same family as the prior probability distribution.

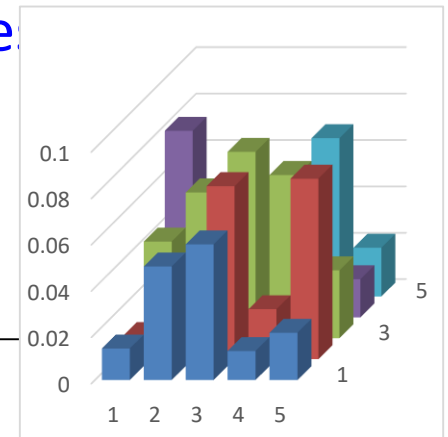
$$p(\mu|x) \propto p(x|\mu)p(\mu)$$

- The prior and posterior are then called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function.
- Dirichlet distribution is a conjugate (prior) distribution to the multinomial distribution.
- Gaussian is a conjugate prior of Gaussian.



Discrete Variables, Dirichlet

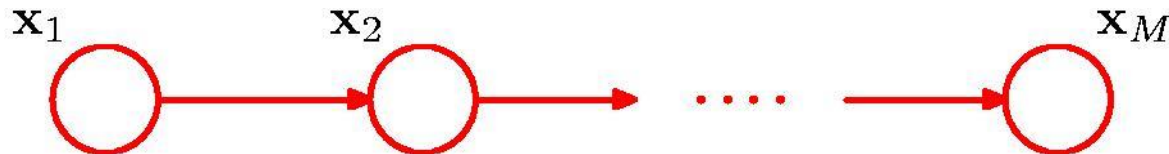
- Posteriori: $p(\mu|x, \alpha) \propto p(x|\mu)p(\mu|\alpha)$
- Mul(K, μ): $p(x_1, \dots, x_K | \mu_1, \dots, \mu_K) = \frac{n!}{x_1! \dots x_K!} \mu_1^{x_1} \dots \mu_K^{x_K}$
- Dir(K, α): $p(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K (\alpha_i - 1))}{\prod_{i=1}^K \Gamma(\alpha_i - 1)} \mu_1^{\alpha_1} \dots \mu_K^{\alpha_K}$
- Parameters: $\alpha_1, \dots, \alpha_K > 0$ (hyper-parameters)
- Support: $\mu_1, \dots, \mu_K \in (0,1)$ where $\sum_{i=1}^K \mu_i = 1$
- Dir($K, c + \alpha$): $p(\mu|x, \alpha) \propto p(x|\mu)p(\mu|\alpha)$
 where $c = (c_1, \dots, c_K)$ is number of occurrence
- $E[\mu_k] = \frac{c_k + \alpha_k}{\sum_{i=1}^K (c_i + \alpha_i)}$



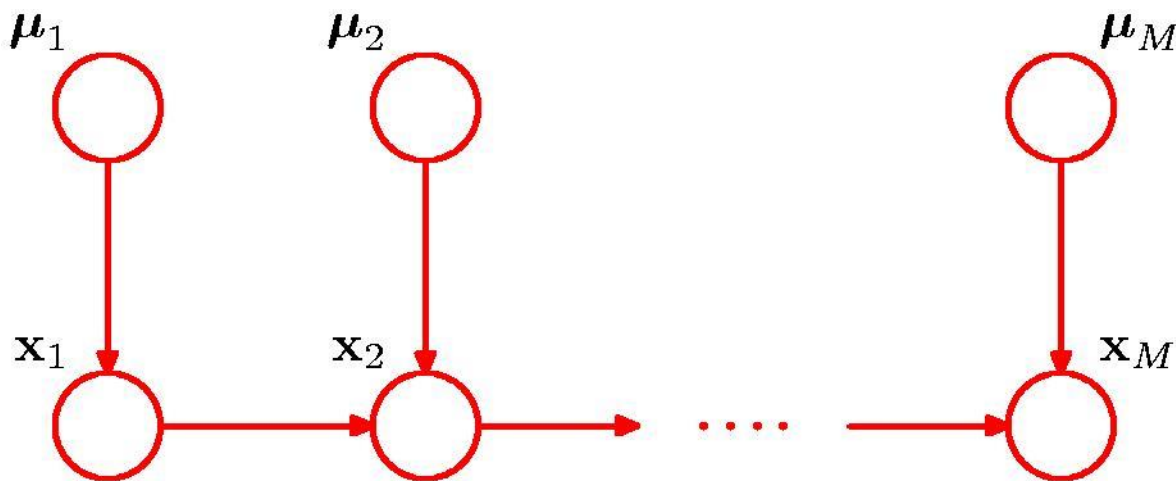
Discrete Variables (2)

