Structure Learning 2

1

Structure Scores

- Searching for highest-scoring network structure is intractable
- Need to resort to heuristic search (eg. hillclimbing)
- Need:
  1. Search space
  2. Scoring function
  3. Search procedure

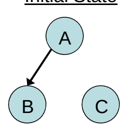
2

Structure Scores

1. Search space

- Start with initial state (eg. disconnected graph or randomly generated one)

Initial State



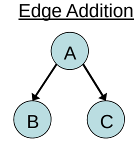
3

Structure Scores

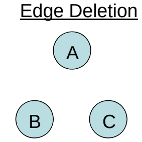
1. Search space

- Move to a neighboring state by applying an operator:

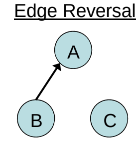
Edge Addition



Edge Deletion



Edge Reversal



Can only perform an operator if it doesn't lead to a cycle!

4

Structure Scores

2. Scoring function:

- Two general classes of scoring functions:
  1. Likelihood scoring functions
  2. Bayesian scoring functions
- More about this in a bit…assume we have a scoring function for now

5

Structure Scores

3. Search procedure

- Greedy search: pick the best scoring neighboring state to move to
- Repeat until convergence
- Converges to a local optimum

Tricks for dealing with this: random restart, simulated annealing, tabu search and data perturbation

6

Structure Scores

7

Likelihood Scores

$$\max_{\mathcal{G},\boldsymbol{\theta}_{\mathcal{G}}} L(\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle : \mathcal{D})$$
$$= \max_{\mathcal{G}} \left[\max_{\boldsymbol{\theta}_{\mathcal{G}}} L(\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle : \mathcal{D}) \right]$$
$$= \max_{\mathcal{G}} \left[L(\langle \mathcal{G}, \hat{\boldsymbol{\theta}}_{\mathcal{G}} \rangle : \mathcal{D}) \right]$$

Graph structure that maximizes the likelihood

Maximum likelihood estimates of parameters

$$score_L(\mathcal{G} : \mathcal{D}) = l(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})$$

Log likelihood

8

Likelihood Scores

Let M be the number of samples. We use the notation $M[x]$ to be the count of x in the data.

Let $\hat{P}$ be the empirical distribution observed in the data. Eg.

- $M[x,y] = M \cdot \hat{P}(x,y)$
- $M[y] = M \cdot \hat{P}(y)$

Note that:

- $\hat{\theta}_{y|x} = \hat{P}(y|x)$
- $\hat{\theta}_y = \hat{P}(y)$

9

Likelihood Scores

Mutual Information

$$\boldsymbol{I}_{\hat{P}}(X;Y) = \sum_{x,y} \hat{P}(x,y) log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$$

$$= \frac{1}{M} \sum_{x,y} M[x,y] \, log \left(\frac{M[x,y]}{M[x]M[y]} \right)$$

10

Likelihood Scores

Claim:

$$score_L(\mathcal{G}:\mathcal{D}) = M \sum_{i=1}^{n} \boldsymbol{I}_{\hat{P}}(X_i; Parents(X_i,\mathcal{G})) \; - M \sum_{i=1}^{n} H_{\hat{P}}(X_i)$$

$$= M \sum_{i=1}^{n} \left[\boldsymbol{I}_{\hat{P}}\big(X_i; Parents(X_i,\mathcal{G})\big) - H_{\hat{P}}(X_i) \right]$$

11

Likelihood Scores

Proof:

$$l\big(\hat{\theta}_{\mathcal{G}}:\mathcal{D}\big) = \sum_{i=1}^{n} \left[\sum_{\boldsymbol{u}_i \in Val(Parents(X_i,\mathcal{G}))} \sum_{x_i} M[x_i, \boldsymbol{u}_i] log \hat{\theta}_{x_i|\boldsymbol{u}_i} \right]$$

$$= M \sum_{i=1}^{n} \left[\frac{1}{M} \sum_{\boldsymbol{u}_i} \sum_{x_i} M[x_i, \boldsymbol{u}_i] log \hat{\theta}_{x_i|\boldsymbol{u}_i} \right]$$

