

## BDe for Gaussian Belief Networks

### Review of learning with complete data

Graph Selection

$$\begin{aligned} P\{G|D\} &\propto P\{D|G\}P\{G\} \\ &= \left( \int_{\Theta} P\{D|G, \Theta\}P\{\Theta|G\}d\Theta \right) P\{G\} \end{aligned}$$

Bayesian Prediction

$$P\{X[m+1] | X[1], \dots, X[m]\}$$

BDe prior:

Prior network plus a number of counts.

## Gaussian Belief Networks

Conditional distribution

$$p(x_i | x_1, \dots, x_{i-1}) = N\left(x; \mu_i + \sum_{j=1}^i b_{ij}(x_j - \mu_j), \frac{1}{v_j}\right)$$

$\mu_i$  is the unconditional mean of  $x_i$

$v_i$  is the conditional variance of  $x_i$  given values for  $x_1, \dots, x_{i-1}$

## Generating Multivariate Normal Density

Technique 1: Use moment form.

Technique 2: [Shachter and Kenley, 88]

Unconditional means have the same form.

Precision:  $W(1) = \frac{1}{v_1}$

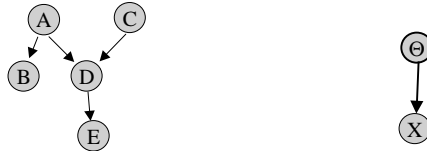
$$W(i+1) = \begin{pmatrix} W(i) + \frac{\vec{b}_{i+1} \vec{b}_{i+1}^T}{v_{i+1}} & -\frac{\vec{b}_{i+1}}{v_{i+1}} \\ -\frac{\vec{b}_{i+1}^T}{v_{i+1}} & \frac{1}{v_{i+1}} \end{pmatrix}$$

$\vec{b}_i$  is the column vector  $(b_{1,i}, \dots, b_{i-1,i})$

## Prior

$$P\{D | G\} = \int_{\Theta} P\{D | G, \Theta\} \underline{P\{\Theta | G\}} d\Theta$$

OK, we now have converted a gaussian belief net into a single multivariate gaussian distribution.



$$P(X | \Theta) = N(X; \mu, W^{-1}) = (2\pi)^{-\frac{n}{2}} |W|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X - \mu)^T W (X - \mu)^T\right)$$

The conjugate prior (\*\*\*) for a multi-variate normal distribution with unknown mean and unknown variance is a normal-Wishart distribution.

## Normal Wishart Prior

Prior over precision:

$$p(W | G) = c(n, \alpha_w) |\Gamma|^{\frac{\alpha_w}{2}} |W|^{\frac{\alpha_w - n - 1}{2}} e^{-\frac{1}{2}tr(\Gamma W)}$$

$$\text{where } c(n, \alpha_w) = \left[ 2^{\frac{\alpha_w n}{2}} \pi^{\frac{n(n-1)}{4}} \prod_{i=1}^n \Gamma\left(\frac{\alpha_w + 1 - i}{2}\right) \right]^{-1}$$

Prior over mean:

$$P(\mu | W, G) = N(\mu; \nu, (\alpha_\mu W)^{-1})$$



## Posterior Distribution

Update to the mean:

$$v' = \frac{\alpha_\mu v + M\bar{x}}{\alpha_\mu + M} \quad \bar{x} = \frac{1}{M} \sum_m x[m]$$

$$\alpha_\mu' = \alpha_\mu + M$$

Update to the precision:

$$T' = T + MS_M + \frac{\alpha_\mu M}{\alpha_\mu + M} (v - \bar{x})(v - \bar{x})^T$$

$$\alpha_w' = \alpha_w + M \quad S_M = \frac{1}{M} \sum_m (x[m] - \bar{x})(x[m] - \bar{x})^T$$

## Bayesian Prediction

Prediction:

$$P\{X[m+1] | X[1] \dots, X[m]\} = \int_{\Theta} P\{X[m+1] | \Theta\} P\{\Theta | X[1] \dots, X[m]\} d\Theta$$

