CS60020: Foundations of Algorithm Design and Machine Learning

Sourangshu Bhattacharya

GAUSSIAN MIXTURE MODELS

Mixture of Gaussians

- $z \in \{0,1\}^K$: be a discrete latent variable, such that $\sum_k z_k = 1$.
- z_k selects the cluster (mixture component) from which the data point is generated.
- There are K Gaussian distributions:

$$\mathcal{N}(x|\mu_1, \Sigma_1)$$
$$\dots$$
$$\mathcal{N}(x|\mu_K, \Sigma_K)$$

Mixture of Gaussians

- Given a data point x:

$$P(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x|\mu_k, \Sigma_k)$$

- Where:

$$\pi_k = P(z_k = 1)$$

Generative Procedure

- Select z from probability distr. π_k.
- Hence: $P(z) = \prod_{k=1}^{K} \pi_k^{z_k}$.
- Given z, generate x according to the conditional distr.:

$$P(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

- Hence:

$$P(x|z) = \prod_{k=1}^{K} \left(\mathcal{N}(x|\mu_k, \Sigma_k)\right)^{z_k}$$

Generative Procedure

- Joint distr.:

$$P(x,z) = p(z)p(x|z)$$
$$= \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)\right)^{z_k}$$

- Marginal:

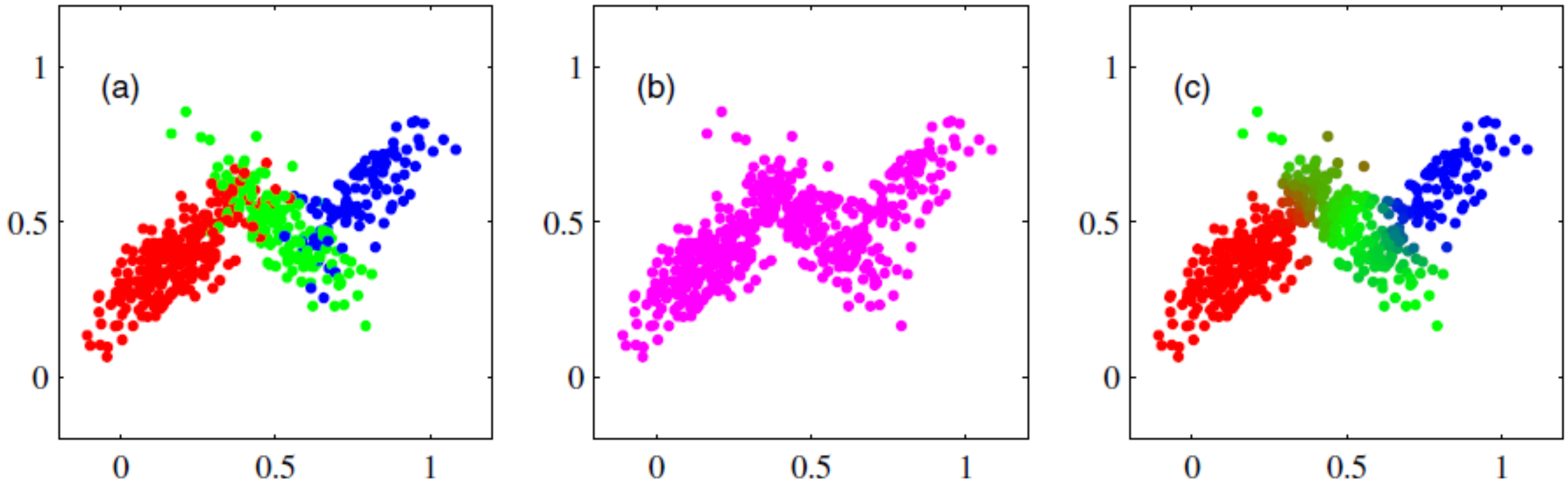
$$p(x) = \sum_{z} p(x,z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Posterior distribution

- $z_k = 1$ given x:

$$
\begin{aligned}
\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\
&= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.
\end{aligned}
$$

Example



Max-likelihood

- Let $D = \{x_1, \ldots, x_N\}$
- Likelihood function:

$$P(D|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

- Log likelihood:

$$\ln\big(P(D|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\big) = \sum_{n=1}^{N} \ln(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k))$$

- Maximize log-likelihood w.r.t. $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

KKT conditions

- Differentiating w.r.t. μ_k:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

- Multiplying by Σ_k^{-1}:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

- Where:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

KKT conditions

- Similarly, differentiating w.r.t. Σ_k:

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

- Lagrangian w.r.t. π_k:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

KKT conditions

- Minimizing:

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

- Multiplying with π_k and adding over k: $\lambda = -N$.
- Hence: $$\pi_k = \frac{N_k}{N}$$
- Where: $$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

(EM) Algorithm

- Initialize μ_k, Σ_k and π_k.
- E-step:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- M-step:

