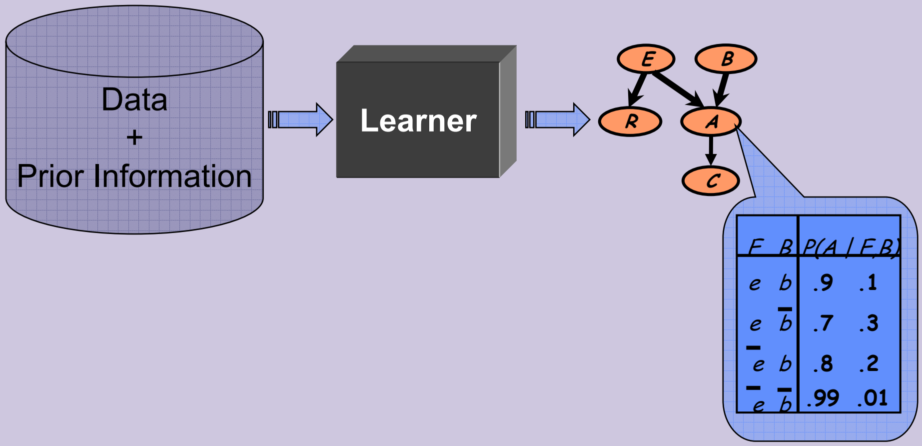
Bayesian Networks Learning – Parameter Estimation

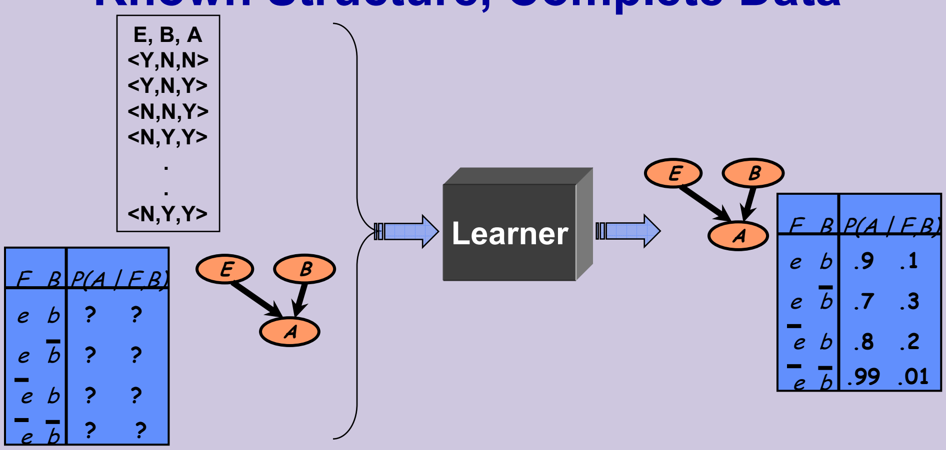
Learning Bayesian networks

- Assume the domain $X=\{X1, \ldots, Xn\}$ is governed by some underlying distribution $P^*(X)$
- P^* is induced by some Bayesian network $B^*=(G,\Theta)$
- Given a data set $D=\{x[1], \ldots, x[M]\}$ of M samples from P^*
- Samples are *i.i.d. - independent and identically distributed*
- The task is to recover the Bayesian network model

