Georgia Institute of Technology

CSIP Center for Signal & Image Processing

ECE 8803

Parameter Learning in Graphical Models

Module 9: Part B

Bayesian Estimation

1. Known Structure &
2. Fully Observed Variables

Faramarz Fekri

Center for Signal and Information Processing

Overview

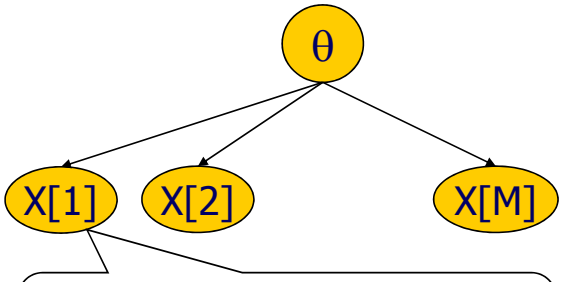
- Bayesian Inference
  - Priors (Beta, Dirichlet)
  - Continues Distributions
    - Single node
      - Gaussian case
    - Two-Nodes (conditional density)
      - Bayesian Estimation of conditional Gaussian
  - Bayesian Estimation in Bayesian Networks

Bayesian Inference

- Assumptions:
  - Parameter(s) θ are random variables
  - Given a fixed θ, the outcomes are independent (see Bayesian graph in below)
  - If θ is unknown variables X[i] are not marginally independent
    - Each X[i] tells us something about θ

$$P(x[1],\ldots,x[M],\theta) = P(x[1],\ldots,x[M]\,|\,\theta)P(\theta)$$
$$= P(\theta)\prod_{i=1}^{M} P(x[i]\,|\,\theta)$$
$$= P(\theta)\theta^{M_H}(1-\theta)^{M_T}$$

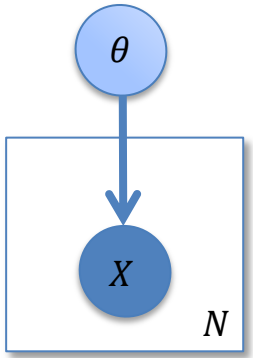
e.g., Coin toss



e.g., Coin: $P(x[m]\,|\,\theta) = \begin{cases} \theta & x[m] = x^1 \\ 1-\theta & x[m] = x^0 \end{cases}$

Bayesian Parameter Estimation

- Bayesian treat the unknown parameters as a random variable, whose distribution can be inferred using Bayes rule:
  - $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$



- The crucial equation can be written in words
  - $Posterior = \frac{likelihood \times prior}{marginal\ likelihood}$

Likelihood Prior

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

Normalizing factor

e.g., Coin toss

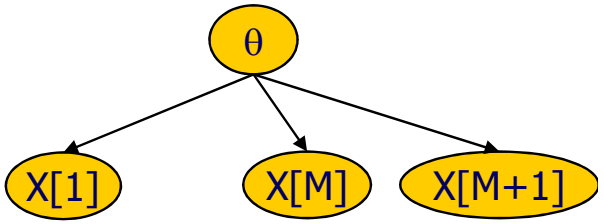
likelihood is $P(D|\theta) = \prod_{i=1}^{N} P(x_i|\theta) = \theta^{M_H}(1-\theta)^{M_T}$

- The prior $P(\theta)$ encodes our prior knowledge on the domain.
  - Different prior $P(\theta)$ results in different estimate for $P(\theta|D)$.
  - 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(θ)=1 (for 0≤θ≤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" →

