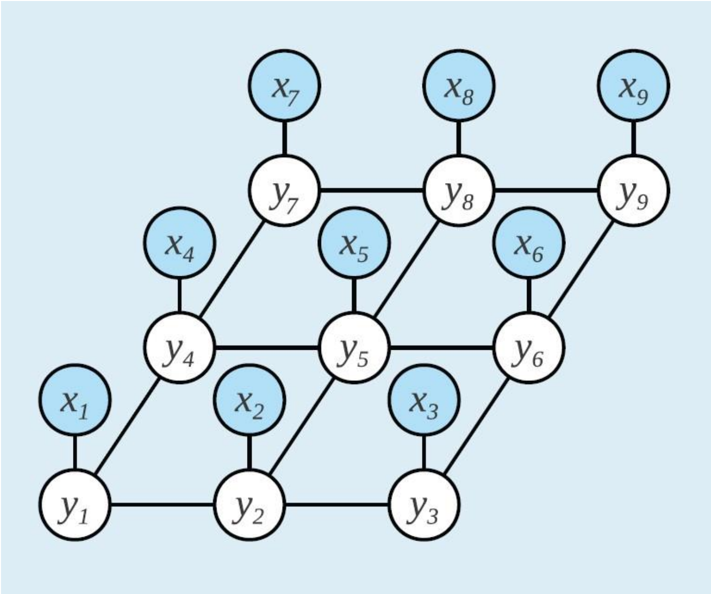
Probabilistic Graphical Models in Bioinformatics

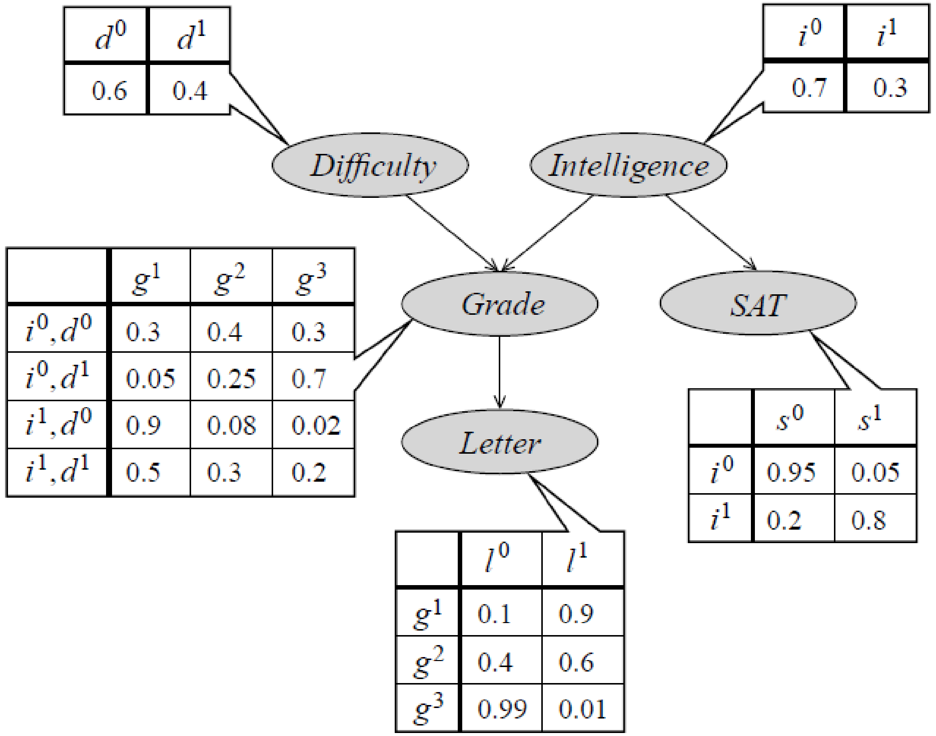
Lecture 4: Conditional probability distributions; Gaussian Bayesian networks



Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Esfand 13, 1397
(March 4, 2019)