Learning Bayesian networks

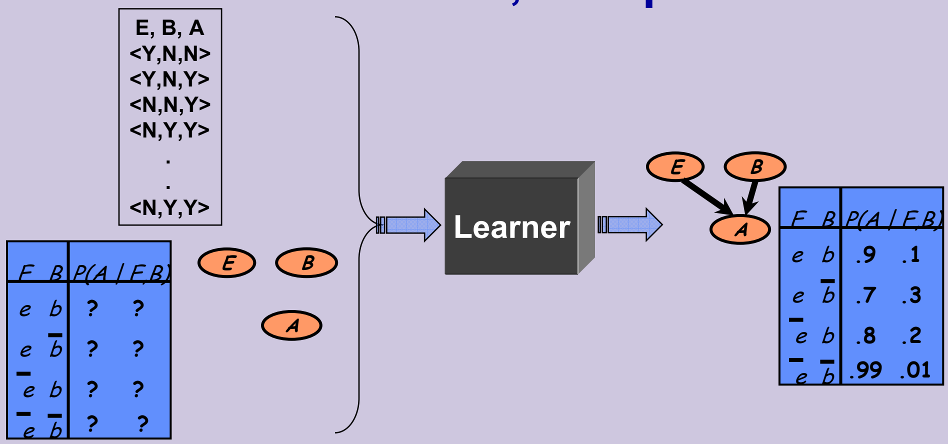


Known Structure, Complete Data



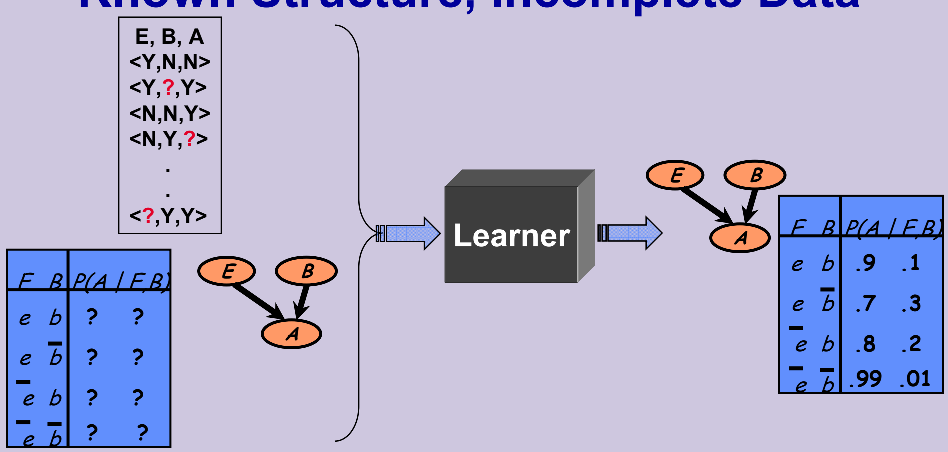
- Network structure is specified
  - Inducer needs to estimate parameters
- Data does not contain missing values

Unknown Structure, Complete Data



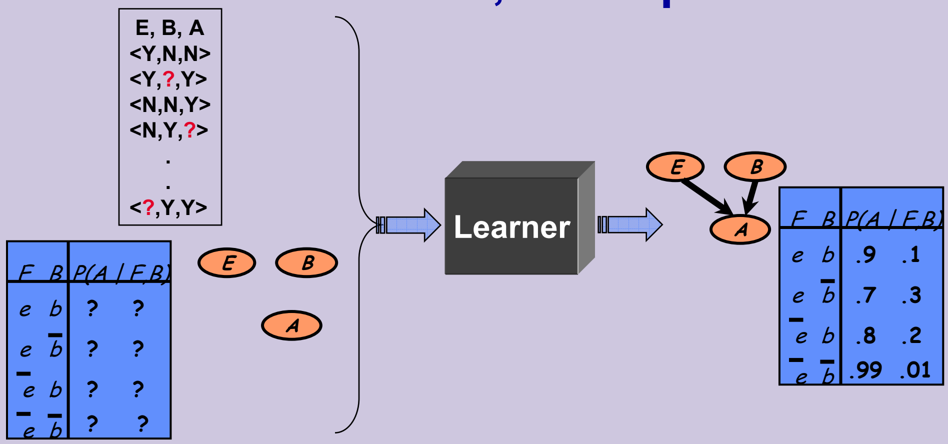
- Network structure is not specified
  - Inducer needs to select arcs & estimate parameters
- Data does not contain missing values

Known Structure, Incomplete Data



- Network structure is specified
- Data contains missing values
  - Need to consider assignments to missing values

Unknown Structure, Incomplete Data



- Network structure is not specified
- Data contains missing values, or hidden variables?
  - Need to consider assignments to missing values

Parameter Estimation

- Network structure is specified
- Data set D consists of fully observed instances of the network variables

  $D=\{x[1], \ldots, x[M]\}$

- Estimate network parameters

Parameter Estimation

- Use a set *D={x[1], …, x[M]}* of training samples drawn independently from a parametric model *P(x : θ)* to estimate the unknown parameter vector θ

- Parameter estimation: a classic problem in statistics
  - Maximum-Likelihood (ML) estimation
  - Bayesian estimation

Maximum-Likelihood Estimation

- IID data samples *D={x[1], …, x[M]}*
- *Likelihood function*

$$L(\theta : D) = P(D \mid \theta) = \prod_{k=1}^{M} P(x[k] : \theta)$$

the likelihood of θ w.r.t. the set of samples

- ML estimate of θ is, by definition, the value $\hat{\theta}$ that maximizes *L(θ:D)*
- it is usually easier to work with the the *log-likelihood function*

$$l(\theta : D) = \log P(D \mid \theta) = \sum_{k=1}^{M} \log P(x[k] : \theta)$$

Example: Discrete Case

- Single binary variable

  P(X=1) = θ,   P(X=0) = 1- θ

$$P(x \mid \theta) = \theta^{x}(1-\theta)^{1-x}$$

Bernoulli distribution

$$L(\theta : D) = \prod_{k=1}^{M} P(x[k] : \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