12

Likelihood Scores

$$\frac{1}{M}\sum_{\boldsymbol{u}_i}\sum_{x_i} M[x_i,\boldsymbol{u}_i] log\hat{\theta}_{x_i|\boldsymbol{u}_i}$$
$$= \sum_{\boldsymbol{u}_i}\sum_{x_i} \hat{P}(x_i,\boldsymbol{u}_i) log\hat{P}(x_i|u_i)$$
$$= \sum_{\boldsymbol{u}_i}\sum_{x_i} \hat{P}(x_i,\boldsymbol{u}_i) log\left(\frac{\hat{P}(x_i,\boldsymbol{u}_i)}{\hat{P}(\boldsymbol{u}_i)}\cdot\frac{\hat{P}(x_i)}{\hat{P}(x_i)}\right)$$
$$= \sum_{\boldsymbol{u}_i}\sum_{x_i} \hat{P}(x_i,\boldsymbol{u}_i) log\left(\frac{\hat{P}(x_i,\boldsymbol{u}_i)}{\hat{P}(\boldsymbol{u}_i)\hat{P}(x_i)}\right) + \sum_{x_i}\left(\sum_{\boldsymbol{u}_i}\hat{P}(x_i,\boldsymbol{u}_i)\right) log\hat{P}(x_i)$$
$$= \boldsymbol{I}_{\hat{P}}\,(X_i;\boldsymbol{U}_i) - \sum_{x_i}\hat{P}(x_i) log\frac{1}{\hat{P}(x_i)}$$
$$= \boldsymbol{I}_{\hat{P}}\,(X_i;\boldsymbol{U}_i) - \boldsymbol{H}_{\hat{P}}(X_i)$$

Note that if $Parents(X_i,\mathcal{G}) = \emptyset$, then $I_{\hat{P}}(X_i;Parents(X_i,\mathcal{G})) = 0$

Likelihood Scores

What are the implications of

$$\underbrace{\boldsymbol{I}_{\hat{P}}(X_i;\boldsymbol{U}_i)} - \underbrace{\boldsymbol{H}_{\hat{P}}(X_i)}$$

Depends on network structure (because $U_i = \text{Parents}(X_i,\mathcal{G})$). Only need to maximize this.

Does not depend on network structure

The likelihood of a network measures how informative $Parents(X_i)$ are about X_i

Likelihood Scores

An alternate representation:

$$\frac{1}{M}score_L(\mathcal{G}:\mathcal{D}) = \underline{H_{\hat{P}}(X_1,\ldots,X_n)}$$
$$\underline{-\sum_{i=1}^{n} I_{\hat{P}}(X_i;\{X_1,\ldots,X_{i-1}\} - Parents(X_i,\mathcal{G})\mid Parents(X_i,\mathcal{G}))}$$

Does not depend on network structure

Depends on network structure

Measures to what extent the Markov properties of the graph are violated in the data (fewer violations ⇒ larger score)

Problems with Likelihood Score

Never prefers a simpler network over a more complex one eg.

$$score_L(\mathcal{G}_1:\mathcal{D}) \geq score_L(\mathcal{G}_0:\mathcal{D})$$

Problems with Likelihood Score

- Exhibits a conditional independence only if it holds exactly in the empirical distribution
  - Due to noise, this almost never happens
- Learns a fully connected graph
  - Overfits the training data and does not generalize well to unseen cases
- Needs a penalty for learning overly complex structures

Bayesian Scoring

Bayesian Score

- Bayesian philosophy: if you are uncertainty about something, put a distribution over it
- In structure learning, we have uncertainty over the structure and the parameters
- We will have two prior distributions:
  - Structure prior $P(\mathcal{G})$
  - Parameter prior $P(\boldsymbol{\theta}_{\mathcal{G}}|\mathcal{G})$

Bayesian Score

Recall: $P(\mathcal{G}|D) = \dfrac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})} = \alpha P(\mathcal{D}|\mathcal{G})P(\mathcal{G})$

$$score_B(\mathcal{G}:\mathcal{D}) = \underbrace{log P(\mathcal{D} \mid \mathcal{G})}_{\text{Marginal Likelihood (dominates the score)}} + \underbrace{\log P(\mathcal{G})}_{\text{Structure prior}}$$

$$P(\mathcal{D}|\mathcal{G}) = \int_{\Theta_{\mathcal{G}}} P(D|\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G}) P(\boldsymbol{\theta}_{\mathcal{G}}|\mathcal{G}) d\boldsymbol{\theta}_{\mathcal{G}}$$

"Averages" out $P(\mathcal{D}|\theta_{\mathcal{G}}, \mathcal{G})$ over the distribution of $\theta_{\mathcal{G}}$. Contrast this with maximum likelihood which finds the $\theta_{\mathcal{G}}$ that maximizes the likelihood of the data