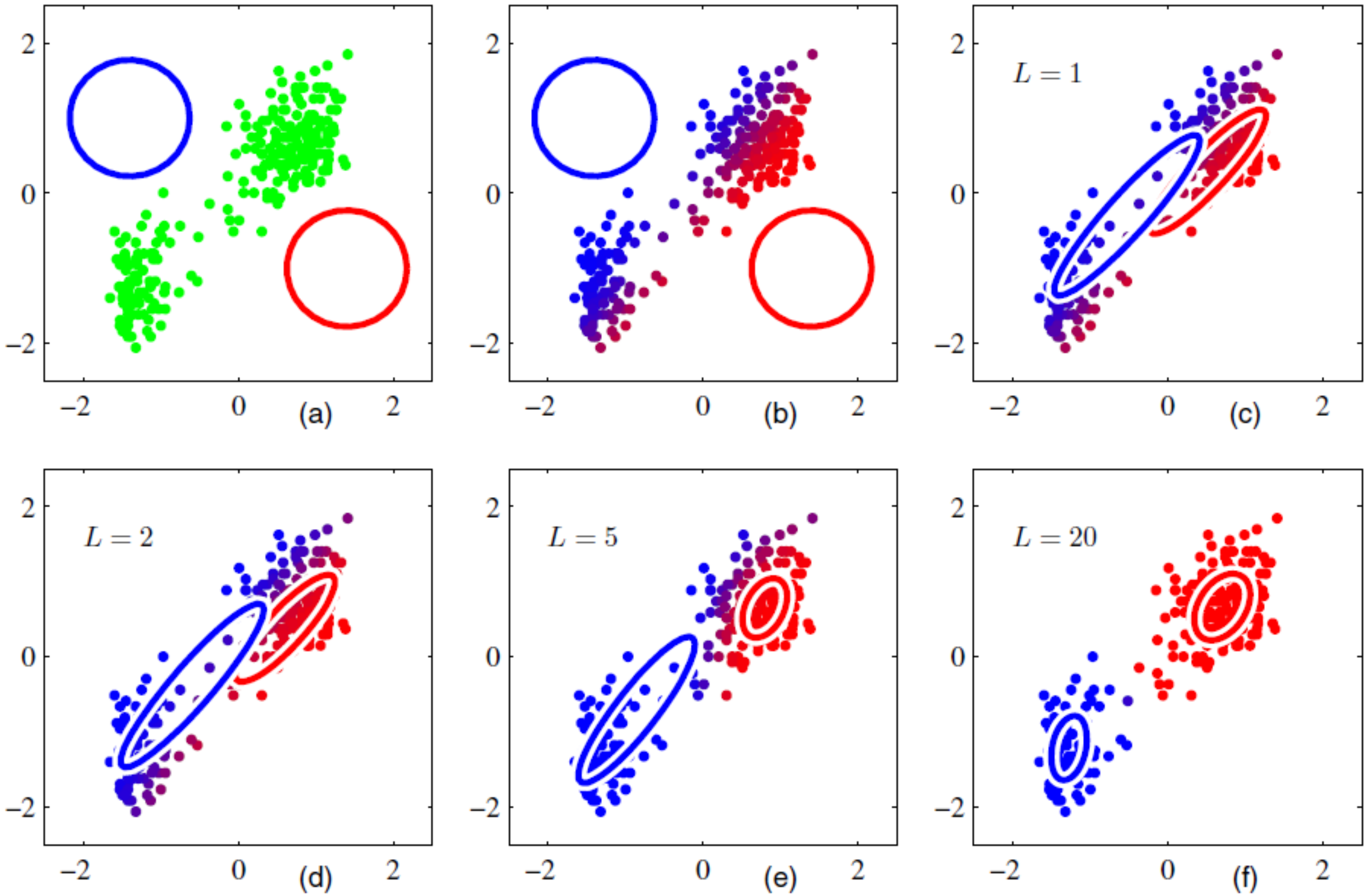
$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

- Repeat above two steps till $\ln\big(P(D|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\big)$ converges.

