# Overview of Course

So far, we have studied

- The concept of Bayesian network
- Independence and Separation in Bayesian networks
- Inference in Bayesian networks

The rest of the course: Data analysis using Bayesian network

- **Parameter learning**: Learn parameters for a given structure.
- **Structure learning**: Learn both structures and parameters
- **Learning latent structures**: Discover latent variables behind observed variables and determine their relationships.

# COMP538: Introduction to Bayesian Networks

## Lecture 6: Parameter Learning in Bayesian Networks

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Fall 2008

# Objective

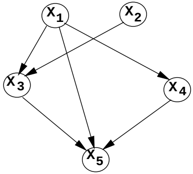
- Objective:
  - Principles for parameter learning in Bayesian networks.
  - Algorithms for the case of complete data.

- Reading: Zhang and Guo (2007), Chapter 7
- Reference: Heckerman (1996) (first half), Cowell *et al* (1999, Chapter 9)

# Outline

# Parameter Learning

- Given:
  - A Bayesian network structure.




  - A data set

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 | 1 |
| : | : | : | : | : |

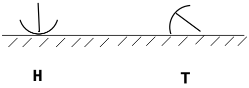
- Estimate conditional probabilities:

$$P(X_1), P(X_2), P(X_3|X_1, X_2), P(X_4|X_1), P(X_5|X_1, X_3, X_4)$$

# Outline

# Single-Node Bayesian Network


X: result of tossing a thumbtack



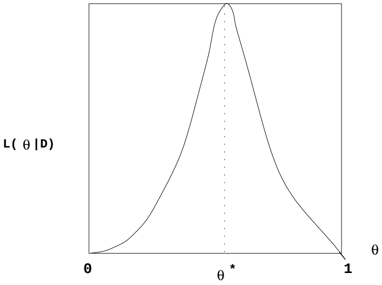

- Consider a Bayesian network with one node $X$, where $X$ is the result of tossing a thumbtack and $\Omega_X = \{H, T\}$.
- Data cases: $D_1 = H$, $D_2 = T$, $D_3 = H$, ..., $D_m = H$
- Data set: $\mathbf{D} = \{D_1, D_2, D_3, \ldots, D_m\}$
- Estimate parameter: $\theta = P(X{=}H)$.

# Likelihood

- Data: $\mathbf{D} = \{H, T, H, T, T, H, T\}$
- As possible values of $\theta$, which of the following is the most likely? Why?
  - $\theta = 0$
  - $\theta = 0.01$
  - $\theta = 10.5$

- $\theta = 0$ contradicts data because $P(\mathbf{D}|\theta = 0) = 0$.It cannot explain the data at all.
- $\theta = 0.01$ almost contradicts with the data. It does not explain the data well. However, it is more consistent with the data than $\theta = 0$ because $P(\mathbf{D}|\theta = 0.01) > P(\mathbf{D}|\theta = 0)$.
- So $\theta = 0.5$ is more consistent with the data than $\theta = 0.01$ because $P(\mathbf{D}|\theta = 0.5) > P(\mathbf{D}|\theta = 0.01)$
  It explains the data the best among the three and is hence the most likely.

# Maximum Likelihood Estimation




- In general, the larger $P(\mathbf{D}|\theta = v)$ is, the more likely $\theta = v$ is.
- Likelihood of parameter $\theta$ given data set:

$$L(\theta|\mathbf{D}) = P(\mathbf{D}|\theta)$$

- The **maximum likelihood estimation (MLE)** $\theta^*$ of $\theta$ is a possible value of $\theta$ such that

$$L(\theta^*|\mathbf{D}) = sup_\theta L(\theta|\mathbf{D}).$$

MLE best explains data or best fits data.

# i.i.d and Likelihood

- Assume the data cases $D_1, \ldots, D_m$ are independent given $\theta$:

$$P(D_1, \ldots, D_m|\theta) = \prod_{i=1}^{m} P(D_i|\theta)$$

- Assume the data cases are identically distributed:

$$P(D_i = H) = \theta, P(D_i = T) = 1-\theta \qquad \text{for all i}$$

(Note: i.i.d means independent and identically distributed)

- Then

$$\begin{aligned} L(\theta|\mathbf{D}) &= P(\mathbf{D}|\theta) = P(D_1, \ldots, D_m|\theta) \\ &= \prod_{i=1}^{m} P(D_i|\theta) = \theta^{m_h}(1-\theta)^{m_t} \end{aligned} \qquad (1)$$

where $m_h$ is the number of heads and $m_t$ is the number of tail. **Binomial likelihood**.