$$= \frac{M_H+1}{M_H+M_T+2}$$

← *"Imaginary counts"*

-Lee

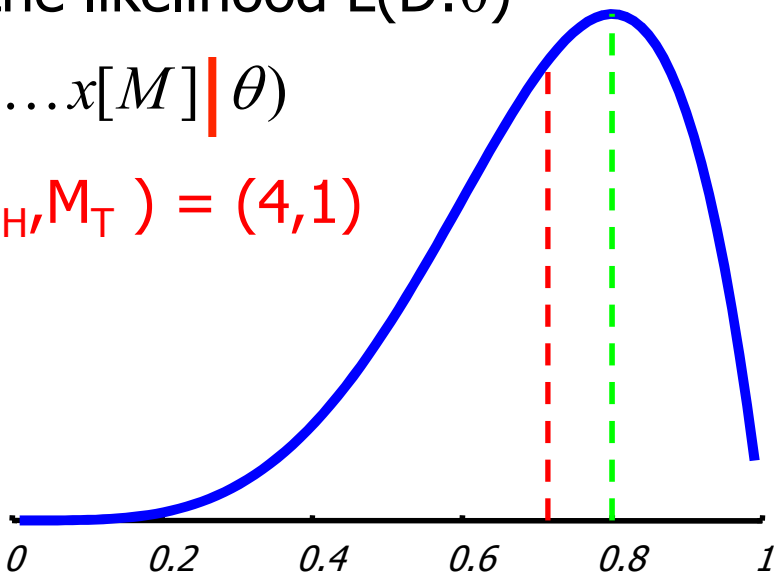
Example: Binomial Data

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

→ P(θ |D) is proportional to the likelihood L(D:θ)

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

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

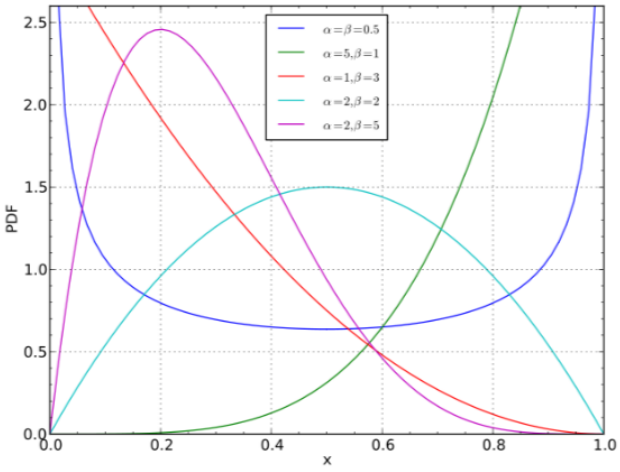


- MLE for P(X=H) is 4/5 = 0.8
- Bayesian prediction is 5/7 = 0.71

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

Bayesian Estimation: Beta Prior

- Prior over θ, Beta distribution
  - $P(\theta;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(a)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
  - When x is discrete $\Gamma(x+1) = x\Gamma(x) = x!$



- Posterior distribution of θ
  - $P(\theta|x_1,\dots,x_N) = \frac{P(x_1,\dots,x_N|\theta)P(\theta)}{P(x_1,\dots,x_N)} \propto$

e.g., Coin

$\theta^{n_h}(1-\theta)^{n_t}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$

  - Posterior is the same type of function as the prior
  - Such a prior is called a conjugate prior
  - α and β are hyperparameters and correspond to the number of "virtual" heads and tails (pseudo counts)

Bayesian Estimation for Bernoulli

- Posterior distribution θ
  - $P(\theta|x_1, \dots, x_N) = \frac{P(x_1,\dots,x_N|\theta)P(\theta)}{P(x_1,\dots,x_N)} \propto$
    $\theta^{n_h}(1-\theta)^{n_t}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}$

- Maximum a posteriori (MAP) estimation:
  - $\theta_{MAP} = argmax_\theta \log P(\theta|x_1, \dots, x_N)$

- Posterior mean estimation:
  - $\theta_{bayes} = \int \theta\, P(\theta|D)d\theta = C\int \theta \times \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1}d\theta = \frac{(n_h+\alpha)}{N+\alpha+\beta}$

- Prior strength: $A = \alpha + \beta$
  - A can be interpreted as an imaginary dataset

