Probabilistic Graphical Models

Lecture Notes
Fall 2009

November 5, 2009

Byoung-Tak Zhang
School of Computer Science and Engineering &
Cognitive Science, Brain Science, and Bioinformatics
Seoul National University
http://bi.snu.ac.kr/~btzhang/

Chapter 6. Bayesian Networks

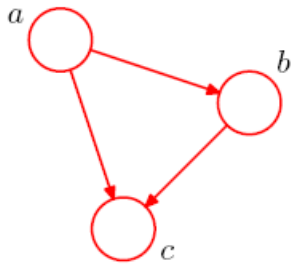
6.1 Bayesian Networks

- Graphical representations of joint distributions using the concept of conditional independence.
- Express dependence relations between variables using DAG (Directed Acyclic Graph)
- Can use prior knowledge on the data (parameters)
- BN = (*S*, *P*) consists of a network structure *S* and a set of local probability distributions *P*
- A directed acyclic graph (DAG) for a compact representation of joint probability distribution between random variables
  - Nodes: random variables
  - Vertices: probabilistic relationships

- The rules of probability
  - Sum rule: $p(X) = \sum_{Y} p(X,Y)$
  - Product rule: $p(X,Y) = p(X)p(Y \mid X)$

- Graphical models for representing the joint distribution over three variables

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



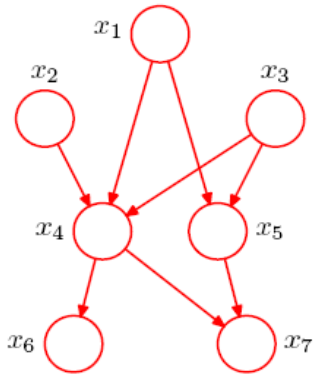
Examples are from [Bishop, 2006]

- In general, for *K* variables

  $p(x_1, \ldots, x_K)$

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

♦ From a directed graph to the corresponding distribution over the variables.



The joint distribution of all 7 variables is given by

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5).$$

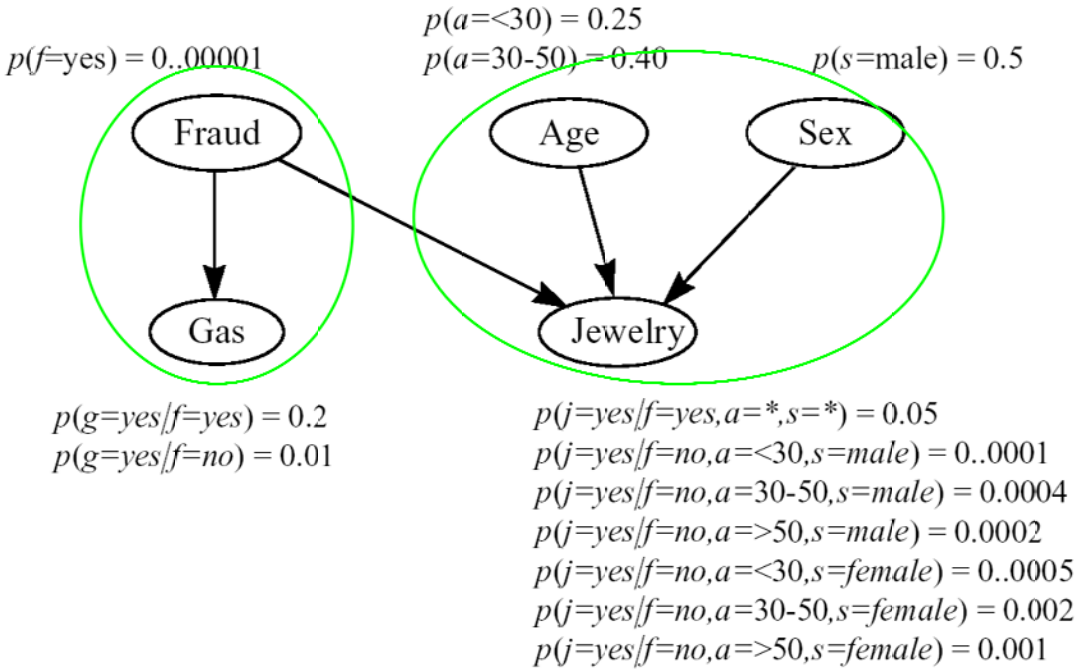
- Generally, we can describe the distributions in not fully connected graphs. The joint distribution defined by a graph is given by the product of a conditional distribution of each node conditioned on their parent nodes.