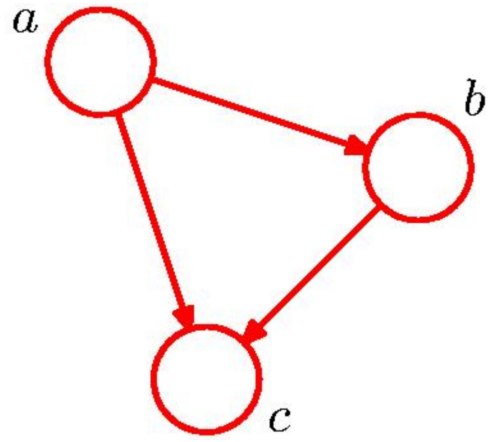
Example



BAYESIAN NETWORKS

Bayesian Networks

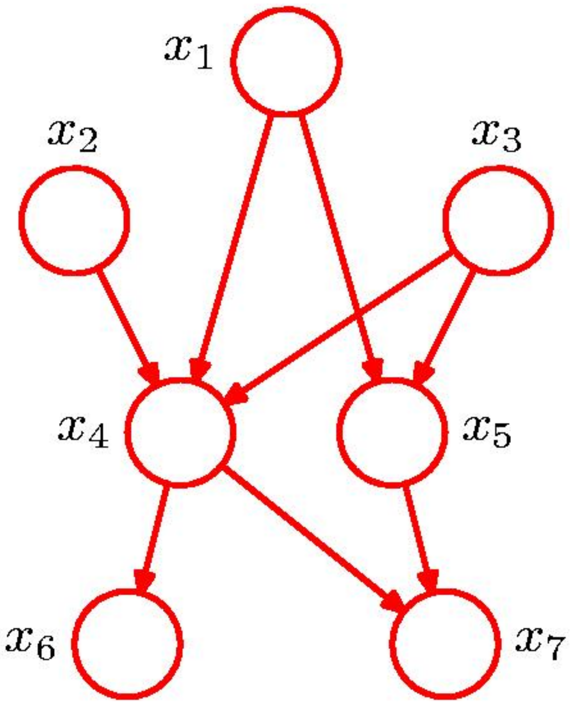
- Directed Acyclic Graph (DAG)



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

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

Bayesian Networks

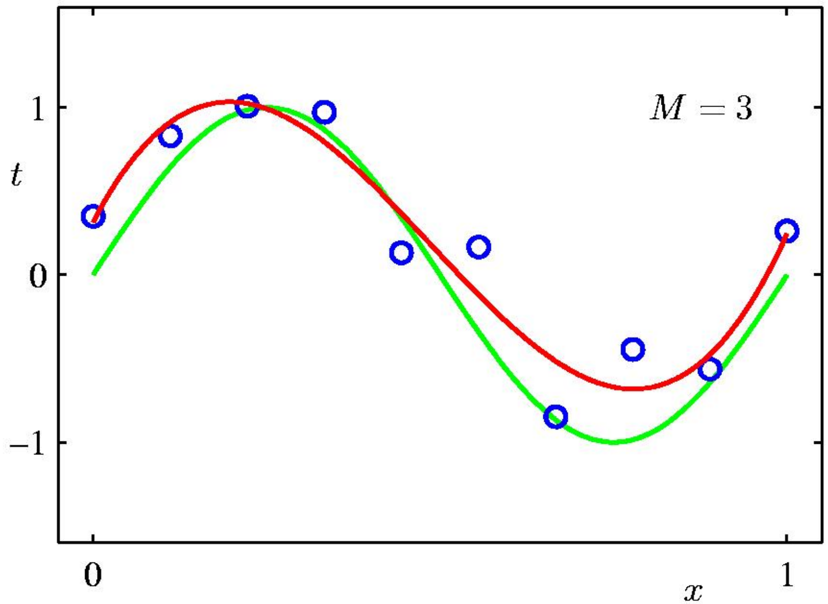


$$
\begin{aligned}
p(x_1, \ldots, x_7) \quad = \quad & p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\
& p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)
\end{aligned}
$$

General Factorization

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

Bayesian Curve Fitting (1)



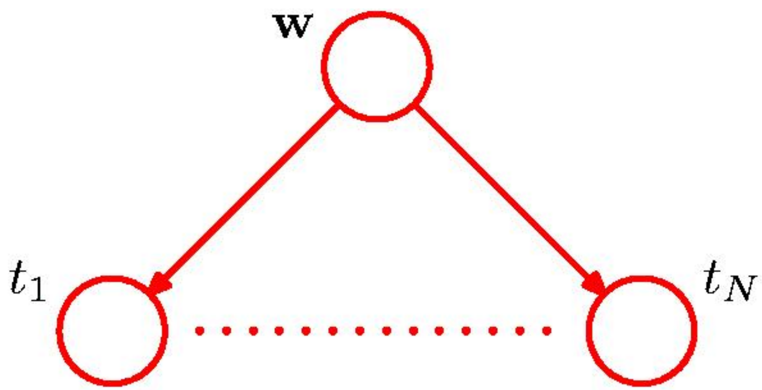
Polynomial

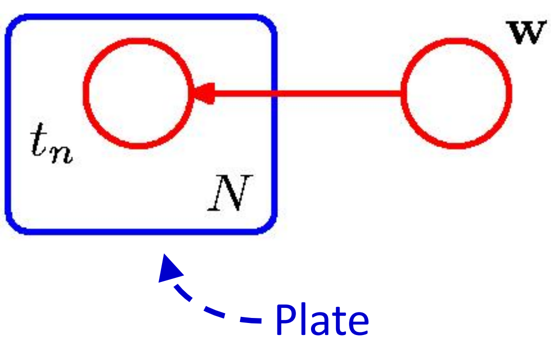
$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n))$$