Factorization and parametrization



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

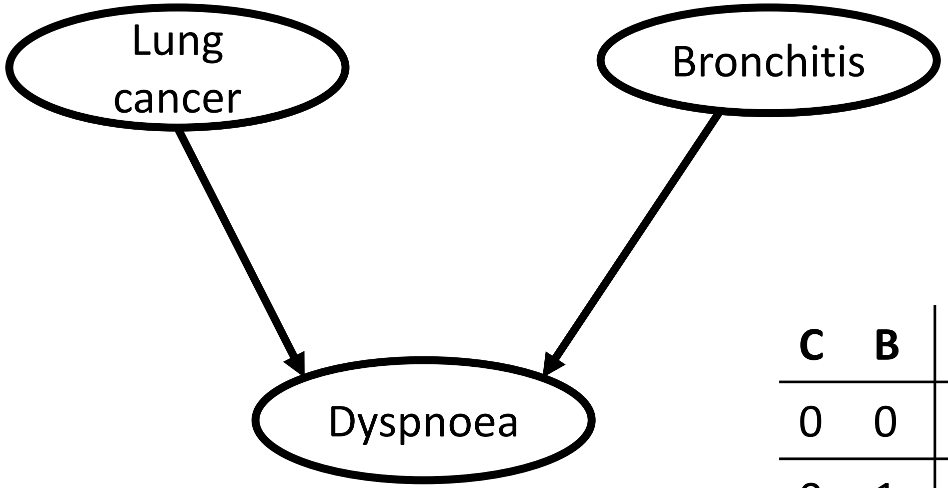
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(I, D, G, S, L) = P(I)\,P(D)\,P(G|I, D)\,P(S|I)\,P(L|G)$$

Conditional probability distributions (CPD)

Tabular CPD

- Encode $P(X|\mathrm{Pa}(\mathrm{X_i}))$ as a table.
- Proper CPD requires all non-negative values and $\sum_x P(X = x|\mathrm{Pa}(\mathrm{X_i})) = 1$



C	B	D=1	D=0
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

- Disadvantages?
    - Limited to discrete values
    - Number of parameters is exponential in the number of parents

General CPD

- CPD $P(X \mid y_1, \ldots, y_k)$ specifies distribution over X for each assignment $y_1, \ldots, y_k$ but does not have to do so by listing each such value explicitly

- Different possibilities: deterministic CPDs, tree-structured CPDs, rule-based CPDs, linear Gaussian, …

Deterministic CPDs

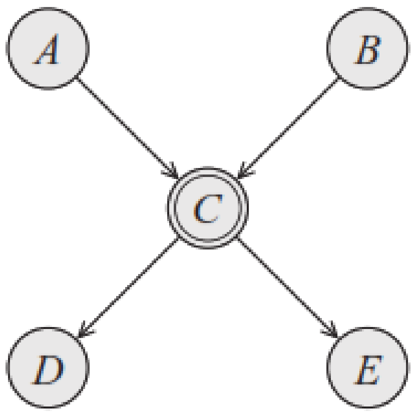
- Simplest type of non-tabular CPD
- X is a deterministic function of its parents Pa_X

$$P(x \mid \mathrm{pa}_X) = \begin{cases} 1 & x = f(\mathrm{pa}_X) \\ 0 & \text{otherwise.} \end{cases}$$

- Examples:
  - Binary variables: X is "or" of its parents, X = Y or Z.
  - Continuous variables: for example a linear function of its parents $X = Y - 3Z + 1$

Deterministic CPDs & independencies

- Need to modify d-separation in the presence of deterministic CPDs

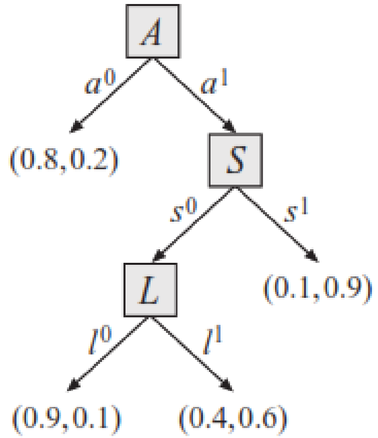


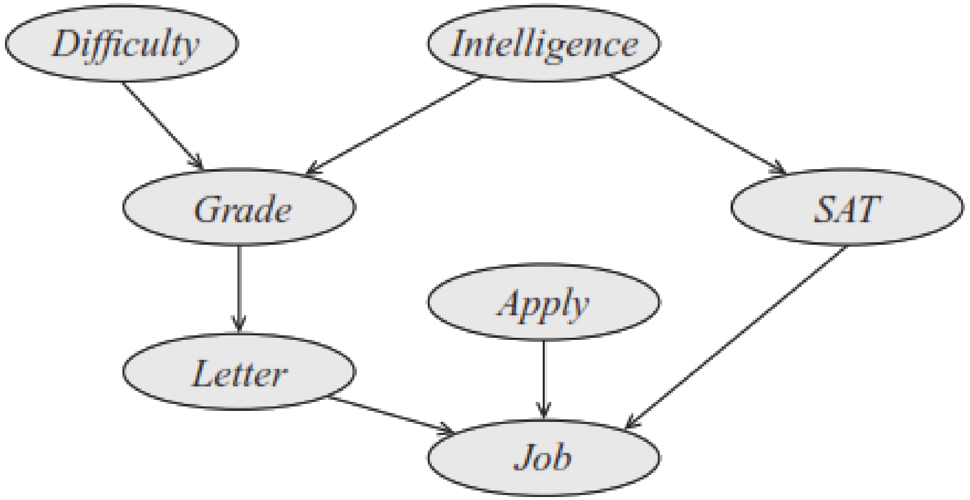
- C is a deterministic function of A and B.
- Given A and B are known, C is known. Hence D and E are independent given A and B ($D \perp E \mid A, B$).
- Not necessarily true if C were a non-deterministic function of its parents.