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid Pa(x_k))$$

$Pa(x_k)$: the set of parents of x_k

Example: Bayesian network for detecting credit card fraud

Case	Fraud	Gas	Jewelry	Age	Sex
1	no	no	no	30-50	female
2	no	no	no	30-50	male
3	yes	yes	yes	>50	male
4	no	no	no	30-50	male
5	no	yes	no	<30	female
6	no	no	no	<30	female
7	no	no	no	>50	male
8	no	no	yes	30-50	female
9	no	yes	no	<30	male
10	no	no	no	<30	female



- Structure can be found by relying on the prior knowledge of causal relationships
- Learning: Finding the structure of the graph and the values of the parameters (conditional probability tables)
- Inference: Predicting the values of variables, given the values of other variables

6.2 Conditional Independence

- Conditional independence simplifies both the structure of a model and the computations

  Ex: The joint distribution of *a* and b are independent *conditioned* on *c* (or given *c*).

$$p(a \mid b, c) = p(a \mid c)$$
$$p(a, b \mid c) = p(a \mid b, c)\, p(b \mid c)$$
$$= p(a \mid c)\, p(b \mid c)$$

  $a \perp\!\!\!\perp b \mid c$: "*a* is conditionally independent of *b* given *c*"

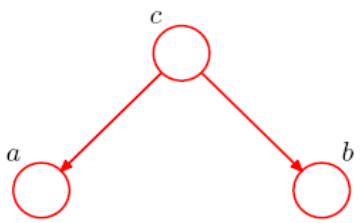
- Conditional independence properties play an important role in using probabilistic models for pattern recognition by simplifying both the structure of a model and the computations needed to perform inference and learning under that model.

- An important and elegant feature of graphical models is that conditional independence properties of the joint distribution can be read directly from the graph without having to perform any analytical manipulations
  - The general framework for this is called *d-separation* ('d' stands for 'directed')

- **Three example graphs for conditional independence**

  - Case 1: tail-to-tail
  - Case 2: head-to-tail
  - Case 3: head-to-head

- **Case 1: tail-to-tail**



Joint distribution: $p(a,b,c) = p(a \mid c)\, p(b \mid c)\, p(c)$

Consider a path from node *a* to node *b* via *c*. The node *c* is said to be tail-to-tail w.r.t. this path.

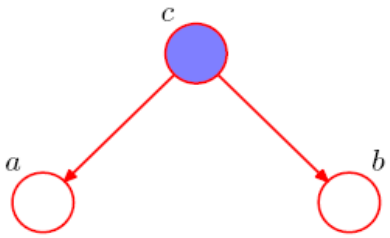
- If none of the variables are observed

Marginalizing both sides of the above equation over *c* we obtain

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

This does not factorize into the product $p(a)\,p(b)$ → $a \not\perp\!\!\!\perp b \mid \emptyset$

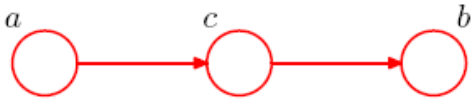
- If the variable *c* is observed



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

  - The conditioned node *c* 'blocks' the path from *a* to *b* and causes *a* and *b* to become (conditionally) independent. → $a \perp\!\!\!\perp b \mid c.$