Bayesian Curve Fitting (2)

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n))$$

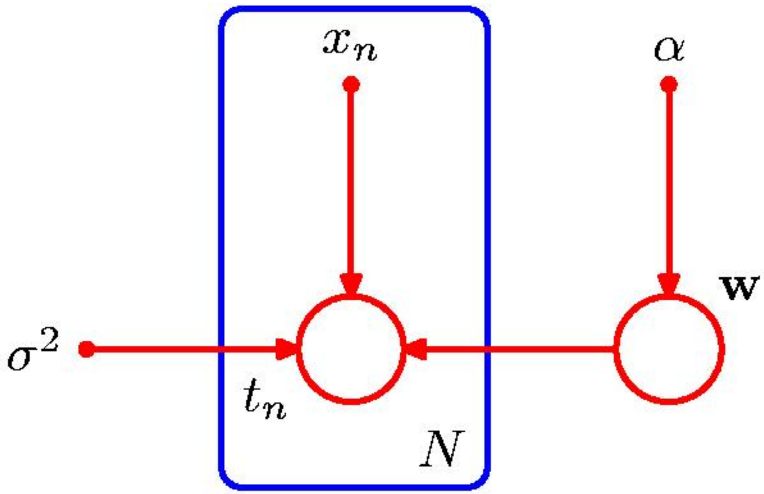




Bayesian Curve Fitting (3)

- Input variables and explicit hyperparameters

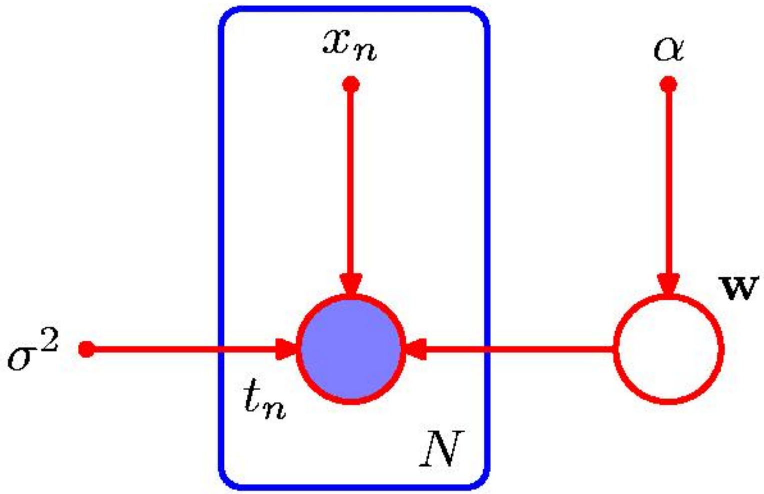
$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2).$$



Bayesian Curve Fitting—Learning

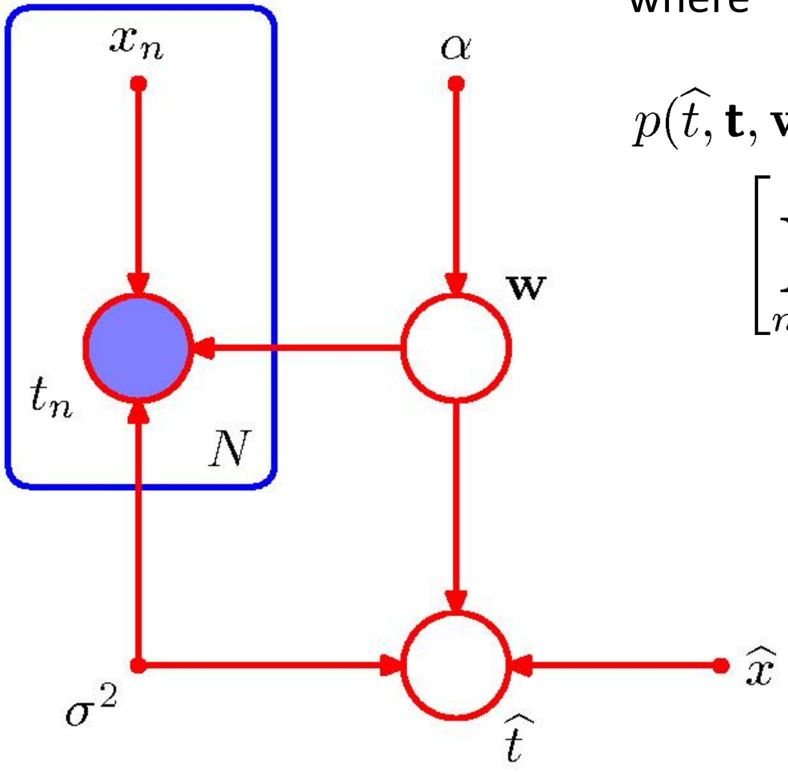
- Condition on data

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$



Bayesian Curve Fitting—Prediction

Predictive distribution: $p(\widehat{t}|\widehat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2)\, \mathrm{d}\mathbf{w}$