# Example of Likelihood Function

- Example: $\mathbf{D} = \{D_1 = H, D_2 T, D_3 = H, D_4 = H, D_5 = T\}$

$$
\begin{aligned}
L(\theta|\mathbf{D}) &= P(\mathbf{D}|\theta) \\
&= P(D_1 = H|\theta)P(D_2 = T|\theta)P(D_3 = H|\theta)P(D_4 = H|\theta)P(D_5 = T|\theta) \\
&= \theta(1-\theta)\theta\theta(1-\theta) \\
&= \theta^3(1-\theta)^2.
\end{aligned}
$$

# Sufficient Statistic

- A **sufficient statistic** is a function $s(\mathbf{D})$ of data that summarizing the relevant information for computing the likelihood. That is

$$s(\mathbf{D}) = s(\mathbf{D}') \Rightarrow L(\theta|\mathbf{D}) = L(\theta|\mathbf{D}')$$

- Sufficient statistics tell us all there is to know about data.
- Since $L(\theta|\mathbf{D}) = \theta^{m_h}(1-\theta)^{m_t}$,
  the pair $(m_h, m_t)$ is a **sufficient statistic**.

# Loglikelihood

- **Loglikelihood**:

$$l(\theta|\mathbf{D}) = logL(\theta|\mathbf{D}) = log\theta^{m_h}(1-\theta)^{m_t} = m_h log\theta + m_t log(1-\theta)$$

  Maximizing likelihood is the same as maximizing loglikelihood. The latter is easier.

- By Corollary 1.1 of Lecture 1, the following value maximizes $l(\theta|\mathbf{D})$:

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}$$

- MLE is intuitive.
- It also has nice properties:
  - E.g. **Consistence**: $\theta^*$ approaches the true value of $\theta$ with probability 1 as $m$ goes to infinity.

# Drawback of MLE

- Thumbtack tossing:
  - $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
  - Reasonable. Data suggest that the thumbtack is biased toward tail.
- Coin tossing:
  - Case 1: $(m_h, m_t) = (3, 7)$. MLE: $\theta = 0.3$.
    - Not reasonable.
    - Our experience (prior) suggests strongly that coins are fair, hence $\theta$=1/2.
    - The size of the data set is too small to convince us this particular coin is biased.
    - The fact that we get (3, 7) instead of (5, 5) is probably due to randomness.
  - Case 2: $(m_h, m_t) = (30,000, 70,000)$. MLE: $\theta = 0.3$.
    - Reasonable.
    - Data suggest that the coin is after all biased, overshadowing our prior.
  - MLE does not differentiate between those two cases. It doe not take prior information into account.

# Two Views on Parameter Estimation

**MLE**:

- Assumes that $\theta$ is unknown but fixed parameter.
- Estimates it using $\theta^*$, the value that maximizes the likelihood function
- Makes prediction based on the estimation: $P(D_{m+1} = H|\mathbf{D}) = \theta^*$

**Bayesian Estimation**:

- Treats $\theta$ as a random variable.
- Assumes a prior probability of $\theta$: $p(\theta)$
- Uses data to get posterior probability of $\theta$: $p(\theta|\mathbf{D})$

# Two Views on Parameter Estimation

**Bayesian Estimation**:

- Predicting $D_{m+1}$

$$
\begin{aligned}
P(D_{m+1} = H|\mathbf{D}) &= \int P(D_{m+1} = H, \theta|\mathbf{D}) d\theta \\
&= \int P(D_{m+1} = H|\theta, \mathbf{D}) p(\theta|\mathbf{D}) d\theta \\
&= \int P(D_{m+1} = H|\theta) p(\theta|\mathbf{D}) d\theta \\
&= \int \theta p(\theta|\mathbf{D}) d\theta.
\end{aligned}
$$