- **Case 2: head-to-tail**



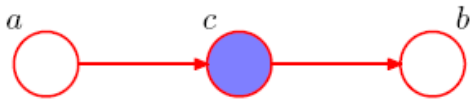
Joint distribution: $p(a,b,c) = p(a)\,p(c\,|\,a)\,p(b\,|\,c)$

- If none of the variables are observed

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

  - p(a)$p(b|a)$ in general does not factorize into $p(a)p(b)$, and so $a \not\perp\!\!\!\perp b \,|\, \emptyset$

- If the variable c is observed



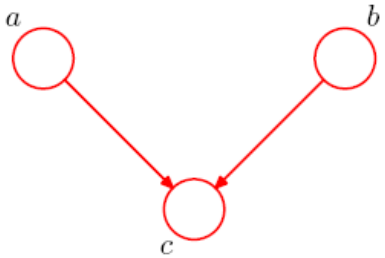
$$\begin{aligned} p(a,b\,|\,c) &= \frac{p(a,b,c)}{p(c)} = \frac{p(a)\,p(c\,|\,a)\,p(b\,|\,c)}{p(c)} \\ &= p(a\,|\,c)\,p(b\,|\,c) \end{aligned}$$

$$\because\ p(c\,|\,a) = \frac{p(a\,|\,c)\,p(c)}{p(a)} \quad \text{Bayes' rule}$$

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

- Observing c blocks the path from a to b and so we obtain the conditional independence property. → $a \not\perp\!\!\!\perp b \,|\, c.$

Case 3: head-to-head



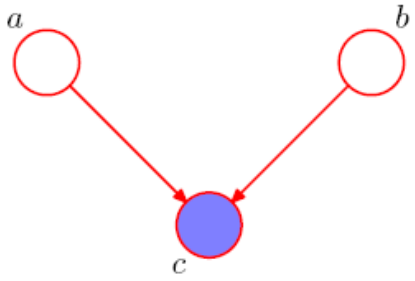
Joint distribution: $p(a,b,c) = p(a)\,p(b)\,p(c\,|\,a,b)$

- If none of the variables are observed

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

and so *a* and *b* are independent with no variables observed, in contrast to the two previous examples. That is, $a \perp\!\!\!\perp b \mid \emptyset$.

- If the variable *c* is observed (if we condition on *c*)



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

  - This in general does not factorize into the product *p*(*a*)*p*(*b*), and so $a \not\perp\!\!\!\perp b \mid c$.

- Thus, the third case (head-to-head) has the opposite behavior from the first two. When node *c* is unobserved, it 'blocks' the path and the variables *a* and *b* are independent.

- Conditioning on c 'unblocks' the path and render *a* and *b* dependent.

Summary

A tail-to-tail node or a head-to-tail node leaves a path unblocked unless it is observed in which case it blocks the path. By contrast, a head-to-head node blocks a path if it is unobserved, but once the node, and/or at least one of its descendants, is observed the path becomes unblocked.

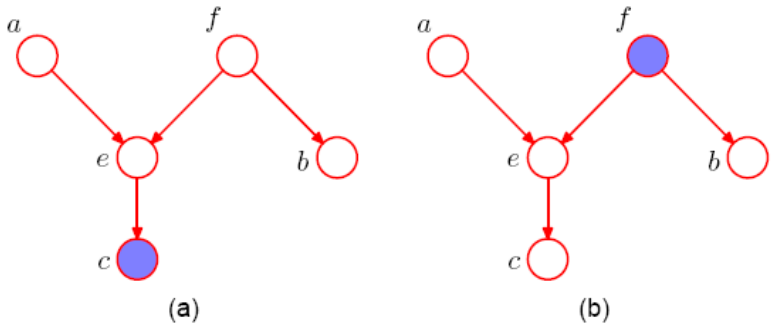
- tail-to-tail & head-to-tail: blocks if observed
- head-to-head: blocks if unobserved

d-Separation

- A, B, C: arbitrary nonintersecting sets of nodes in a general directed graph
- We consider all possible paths from any node in *A* to any node in *B*.
- Any such **path** is said to be ***blocked*** if it includes 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, is in the set *C*.