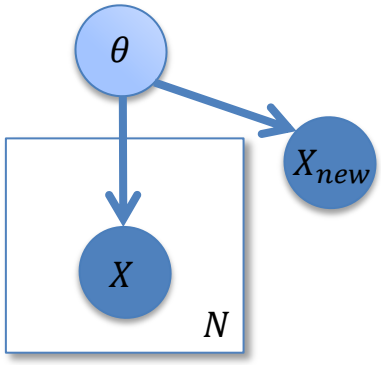
Effect of Prior Strength

- Suppose we have a uniform prior ($\alpha = \beta$), and we observed that $n_h = 2$, and $n_t = 8$
- Weak prior $A = \alpha + \beta = 2$. Posterior prediction:
  - $P(x = h | n_h = 2, n_t = 8, \alpha = 1, \beta = 1) = \frac{2+1}{10+2} = 0.25$
- Strong prior $A = \alpha + \beta = 20$. Posterior prediction:
  - $P(x = h | n_h = 2, n_t = 8, \alpha = 10, \beta = 10) = \frac{2+10}{10+20} = 0.4$
- However if we have enough data, it washes away the prior.
  - E.g. $n_h = 200$, and $n_t = 800$. Then the estimate under weak and strong prior are $\frac{200+1}{1000+2}, \frac{200+10}{1000+10}$ respectively. Both close to 0.2

How estimators should be used?

- θ_{MAP} is not Bayesian (although MAP uses a prior) because it is a point estimate
- Consider predicting the future:
  - A sensible way is to combine predictions based on all possible value of θ, weighted by their posterior probability, this is called Bayesian prediction:

$$P(x_{new}|D) = \int P(x_{new}, \theta|D)d\theta$$
$$= \int P(x_{new}|\theta, D)P(\theta|D)d\theta$$
$$= \int P(x_{new}|\theta)P(\theta|D)d\theta$$



- A frequentist prediction will typically use a "plug-in" estimator such as ML/MAP

$$P(x_{new}|D) = P(x_{new}|\,\theta_{ML}) \; or \; P(x_{new}|D) = P(x_{new}|\,\theta_{MAP})$$

Dirichlet Priors

- Dirichlet prior is specified by a set of (non-negative) hyper-parameters $\alpha_1,\ldots\alpha_k$ such that

$$\theta=[\theta_1,\ldots,\theta_k] \sim \text{Dirichlet}(\alpha_1,\ldots\alpha_k) \text{ if}$$

$$p(\theta)=\frac{1}{Z}\prod_k \theta_k^{\alpha_k-1} \quad \text{where} \quad \sum_k \theta_k = 1 , \quad \Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$$

$$\text{and} \quad Z=\frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)} .$$

- Equivalent sample size = $\alpha_1+\ldots+\alpha_K$
  The larger the equivalent sample size, the more confident the prior.

- We may interpret hyper-parameters as the number of imaginary counts before starting the experiment.
- Dirichlet priors have the property that the posterior is also Dirichlet:
  - Data counts $M_1,\ldots,M_k$

⇨ Posterior is Dir($\alpha_1+M_1,\ldots\alpha_k+M_k$) $\quad p(\theta \mid D)=\frac{1}{Z'}\prod_k \theta_k^{\alpha_k+M_k-1}$

General Formulation

- Joint distribution of D, θ:

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

- Posterior distribution of parameters:

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

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

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

- Recall that likelihood can be described compactly using sufficient statistics
- We want conditions by which posterior is also compact
  - For example: 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. That is:

$$P(\theta:\alpha') \propto \underbrace{P(D|\theta)P(\theta:\alpha)}_{\text{posterior}}$$

  - Posterior has the same parametric form as the prior
  - Dirichlet prior is a conjugate family for the multinomial likelihood

- Conjugate families are useful:
  - Many distributions can be represented with hyperparameters
  - They allow for sequential update within the same representation
  - Closed-form solutions for prediction exist in many cases.