Context-specific CPDs

- Two common choices
  - Tree CPDS
  - Rule CPDs

- A tree-CPD for $P(J \mid A, S, L)$



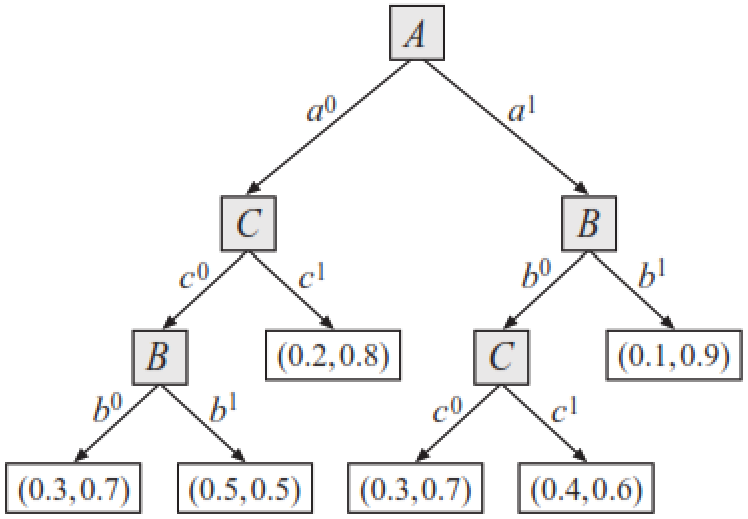


$$J \perp_c L, S \mid A = a^0$$

Context-specific CPDs-Rule CPDs

- A rule-based CPD for $P(X \mid A, B, C)$

$$\rho_1{:}\langle a^1, b^1, x^0; 0.1\rangle \qquad \rho_2{:}\langle a^1, b^1, x^1; 0.9\rangle$$
$$\rho_3{:}\langle a^0, c^1, x^0; 0.2\rangle \qquad \rho_4{:}\langle a^0, c^1, x^1; 0.8\rangle$$
$$\rho_5{:}\langle b^0, c^0, x^0; 0.3\rangle \qquad \rho_6{:}\langle b^0, c^0, x^1; 0.7\rangle$$
$$\rho_7{:}\langle a^1, b^0, c^1, x^0; 0.4\rangle \qquad \rho_8{:}\langle a^1, b^0, c^1, x^1; 0.6\rangle$$
$$\rho_9{:}\langle a^0, b^1, c^0; 0.5\rangle$$



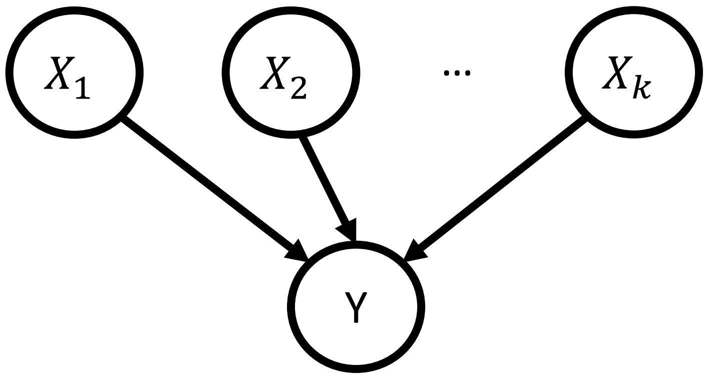
Corresponding tree-CPD

- Results in the following CPD

X	$a^0b^0c^0$	$a^0b^0c^1$	$a^0b^1c^0$	$a^0b^1c^1$	$a^1b^0c^0$	$a^1b^0c^1$	$a^1b^1c^0$	$a^1b^1c^1$
x^0	0.3	0.2	0.5	0.2	0.3	0.4	0.1	0.1
x^1	0.7	0.8	0.5	0.8	0.7	0.6	0.9	0.9

