Readings: K&F 17.1, 17.2, 17.3, 17.4

Parameter Estimation

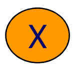
Lecture 9 – Apr 25, 2011
CSE 515, Statistical Methods, Spring 2011

Instructor: Su-In Lee
University of Washington, Seattle

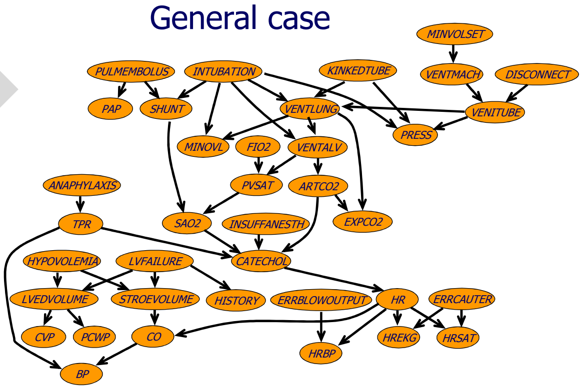
Parameter estimation

- Maximum likelihood estimation (MLE)
  - Parameter estimation based on observations
- Bayesian approach
  - Incorporate our prior knowledge

A single variable Bayesian network

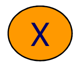


General case



Maximum Likelihood Estimator

- The *Coin* example – general case
  - X: result of a coin toss (head or tail)
  - Training data (instances) D=<x[1],...x[m]> (M_H heads and M_T tails)
  - Parameters: P(X=h)= θ



- Goal: find $\theta \in [0,1]$ that predicts the data well
  - "Predicts the data well" = likelihood of the data given θ

    $L(\theta : D) = P(D : \theta) = P(x[1],\ldots,x[m] : \theta)$
  - MLE: Find θ maximizing likelihood

    $L(\theta : D) = \prod_{i=1}^{m} P(x[i] \mid x[1],\ldots,x[i-1],\theta) = \prod_{i=1}^{m} P(x[i] \mid \theta) = \theta^{M_H}(1-\theta)^{M_T}$
  - Equivalent to maximizing log-likelihood

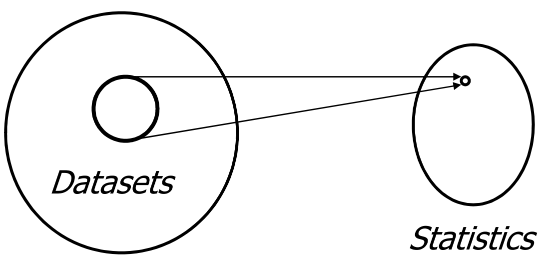
    $l(\theta : D) = \log P(D : \theta) = M_H \log\theta + M_T \log(1-\theta)$
  - Differentiating the log-likelihood and solving for θ, we get that the maximum likelihood parameter: $\theta_{mle} = \arg\max l(\theta : D) = \frac{M_H}{M_H + M_T}$

Sufficient Statistics

- For computing the parameter θ of the coin toss example, we only needed M_H and M_T since

$$L(\theta : D) = P(D : \theta) = \theta^{M_H}(1-\theta)^{M_T}$$

→ M_H and M_T are sufficient statistics

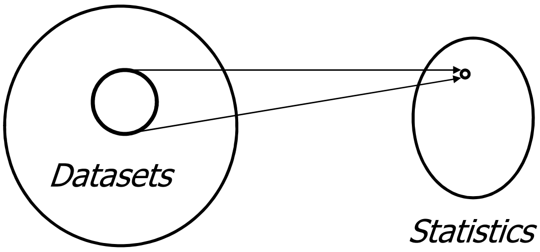


Sufficient Statistics

- A function s(D) is a sufficient statistic from instances to a vector in $\Re^k$ if, for any two datasets D and D′ and any $\theta \in \Theta$, we have

$$\sum_{x[i]\in D} s(x[i]) = \sum_{x[i]\in D'} s(x[i]) \quad \Rightarrow \quad L(D:\theta) = L(D':\theta)$$

- We often refer to the tuple $\sum_{x[i]\in D} s(x[i])$ as the sufficient statistics of the data set D.
  - In coin toss experiment, M_H and M_T are sufficient statistics