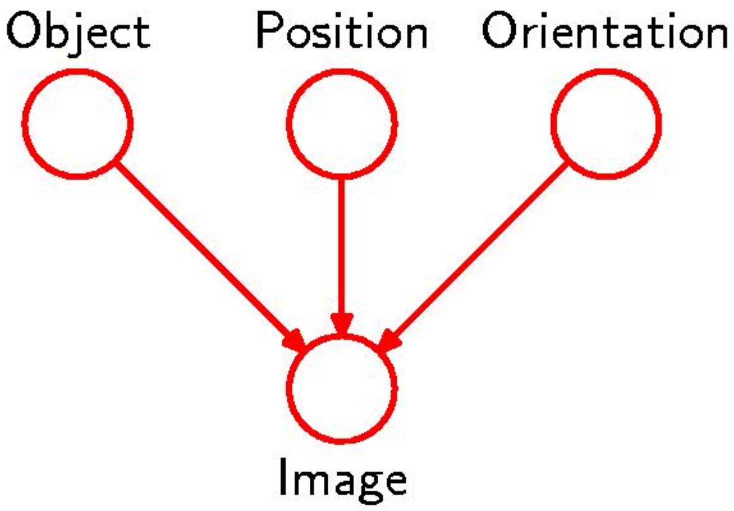


where

$$p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha) p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2)$$

Generative Models

- Causal process for generating images



Discrete Variables (1)

- 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

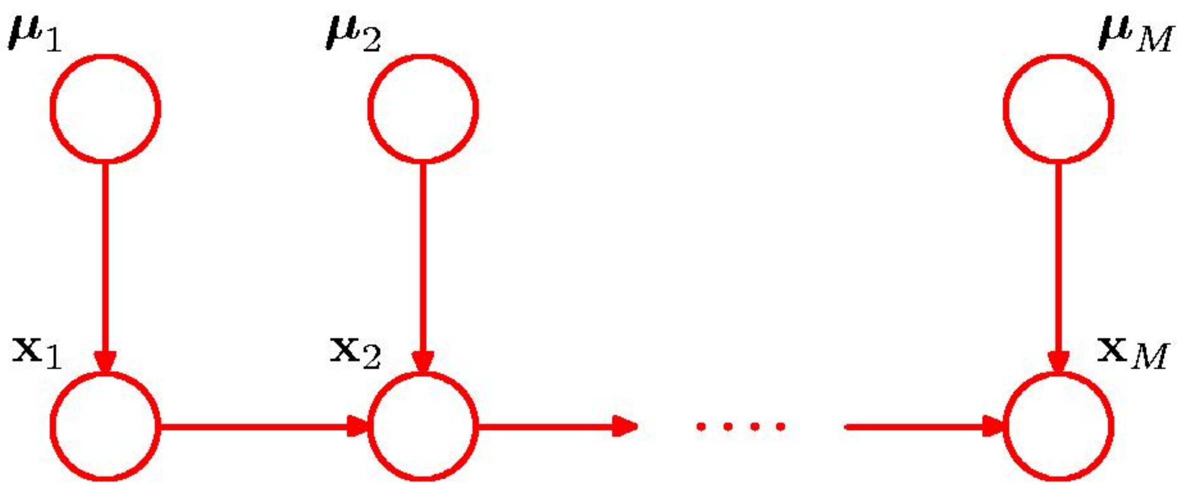
$$\hat{p}(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_{1k}^{x_{1k}} \prod_{l=1}^{K} \mu_{2l}^{x_{2l}}$$

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

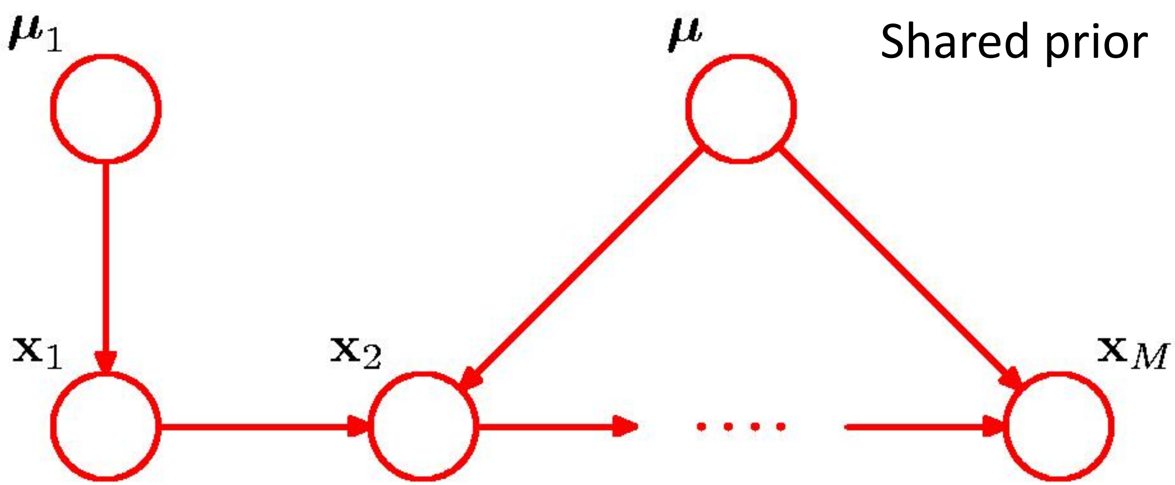
Discrete Variables: Bayesian Parameters (1)