- General joint distribution over M variables: $K^M - 1$ parameters
- M -node Markov chain: $K - 1 + (M - 1)K(K - 1)$ parameters

$$p(x_1, x_2, \dots, x_M) = p(x_1)p(x_2|x_1)p(x_3|x_2)\dots p(x_M|x_{M-1})$$



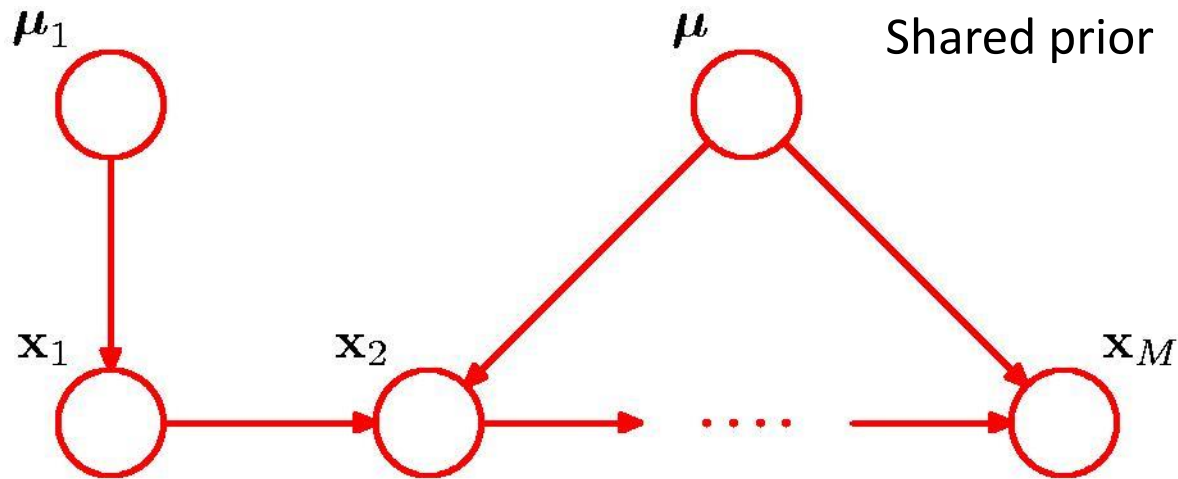
Discrete Variables: Bayesian Parameters (1)



$$p(\{\mathbf{x}_m, \boldsymbol{\mu}_m\}) = p(\mathbf{x}_1 | \boldsymbol{\mu}_1) p(\boldsymbol{\mu}_1) \prod_{m=2}^M p(\mathbf{x}_m | \mathbf{x}_{m-1}, \boldsymbol{\mu}_m) p(\boldsymbol{\mu}_m)$$

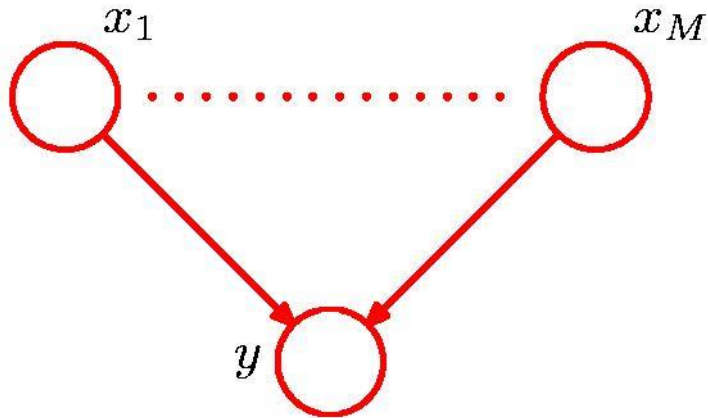
$$p(\boldsymbol{\mu}_m) = \text{Dir}(\boldsymbol{\mu}_m | \boldsymbol{\alpha}_m)$$