  **Full Bayesian**: Take expectation over $\theta$.

- **Bayesian MAP**:

$$P(D_{m+1} = H|\mathbf{D}) = \theta^* = \arg\max p(\theta|\mathbf{D})$$

# Calculating Bayesian Estimation

- Posterior distribution:

$$\begin{aligned} p(\theta|\mathbf{D}) &\propto p(\theta)L(\theta|\mathbf{D}) \\ &= \theta^{m_h}(1-\theta)^{m_t}p(\theta) \end{aligned}$$

  where the equation follows from (1)

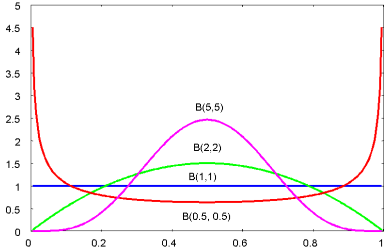
- To facilitate analysis, assume prior has **Beta distribution** $B(\alpha_h, \alpha_t)$

$$p(\theta) \propto \theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$$

- Then

$$p(\theta|\mathbf{D}) \propto \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1} \quad (2)$$

# Beta Distribution




- The normalization constant for the Beta distribution $B(\alpha_h, \alpha_t)$

$$\frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}$$

where $\Gamma(.)$ is the **Gamma function**. For any integer $\alpha$, $\Gamma(\alpha) = (\alpha - 1)!$. It is also defined for non-integers.

- Density function of prior Beta distribution $B(\alpha_h, \alpha_t)$,

$$p(\theta) = \frac{\Gamma(\alpha_t + \alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)} \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

- The **hyperparameters** $\alpha_h$ and $\alpha_t$ can be thought of as "imaginary" counts from our prior experiences.
- Their sum $\alpha = \alpha_h + \alpha_t$ is called **equivalent sample size**.
- The larger the equivalent sample size, the more confident we are in our prior.

# Conjugate Families

- Binomial Likelihood: $\theta^{m_h}(1-\theta)^{m_t}$
- Beta Prior: $\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$
- Beta Posterior: $\theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}$.
- Beta distributions are hence called a **conjugate family** for Binomial likelihood.
- Conjugate families allow closed-form for posterior distribution of parameters and closed-form solution for prediction.

# Calculating Prediction

- We have

$$P(D_{m+1} = H|\mathbf{D}) = \int \theta p(\theta|\mathbf{D})d\theta$$
$$= c\int \theta\theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}d\theta$$
$$= \frac{m_h+\alpha_h}{m+\alpha}$$

where $c$ is the normalization constant, $m{=}m_h{+}m_t$, $\alpha = \alpha_h{+}\alpha_t$.

- Consequently,

$$P(D_{m+1} = T|\mathbf{D}) = \frac{m_t+\alpha_t}{m+\alpha}$$

- After taking data **D** into consideration, now our **updated belief** on $X{=}T$ is $\frac{m_t+\alpha_t}{m+\alpha}$.

# MLE and Bayesian estimation

- As $m$ goes to infinity, $P(D_{m+1} = H|\mathbf{D})$ approaches the MLE $\frac{m_h}{m_h+m_t}$,which approaches the true value of $\theta$ with probability 1.
- Coin tossing example revisited:
  - Suppose $\alpha_h = \alpha_t = 100$. Equivalent sample size: 200
  - In case 1,

$$P(D_{m+1} = H|\mathbf{D}) = \frac{3+100}{10+100+100} \approx 0.5$$

    Our prior prevails.
  - In case 2,

$$P(D_{m+1} = H|\mathbf{D}) = \frac{30,000+100}{100,0000+100+100} \approx 0.3$$

    Data prevail.

# Variable with Multiple Values

Bayesian networks with a single multi-valued variable.