For normal Wishart, prediction is a t-distribution with  $n$  dimensions and  $\alpha_w - n + 1$  degrees of freedom:

$$P\{X[m+1] | X[1] \dots, X[m]\} = T[X; \alpha_w - n + 1, T']$$

$$= \frac{\Gamma\left(\frac{\alpha_w + 1}{2}\right) T'^{\frac{1}{2}}}{\Gamma\left(\frac{\alpha_w - n + 1}{2}\right) (\alpha_w - n + 1)^n \pi^{\frac{n}{2}}} \left(1 + \frac{1}{(\alpha_w - n + 1)} (X[m] - \mu)^T T' (X[m] - \mu)\right)^{-\frac{\alpha_w + 1}{2}} \quad T'' = \frac{\alpha_\mu (\alpha_w - n + 1)}{\alpha_\mu + 1} T'^{-1}$$

ICK! Would like a gaussian MAP

## Bayesian Prediction (approx)

Match first two moments and use a gaussian.

$$P(X[m+1] | X[1], \dots, X[m]) \approx N(X[m+1]; \mu, W^{-1})$$

$$\mu = \nu$$

$$W^{-1} = \frac{\alpha_\mu + 1}{\alpha_\mu (\alpha_w - n - 1)} T$$

## Marginal Likelihood of Data

$$P\{D | G\} = (2\pi)^{-nM/2} \left( \frac{\alpha_\mu}{\alpha_\mu'} \right)^{n/2} \frac{c(n, \alpha_w)}{c(n, \alpha_w')} |T|^{n/2} |T'|^{n/2}$$

where

$$c(n, \alpha) = \left[ 2^{n/2} \pi^{n(n-1)/4} \prod_{i=1}^n \Gamma\left(\frac{\alpha+1-i}{2}\right) \right]^{-1}$$

## BDe Prior

Construct a gaussian belief network.

Assess an equivalent number of counts for estimate for variance and mean,  $\alpha_{\mu,0}, \alpha_{W,0}$  .

$$\nu = \mu$$

$$T^{-1} = \frac{\alpha_{\mu,0} (\alpha_{W,0} - n - 1)}{\alpha_{\mu,0} + 1} W$$

Approximate since the marginal over X is distributed as a T-distribution.

## Motivation for Normal-Wishart Prior

Given assumptions about the nature of the prior

Marginal independence of parameters,

Parameter and likelihood modularity,

Regularity assumptions, etc

Then the prior must be a normal-Wishart

[Geiger + Heckerman, Parameter Priors for Directed Acyclic Models and the Characterization of Several Probability Distributions, 1999] the prior must be a normal-Wishart

Similar set of assumptions to demonstrate Dirichlet prior for a multinomial distribution.

## Prior Assumptions

Complete = no missing arcs (no independence)

A1 (Regularity):

If two complete models represent the same distribution over  $X$ , there exists a one to one mapping between the parameters of the two models such that

$$P(x | \Theta_{G_1}, G_1) = P(x | \Theta_{G_2}, G_2)$$

A2 (Likelihood Modularity):

For every two DAG models  $m_1$  and  $m_2$  for  $X$  such that  $X_i$  has the same parents in  $m_1$  and  $m_2$ , the local distributions for  $x_i$  in both models are the same, namely

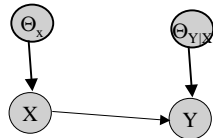
$$P\{x_i | pa_i, \theta_i, m_1\} = P\{x_i | pa_i, \theta_i, m_2\}$$

A3 (Prior Modularity):

For every two DAG models  $m_1$  and  $m_2$  for  $X$  such that  $X_i$  has the same parents in  $m_1$  and  $m_2$ ,  $P\{\theta_i | m_1\} = P\{\theta_i | m_2\}$

## Prior Assumptions

A4 (Global Parameter Equivalence):



Parameters for different conditional distributions are marginally independent.

## If Prior Assumptions, then...

[Geiger + Heckerman] Parameter Priors for Directed Acyclic Models and the Characterization of Several Probability Distributions, 1999 (to appear in *Annals of Statistics*)