Generalized linear models

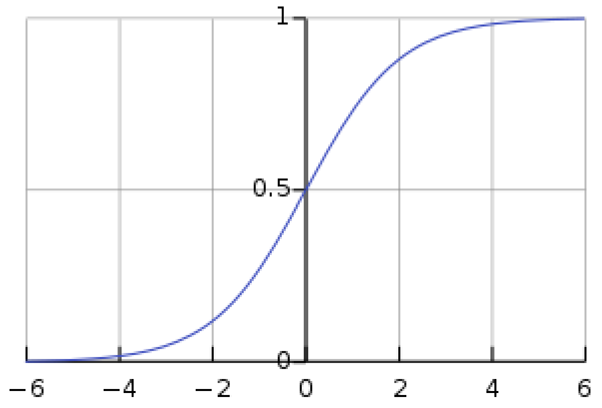
- Independence of causal influence
  - The combined influence of the X_i's on Y is a simple combination of the influence of each of the X_i's on Y in isolation.



- Let *Y* be a binary-valued with *k* parents $X_1, \ldots, X_k$

$$P(y^1 \mid X_1, \ldots, X_k) \quad = \quad \text{sigmoid}(w_0 + \sum_{i=1}^{k} w_i X_i).$$

$$\text{sigmoid}(z) = \frac{e^z}{1 + e^z}.$$

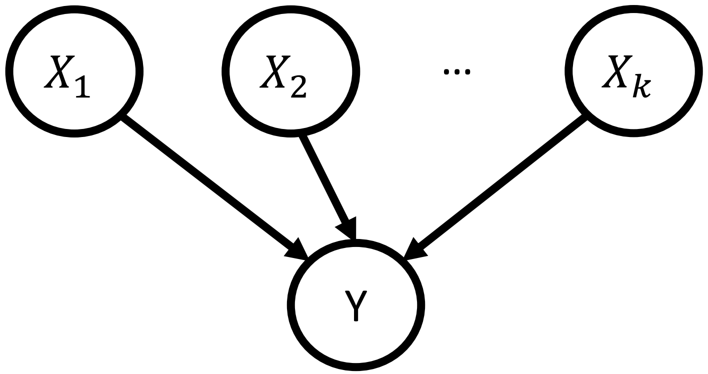


- Easily extendable to multivariate Y

Continuous variables

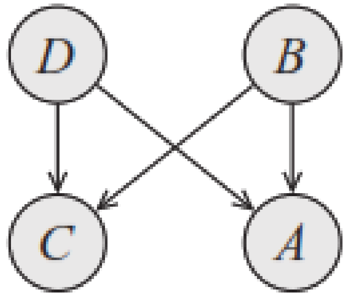
Linear Gaussian model

- Examples of continuous variables: weight, blood pressure, glucose level
- Nothing in formulation of Bayesian networks requires restricting attention to discrete variables
- Linear Gaussian model
  - Continuous variable Y with continuous parents $X_1, \ldots, X_k$



$$p(Y \mid x_1, \ldots, x_k) = N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k; \sigma_Y^2)$$

Question



- $p(B) = N(\beta_{B,0}; \sigma_B^2)$
- $p(D) = N(\beta_{D,0}; \sigma_D^2)$
- $p(A \mid B, D) = N(\beta_{A,0} + \beta_{A,1} b + \beta_{A,2} d; \sigma_A^2)$
- $p(C \mid B, D) = N(\beta_{C,0} + \beta_{C,1} b + \beta_{C,2} d; \sigma_C^2)$
- How many parameters?
- Inference: e.g. $P(-3 < C < 3 \mid A = 2)$

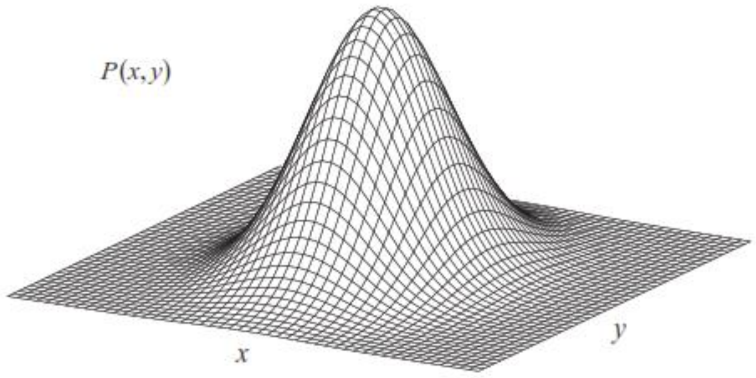
Multivariate Gaussians

- Univariate Gaussian is defined in terms of two parameters: a mean and a variance $N(\mu, \sigma^2)$.

- A multivariate Gaussian distribution over $X_1, \dots, X_n$ is characterized by an n-dimensional mean vector $\boldsymbol{\mu}$ and a symmetric $n \times n$ covariance matrix Σ.