- $\Omega_X = \{x_1, x_2, \ldots, x_r\}$.
- Let $\theta_i = P(X = x_i)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_r)$.
- Note that $\theta_i \geq 0$ and $\sum_i \theta_i = 1$.
- Suppose in a data set $\mathbf{D}$, there are $m_i$ data cases where $X$ takes value $x_i$.
- Then

$$L(\theta|\mathbf{D}) = P(\mathbf{D}|\theta) = \prod_{j=1}^{N} P(D_j|\theta) = \prod_{i=1}^{r} \theta_i^{m_i}$$

**Multinomial likelihood.**

# Dirichlet distributions

- Conjugate family for multinomial likelihood: **Dirichlet distributions**.
  - A Dirichlet distribution is parameterized by $r$ parameters $\alpha_1$, $\alpha_2$, ..., $\alpha_r$.
  - Density function given by

$$\frac{\Gamma(\alpha)}{\prod_{i=1}^{r}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

    where $\alpha = \alpha_1 + \alpha_2 + \ldots + \alpha_r$.
  - Same as Beta distribution when $r$=2.
  - Fact: For any $i$:

$$\int \theta_i \frac{\Gamma(\alpha)}{\prod_{i=1}^{r}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1} d\theta_1 d\theta_2 \ldots d\theta_r = \frac{\alpha_i}{\alpha}$$

# Calculating Parameter Estimations

- If the prior probability is a Dirichlet distribution $Dir(\alpha_1, \alpha_2, \ldots, \alpha_r)$, then the posterior probability $p(\theta|D)$ is a given by

$$p(\theta|D) \propto \prod_{i=1}^{r} \theta_i^{m_i+\alpha_i-1}$$

- So it is Dirichlet distribution $Dir(\alpha_1 + m_1, \alpha_2 + m_2, \ldots, \alpha_r + m_r)$,
- Bayesian estimation has the following closed-form:

$$P(D_{m+1}{=}x_i|\mathbf{D}) = \int \theta_i p(\theta|\mathbf{D}) d\theta = \frac{\alpha_i + m_i}{\alpha + m}$$

- MLE: $\theta_i^* = \frac{m_i}{m}$. (Exercise: Prove this.)

# Outline

# The Parameters

- $n$ variables: $X_1$, $X_2$, ..., $X_n$.
- Number of states of $X_i$: 1, 2, ..., $r_i{=}|\Omega_{X_i}|$.
- Number of configurations of parents of $X_i$ : 1, 2, ..., $q_i{=}|\Omega_{pa(X_i)}|$.
- Parameters to be estimated:

$$\theta_{ijk} = P(X_i = j | pa(X_i) = k), \qquad i = 1, \ldots, n; j = 1, \ldots, r_i; k = 1, \ldots, q_i$$

- Parameter vector: $\theta = \{\theta_{ijk} | i = 1, \ldots, n; j = 1, \ldots, r_i; k = 1, \ldots, q_i\}$.
  Note that $\sum_j \theta_{ijk} = 1 \forall i, k$
- $\theta_{i..}$: Vector of parameters for $P(X_i|pa(X_i))$

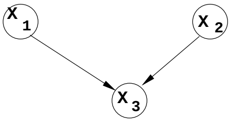
$$\theta_{i..} = \{\theta_{ijk} | j = 1, \ldots, r_i; k = 1, \ldots, q_i\}$$

- $\theta_{i.k}$: Vector of parameters for $P(X_i|pa(X_i){=}k)$

$$\theta_{i.k} = \{\theta_{ijk} | j = 1, \ldots, r_i\}$$

# The Parameters

- Example: Consider the Bayesian network shown below. Assume all variables are binary, taking values 1 and 2.




$$
\begin{aligned}
\theta_{111} &= P(X_1{=}1), \theta_{121} = P(X_1{=}2) \\
\theta_{211} &= P(X_2{=}1), \theta_{221} = P(X_2{=}2) \\
pa(X_3) = 1 : \theta_{311} &= P(X_3{=}1|X_1 = 1, X_2 = 1), \theta_{321} = P(X_3{=}2|X_1 = 1, X_2 = 1) \\
pa(X_3) = 2 : \theta_{312} &= P(X_3{=}1|X_1 = 1, X_2 = 2), \theta_{322} = P(X_3{=}2|X_1 = 1, X_2 = 2) \\
pa(X_3) = 3 : \theta_{313} &= P(X_3{=}1|X_1 = 2, X_2 = 1), \theta_{323} = P(X_3{=}2|X_1 = 2, X_2 = 1) \\
pa(X_3) = 4 : \theta_{314} &= P(X_3{=}1|X_1 = 2, X_2 = 2), \theta_{324} = P(X_3{=}2|X_1 = 2, X_2 = 2)
\end{aligned}
$$