Sufficient Statistics for Multinomial

- Y: multinomial, k values (e.g. result of a dice throw)
- A sufficient statistics for a dataset D over Y is the tuple of counts $<M_1,\ldots M_k>$ such that M_i is the number of times that the $Y=y^i$ in D
- Likelihood function: $L(D:\theta) = \prod_{i=1}^{k} \theta_i^{M_i}$ where $\theta_i = P(Y = y^i)$
- MLE Principle: Choose Θ that maximize $L(D:\Theta)$
- Multinomial MLE: $\theta^i = \frac{M_i}{\sum_{i=1}^{m} M_i}$

Sufficient Statistic for Gaussian

- Gaussian distribution: $X \sim N(\mu, \sigma^2)$
  - Probability density function (pdf): $p(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Rewrite as $p(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-x^2 \frac{1}{2\sigma^2} + x\frac{\mu}{\sigma^2} - \frac{\mu^2}{\sigma^2}\right)$

  → sufficient statistics for Gaussian: <M,Σ_mx[m], Σ_mx[m]^2>

- MLE Principle: Choose Θ that maximize L(D:Θ)
- Multinomial MLE: $\mu = \frac{1}{M}\sum_m x[m]$

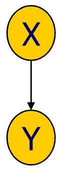
$$\sigma = \sqrt{\frac{1}{M}\sum_m (x[m]-\mu)^2}$$

MLE for Bayesian Networks

- Parameters
  - θ_{x0}, θ_{x1}
  - $\theta_{y0|x0}$, $\theta_{y1|x0}$, $\theta_{y0|x1}$, $\theta_{y1|x1}$
- Training data:
  - tuples <x[m],y[m]> m=1,…,M
- Likelihood function:

$$L(D:\theta) = \prod_{m=1}^{M} P(x[m], y[m]:\theta)$$
$$= \prod_{m=1}^{M} P(x[m]:\theta_X) P(y[m] \mid x[m]:\theta_{Y|X})$$
$$= \left(\prod_{m=1}^{M} P(x[m]:\theta_X)\right)\left(\prod_{m=1}^{M} P(y[m] \mid x[m]:\theta_{Y|X})\right)$$

X	
x^0	x^1
0.7	0.3



X	Y: y^0	y^1
x^0	0.95	0.05
x^1	0.2	0.8

→ Likelihood decomposes into two separate terms, one for each variable ("decomposability of the likelihood function")

MLE for Bayesian Networks

- Terms further decompose by CPDs:

$$\begin{aligned}\prod_{m=1}^{M} P(y[m] \mid x[m]:\theta) &= \prod_{m:x[m]=x^0} P(y[m] \mid x[m]:\theta_{Y|X}) \prod_{m:x[m]=x^1} P(y[m] \mid x[m]:\theta_{Y|X}) \\ &= \prod_{m:x[m]=x^0} P(y[m] \mid x[m]:\theta_{Y|x^0}) \prod_{m:x[m]=x^1} P(y[m] \mid x[m]:\theta_{Y|x^1})\end{aligned}$$

- By sufficient statistics

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

where $M[x^1,y^1]$ is the number of data instances in which X takes the value x^1 and Y takes the value y^1

- MLE

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

9

MLE for Bayesian Networks

- Likelihood for Bayesian network

$$\begin{aligned}L(\Theta : D) &= \prod_m P(x[m]:\Theta) \\ &= \prod_m \prod_i P(x_i[m] \mid Pa_i[m]:\Theta_i) \\ &= \prod_i \left[\prod_m P(x_i[m] \mid Pa_i[m]:\Theta_i)\right] \\ &= \prod_i L_i(\boldsymbol{\theta}_{x_i|Px_i} : X_i, Pa_i)\end{aligned}$$

Conditional likelihood or "Local likelihood"

→ if $\theta_{X_j|Pa(X_j)}$ are disjoint then MLE can be computed by maximizing each local likelihood separately

10

MLE for Table CPD BayesNets

- Multinomial CPD

$$L_Y(D:\theta_{Y|\mathbf{X}}) = \prod_m \theta_{y[m]|\mathbf{X}[m]}$$
$$= \prod_{\mathbf{x}\in Val(\mathbf{X})} \left[\prod_{y \in Val(Y)} \theta_{y|\mathbf{x}}{}^{M[\mathbf{x},y]} \right]$$