Discrete Variables: Bayesian Parameters (2)



$$p(\{\mathbf{x}_m\}, \mu_1, \mu) = p(\mathbf{x}_1 | \mu_1) p(\mu_1) \prod_{m=2}^M p(\mathbf{x}_m | \mathbf{x}_{m-1}, \mu) p(\mu)$$

Parameterized Conditional Distributions



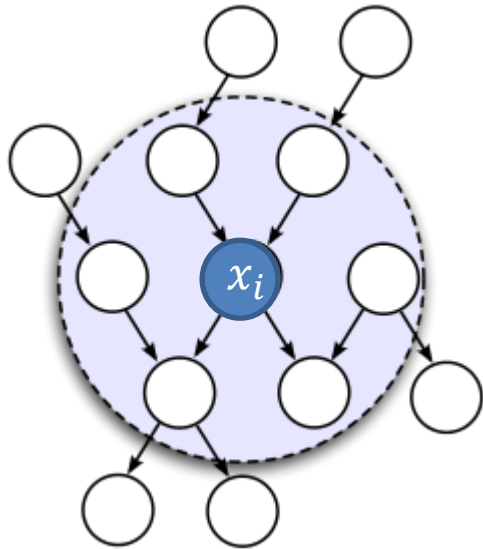
If x_1, \dots, x_M are discrete, K -state variables, $p(y = 1|x_1, \dots, x_M)$ in general has $O(K^M)$ parameters because $p(x_1, \dots, x_M|y = 1)$ requires $K^M - 1$ parameters.

The parameterized form

$$p(y = 1|x_1, \dots, x_M) = \sigma \left(w_0 + \sum_{i=1}^M w_i x_i \right) = \sigma(\mathbf{w}^T \mathbf{x})$$

requires only $M + 1$ parameters (actually this can not model a probability distribution).

The Markov Blanket



$$\begin{aligned} p(\mathbf{x}_i | \mathbf{x}_{\{j \neq i\}}) &= \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_M)}{\int p(\mathbf{x}_1, \dots, \mathbf{x}_M) d\mathbf{x}_i} \\ &= \frac{\prod_k p(\mathbf{x}_k | pa_k)}{\int \prod_k p(\mathbf{x}_k | pa_k) d\mathbf{x}_i} \\ &= \prod_{k \in MB} p(x_k | pa_k) \end{aligned}$$

Any factor $p(x_k | pa_k)$ that does not have any functional dependence on x_i can be taken outside the integral over x_i , and will therefore cancel between numerator and denominator.

LDA Model (Topic Modelling)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

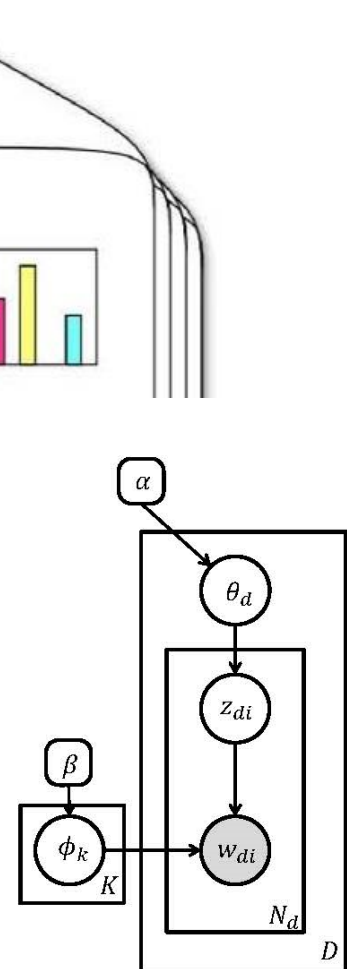
COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions** "are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a biologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