- The joint density function:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\underbrace{|\Sigma|}_{\text{the determinant of } \Sigma}^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$



An example of a bivariate Gaussian

Example

$$\mu = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}$$

- X_3 is negatively correlated with X_1: when X_1 goes up, X_3 goes down (and similarly for X_3 and X_2).

Alternative parametrization of multivariate Gaussians

- Information matrix (or precision matrix) is defined as inverse covariance matrix $J = \Sigma^{-1}$.

$$
\begin{aligned}
-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T J(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}\left[\boldsymbol{x}^T J\boldsymbol{x} - 2\boldsymbol{x}^T J\boldsymbol{\mu} + \boldsymbol{\mu}^T J\boldsymbol{\mu}\right]
\end{aligned}
$$

- The formulation of Gaussian density in *information form*

$$
p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^T J\boldsymbol{x} + (J\boldsymbol{\mu})^T\boldsymbol{x}\right]
$$

Marginalization

- Given joint Gaussian distribution over $\{X, Y\}$ where $X \in \mathcal{R}^n$ and $Y \in \mathcal{R}^m$:

$$p(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\left(\left(\begin{array}{c} \mu_{\boldsymbol{X}} \\ \mu_{\boldsymbol{Y}} \end{array}\right); \left[\begin{array}{cc} \Sigma_{\boldsymbol{X}\boldsymbol{X}} & \Sigma_{\boldsymbol{X}\boldsymbol{Y}} \\ \Sigma_{\boldsymbol{Y}\boldsymbol{X}} & \Sigma_{\boldsymbol{Y}\boldsymbol{Y}} \end{array}\right]\right)$$

where $\Sigma_{\boldsymbol{XX}}$ is a matrix of size $n \times n$, $\Sigma_{\boldsymbol{XY}}$ is a matrix of size $n \times m$, $\Sigma_{\boldsymbol{YY}}$ is a matrix of size $\mathrm{m} \times m$.

- Marginal distribution over Y is a normal distribution $N(\mu_{\boldsymbol{Y}}; \Sigma_{\boldsymbol{YY}})$.

Independencies in Gaussians

Marginal independencies:

- Let $\boldsymbol{X} = X_1, \dots, X_n$ have a joint normal distribution $N(\mu; \Sigma)$. Then X_i and X_j are independent if and only if $\Sigma_{\mathrm{i,j}} = 0$.

- This property does not hold in general. For a non-Gaussian distribution, it is possible $\mathrm{Cov}(X, Y) = 0$ while X and Y are dependent.

Conditional independencies:

- Let $\mathrm{J} = \Sigma^{-1}$ be the information matrix. Then $J_{i,j} = 0$ if and only if $X_i \perp X_j \mid \mathcal{X} - \{X_i, X_j\}$

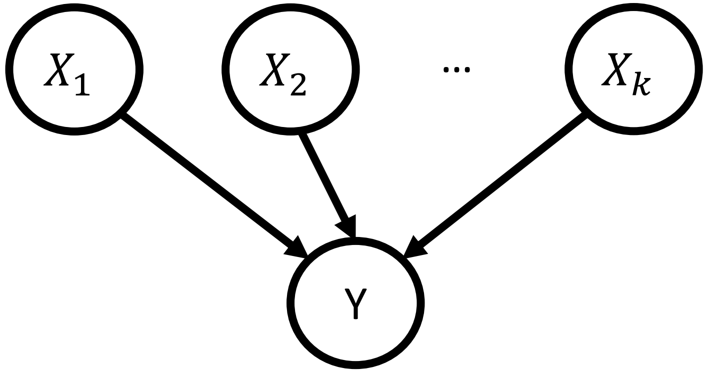
- Example: $J_{1,3} = 0$ indicates $X_1 \perp X_3 \mid X_2$.

$$J = \begin{pmatrix} 0.3125 & -0.125 & 0 \\ -0.125 & 0.5833 & 0.3333 \\ 0 & 0.3333 & 0.3333 \end{pmatrix}$$

From linear Gaussian models to multivariate Gaussians

Theorem:

- Given:
  - Y be a linear Gaussian of its parents $X_1, \dots, X_k$
  - $X_1, \dots, X_k$ are jointly Gaussian with distribution $N(\boldsymbol{\mu}; \Sigma)$



$P(Y \mid x) = N(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}; \sigma^2)$

- Then:
  - $P(Y) = N(\mu_Y; \sigma_Y^2)$ where
    - $\mu_Y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu}$
    - $\sigma_Y^2 = \sigma^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}$.
  - The joint distribution over $\{\boldsymbol{X}, Y\}$ is a normal distribution where:
    - $Cov[X_i; Y] = \sum_{j=1}^{k} \beta_j \Sigma_{i,j}$

