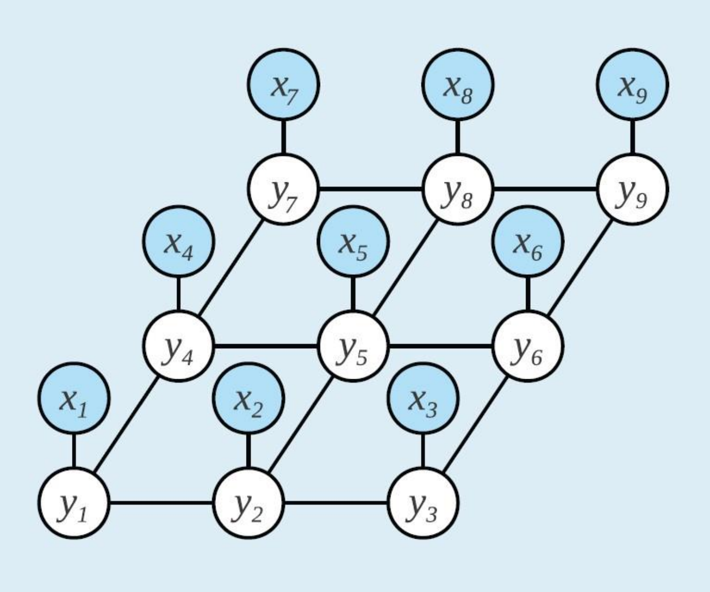
Probabilistic Graphical Models in Bioinformatics