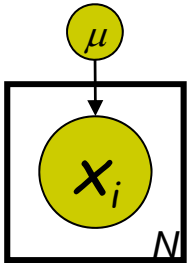
Bayesian Estimation in Gaussian Distributions

- Bayesian estimation for Gaussian:
  - Known σ, unknown μ
  - Known μ, unknown σ
  - Unknown μ and unknown σ

Bayesian estimation: unknown μ, known σ

- Conjugate Normal Prior on mean:

$$P(\mu) = \left(2\pi\tau^2\right)^{-1/2} \exp\left\{-(\mu-\mu_0)^2/2\tau^2\right\}$$



  - Where μ_0 and τ define a Gaussian prior on μ.
  - Then we have:

$$P(x,\mu) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}\left(2\pi\tau^2\right)^{-1/2} \exp\left\{-(\mu-\mu_0)^2/2\tau^2\right\}$$

$$P(\mu \mid x) = \left(2\pi\tilde{\sigma}^2\right)^{-1/2} \exp\left\{-(\mu-\tilde{\mu})^2/2\tilde{\sigma}^2\right\} \quad \text{posterior}$$

Here:

$$\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2+1/\tau^2}\bar{x} + \frac{1/\tau^2}{N/\sigma^2+1/\tau^2}\mu_0, \quad \text{and} \quad \tilde{\sigma}^2 = \left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

Sample average

Bayesian Estimation for Gaussian: Other cases

- Known μ, unknown $\lambda = 1/\sigma_2$
  - The conjugate prior for λ is a Gamma with shape a_0 and rate (inverse scale) b_0

$$p(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

  - The conjugate prior for σ^2 is Inverse-Gamma

$$IG(\sigma^2|a,b) = \frac{1}{\Gamma(a)} b^a (\sigma^2)^{-(a+1)} \exp(-b/(\sigma^2))$$

- Unknown μ and unknown σ_2
  - The conjugate prior is

    Normal-Inverse-Gamma

$$\begin{aligned} P(\mu,\sigma^2) &= P(\mu|\sigma^2)P(\sigma^2) \\ &= \mathcal{N}(\mu|m,\sigma^2 V)\ IG(\sigma^2|a,b) \end{aligned}$$

  - Semi conjugate prior

- Multivariate case:
  - The conjugate prior is

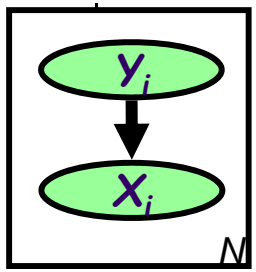
    Normal-Inverse-Wishart

$$\begin{aligned} P(\mu,\Sigma) &= P(\mu|\Sigma)P(\Sigma) \\ &= \mathcal{N}(\mu|\mu_0,\frac{1}{\kappa_0}\Sigma)\ \mathcal{IW}(\Sigma|\Lambda_0^{-1},\nu_0) \end{aligned}$$

-Xing

Bayesian Estimation for Conditional Gaussian (I)

- Generative Model: Need to specify a suitable form for class-conditional density p(x|y = c, θ), which defines what kind of data we expect to see in each class.
- $D=\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_N, y_N)\}$
- Y is a multi-class indicator whereas X is a conditional Gaussian variable with a class-specific mean:



$$p(y_n) = \text{multi}(y_n : \pi) = \prod_k \pi_k^{y_{n,k}}$$

$$p(x_n \mid y_{n,k} = 1, \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_n - \mu_k)^2 \right\}$$

$$p(x \mid y, \mu, \sigma) = \prod_n \left(\prod_k N(x_n : \mu_k, \sigma)^{y_{n,k}} \right)$$

- In Generative Model, we maximize joint log likelihood:

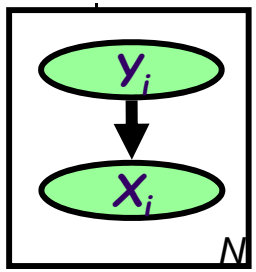
$$\ell(\theta; D) = \log \prod_n p(y_n \mid \pi) p(x_n \mid y_n, \mu, \sigma)$$

$$\hat{\pi}_{k,MLE} = \sum_n y_{n,k} \Big/ N = n_k / N$$

$$\hat{\mu}_{k,MLE} = \frac{\sum_n y_{n,k} x_n}{\sum_n y_{n,k}} = \frac{\sum_n y_{n,k} x_n}{n_k}$$

Average of samples in class k

Bayesian Estimation for Conditional Gaussian (II)



- **Bayesian estimation of conditional Gaussian:**
- **Prior**

$$P(\vec{\pi} \mid \vec{\alpha}) = \mathrm{Dir}(\vec{\pi} : \vec{\alpha})$$

$$P(\mu_k \mid \nu) = \mathrm{Normal}(\mu_k : \nu, \tau)$$

- Resulting posterior estimate mean:

$$\pi_{k,Bayes} = \frac{N}{N+|\alpha|}\hat{\pi}_{k,ML} + \frac{|\alpha|}{N+|\alpha|}\frac{\alpha_k}{|\alpha|} = \frac{n_k+\alpha_k}{N+|\alpha|}$$

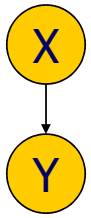
$$\mu_{k,Bayes} = \frac{n_k/\sigma^2}{n_k/\sigma^2+1/\tau^2}\hat{\mu}_{k,ML} + \frac{1/\tau^2}{n_k/\sigma^2+1/\tau^2}\nu$$

$$\sigma^2_{Bayes} = \left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

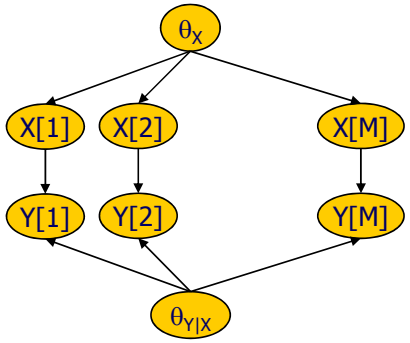
- Unlike generative models, when fitting a discriminative model, we maximize the conditional log likelihood. (See Section 8.6 of Murphy)

Bayesian Estimation in Bayesian Networks

Bayesian network



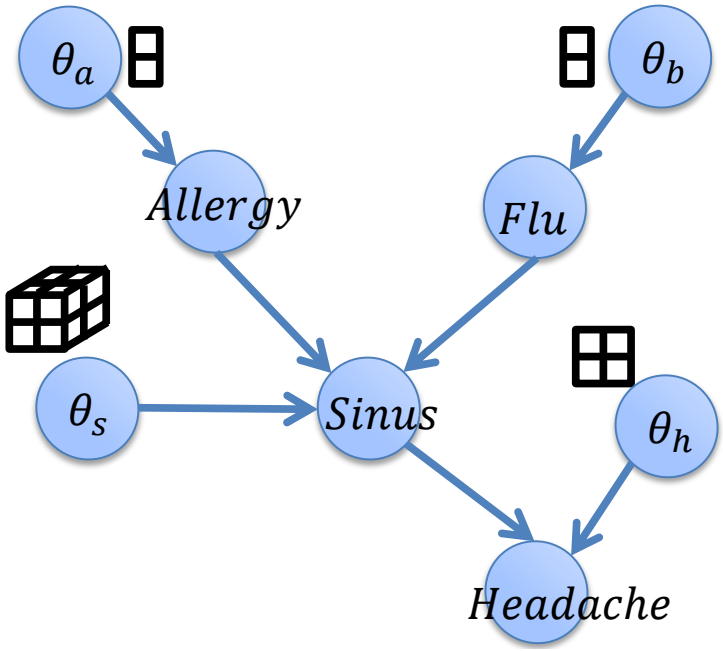
Bayesian network for parameter estimation



- 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 $P(\theta) = \prod_i P(\theta_{X_i|Pa(X_i)})$
- 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