- *Sufficient statistics*:

  N_1: number of 1's in D, N_0: number of 0's in D

- Log-likelihood

  $l(\theta:D) = N_1 \ln \theta + N_0 \ln (1 - \theta)$

- ML estimation

$$\hat{\theta}_{ML} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{M}$$

◆ Multi-valued discrete random variables {1, …, K}

$$\theta_i = P(X = i)$$

$$P(x \mid \boldsymbol{\theta}) = \prod_{i=1}^{K} \theta_i^{\delta_{xi}}$$

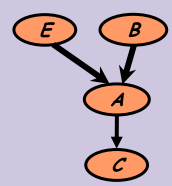
$$L(\boldsymbol{\theta} : D) = \prod_{j=1}^{M} P(x[j] : \theta) = \prod_{i=1}^{K} \theta_i^{N_i}$$

- Sufficient statistics N_i: the # of times i appears in D

$$\hat{\theta}_{iML} = \frac{N_i}{\sum_j N_j} = \frac{N_i}{M}$$

MLE for Bayesian Networks

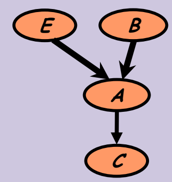
◆ Training data has the form:



$$D = \begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

Likelihood Function

- By definition of network, we get

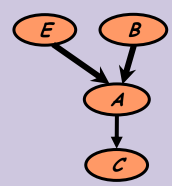


$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \prod_m \begin{pmatrix} P(E[m] : \Theta) \\ P(B[m] : \Theta) \\ P(A[m] \mid B[m], E[m] : \Theta) \\ P(C[m] \mid A[m] : \Theta) \end{pmatrix}$$

Likelihood Function

- Rewriting terms, we get

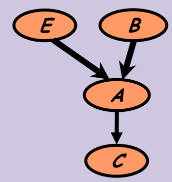


$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \begin{array}{l} \prod_m P(E[m] : \Theta) \\ \prod_m P(B[m] : \Theta) \\ \prod_m P(A[m] \mid B[m], E[m] : \Theta) \\ \prod_m P(C[m] \mid A[m] : \Theta) \end{array}$$

Likelihood Function

◆ Rewriting terms, we get



$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \begin{array}{l} \prod_m P(E[m] : \Theta_E) \\ \prod_m P(B[m] : \Theta_B) \\ \prod_m P(A[m] \mid B[m], E[m] : \Theta_{A|BE}) \\ \prod_m P(C[m] \mid A[m] : \Theta_{C|A}) \end{array}$$

General Bayesian Networks

Generalizing for any Bayesian network:

$$L(\Theta : D) = \prod_m P(x_1[m], \ldots, x_n[m] : \Theta)$$

$$= \prod_i \prod_m P(x_i[m] \mid Pa_i[m] : \Theta_i)$$

$$= \prod_i L_i(\Theta_i : D) \qquad \rightarrow \text{local likelihood function}$$

Global Decomposition of the likelihood function
⇒ Independent estimation problems

MLE

- Assuming discrete variables (CPTs) leads to further decomposition → *local decomposition of the likelihood function*

$$L_i(\Theta_i : D) = \prod_m P(x_i[m] \mid Pa_i[m] : \Theta_i)$$

$$= \prod_{pa_i} \prod_{m, Pa_i[m] = pa_i} P(x_i[m] \mid pa_i : \Theta_{X_i \mid pa_i})$$

$$= \prod_{pa_i} \prod_{x_i} \theta_{x_i \mid pa_i}^{N(x_i, pa_i)}$$

$$\hat{\theta}_{x_i \mid pa_i} = \frac{N(x_i, pa_i)}{N(pa_i)}$$

Bayesian Inference

- Represent uncertainty about parameters using a probability distribution over parameters
- Learning using Bayes rule

Likelihood Prior

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

Posterior Probability of data

Example: Discrete Variable

- Single binary variable

  P(X=1) = θ,   P(X=0) = 1- θ

$$L(\theta : D) = \theta^{N_1}(1-\theta)^{N_0}$$

- What prior *p(θ)* to use?