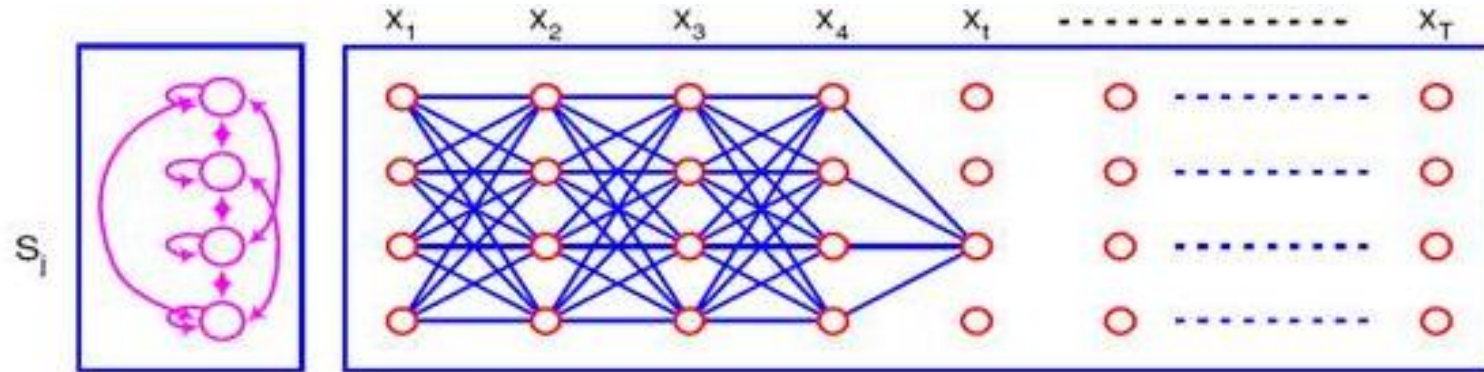


Application: Traffic Pattern Analysis

- Surveillance in crowded scenes

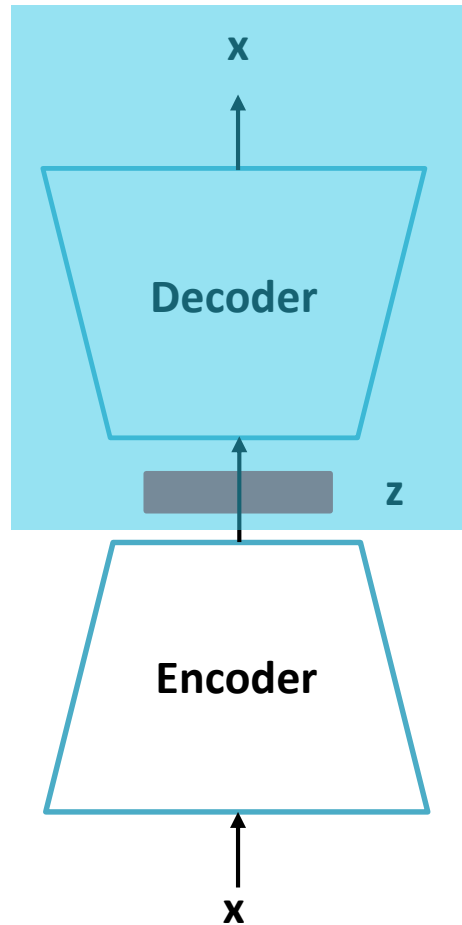


Hidden Markov Model



$$p(x_1, \dots, x_T) = p(x_1)p(x_2|x_1) p(x_3|x_2) \dots \dots \dots p(x_T|x_{T-1})$$

Variational Auto-encoder (VAE)



Reconstruction Loss

$$Loss = -\log P_{\theta}(x|z) + D_{KL}(q_{\phi}(z|x) || P_{\theta}(z))$$

Variational Inference

$p_{\theta}(x|z)$: a multivariate Gaussian (real-valued data)

a Bernoulli (binary-valued data)



Interim Summary

- Bayesian Networks
 - Directed Acyclic Graph
 - Conditional Independence
 - D-separation
 - Bayesian Parameters
 - Parameterized Conditional Distributions
 - Multinomial, Dirichlet Distribution, Conjugate Prior
 - Markov Blanket
 - Applications:
 - Topic Model (Document Analysis, Traffic Pattern Analysis)
 - Hidden Markov Model
 - Generative Models
-