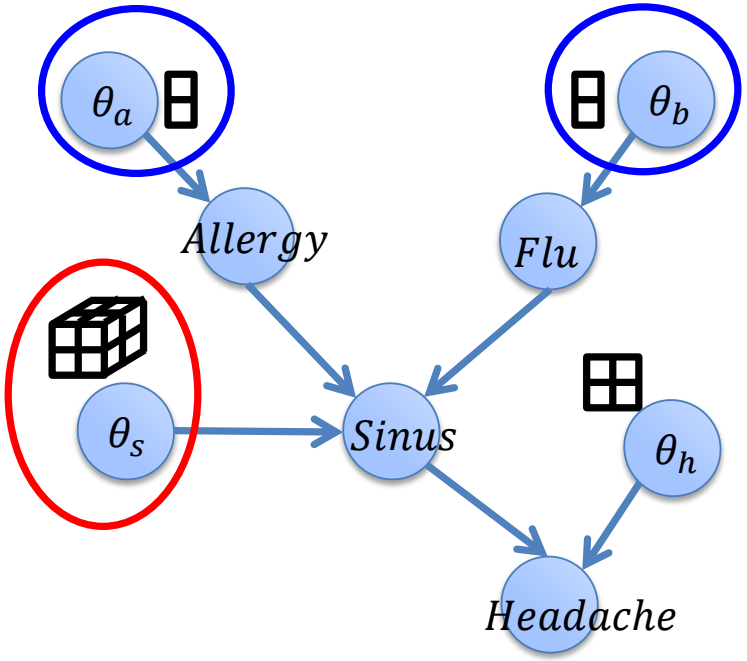
Bayesian Estimator for Tabular CPTs

- Factorization $P(X = x) = \prod_i P(x_i \,|pa_{X_i}, \theta_i)$
- Local CPT: multinomial distribution $P\left(X_i = k \middle| Pa_{X_i} = j\right) = \theta_{kj}$
- Put prior distribution over parameters $P(\theta_a, \theta_b, \theta_s\,, \theta_h)$



Global and Local Parameter Independence

- Global parameter independence
  - Parameters for all nodes in a GM
  - $P(\theta_a, \theta_b, \theta_s, \theta_h) = P(\theta_a)P(\theta_b)P(\theta_s)P(\theta_h)$



- Local Parameter Independence
  - Parameters in each node
  - $P(X_i = k | Pa_{X_i} = j) = \theta_{kj}$
  - $P(\theta_a) = \prod_{kj} P(\theta_{kj})$

Given data D

- For multinomial $\theta_{X_i|pa_i}$ posterior is Dirichlet with parameters $(\alpha_{X_i=1|pa_i}+M[X_i=1|pa_i]),\ldots, (\alpha_{X_i=k|pa_i}+M[X_i=k|pa_i])$
- $$P(X_i[M+1]=x_i \mid Pa_i[M+1]=pa_i, D) = \frac{\alpha_{x_i|pa_i} + M[x_i, pa_i]}{\sum \alpha_{x_i|pa_i} + M[\quad pa_i]}$$

Parameter Estimation Summary

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

$$\hat{\theta}_{x_i|pa_i} = \frac{M[x_i, pa_i]}{M[pa_i]}$$

MLE

$$P(x_i \mid pa_i, D) = \frac{\alpha_{x_i, pa_i} + M[x_i, pa_i]}{\alpha_{pa_i} + M[pa_i]}$$

Bayesian (Dirichlet)

- Bayesian methods also require choice of priors

Frequentist vs. Bayesian

- Pros of Bayesian approach:
  - Mathematically sophisticated
  - Works well when amount of data is less relative to the number of parameters
  - Easy for incremental (sequential) learning
  - Can be used for model selection (max likelihood will always pick the most complex model)
- Pros of frequentist approach:
  - Mathematically/computationally simpler
  - "objective", unbiased, invariant to reparametrization
- As data becomes large ($D \rightarrow \infty$), the two methods become identical