Beta distribution

$$\text{Beta}\ (\theta \mid \alpha_1, \alpha_0) = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

$$0 \le \theta \le 1$$

$$E(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

- The parameters α_1 and α_2 are positive reals, often called *hyperparameters*
- Gamma Function

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(1) = 1, \Gamma(x) = (x-1)! \quad \text{for interger} \quad x$$

Assume the prior is a Beta distribution

$$p(\theta) = Beta(\theta \mid \alpha_1, \alpha_0) = c\,\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

The posterior density p(θ |D)

$$p(\theta \mid \mathrm{D}) = c' \cdot p(\mathrm{D} \mid \theta)\, p(\theta)$$
$$= Beta(\theta \mid N_1 + \alpha_1, N_0 + \alpha_0)$$

- The property that the posterior distribution follows the same parametric form as the prior distribution is called conjugacy
- Beta prior is a conjugate family for the binomial distribution

$$P(X=1 \mid D) = \int P(X=1 \mid \theta)\, p(\theta \mid D)\, d\theta$$

$$= \int \theta p(\theta \mid D)\, d\theta = \frac{N_1 + \alpha_1}{N_1 + N_0 + \alpha_1 + \alpha_0} \equiv \hat{\theta}_{BE}$$

- It can be proved that:

  If the prior is well-behaved – i.e. does not assign 0 density to any *feasible* parameter value, then both MLE and Bayesian estimate converge to the same value in the limit
- Both *almost surely* converge to the underlying distribution *P*(X)
- But the ML and Bayesian approaches behave differently when the number of samples is small

- Multi-valued discrete random variables {1, …, k}

$$\theta_i = P(X = i)$$

$$P(D \mid \theta) = \prod_{i=1}^{k} \theta_i^{N_i}$$

Sufficient statistics N_i: the # of times i appears in D

- Assume the prior p(θ) is a Dirichlet distribution Dir(**θ**|α)

Dirichlet distribution with hyperparameters α_i's

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

$$0 \le \theta_i \le 1, \quad \sum_{i=1}^{k} \theta_i = 1, \quad \alpha = \sum_{i=1}^{k} \alpha_i$$

$$E(\theta_i) = \frac{\alpha_i}{\alpha}$$

- Multi-valued discrete random variables {1, …, k}
- Assume the prior p(θ) is a Dirichlet distribution Dir(θ|α) with *hyperparameters* α_i's
- Then the posterior density p(θ |D) is also a Dirichlet distribution with hyperparameters $\alpha_1+N_1,\dots,\alpha_k+N_k$

$$p(\boldsymbol{\theta} \mid D) = c \cdot P(\mathsf{D} \mid \theta)\, p(\boldsymbol{\theta})$$
$$= Dir(\boldsymbol{\theta} \mid \mathbf{N} + \boldsymbol{\alpha})$$

- Dirichlet prior is a conjugate family for the multinomial distribution

- Bayesian estimates

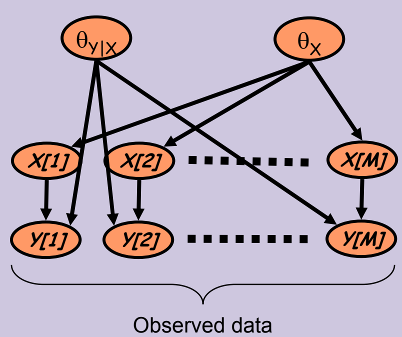
$$P(X = i \mid D) = \int P(X = i \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$
$$= \int \theta_i Dir(\boldsymbol{\theta} \mid \mathbf{N} + \boldsymbol{\alpha})\, d\boldsymbol{\theta} = \frac{N_i + \alpha_i}{M + \alpha} \equiv \hat{\theta}_{iBE}$$

- The hyperparameters α_i can be thought of as "*imaginary*" counts from our prior experience
- α: imaginary *equivalent sample size*
- Let p_i be prior belief about θ_i : $\alpha_i = \alpha\, p_i$
- The larger the equivalent sample size, the more confident we are in our prior
- *Laplace estimates:* α*=k,* α_i*= 1*

Summary of Bayesian estimation

- Treat the unknown parameters as random variables
- Assume a prior distribution for the unknown parameters
- Update the distribution of the parameters based on data
- Finally compute *p(x|D)*

Bayesian Estimation in BNs



- **Meta-network** for P(Θ,D)
- Priors for each parameter group are independent
- Data instances are independent given the unknown parameters