- If *all paths are* ***blocked***, then ***A* is said to be *d*-separated** from *B* by *C*, and the joint distribution over all of the variables in the graph will satisfy $A \perp B \mid C$.

♦ **Example**



(a) *a* is dependent on *b* given *c*
- Head-to-head node *e* is blocked
- Tail-to-tail node *f* is blocked

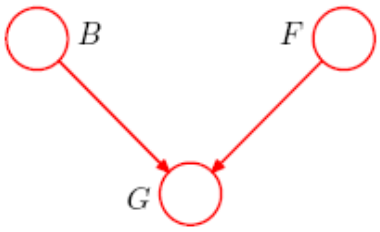
(b) a is independent of *b* given
- Head-to-head node *e* is unblocked, because a descendant *c* is in the conditioning set
- Tail-to-tail node *f* is unblocked

6.3 Inference in Bayesian Networks

- **Example: Fuel gauge**

  A concrete example for case 3 (head-to-head)

  B – Battery F – fuel G - electric fuel gauge

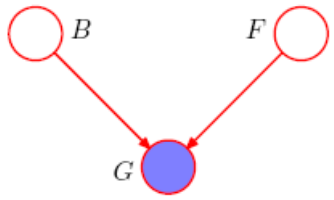


$$p(B=1) = 0.9$$
$$p(F=1) = 0.9.$$

$$p(G=1|B=1,F=1) = 0.8$$
$$p(G=1|B=1,F=0) = 0.2$$
$$p(G=1|B=0,F=1) = 0.2$$
$$p(G=1|B=0,F=0) = 0.1$$

From $p(F=1) = 0.9$, we know that $p(F=0) = 0.1$ since $p(F=1) + p(F=0)$ should be 1.
From $p(G=1 \mid B=0, F=0) = 0.1$, we have $p(G=0 \mid B=0, F=0) = 0.0$.

- **Suppose that we observe the fuel gauge and discover that *G* = 0.**



- **We want to compute *p*(*F*=0|*G*=0),** i.e. want to know if checking the fuel gauge makes sense.

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

$$\begin{aligned}
&p(G=0\,|\,F=0)\\
&=\sum_{B\in\{0,1\}} p(G=0,B\,|\,F=0)\\
&=\sum_{B\in\{0,1\}} p(G=0\,|\,B,F=0)\,p(B)\\
&=p(G=0\,|\,B=0,F=0)\,p(B=0)+p(G=0\,|\,B=1,F=0)\,p(B=1)\\
&=(0.9\times 0.1)+(0.8\times 0.9)=0.81
\end{aligned}$$

$$\begin{aligned}
&p(G=0)\\
&=\sum_{B\in\{0,1\}}\sum_{F\in\{0,1\}} p(G=0,B,F)\\
&=\sum_{B\in\{0,1\}}\sum_{F\in\{0,1\}} p(G=0\,|\,B,F)\,p(B)\,p(F)\\
&=p(G=0\,|\,B=0,F=0)\,p(B=0)\,p(F=0)\\
&\quad+p(G=0\,|\,B=0,F=1)\,p(B=0)\,p(F=1)\\
&\quad+p(G=0\,|\,B=1,F=0)\,p(B=1)\,p(F=0)\\
&\quad+p(G=0\,|\,B=1,F=1)\,p(B=1)\,p(F=1)\\
&=(0.9\times 0.1\times 0.1)+(0.8\times 0.1\times 0.9)+(0.8\times 0.9\times 0.1)+(0.2\times 0.9\times 0.9)\\
&=0.315
\end{aligned}$$