Lecture 5: Learning in Bayesian networks

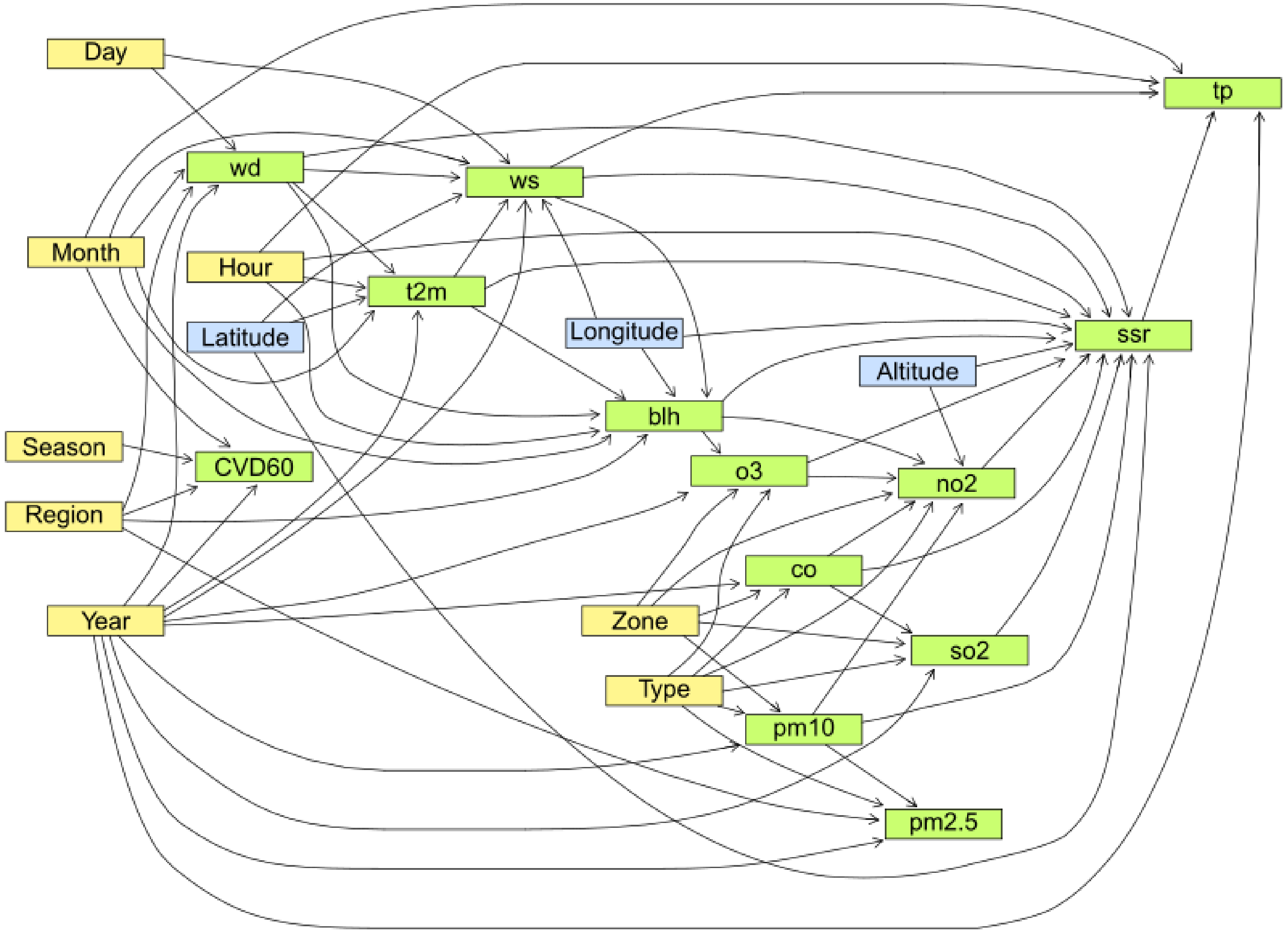


Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Esfand 20, 1397
(March 11, 2019)

Review by example: analysis of pollution, climate and health data, 2018.

- 50 million observations
- 24 variables
    - various air pollutants (O3, $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO)
    - geography (latitude, longitude, latitude, region and zone type)
    - climate (wind speed and direction, temperature, rainfall, solar radiation)
    - demography and mortality rates.

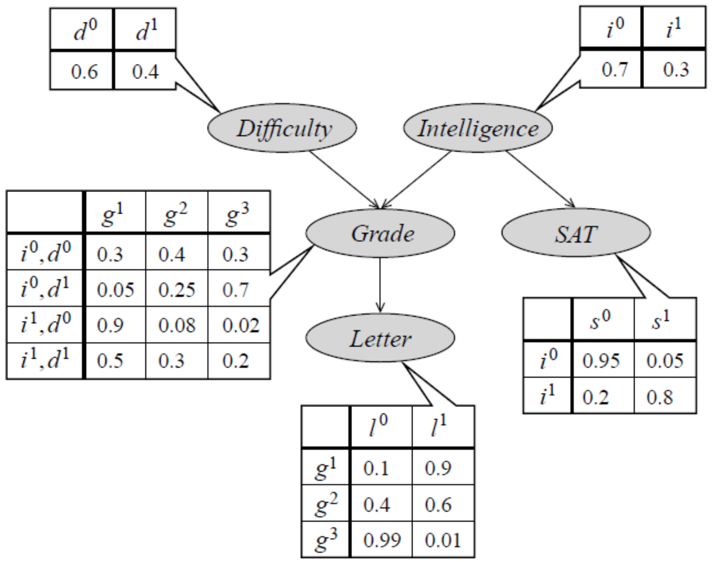


Yellow nodes: discrete
Blue nodes: Gaussian
Green nodes: conditional linear Gaussians

Learning Bayesian networks

Motivation for learning

Data → Learning process → A learned PGM → structure, parameters



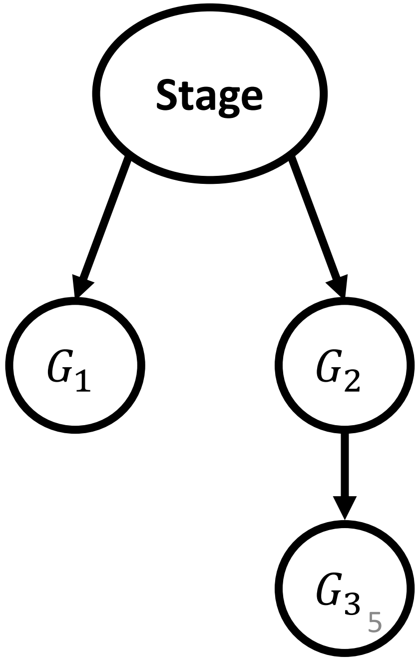
Learning tasks

- parameter estimation
- structure learning

Learning general Bayesian networks

	Known structure	Unknown structure
Fully observable	Easy Global decomposition	?
Partially observable	?	?

Question: examples of fully and partially observable data for the following network?