- For each value $\mathbf{x}\in\mathbf{X}$ we get an independent multinomial problem where the MLE is

$$\theta_{y^i|\mathbf{x}} = \frac{M[\mathbf{x},y^i]}{M[\mathbf{x}]}$$

MLE for Tree CPDs

- Assume tree CPD with known tree structure

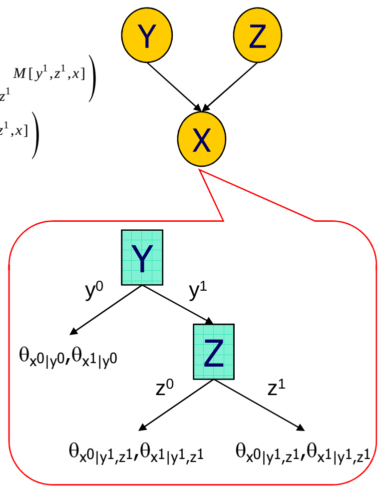
$$L(D:\theta_{X|Y,Z})$$
$$= \prod_y \prod_z \prod_x P(x \mid y,z,\theta_{X|Y,Z})^{M[y,z,x]}$$
$$= \prod_x \left(\theta_{x|y^0}{}^{M[y^0,z^0,x]}\right)\cdot\left(\theta_{x|y^0}{}^{M[y^0,z^1,x]}\right)\cdot\left(\theta_{x|y^1,z^0}{}^{M[y^1,z^0,x]}\right)\cdot\left(\theta_{x|y^1,z^1}{}^{M[y^1,z^1,x]}\right)$$
$$= \prod_x \left(\theta_{x|y^0}{}^{M[y^0,z^0,x]+M[y^0,z^1,x]}\right)\cdot\left(\theta_{x|y^1,z^0}{}^{M[y^1,z^0,x]}\right)\cdot\left(\theta_{x|y^1,z^1}{}^{M[y^1,z^1,x]}\right)$$
$$= \prod_x \left(\theta_{x|y^0}{}^{M[y^0,x]}\right)\cdot\left(\theta_{x|y^1,z^0}{}^{M[y^1,z^0,x]}\right)\cdot\left(\theta_{x|y^1,z^1}{}^{M[y^1,z^1,x]}\right)$$

$\downarrow$

Terms for <y^0,z^0> and <y^0,z^1> can be combined

$\downarrow$

Optimization can be done by leaves



MLE for Tree CPD BayesNets

- Tree CPD T, leaves l

$$
\begin{aligned}
L_Y(D:\theta_{Y|\mathbf{X}}) &= \prod_m P(y[m] \mid \mathbf{x}[m] : \theta_{Y|\mathbf{X}}) \\
&= \prod_m \theta_{y[m] \mid l(\mathbf{x}[m])} \\
&= \prod_{l \in Leaves(T)} \left[\prod_{y \in Val(Y)} \theta_{y|l}^{\,M[c_l, y]} \right]
\end{aligned}
$$

- For each value l∈Leaves(T) we get an independent multinomial problem where the MLE is

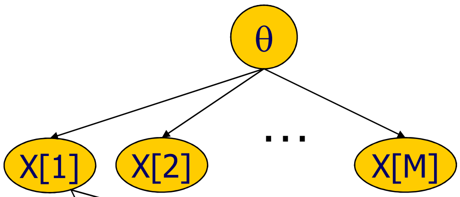
$$
\theta_{y^i|l} = \frac{M[c_l, y^i]}{M[c_l]} \qquad M[c_l] = \sum_{x:l(x)=l} M[\mathbf{x}, y^i]
$$

Limitations of MLE

- A thumbtack is tossed 10 times, and comes out 'head' 3 of the 10 tosses → Probability of head = 0.3
- A coin is tossed 10 times, and comes out 'head' 3 of the 10 tosses → Probability of head = 0.3
- A coin is tossed 1,000,000 times, and comes out 'head' 300,000 of the 1,000,000 tosses → Probability of head = 0.3
- Would you place the same bet on the next thumbtack toss as you would on the next coin toss?
- We need to incorporate prior knowledge
  - Prior knowledge should only be used as a guide