Hence,

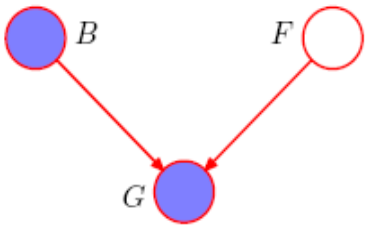
$$\begin{aligned}
&p(F=0\,|\,G=0)\\
&=\frac{p(G=0\,|\,F=0)\,p(F=0)}{p(G=0)}\\
&=\frac{0.81\times 0.1}{0.315}\\
&\simeq 0.257
\end{aligned}$$

Note

$$p(F=0\,|\,G=0)=0.257\ \geq\ p(F=0)=0.1$$

- Thus, observing that the gauge reads empty makes it more likely that the tank is indeed empty.

- **Next suppose that we also check the state of the battery and find that it is flat, i.e., $B = 0$.**
  - Checking if the battery also makes sense?



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

Note
$$p(F=0 \mid G=0, B=0) \;\leq\; p(F=0 \mid G=0) = 0.257$$

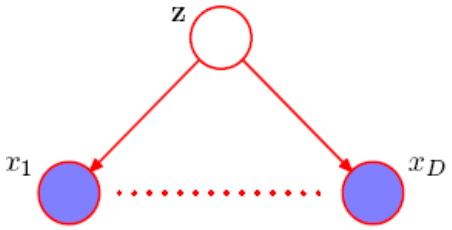
- Thus, additional observation of the state of the battery decreased the probability of fuel being empty.

- In the equation above, the prior probability $p(B=0)$ has cancelled between numerator and denominator. Thus the probability that the tank is empty has decreased (from 0.257 to 0.111) as a result of the observation of the state of the battery. We say that finding out that the battery is flat ***explains away*** the observation that the fuel gauge reads empty.

6.4 Variants of Bayesian Networks

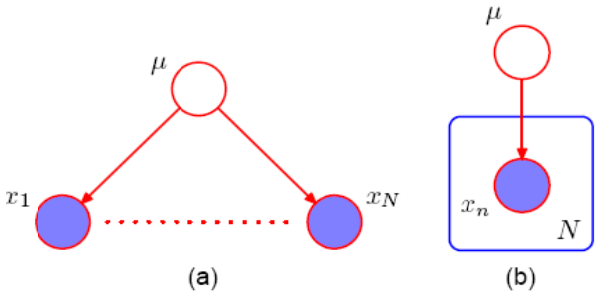
6.4.1 Naïve Bayes Models

- A special case of latent variable models (Chap. 7)
- Key assumption: conditioned on the class z, the distribution of the input variables $x_1, \ldots, x_D$ are independent.
- Input $\{x_1,\ldots,x_N\}$ with their class labels, then we can fit the naïve Bayes model to the training data using maximum likelihood assuming that the data are drawn independently from the model.



- **Example**
  - Problem: finding posterior distribution for the mean of a univariate Gaussian distribution

- Every path is blocked and so the observations $D = \{x_1, \ldots, x_N\}$ are independent given



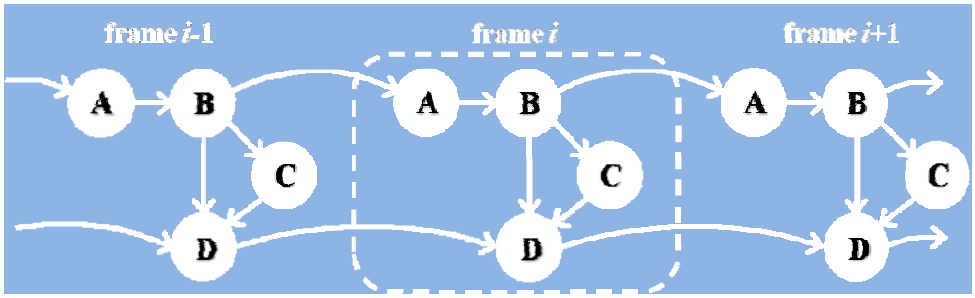
$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu).$$

independent

$$p(\mathcal{D}) = \int_0^\infty p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n).$$

The observations are in general no longer independent!