Parameter estimation

- Assumptions in this section:
  - the network structure is fixed.
  - Data set is fully observed
- Two main approaches for parameter estimation task
  - Maximum likelihood estimation
  - Bayesian estimation (don't confuse with Bayesian networks)

The thumbtack example

- Possible outcomes: head/tail

- Suppose X follows Bernoulli distribution with an unknown parameter $P(X = H) = \theta$.
- Data: out of M=100 tosses, 35 come up heads.
- Our intuition suggests the best estimate 0.35.
- **Question:** why $\theta = 0.1$ does not seem plausible?

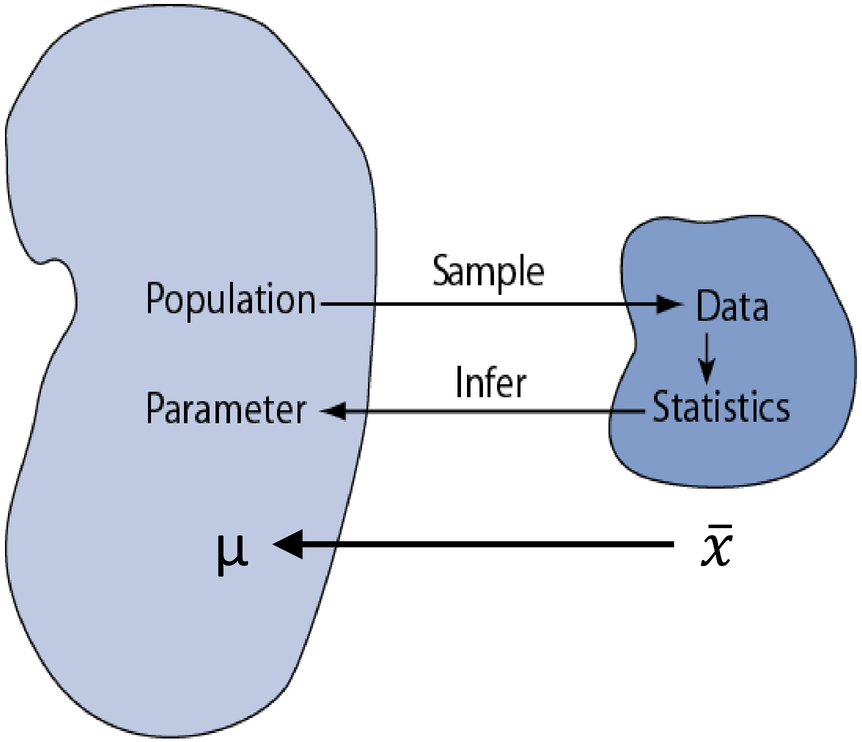
- How to formalize this intuition?
  - Optimize for θ which makes data more likely

Maximum likelihood estimate: $\hat{\theta}_{ML} = \arg\max_{\theta} P(D : \theta)$

Classical statistical inference

- **Statistical inference** is the act of generalizing from a **sample** to a **population** with calculated degree of certainty.

We want to learn about population *parameters…*



…but we can only calculate *sample statistics*

Parameters and Statistics

- It is essential that we draw distinctions between parameters and statistics

	Parameters	**Statistics**
Source	Population	Sample
Calculated?	No	Yes
Constants?	Yes	No
Examples	μ, σ, p	$\bar{x}, s, \hat{p}$

Maximum likelihood principle

- Likelihood function:
  - $\mathcal{L}(\theta : D) = P(D : \theta)$

$\mathcal{L}(\theta : D)$ is denoted differently than $P(D:\theta)$ to emphasize 1) likelihood is a function of θ (so data D is fixed in the likelihood function) 2) likelihood is not density or probability mass function.

- Maximum likelihood estimate:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}(\theta : D) = \arg\max_{\theta} P(D : \theta)$$

- Often much easier to work with log-likelihood function, $\ell(\theta : D) = \log \mathcal{L}(\theta : D)$