# Data

- A complete case $D_l$: a vector of values, one for each variable.
- Example: $D_l = (X_1 = 1, X_2 = 2, X_3 = 2)$
- Given: A set of complete cases: $\mathbf{D} = \{D_1, D_2, \ldots, D_m\}$.
- Example:

| $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 1 | 1 |
| 1 | 1 | 2 | 2 | 1 | 2 |
| 1 | 1 | 2 | 2 | 2 | 1 |
| 1 | 2 | 2 | 2 | 2 | 1 |
| 1 | 2 | 2 | 2 | 2 | 2 |
| 1 | 2 | 2 | 2 | 2 | 2 |
| 2 | 1 | 1 | 2 | 2 | 2 |
| 2 | 1 | 1 | 2 | 2 | 2 |

- Find: The ML estimates of the parameters $\theta$.

# The Loglikelihood Function

- Loglikelihood:

$$l(\theta|D) = logL(\theta|D) = logP(D|\theta) = log \prod_l P(D_l|\theta) = \sum_l logP(D_l|\theta).$$

- The term $logP(D_l|\theta)$:
  - $D_4 = (1, 2, 2)$,

$$\begin{aligned} logP(D_4|\theta) &= logP(X_1 = 1, X_2 = 2, X_3 = 2) \\ &= logP(X_1{=}1|\theta)P(X_2{=}2|\theta)P(X_3{=}2|X_1{=}1, X_2{=}2, \theta) \\ &= log\theta_{111} + log\theta_{221} + log\theta_{322}. \end{aligned}$$

Recall:
$\theta = \{\theta_{111}, \theta_{121}; \theta_{211}, \theta_{221}; \theta_{311}, \theta_{312}, \theta_{313}, \theta_{314}, \theta_{321}, \theta_{322}, \theta_{323}, \theta_{324}\}$

# The Loglikelihood Function

- Define the **characteristic function** of case $D_l$:

$$\chi(i,j,k:D_l) = \begin{cases} 1 & \text{if } X_i = j,\ pa(X_i) = k \text{ in } D_l \\ 0 & \text{otherwise} \end{cases}$$

- When $l$=4, $D_4 = (1,2,2)$.

$$\chi(1,1,1:D_4) = \chi(2,2,1:D_4) = \chi(3,2,2:D_4) = 1$$

$$\chi(i,j,k:D_4) = 0 \text{ for all other i, j, k}$$

- So, $logP(D_4|\theta) = \sum_{ijk} \chi(i,j,k;D_4) log\theta_{ijk}$
- In general,

$$logP(D_l|\theta) = \sum_{ijk} \chi(i,j,k:D_l) log\theta_{ijk}$$

# The Loglikelihood Function

- Define

$$m_{ijk} = \sum_{l} \chi(i, j, k : D_l).$$

It is the number of data cases where $X_i = j$ and $pa(X_i) = k$.

- Then

$$\begin{aligned} l(\theta|\mathbf{D}) &= \sum_{l} logP(D_l|\theta) \\ &= \sum_{l} \sum_{i,j,k} \chi(i, j, k : D_l) log\theta_{ijk} \\ &= \sum_{i,j,k} \sum_{l} \chi(i, j, k : D_l) log\theta_{ijk} \\ &= \sum_{ijk} m_{ijk} log\theta_{ijk} \\ &= \sum_{i,k} \sum_{j} m_{ijk} log\theta_{ijk}. \quad (4) \end{aligned}$$

# MLE

- Want:

$$\arg\max_{\theta} l(\theta|\mathbf{D}) = \arg\max_{\theta_{ijk}} \sum_{i,k} \sum_{j} m_{ijk} log \theta_{ijk}$$

- Note that $\theta_{ijk} = P(X_i{=}j|pa(X_i){=}k)$ and $\theta_{i'j'k'} = P(X_{i'}{=}j'|pa(X_{i'}){=}k')$ are not related if either $i{\neq}i'$ or $k{\neq}k'$.