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

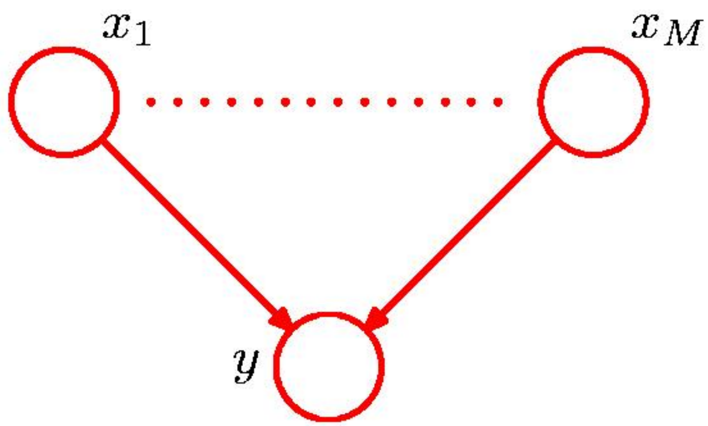
$$p(\boldsymbol{\mu}_m) = \mathrm{Dir}(\boldsymbol{\mu}_m | \boldsymbol{\alpha}_m)$$

Discrete Variables: Bayesian Parameters (2)



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

Parameterized Conditional Distributions



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

The parameterized form

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

requires only M + 1 parameters

Linear-Gaussian Models

- Directed Graph

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

Each node is Gaussian, the mean is a linear function of the parents.

– Vector-valued Gaussian Nodes

$$p(\mathbf{x}_i|\mathrm{pa}_i) = \mathcal{N}\left(\mathbf{x}_i \,\middle|\, \sum_{j\in\mathrm{pa}_i} \mathbf{W}_{ij}\mathbf{x}_j + \mathbf{b}_i, \mathbf{\Sigma}_i\right)$$