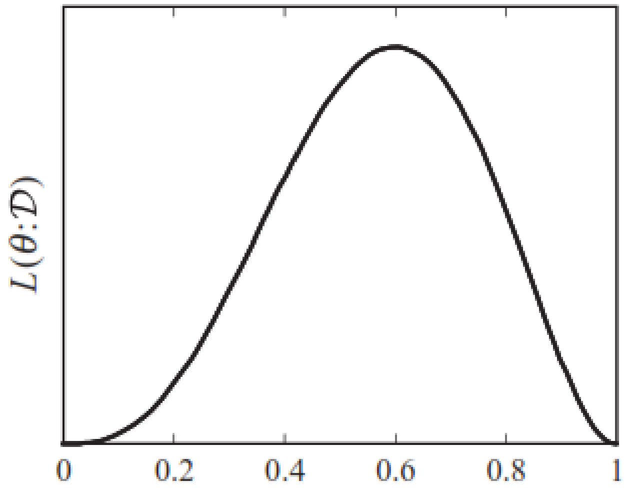
$$\hat{\theta}_{ML} = \arg\max_{\theta} \ell(\theta : D)$$

The thumbtack example-continued

- Data: we observe the sequence of outcomes H, T, T, H, H.

- Likelihood function:

$$L(\theta : \langle H, T, T, H, H \rangle) = P(\langle H, T, T, H, H \rangle : \theta) = \theta^3(1-\theta)^2.$$



$$\hat{\theta} = 0.6 = 3/5$$

ML- Bernoulli distribution

- Data: M[1] heads & M[0] tails
- Likelihood function:
  - $\mathcal{L}(\theta : D) = \theta^{M[1]}(1-\theta)^{M[0]}$
- Log-likelihood:
  - $\ell(\theta : D) = M[1]\log\theta + M[0]\log(1-\theta)$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{M[1]}{\theta} - \frac{M[0]}{1-\theta} = 0 \text{ results in } \hat{\theta}_{ML} = \frac{M[1]}{M[1]+M[0]}$$

Maximum Likelihood-Gaussian distribution

- Data: given observations $x_1, \ldots, x_n$ are normally distributed with the mean μ and the variance σ^2:

$$p(x : \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Likelihood function

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} p(x_i : \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- The log-likelihood is

$$\ell(\mu, \sigma) = \sum_{i=1}^{n} \log p(x_i : \mu, \sigma) = -\frac{1}{2} n \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

MLE solution:

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Maximum Likelihood-Multinomial distribution

- Data: suppose that X is a multinomial random variable that can take values $x^1, \dots, x^K$. The multinomial distribution has k parameters $\Theta = (\theta_1, \dots, \theta_K)$ such that

$$P(x : \Theta) = \theta_k \quad \text{if } x = x_k$$

$$\text{s.t. } \sum_k \theta_k = 1$$

- Data: $\langle M[1], \dots, M[K] \rangle$, such that $M[k]$ is number of times the value x^k appears in the data.

- **Homework:** prove the MLE estimate for θ_k is $\hat{\theta}_k = \frac{M[k]}{M}$ where $M = \sum_k M[k]$.

ML for Bayesian networks

University of TEHRAN

- Simple case: $X \to Y$

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_{m=1}^{M} P(x[m], y[m] : \boldsymbol{\theta}).$$

Question: write down all parameters.

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_{m} P(x[m] : \boldsymbol{\theta}) P(y[m] \mid x[m] : \boldsymbol{\theta}).$$

$$L(\boldsymbol{\theta} : \mathcal{D}) = \left(\prod_{m} P(x[m] : \boldsymbol{\theta}) \right) \left(\prod_{m} P(y[m] \mid x[m] : \boldsymbol{\theta}) \right)$$

Local likelihoods

ML for Bayesian networks

- Simple case: $X \rightarrow Y$

$$\prod_m P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|X})$$

$$= \prod_{m:x[m]=x^0} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|X}) \cdot \prod_{m:x[m]=x^1} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|X})$$

$$= \boxed{\prod_{m:x[m]=x^0} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|x^0})} \cdot \prod_{m:x[m]=x^1} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|x^1}).$$

Decomposability of the likelihood function

$$\prod_{m:x[m]=x^0} P(y[m] \mid x[m] : \boldsymbol{\theta}_{Y|x^0}) = \theta_{y^1|x^0}^{M[x^0,y^1]} \cdot \theta_{y^0|x^0}^{M[x^0,y^0]}.$$

ML estimate: $$\theta_{y^1|x^0} = \frac{M[x^0,y^1]}{M[x^0,y^1] + M[x^0,y^0]} = \frac{M[x^0,y^1]}{M[x^0]},$$