- Consequently, we can separately maximize each term in the summation $\sum_{i,k}[\ldots]$

$$\arg\max_{\theta_{ijk}} \sum_{j} m_{ijk} log \theta_{ijk}$$

# MLE

- By Corollary 1.1 , we get

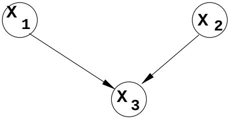
$$\theta^*_{ijk} = \frac{m_{ijk}}{\sum_j m_{ijk}}$$

- In words, the MLE estimate for $\theta_{ijk} = P(X_i{=}j|pa(X_i){=}k)$ is:

$$\theta^*_{ijk} = \frac{\text{number of cases where } X_i{=}j \text{ and } pa(X_i){=}k}{\text{number of cases where } pa(X_i){=}k}$$

# Example

Example:



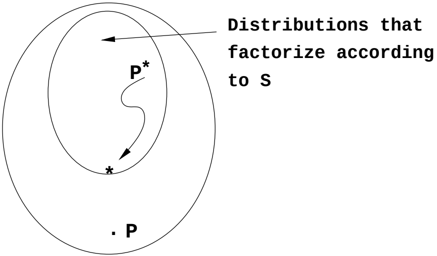

| $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 1 | 1 |
| 1 | 1 | 2 | 2 | 1 | 2 |
| 1 | 1 | 2 | 2 | 2 | 1 |
| 1 | 2 | 2 | 2 | 2 | 1 |
| 1 | 2 | 2 | 2 | 2 | 2 |
| 1 | 2 | 2 | 2 | 2 | 2 |
| 2 | 1 | 1 | 2 | 2 | 2 |
| 2 | 1 | 1 | 2 | 2 | 2 |

- MLE for $P(X_1{=}1)$ is: $6/16$
- MLE for $P(X_2{=}1)$ is: $7/16$
- MLE for $P(X_3{=}1|X_1{=}2, X_2{=}2)$ is: $2/6$
- ...

# A Question

- Start from a joint distribution $P(\mathbf{X})$ (**Generative Distribution**)
- **D**: collection of data sampled from $P(\mathbf{X})$.
- Let $S$ be a BN structrue (DAG) over variables **X**.
- Learn parameters $\theta^*$ for BN structure $S$ from **D**.
- Let $P^*(\mathbf{X})$ be the joint probability of the BN $(S, \theta^*)$.
  - Note: $\theta^*_{ijk} = P^*(X_i{=}j|pa_S(X_i){=}k)$
- How is $P^*$ related to $P$?

# MLE in General Bayesian Networks with Complete Data




- $P^*$ factorizes according to $S$.
- $P$ does not necessarily factorize according to $S$.

- We will show that, with probability 1, $P^*$ converges to the distribution that
  - Factorizes according to $S$,
  - Is closest to $P$ under KL divergence among all distributions that factorize according to $S$.
- If $P$ factorizes according to $S$, $P^*$ converges to $P$ with probability 1. (MLE is **consistent**.)

# The Target Distribution

- Define

$$\theta_{ijk}^{S} = P(X_i{=}j|pa_S(X_i) = k))$$

- Let $P^S(\mathbf{X})$ be the joint distribution of the BN $(S, \theta^S)$
- $P^S$ factorizes according to $S$ and for any $X \in \mathbf{X}$,

$$P^S(X|pa(X)) = P(X|pa(X))$$

- If $P$ factorizes according to $S$, then $P$ and $P^S$ are identical.
- If $P$ does not factorize according to $S$, then $P$ and $P^S$ are different.

# First Theorem

**Theorem (6.1)**

*Among all distributions Q that factorizes according to S, the KL divergence $KL(P, Q)$ is minimized by $Q{=}P^S$.*
*$P^S$ is the closest to P among all those that factorize according to S.*

**Proof**:

- Since

$$KL(P, Q) = \sum_{\mathbf{X}} P(\mathbf{X}) log \frac{P(\mathbf{X})}{Q(\mathbf{X})}$$

- It suffices to show that

*Proposition:* $Q{=}P^S$ *maximizes* $\sum_{\mathbf{X}} P(\mathbf{X}) log Q(\mathbf{X})$

- We show the claim by induction on the number of nodes.
- When there is only one node, the proposition follows from property of KL divergence (Corollary 1.1).