Bayesian Inference

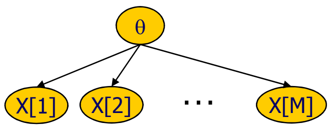
- Assumptions
  - Given a fixed θ tosses are independent
  - If θ is unknown tosses are not marginally independent – each toss tells us something about θ
- The following network captures our assumptions



$$P(x[m] \mid \theta) = \begin{cases} \theta & x[m] = x^1 \\ 1-\theta & x[m] = x^0 \end{cases}$$

Bayesian Inference

- Joint probabilistic model



$$\begin{aligned} P(x[1],\ldots,x[M],\theta) &= P(x[1],\ldots,x[M] \mid \theta)P(\theta) \\ &= P(\theta)\prod_{i=1}^{M} P(x[i] \mid \theta) \\ &= P(\theta)\theta^{M_H}(1-\theta)^{M_T} \end{aligned}$$

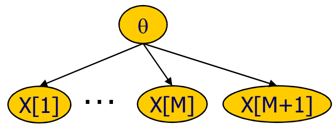
- Posterior probability over θ

$$P(\theta \mid x[1],\ldots,x[M]) = \frac{\overbrace{P(x[1],\ldots,x[M] \mid \theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(x[1],\ldots,x[M])}_{\text{Normalizing factor}}}$$

For a uniform prior, posterior is the normalized likelihood

Bayesian Prediction

- Predict the data instance from the previous ones



$$P(x[M+1] \mid x[1],\ldots,x[M])$$
$$= \int_{\theta} P(x[M+1],\theta \mid x[1],\ldots,x[M])d\theta$$
$$= \int_{\theta} P(x[M+1] \mid x[1],\ldots,x[M],\theta)P(\theta \mid x[1],\ldots,x[M])d\theta$$
$$= \int_{\theta} P(x[M+1] \mid \theta)P(\theta \mid x[1],\ldots,x[M])d\theta$$

- Solve for uniform prior $P(\theta)=1$ (for $0 \le \theta \le 1$) and binomial variable

$$P(x[M+1]=x^1 \mid x[1],\ldots,x[M]) = \frac{1}{P(x[1],\ldots,x[M])}\int_{\theta} \theta \cdot \theta^{M_H} \cdot (1-\theta)^{M_T}$$

"Bayesian estimate" $\rightarrow$ $= \dfrac{M_H+1}{M_H+M_T+2}$ $\leftarrow$ "Imaginary counts"

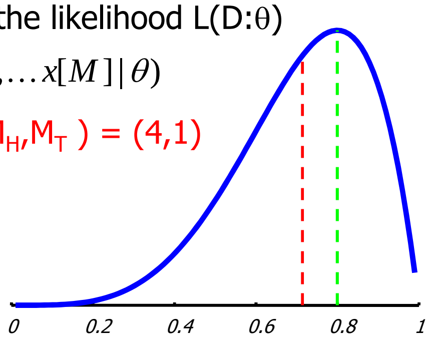
Example: Binomial Data

- Prior: uniform for θ in [0,1]
  - $P(\theta) = 1$

  → $P(\theta \mid D)$ is proportional to the likelihood $L(D:\theta)$

  $$P(\theta \mid x[1],\ldots x[M]) \propto P(x[1],\ldots x[M] \mid \theta)$$

$(M_H, M_T) = (4,1)$



- MLE for $P(X=H)$ is $4/5 = 0.8$
- Bayesian prediction is $5/7 = 0.71$

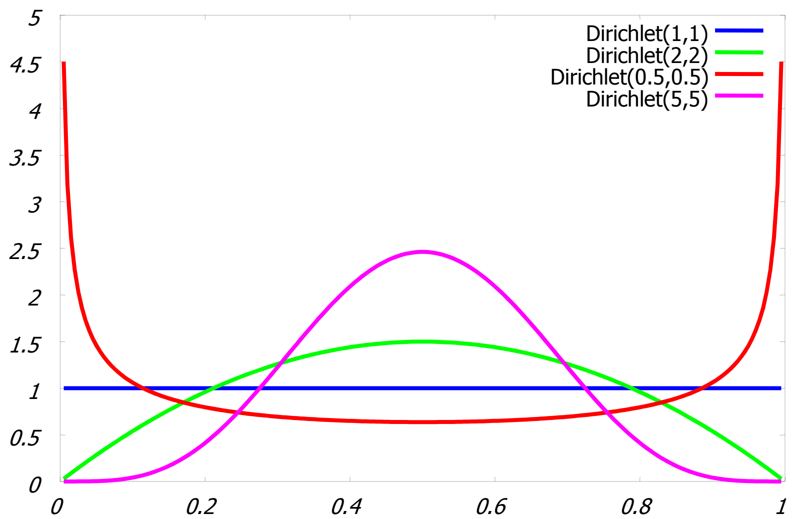
$$P(x[M+1]=H \mid D) = \int \theta \cdot P(\theta \mid D)d\theta = \frac{5}{7} = 0.7142\ldots$$