Example: Consider the linear Gaussian network $X_1 \to X_2 \to X_3$, where

$$\begin{aligned} p(X_1) &= \mathcal{N}(1; 4) \\ p(X_2 \mid X_1) &= \mathcal{N}(0.5X_1 - 3.5; 4) \\ p(X_3 \mid X_2) &= \mathcal{N}(-X_2 + 1; 3). \end{aligned}$$

$P(Y) = N(\mu_Y; \sigma_Y^2)$ where

$$\mu_Y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu}$$

$$\sigma_Y^2 = \sigma^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}.$$

$$Cov[X_i; Y] = \sum_{j=1}^{k} \beta_j \Sigma_{i,j}$$

- **Goal:** computing the joint Gaussian distribution $p(X_1, X_2, X_3)$.

Computing the mean of X_2 and X_3:

$$\begin{aligned} \mu_2 &= 0.5\mu_1 - 3.5 = 0.5 \cdot 1 - 3.5 = -3 \\ \mu_3 &= (-1)\mu_2 + 1 = (-1) \cdot (-3) + 1 = 4. \end{aligned}$$

Computing the variance of X_2 and X_3:

$$\begin{aligned} \Sigma_{22} &= 4 + (1/2)^2 \cdot 4 = 5 \\ \Sigma_{33} &= 3 + (-1)^2 \cdot 5 = 8. \end{aligned}$$

Computing the covariance:

$$\begin{aligned} \Sigma_{12} &= (1/2) \cdot 4 = 2 \\ \Sigma_{23} &= (-1) \cdot \Sigma_{22} = -5 \\ \Sigma_{13} &= (-1) \cdot \Sigma_{12} = -2. \end{aligned}$$

$p(X_1, X_2, X_3)$:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}$$

Probability query

Theorem 7.4 *Let $\{\boldsymbol{X}, Y\}$ have a joint normal distribution defined in equation (7.3). Then the conditional density*

$$p(Y \mid \boldsymbol{X}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}; \sigma^2\right),$$

is such that:

$$\begin{array}{rcl} \beta_0 & = & \mu_Y - \Sigma_{Y\boldsymbol{X}}\Sigma_{\boldsymbol{XX}}^{-1}\boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\beta} & = & \Sigma_{\boldsymbol{XX}}^{-1}\Sigma_{Y\boldsymbol{X}} \\ \sigma^2 & = & \Sigma_{YY} - \Sigma_{Y\boldsymbol{X}}\Sigma_{\boldsymbol{XX}}^{-1}\Sigma_{\boldsymbol{X}Y}. \end{array}$$

$$p(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\mu}_{\boldsymbol{Y}} \end{pmatrix}; \begin{bmatrix} \Sigma_{\boldsymbol{XX}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{YY}} \end{bmatrix}\right)$$

Example: $P(\,X_1 \mid X_3 = 2\,)$?

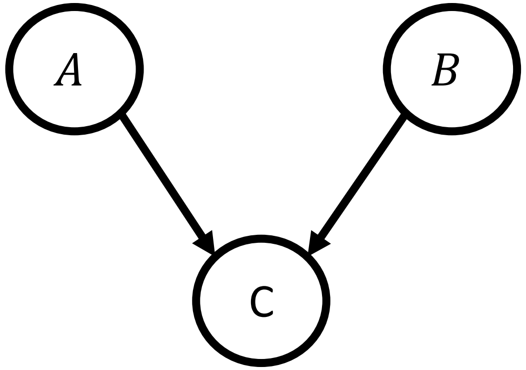
$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}$$

Hybrid Models

Hybrid Models

- Incorporate both discrete and continuous variables
- We need to address two types of dependencies:
  - **Case 1:** the continuous variable X with continuous parents Y and discrete parents U.
  - Simplest solution: define different set of parameters for every value $u \in Val(U)$

$$p(X \mid \boldsymbol{u}, \boldsymbol{y}) = \mathcal{N}\left(a_{\boldsymbol{u},0} + \sum_{i=1}^{k} a_{\boldsymbol{u},i} y_i; \sigma_{\boldsymbol{u}}^2\right)$$