# First Theorem

- Suppose the proposition is true for the case of $n$ nodes. Consider the case of $n+1$ nodes.
- Let $X$ be a leaf node and $\mathbf{X}'=\mathbf{X} \setminus \{X\}$. $S'$ be the obtained from $S$ by removing $X$.
- Then

$$\sum_{\mathbf{X}} P(\mathbf{X}) logQ(\mathbf{X}) = \sum_{\mathbf{X}'} P(\mathbf{X}') logQ(\mathbf{X}') + \sum_{pa(X)} P(pa(X)) \sum_{X} P(X|pa(X)) logQ(X|pa(X))$$

- By the induction hypothesis, the first term is maximized by $P^{S'}$.
- By Corollary 1.1, the second term is maximized if $Q(X|pa(X)) = P(X|pa(X))$.
- Hence the sum is maximized by $P^{S}$.

# Second Theorem

**Theorem (6.2)**

$$\lim_{N \to \infty} P^*(\mathbf{X}=\mathbf{x}) = P^S(\mathbf{X}=\mathbf{x}) \textit{ with probability 1}$$

*where $N$ is the sample size, i.e. number of cases in* $\mathbf{D}$.

**Proof**:

- Let $\hat{P}(\mathbf{X})$ be the **empirical distribution**:

$$\hat{P}(\mathbf{X}=\mathbf{x}) = \text{ fraction of cases in } \mathbf{D} \text{ where } \mathbf{X}=\mathbf{x}$$

- It is clear that

$$P^*(X_i=j|pa_S(X_i)=k) = \theta^*_{ijk} = \hat{P}(X_i=j|pa_S(X_i)=k)$$

# Second Theorem

- On the other hand, by the law of large numbers, we have

$$\lim_{N\to\infty} \hat{P}(\mathbf{X}{=}\mathbf{x}) = P(\mathbf{X}{=}\mathbf{x}) \text{ with probability 1}$$

- Hence

$$\begin{aligned}
\lim_{N\to\infty} P^*(X_i{=}j|pa_S(X_i){=}k) &= \lim_{N\to\infty} \hat{P}(X_i{=}j|pa_S(X_i){=}k) \\
&= P(X_i{=}j|pa_S(X_i){=}k) \text{ with probability 1} \\
&= P^S(X_i{=}j|pa_S(X_i){=}k)
\end{aligned}$$

- Because both $P^*$ and $P^S$ factorizes according to $S$, the theorem follows. Q.E.D.

# A Corollary

**Corollary**

*If $P$ factorizes according to $S$, then*

$$\lim_{N \to \infty} P^*(\mathbf{X}{=}\mathbf{x}) = P(\mathbf{X}{=}\mathbf{x}) \text{ with probability 1}$$

# Bayesian Estimation

- View $\theta$ as a vector of random variables with prior distribution $p(\theta)$.
- Posterior:

$$\begin{aligned} p(\theta|\mathbf{D}) &\propto p(\theta)L(\theta|\mathbf{D}) \\ &= p(\theta)\prod_{i,k}\prod_{j}\theta_{ijk}^{m_{ijk}} \end{aligned}$$

where the equation follows from (4).

- Assumptions need to be made about prior distribution.

# Assumptions

- **Global independence** in prior distribution:

$$p(\theta) = \prod_i p(\theta_{i..})$$

- **Local independence** in prior distribution: For each $i$

$$p(\theta_{i..}) = \prod_k p(\theta_{i.k})$$

- **Parameter independence** = global independence + local independence:

$$p(\theta) = \prod_{i,k} p(\theta_{i.k})$$

# Assumptions

- Further assume that $p(\theta_{i.k})$ is Dirichlet distribution $Dir(\alpha_{i0k}, \alpha_{i1k}, \ldots, \alpha_{ir_ik})$:

$$p(\theta_{i.k}) \propto \prod_j \theta_{ijk}^{\alpha_{ijk}-1}$$

- Then,

$$p(\theta) = \prod_{i,k} \prod_j \theta_{ijk}^{\alpha_{ijk}-1}$$

**product Dirichlet distribution**.

# Bayesian Estimation

- Posterior:

$$
\begin{aligned}
p(\theta|\mathbf{D}) &\propto p(\theta)\prod_{i,k}\prod_{j}\theta_{ijk}^{m_{ijk}} \\
&= [\prod_{i,k}\prod_{j}\theta_{ijk}^{\alpha_{ijk}-1}]\prod_{i,k}\prod_{j}\theta_{ijk}^{m_{ijk}} \\
&= \prod_{i,k}\prod_{j}\theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}
\end{aligned}
$$

- It is also a product product Dirichlet distribution.(Think: What does this mean?)