Dirichlet Priors

- A Dirichlet prior is specified by a set of (non-negative) hyper-parameters $\alpha_1,\ldots\alpha_k$ so that $\theta=[\theta_1,\ldots,\theta_k] \sim \text{Dirichlet}(\alpha_1,\ldots\alpha_k)$ if
  - $p(\theta) = \frac{1}{Z}\prod_k \theta_k^{\alpha_k-1}$ where $\sum_k \theta_k = 1$, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ and $Z = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$.
  - Intuitively, hyper-parameters correspond to the number of imaginary counts before starting the coin toss experiment

19

Dirichlet Priors – Example



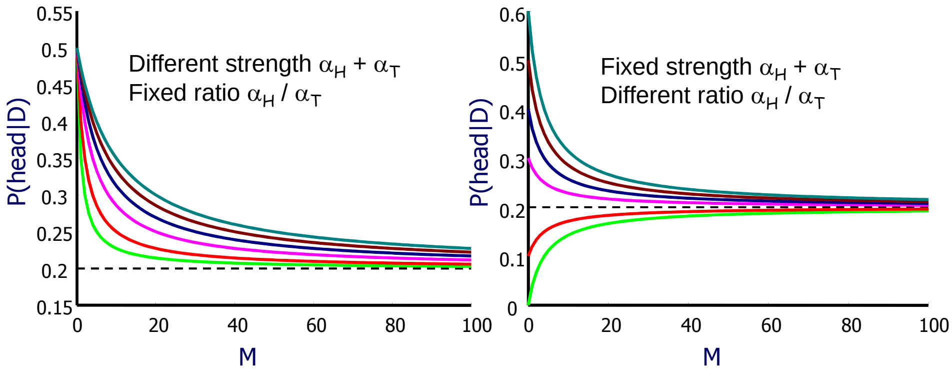
20

Dirichlet Priors

- Dirichlet priors have the property that the posterior is also Dirichlet
  - Prior is Dir($\alpha_1,\ldots\alpha_k$) $p(\theta)=\frac{1}{Z}\prod_k \theta_k^{\alpha_k-1}$
  - Data counts are $M_1,\ldots,M_k$
  - Posterior is Dir($\alpha_1+M_1,\ldots\alpha_k+M_k$) $p(\theta\mid D)=\frac{1}{Z'}\prod_k \theta_k^{\alpha_k+M_k-1}$
- The hyperparameters $\alpha_1,\ldots,\alpha_K$ can be thought of as "imaginary" counts from our prior experience
- Equivalent sample size = $\alpha_1+\ldots+\alpha_K$
  - The larger the equivalent sample size the more confident we are in our prior

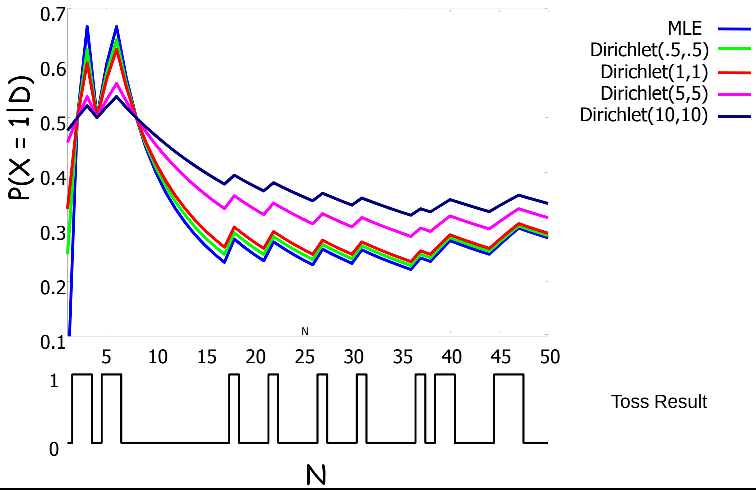
Effect of Priors

- Prediction of P(X=H) after seeing data with M_H=0.2M, M_T=0.8M as a function of the sample size



Effect of Priors (cont.)