Global Likelihood decomposition

$$
\begin{aligned}
L(\boldsymbol{\theta} : \mathcal{D}) &= \prod_m P_{\mathcal{G}}(\xi[m] : \boldsymbol{\theta}) \\
&= \prod_m \prod_i P(x_i[m] \mid \mathrm{pa}_{X_i}[m] : \boldsymbol{\theta}) \\
&= \prod_i \left[\prod_m P(x_i[m] \mid \mathrm{pa}_{X_i}[m] : \boldsymbol{\theta}) \right]
\end{aligned}
$$

Only depends on $\theta_{X_i|pa_{X_i}}$

Let's define the local likelihood as:

$$
L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}} : \mathcal{D}) = \prod_m P(x_i[m] \mid \mathrm{pa}_{X_i}[m] : \boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}).
$$

then

$$
L(\boldsymbol{\theta} : \mathcal{D}) = \prod_i L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}} : \mathcal{D}),
$$

The likelihood decomposes a product of independent terms, one for each CPD in the network!

Conclusion: we can maximize each local likelihood independently.

Table CPDs

- The choice of parameters determines how we maximize each of the likelihood functions.

- By table CPDs, we will have $\theta_{x|\boldsymbol{u}}$ for each combination of $x \in Val(X)$ and $\boldsymbol{u} \in Val(U)$.

- For each choice of value for the parents U, we have the following constraint:

$$\sum \theta_{x|\boldsymbol{u}} = 1 \quad \text{for all } \boldsymbol{u}.$$

The MLE parameters: $\hat{\theta}_{x|\boldsymbol{u}} = \frac{M[\boldsymbol{u}, x]}{M[\boldsymbol{u}]}$ where $M[\boldsymbol{u}] = \sum_x M[\boldsymbol{u}, x]$.

Data fragmentation & overfitting

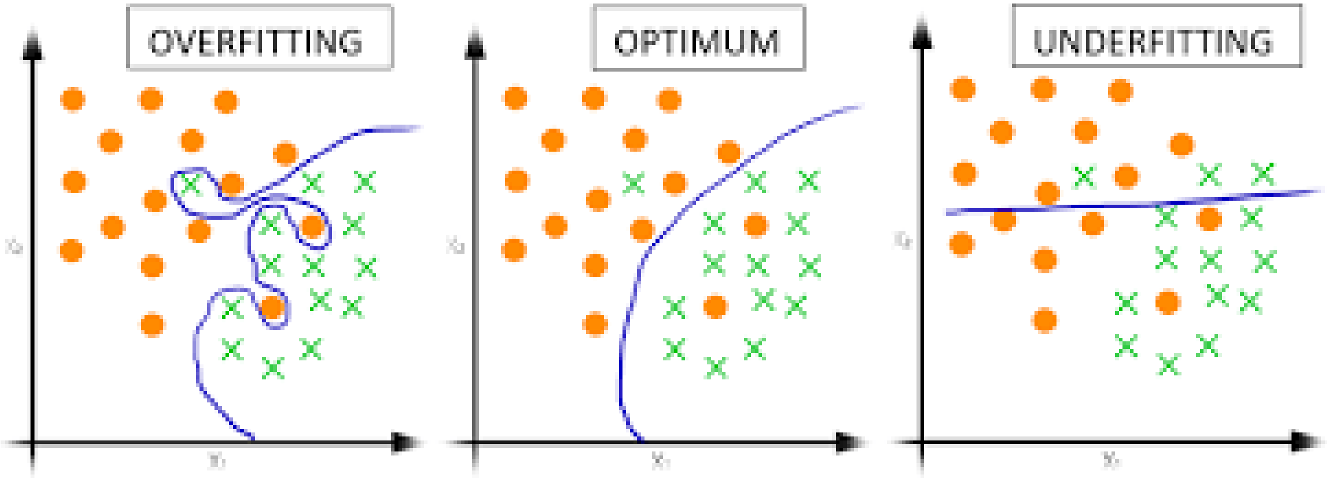
- Number of data points used to estimate parameter $\hat{\theta}_{x|u}$ is M[u]
  - Estimated from samples with parent value u

- Data points that do not agree with the parent assignment u play no role in this computation.

- As the number of parents U grows, number of parent assignment grows exponentially:
  - Hence, we may have a very small number of data instances, M[u], to estimate a parameter *(data fragmentation)*
  - Might results in *overfitting.*

Overfitting

- Overfitting refers to a model that models the training data too well.
- It happens when the model is too complex.