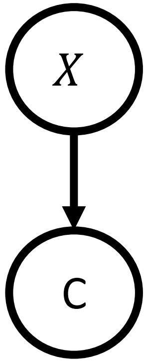


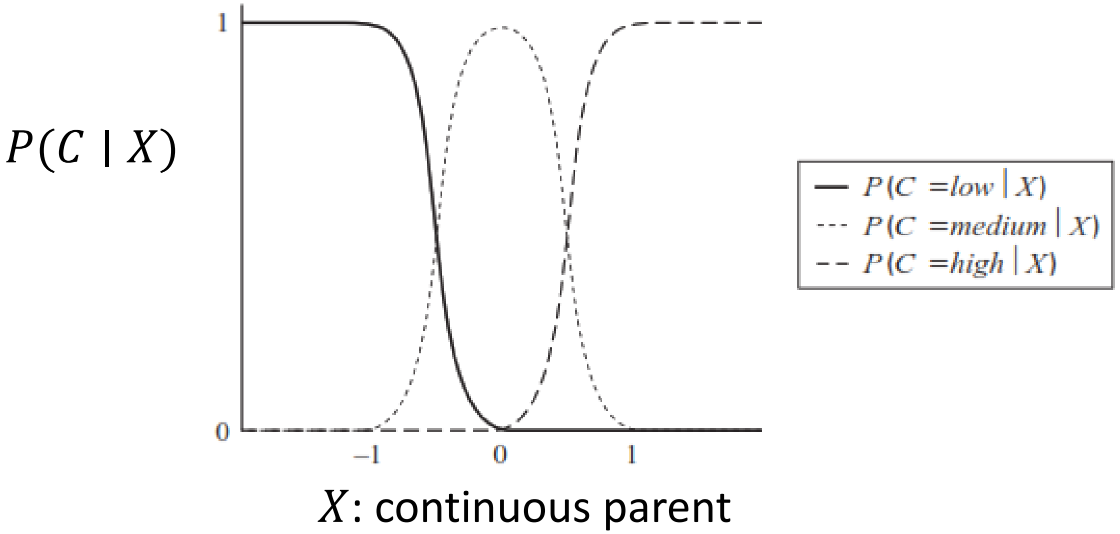
$$p(C) = \begin{cases} N(10 + 2b; 1) & A = 0 \\ N(20 - 3b; 4) & A = 1 \end{cases}$$

Hybrid Models

- Case 2: discrete child with continuous parents
  - One possibility: generalized linear models

  - Example: a sensor has three values: low, medium, high
  - It depends on a continuous parent X.





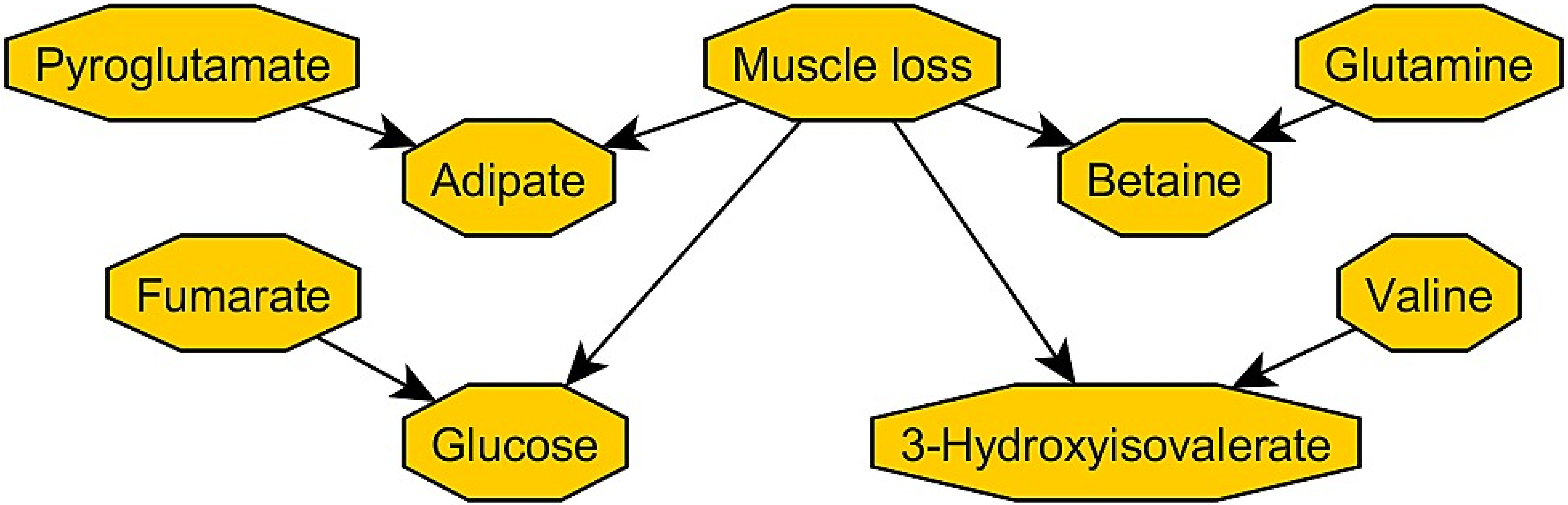
Conditional Gaussian Bayesian Network

- A special case of hybrid models
- Also referred to as *conditional linear Gaussian (CLG) model.*