- In real data, Bayesian estimates are less sensitive to noise in the data



General Formulation

- Joint distribution over D,θ

$$P(D,\theta) = P(D|\theta)P(\theta)$$

- Posterior distribution over parameters

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- P(D) is the marginal likelihood of the data

$$P(D) = \int_\theta P(D|\theta)P(\theta)d\theta$$

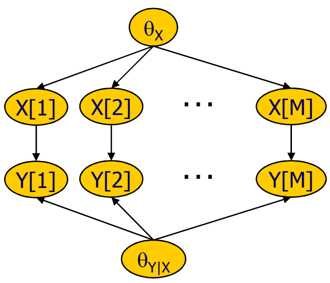
- As we saw, likelihood can be described compactly using sufficient statistics
- We want conditions in which posterior is also compact
  - E.g. Dirichlet priors

Conjugate Families

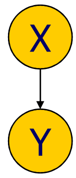
- A family of priors $P(\theta:\alpha)$ is conjugate to a model $P(\xi|\theta)$ if for any possible dataset D of i.i.d samples from $P(\xi|\theta)$ and choice of hyperparameters α for the prior over θ, there are hyperparameters α' that describe the posterior, i.e.,
  $P(\theta:\alpha') \propto P(D|\theta)P(\theta:\alpha)$
  - Posterior has the same parametric form as the prior
  - Dirichlet prior is a conjugate family for the multinomial likelihood
- Conjugate families are useful since:
  - Many distributions can be represented with hyperparameters
  - They allow for sequential update within the same representation
  - In many cases we have closed-form solutions for prediction

Bayesian Estimation in BayesNets

Bayesian network for parameter estimation



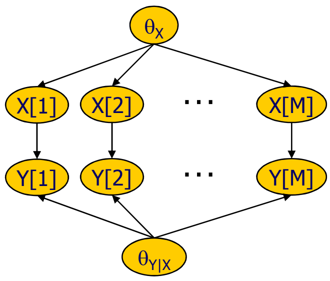
Bayesian network



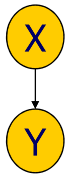
- Instances are independent given the parameters
  - (x[m'],y[m']) are d-separated from (x[m],y[m]) given θ
- Priors for individual variables are a priori independent
  - Global independence of parameters $P(\theta)=\prod_i P(\theta_{X_i|Pa(X_i)})$

Bayesian Estimation in BayesNets

Bayesian network for parameter estimation



Bayesian network

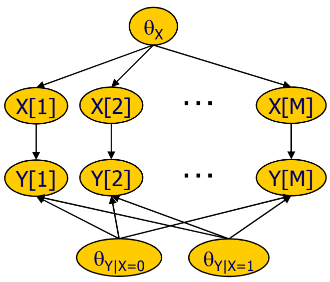


- Posteriors of θ are independent given complete data
  - Complete data d-separates parameters for different CPDs
  - $P(\theta_X, \theta_{Y|X} \mid D) = P(\theta_X \mid D) P(\theta_{Y|X} \mid D)$
  - As in MLE, we can solve each estimation problem separately

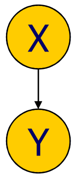
27

Bayesian Estimation in BayesNets

Bayesian network for parameter estimation



Bayesian network

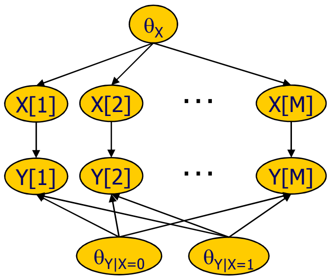


- Posteriors of θ are independent given complete data
  - Also holds for parameters within families
  - Note context specific independence between $\theta_{Y|X=0}$ and $\theta_{Y|X=1}$ when given both X and Y