# Prediction

- Predicting $D_{m+1} = \{X_1^{m+1}, X_2^{m+1}, \ldots, X_n^{m+1}\}$. Random variables.
- For notational simplicity, simply write $D_{m+1} = \{X_1, X_2, \ldots, X_n\}$.
- First, we have:

$$P(D_{m+1}|\mathbf{D}) = P(X_1, X_2, \ldots, X_n|\mathbf{D}) = \prod_i P(X_i|pa(X_i), \mathbf{D})$$

# Proof

$$P(D_{m+1}|\mathbf{D}) = \int P(D_{m+1}|\theta)p(\theta|\mathbf{D})d\theta$$

$$\begin{aligned} P(D_{m+1}|\theta) &= P(X_1, X_2, \ldots, X_n|\theta) \\ &= \prod_i P(X_i|pa(X_i), \theta) \\ &= \prod_i P(X_i|pa(X_i), \theta_{i..}) \end{aligned}$$

$$p(\theta_i|\mathbf{D}) = \prod_i p(\theta_{i..}|\mathbf{D})$$

Hence

$$\begin{aligned} P(D_{m+1}|\mathbf{D}) &= \prod_i \int P(X_i|pa(X_i), \theta_{i..})p(\theta_{i..}|\mathbf{D})d\theta_{i..} \\ &= \prod_i P(X_i|pa(X_i), \mathbf{D}) \end{aligned}$$

# Prediction

- Further, we have

$$\begin{aligned} P(X_i{=}j|pa(X_i){=}k, \mathbf{D}) &= \int P(X_i{=}j|pa(X_i) = k, \theta_{ijk})p(\theta_{ijk}|\mathbf{D})d\theta_{ijk} \\ &= \int \theta_{ijk}p(\theta_{ijk}|\mathbf{D})d\theta_{ijk} \end{aligned}$$

- Because

$$p(\theta_{i.k}|\mathbf{D}) \propto \prod_j \theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}$$

- We have

$$\int \theta_{ijk}p(\theta_{ijk}|\mathbf{D})d\theta_{ijk} = \frac{m_{ijk}+\alpha_{ijk}}{\sum_j(m_{ijk}+\alpha_{ijk})}$$

# Prediction

- Conclusion:

$$P(X_1, X_2, \ldots, X_n|\mathbf{D}) = \prod_i P(X_i|pa(X_i), \mathbf{D})$$

  where

$$P(X_i{=}j|pa(X_i){=}k, \mathbf{D}) = \frac{m_{ijk}+\alpha_{ijk}}{m_{i*k}+\alpha_{i*k}}$$

  where $m_{i*k} = \sum_j m_{ijk}$ and $\alpha_{i*k} = \sum_j \alpha_{ijk}$

- Notes:
  - Conditional independence or structure preserved after absorbing **D**.
  - Important property for sequential learning where we process one case at a time.
  - The final result is independent of the order by which cases are processed.
  - Comparison with MLE estimation:

$$\theta^*_{ijk} = \frac{m_{ijk}}{\sum_j m_{ijk}}$$

# Summary

- $\theta$: random variable.
- Prior $p(\theta)$: product Dirichlet distribution

$$p(\theta) = \prod_{i,k} p(\theta_{i.k}) \propto \prod_{i,k} \prod_{j} \theta_{ijk}^{\alpha_{ijk}-1}$$

- Posterior $p(\theta|\mathbf{D})$: also product Dirichlet distribution

$$p(\theta|\mathbf{D}) \propto \prod_{i,k} \prod_{j} \theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}$$

- Prediction:

$$P(D_{m+1}|\mathbf{D}) = P(X_1, X_2, \ldots, X_n|\mathbf{D}) = \prod_{i} P(X_i|pa(X_i), \mathbf{D})$$

where

$$P(X_i{=}j|pa(X_i){=}k, \mathbf{D}) = \frac{m_{ijk}+\alpha_{ijk}}{m_{i*k}+\alpha_{i*k}}$$