Conditional Independence

- a is independent of b given c

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

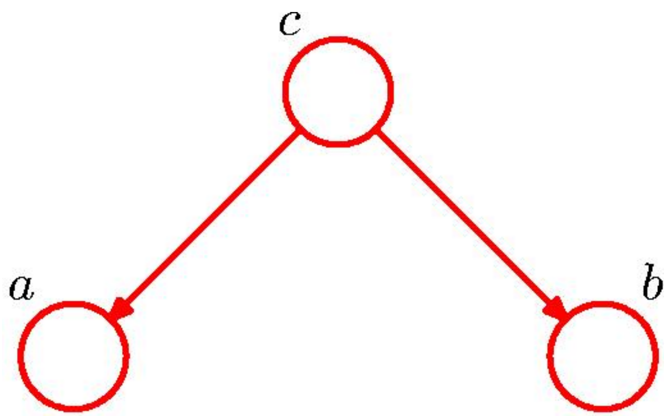
- Equivalently

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

- Notation

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

Conditional Independence: Example 1

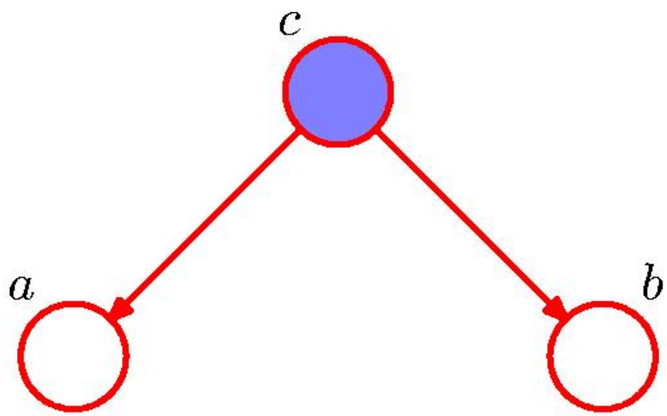


$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

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

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

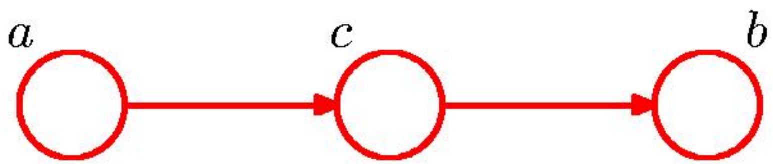
Conditional Independence: Example 1



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

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

Conditional Independence: Example 2

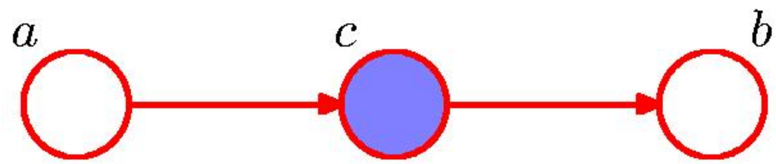


$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

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

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

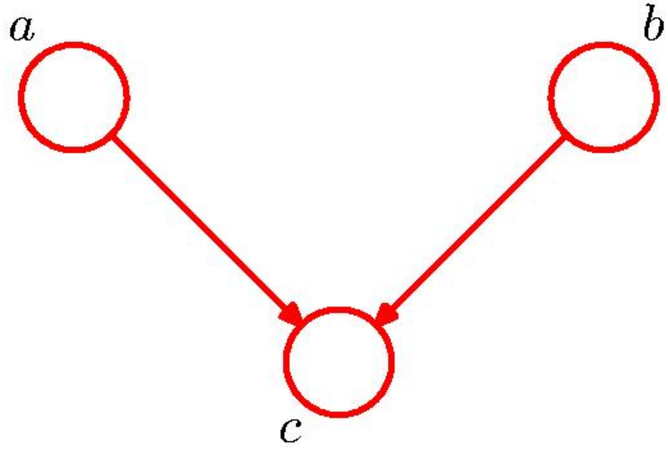
Conditional Independence: Example 2



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

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

Conditional Independence: Example 3



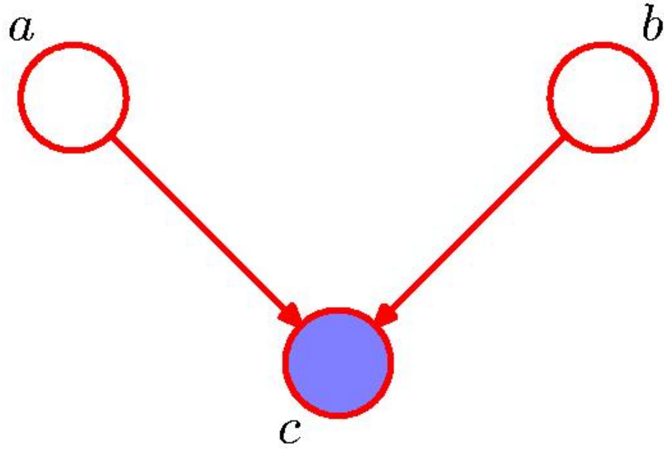
$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

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

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

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

Conditional Independence: Example 3



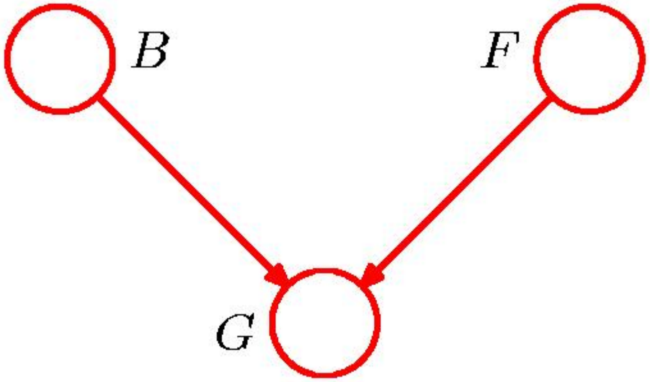
$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b)p(c|a, b)}{p(c)}$$

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

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

“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8\\
p(G=1|B=1,F=0) &= 0.2\\
p(G=1|B=0,F=1) &= 0.2\\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$



$$
\begin{aligned}
p(B=1) &= 0.9\\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