6.4.2 Dynamic Bayesian Networks (Chap. 8)



- Dynamic Bayesian networks are Bayesian networks consisting of a structure that repeats an indefinite (or dynamic) number of times
- **A special case of temporal graphical models (Chap. 8)**

$$\lambda_{DBN} = \left(G_1, G_{tr}, \{\Pi_i\}_{i\in\{1,\ldots,I\}}, \{CPT_j\}_{j\in\{1,\ldots,J\}}\right)$$

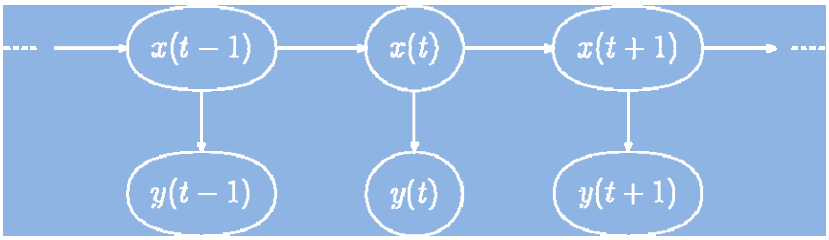
  - a directed, acyclic graph of starting nodes (initial probability distribution)
  - a directed, acyclic graph of transition nodes (transition probabilities between time slices)
  - starting vectors of observable as well as hidden random variables
  - transition matrices regarding observable as well as hidden random variables

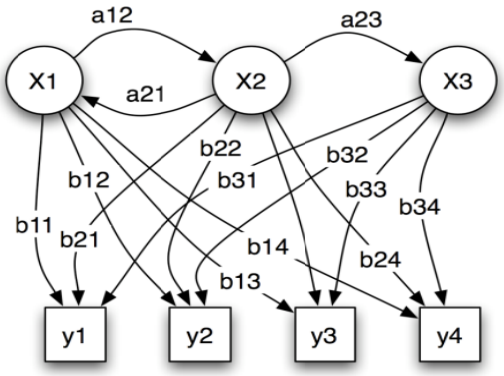
- Time-invariant:
  - the term 'dynamic' means that we are modeling a dynamic model, not that the networks change over time

- A general form of Hidden Markov Models (HMMs)
  - Represent the hidden and observed states in terms of state variables of complex interdependencies

6.4.3 Hidden Markov Models

- Temporal Bayesian networks, i.e. a specific kind of temporal graphical model **(Chap. 8)**
- A dynamic Bayesian network for discrete random variables





- ♦ Structure
  - One discrete hidden node (*X*: hidden variables)
  - One discrete or continuous observed node per time slice (*Y*: observations)
- ♦ Parameters
  - Initial state distribution $P(X_1)$
  - Transition model $P(X_t \mid X_{t-1})$
  - Observation model $P(Y_t \mid X_t)$
- ♦ Features
  - A discrete state variable with arbitrary dynamics and arbitrary measurements
  - Structures and parameters remain same over time
- Applications
  - Widely applied to dynamic system modeling
  - Sequential data modeling (speech, bio-sequence, gesture, language)
  - Time series modeling (gesture, stock price)

6.5 Discrete vs. Continuous Variables

6.5.1 Discrete Variables (Multinomials)

- The framework of graphical models is very useful in expressing the way in which “building blocks” are linked together.
- Choose the relationship between each parent-child pair in a directed graph to be *conjugate*.
- The followings can extend hierarchically to construct arbitrarily complex DAGs.
  - ♦ Discrete variables
  - ♦ Gaussian variables
- Examples of conjugate priors