ML-Gaussian Bayesian networks

- Consider a variable X with parents $\mathbf{U} = \{U_1, \dots, U_k\}$ with a linear Gaussian CPD:

$$P(X \mid \boldsymbol{u}) = \mathcal{N}\left(\beta_0 + \beta_1 u_1 + \dots, \beta_k u_k; \sigma^2\right).$$

- **Goal:** to learn the parameters $\boldsymbol{\theta}_{X|U} = \langle \beta_0, \dots, \beta_k, \sigma \rangle$.

- Log-likelihood:

$$\begin{aligned} \ell_X(\boldsymbol{\theta}_{X|\boldsymbol{U}} : \mathcal{D}) &= \log L_X(\boldsymbol{\theta}_{X|\boldsymbol{U}} : \mathcal{D}) \\ &= \sum_m \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{1}{\sigma^2}\left(\beta_0 + \beta_1 u_1[m] + \dots + \beta_k u_k[m] - x[m]\right)^2 \right] \end{aligned}$$

- We need to solve the following equations to obtain the MLE solution

$$\frac{\partial}{\partial \beta_i}\ell(\boldsymbol{\theta}_{X|U} : D) = 0, \quad \frac{\partial}{\partial \sigma^2}\ell(\boldsymbol{\theta}_{X|U} : D) = 0$$

ML-Gaussian Bayesian networks 2

- The ML solution:
  1. We estimate means of X and U and covariance matrix of $\{X\} \cup U$ from the data.
  2. This is the ML estimate of the joint Gaussian.
  3. Then theorem 7.4 shows the solutions for the equations in the previous slide

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:

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

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

Example- estimation of $P(X \mid U)$

Data:

U	X
80.2	17.0
83.1	45.1
92.5	39.7
85.8	36.5
76.9	43.5
76.1	35.3
83.8	70.2
92.4	67.8
82.4	53.3

$$\hat{\mu} = \begin{pmatrix} 83.69 \\ 45.38 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} 30.65 & 34.16 \\ 34.16 & 244.80 \end{pmatrix}$$

Use theorem 7.4

$$\hat{\beta}_0 = -47.91$$

$$\hat{\beta}_1 = 1.11$$

$$\hat{\sigma} = 14.38$$

Sketch of the proof

Goal: to find the ML parameters for $P(X \mid \boldsymbol{u}) = \mathcal{N}\left(\beta_0 + \beta_1 u_1 + \ldots, \beta_k u_k; \sigma^2\right).$

By definition of the Gaussian distribution :

$$\begin{aligned}\ell_X(\boldsymbol{\theta}_{X|\boldsymbol{U}} : \mathcal{D}) &= \log L_X(\boldsymbol{\theta}_{X|\boldsymbol{U}} : \mathcal{D}) \\ &= \sum_m \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{1}{\sigma^2}\left(\beta_0 + \beta_1 u_1[m] + \ldots + \beta_k u_k[m] - x[m]\right)^2\right].\end{aligned}$$

We consider the gradient of the log-likelihood with respect to β_0

$$\begin{aligned}\frac{\partial}{\partial \beta_0}\ell_X(\boldsymbol{\theta}_{X|\boldsymbol{U}} : \mathcal{D}) &= \sum_m -\frac{1}{\sigma^2}\left(\beta_0 + \beta_1 u_1[m] + \ldots + \beta_k u_k[m] - x[m]\right) \\ &= -\frac{1}{\sigma^2}\left(M\beta_0 + \beta_1\sum_m u_1[m] + \ldots + \beta_k\sum_m u_k[m] - \sum_m x[m]\right).\end{aligned}$$

By equating the equation to 0, and multiplying both sides with $\frac{\sigma^2}{M}$, we get

$$\frac{1}{M}\sum_m x[m] = \beta_0 + \beta_1 \boxed{\frac{1}{M}\sum_m u_1[m]} + \ldots + \beta_k \boxed{\frac{1}{M}\sum_m u_k[m]}.$$

Average value of each variable in the data

New notation: $\boldsymbol{E}_{\mathcal{D}}[X] = \frac{1}{M}\sum_m x[m]$

Hence $\boldsymbol{E}_{\mathcal{D}}[X] = \beta_0 + \beta_1 \boldsymbol{E}_{\mathcal{D}}[U_1] + \ldots + \beta_k \boldsymbol{E}_{\mathcal{D}}[U_k].$