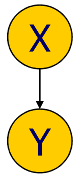
28

Bayesian Estimation in BayesNets

Bayesian network for parameter estimation



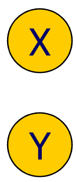
Bayesian network



- Posteriors of θ can be computed independently
  - For multinomial $\theta_{X_i|pa_i}$ posterior is Dirichlet with parameters $(\alpha_{X_i=1|pa_i}+M[X_i=1|pa_i]),\ldots,(\alpha_{X_i=k|pa_i}+M[X_i=k|pa_i])$
  - $P(X_i[M+1]=x_i \mid Pa_i[M+1]=pa_i, D) = \dfrac{\alpha_{x_i|pa_i} + M[x_i, pa_i]}{\sum_i \alpha_{x_i|pa_i} + M[x_i, pa_i]}$

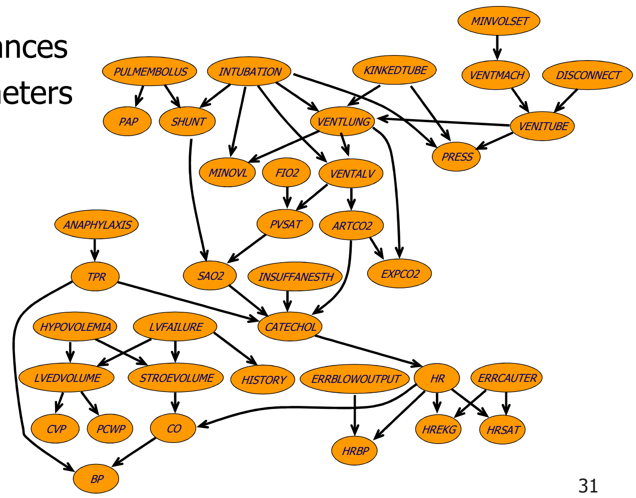
Assessing Priors for BayesNets

- We need the $\alpha(x_i,pa_i)$ for each node x_i
- We can use initial parameters Θ_0 as prior information
  - Need also an equivalent sample size parameter M'
  - Then, we let $\alpha(x_i,pa_i) = M' \cdot P(x_i,pa_i|\Theta_0)$
- This allows to update a network using new data
  - Example network for priors
    - $P(X=0)=P(X=1)=0.5$
    - $P(Y=0)=P(Y=1)=0.5$
    - $M'=1$
    - Note: $\alpha(x_0)=0.5$ $\alpha(x_0,y_0)=0.25$



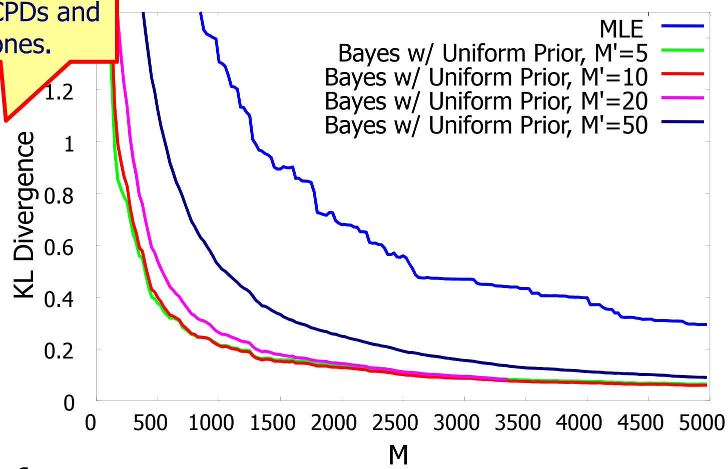
Case Study: ICU Alarm Network

- The "Alarm" network
  - 37 variables
- Experiment
  - Sample instances
  - Learn parameters
    - MLE
    - Bayesian



Case Study: ICU Alarm Network

The distance between the original CPDs and the learned ones.



- MLE performs worst
- Prior M'=5 provides best smoothing

Parameter Estimation Summary

- Estimation relies on sufficient statistics
  - For multinomials these are of the form $M[x_i, pa_i]$
  - Parameter estimation

$$\hat{\theta}_{x_i|pa_i} = \frac{M[x_i, pa_i]}{M[pa_i]} \qquad P(x_i \mid pa_i, D) = \frac{\alpha_{x_i, pa_i} + M[x_i, pa_i]}{\alpha_{pa_i} + M[pa_i]}$$

MLE — Bayesian (Dirichlet)

- Bayesian methods also require choice of priors
- MLE and Bayesian are asymptotically equivalent
- Both can be implemented in an online manner by accumulating sufficient statistics

Acknowledgement

- These lecture notes were generated based on the slides from Prof Eran Segal.