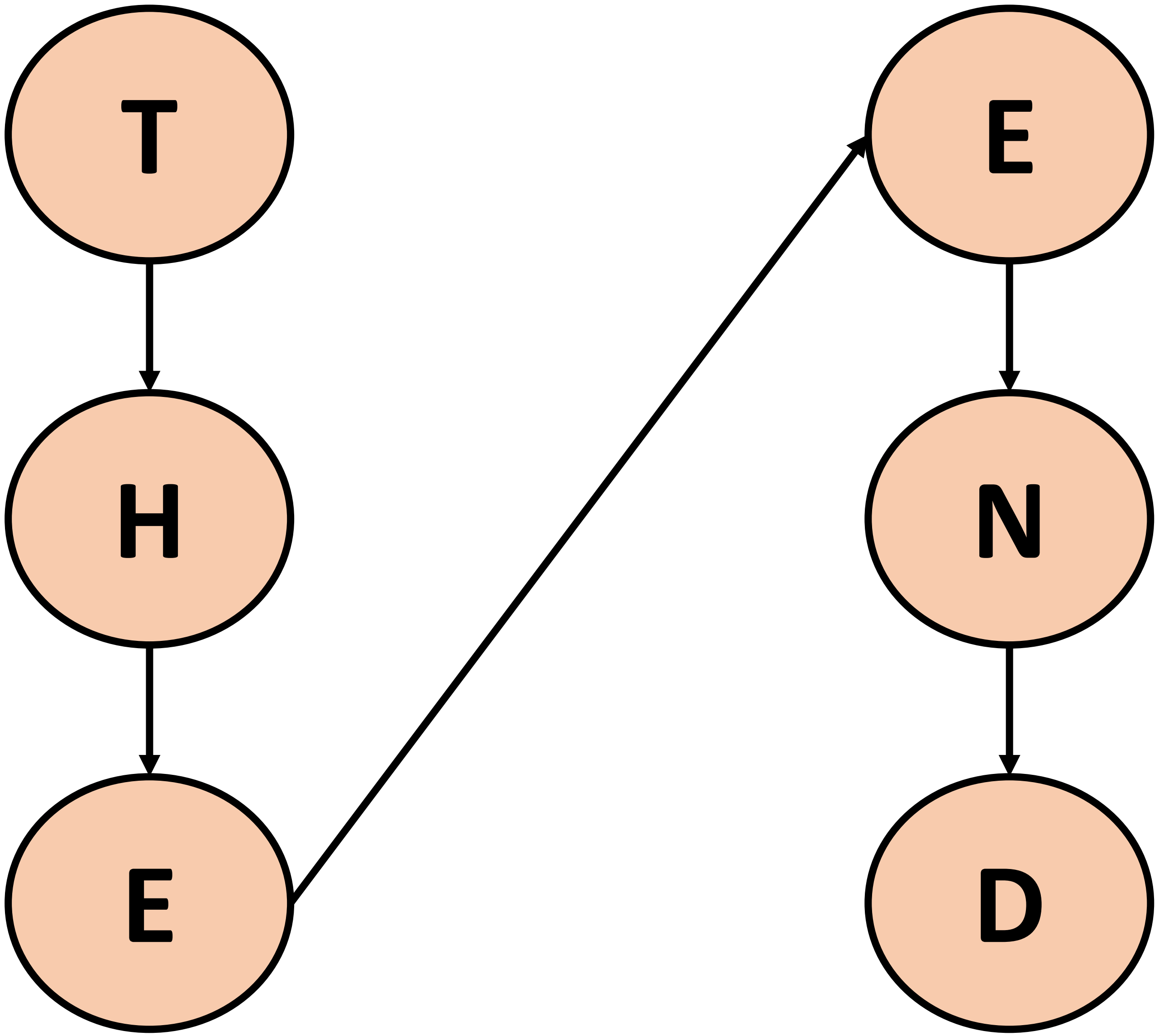
- Important: in this model, continuous variables cannot have discrete children.

- Distribution is a mixture of Gaussians
  - One component for each instantiation of discrete variables.

Conditional Gaussian Bayesian Network of Cachexia



Markov blanket of “Muscle loss” in an estimated conditional Gaussian BN [McGeachie, 2014]


T
H
E
E
N
D