Sketch of the proof-2

$$\boxed{\boldsymbol{E}_{\mathcal{D}}[X] = \beta_0 + \beta_1 \boldsymbol{E}_{\mathcal{D}}[U_1] + \ldots + \beta_k \boldsymbol{E}_{\mathcal{D}}[U_k].}$$

Similarly, the equation $0 = \frac{\partial}{\partial \beta_i} \ell_X(\boldsymbol{\theta}_{X|\boldsymbol{U}} : \mathcal{D})$ can be formulated as

$$\boxed{\boldsymbol{Cov}_{\mathcal{D}}[X; U_i] = \beta_1 \boldsymbol{Cov}_{\mathcal{D}}[U_1; U_i] + \ldots + \beta_k \boldsymbol{Cov}_{\mathcal{D}}[U_k; U_i].}$$

Finally, we get the following equation for $\frac{\partial}{\partial \sigma^2} \ell(\boldsymbol{\theta}_{X|U} : D) = 0$

$$\boxed{\sigma^2 = \boldsymbol{Cov}_{\mathcal{D}}[X; X] - \sum_i \sum_j \beta_i \beta_j \boldsymbol{Cov}_{\mathcal{D}}[U_i; U_j]}$$

Formulas in theorem 7.4 give the solution to the system of linear equations

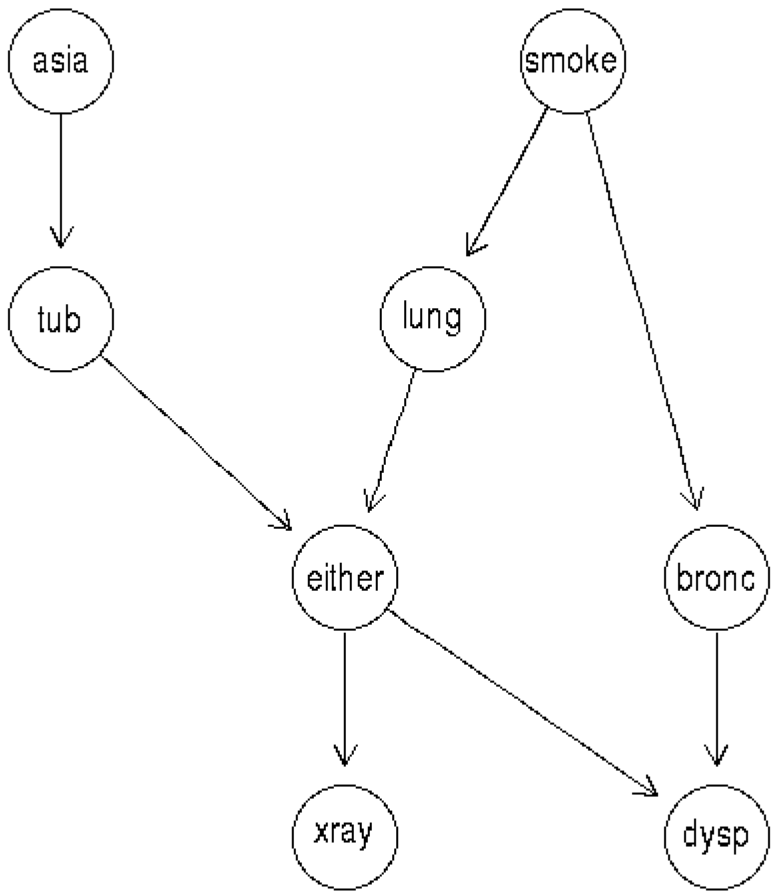
Theorem 7.4 *Let $\{\boldsymbol{X}, Y\}$ have a joint normal distribution defined in equation (7.3). Then the conditional density*

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

is such that:

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

Quiz: given the following BN:



1. Write the joint distribution

2. Explain how you compute P(bronc | tub)

3. True or False
   a) $xray \perp bronc$
   b) $asia \perp smoke \mid bronc$
   c) $tub \perp bronc \mid xray$

References

[1] Probabilistic Graphical Models by Daphne Koller & Nir Friedman, chapter 17.

[2] Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A. and Russell, A., 2018. Modeling air pollution, climate, and health data using Bayesian Networks: A case study of the English regions. Earth and Space Science, 5(4), pp.76-88.