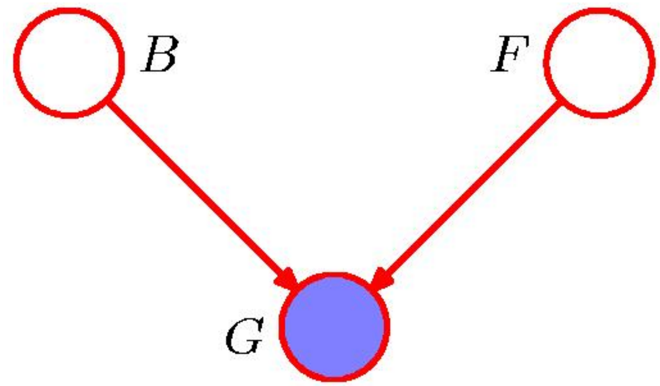
$$p(F=0) = 0.1$$

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

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

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

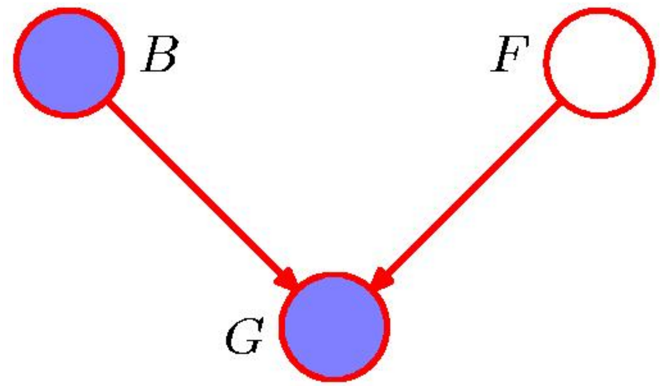
“Am I out of fuel?”



$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \simeq 0.257$$

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

“Am I out of fuel?”



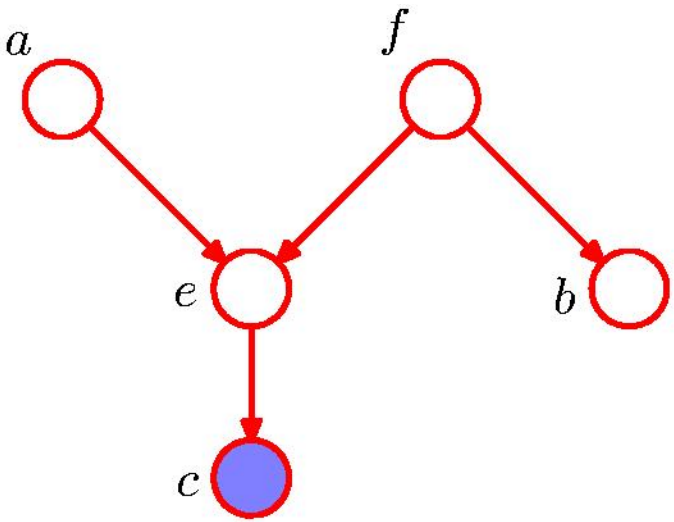
$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \simeq 0.111$$

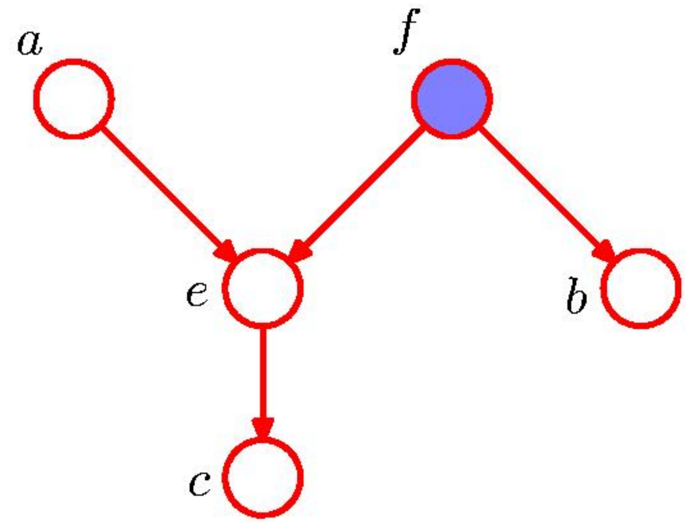
Probability of an empty tank reduced by observing B = 0.
This referred to as “explaining away”.

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
  a) 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
  b) 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 $A \perp\!\!\!\perp B \mid C$

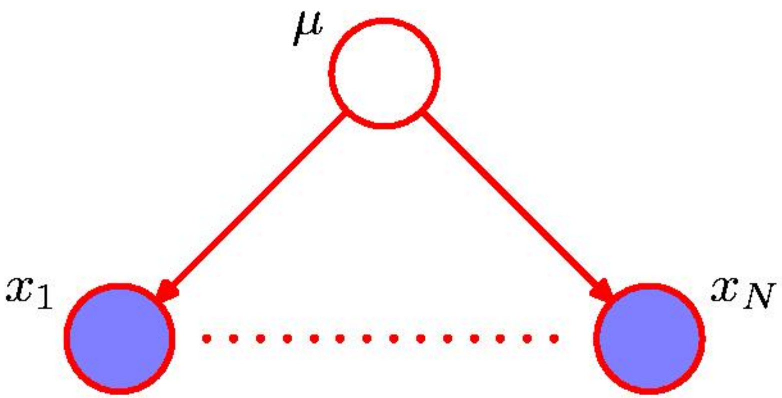
D-separation: Example





$a \not\perp\!\!\!\perp b \mid c$ $\qquad$ $a \perp\!\!\!\perp b \mid f$

D-separation: I.I.D. Data



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

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