- Dirichlet for multinomial data
- Normal-Wishart for normal data

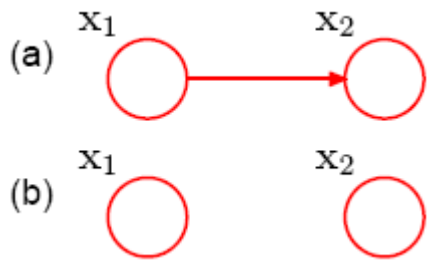
- Given a node with k discrete states
  - # of parameters: (k-1) because of $\sum_k \mu_k = 1$

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

- Given two nodes x_1, x_2 and each has k discrete states,
  - If x_1 and x_2 are dependent,
    - # of parameters on $p(x_1,x_2)$: $k^2 - 1$
    - Given M variables: $k^M - 1$ (fully connected

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

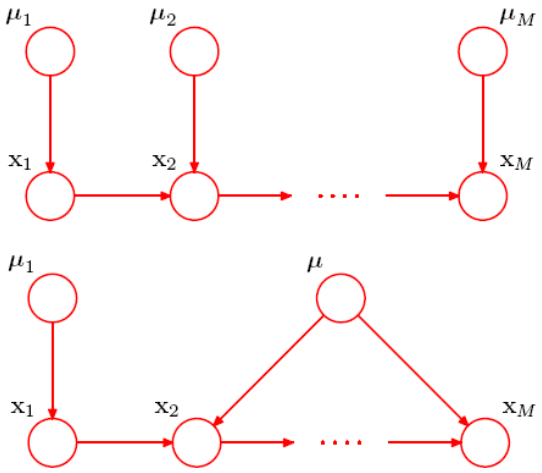
  - If x_1 and x_2 are independent,
    - # of parameters on $p(x_1,x_2)$: 2(k-1)
    - Given M variables: M(k-1)



- Given a chain of M discrete nodes (not fully connected), each having K states, requires the specification of K-1 + (M-1)K(K-1)

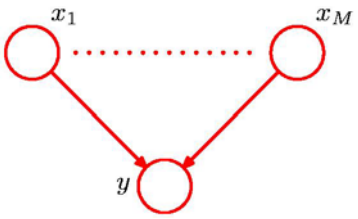
- Sharing parameters
  - For the chain, the number of parameters is K^2-1 when $p(x_i \mid x_{i-1})$ are governed by the same set of K(K-1) parameters

- A graph over discrete variables into a Bayesian model by introducing *Dirichlet priors* to parameters
  - Each node acquires an additional parent representing the Dirichlet distribution over the parameters



- Parameterized models for the conditional distributions
  - Require 2^M parameters the probability $p(y=1)$ over $p(y|x_1,\ldots,x_M)$
  - The number of parameters grows linearly with M

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



6.5.2 Continuous Variables (Linear Gaussians)

- A multivariate Gaussian can be expressed as a directed graph corresponding to a linear-Gaussian model over the component variables.
- Graph G with D variables $\mathbf{X} = \{x_1,\ldots x_D\}$, 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 of node x_i

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

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

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

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

$(.)^2$ term: A quadratic form of the components of **x** → The joint distribution is a multivariate Gaussian

- Can evaluate the mean and covariance of the joint distribution recursively.

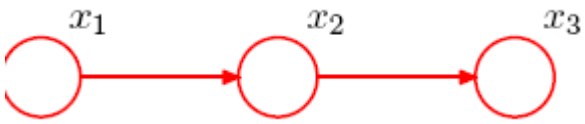
$$x_i = \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i + \sqrt{v_i}\epsilon_i$$

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

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

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

- Consider two cases
  - ♦ No links
    - ■ The mean of $p(x)$ is given b $(b_1, \ldots, b_D)^{\mathrm{T}}$
    - ■ The covariance matrix is diagonal of the form $\mathrm{diag}(v_1, \ldots, v_D)$
    - ■ The joint distribution represents a set of D independent univariate Gaussian distributions
  - ♦ Graph with one missing link
    - ■ Multivariate Gaussian distribution



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

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

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