Bayesian Estimation in BNs

- *Global parameter independence* assumption

$$P(\Theta) = \prod_i P(\Theta_{X_i|Pa_i})$$

- Global Decomposition of the likelihood function

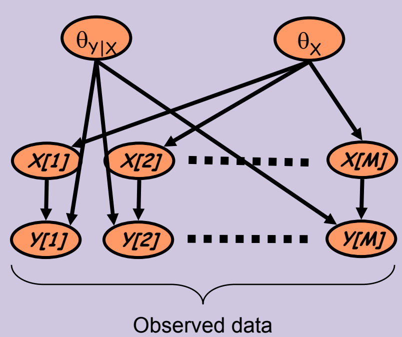
$$L(\Theta : D) = P(D \mid \Theta) = \prod_i L_i(\Theta_{X_i|Pa_i} : D)$$

- Posterior parameter independence

$$P(\Theta \mid D) = \prod_i P(\Theta_{X_i|Pa_i} \mid D)$$

- We can solve the prediction problem for each CPD independently

Bayesian Estimation in BNs

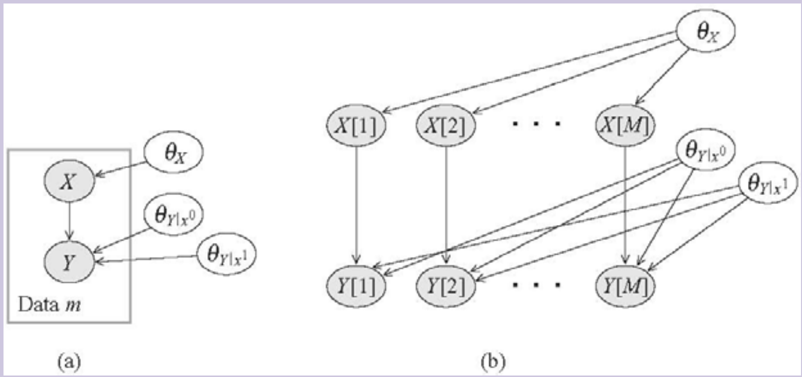


- This can be "read" from the network by d-separation

  Complete data ⇒
  posteriors on parameters are independent

- Can compute posterior over parameters separately!

Bayesian Estimation in BNs - CPTs



- Local parameter independence assumption

$$P(\Theta_{X_i|Pa_i}) = \prod_{pa_i} P(\Theta_{X_i|pa_i})$$

Bayesian Nets & Bayesian Prediction

- Posterior parameter independence

$$P(\Theta \mid D) = \prod_i \prod_{pa_i} P(\Theta_{X_i|pa_i} \mid D)$$

- Assume Dirichlet prior

$$P(\Theta_{X_i|pa_i}) = Dir(\Theta_{X_i|pa_i} \mid \alpha_{x_i|pa_i}, \ldots)$$

- Then

$$P(\Theta_{X_i|pa_i} \mid D) = Dir(\Theta_{X_i|pa_i} \mid \alpha_{x_i|pa_i} + N(x_i, pa_i), \ldots)$$

Bayesian Nets & Bayesian Prediction

- Bayesian estimation

$$\tilde{\theta}_{x_i|pa_i} = P(X_i = x_i \mid Pa_i = pa_i, D)$$
$$= \frac{\alpha_{x_i|pa_i} + N(x_i, pa_i)}{\alpha_{pa_i} + N(pa_i)}$$

Assessing Priors for Bayesian Nets

The BDe prior

- Introduce an equivalent sample size *α* and a prior distribution P'(X), and set

$$\alpha_{x_i|pa_i} = \alpha P'(x_i, pa_i)$$

- We can represent P' as a BN (G_0, Θ_0), and set

$$\alpha_{x_i|pa_i} = \alpha P(x_i, pa_i^G \mid G_0, \Theta_0)$$

  Use BN inference to compute this

- E.g., empty BN with uniform distribution

$$\alpha_{x_i|pa_i} = \frac{\alpha}{|X_i||Pa_i|}$$

Learning Parameters: Summary

- Estimation relies on **sufficient statistics**
  - For multinomials: counts $N(x_i, pa_i)$
  - Parameter estimation

$$\hat{\theta}_{x_i|pa_i} = \frac{N(x_i, pa_i)}{N(pa_i)}$$

MLE

$$\tilde{\theta}_{x_i|pa_i} = \frac{\alpha_{x_i|pa_i} + N(x_i, pa_i)}{\alpha_{pa_i} + N(pa_i)}$$

Bayesian (Dirichlet)

- Both are asymptotically equivalent and consistent
- Both can be implemented in an on-line manner by accumulating sufficient statistics