Bayesian score

- How does the Bayesian score improve over the likelihood score?
  - By avoiding overfitting
- Likelihood score commits to a single $\hat{\boldsymbol{\theta}}$ value
- Bayesian score works with a distribution of $\boldsymbol{\theta}_{\mathcal{G}}$ and averages $P(\mathcal{D}|\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G})$ over this distribution
  - Results in an expected likelihood

Marginal Likelihood (Single Variable case)

- Suppose we have a single binary random variable X
- Let the prior distribution over the parameters of X be $Dirichlet(\alpha_1, \alpha_0)$
- Let the data $\mathcal{D} = \{x[1], \ldots, x[M]\}$ have $M[1]$ heads and $M[0]$ tails
- Maximum likelihood value given D is: $P(\mathcal{D}|\hat{\theta}) = \left(\frac{M[1]}{M}\right)^{M[1]} \cdot \left(\frac{M[0]}{M}\right)^{M[0]}$

Marginal Likelihood (Single Variable case)

What about the marginal likelihood?

$$P(\mathcal{D}|\mathcal{G}) = \int_{\Theta_{\mathcal{G}}} \underbrace{P(\mathcal{D}|\boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G})}_{\left(\frac{M[1]}{M}\right)^{M[1]} \cdot \left(\frac{M[0]}{M}\right)^{M[0]}} \underbrace{P(\boldsymbol{\theta}_{\mathcal{G}}|\mathcal{G})}_{Dirichlet(\alpha_1, \alpha_0)} d\boldsymbol{\theta}_{\mathcal{G}}$$

Shorthand: let $p_i = \frac{M[i]}{M}$ and $\alpha = \alpha_0 + \alpha_1$

Marginal Likelihood (Single Variable case)

$$P(\mathcal{D}|\mathcal{G}) = \int_{\Theta_{\mathcal{G}}} p_1^{\;M[1]} p_0^{\;M[0]} \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} p_1^{(\alpha_1 - 1)} p_0^{(\alpha_0 - 1)} d\boldsymbol{\theta}_{\mathcal{G}}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \int_{\Theta_{\mathcal{G}}} p_1^{\;(M[1]+\alpha_1-1)} p_0^{\;(M[0]+\alpha_0-1)} d\boldsymbol{\theta}_{\mathcal{G}}$$

Note: $\int_{\Theta_{\mathcal{G}}} Beta(\mathrm{M}[1] + \alpha_1, M[0] + \alpha_0) d\boldsymbol{\theta}_{\mathcal{G}} = 1$

$$\Rightarrow \int_{\Theta_{\mathcal{G}}} \frac{\Gamma(\alpha + M)}{\Gamma(M[1] + \alpha_1)\Gamma(M[0] + \alpha_0)} p_1^{\;(M[1]+\alpha_1-1)} p_0^{\;(M[0]+\alpha_0-1)} d\boldsymbol{\theta}_{\mathcal{G}} = 1$$

$$\Rightarrow \int_{\Theta_{\mathcal{G}}} p_1^{\;(M[1]+\alpha_1-1)} p_0^{\;(M[0]+\alpha_0-1)} d\boldsymbol{\theta}_{\mathcal{G}} = \frac{\Gamma(\alpha + M)}{\Gamma(M[1] + \alpha_1)\Gamma(M[0] + \alpha_0)}$$

Marginal Likelihood (Single Variable case)

$$P(\mathcal{D}|\mathcal{G}) = \int_{\Theta_\mathcal{G}} p_1^{\;M[1]} p_0^{\;M[0]} \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} p_1^{(\alpha_1-1)} p_0^{(\alpha_0-1)} d\boldsymbol{\theta}_\mathcal{G}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \int_{\Theta_\mathcal{G}} p_1^{\;(M[1]+\alpha_1-1)} p_0^{\;(M[0]+}$$

Note that the Gamma function is as follows:
$\Gamma(1) = 1$
$\Gamma(x+1) = x\Gamma(x)$

ie. it is a continuous generalization of the factorial:
$\Gamma(n+1) = n!$

Note:

$$\int_{\Theta_\mathcal{G}} Beta(\mathrm{M}[1]+\alpha_1, M[0]+\alpha_0) d\boldsymbol{\theta}_\mathcal{G} = 1$$

$$\Rightarrow \int_{\Theta_\mathcal{G}} \frac{\Gamma(\alpha+M)}{\Gamma(M[1]+\alpha_1)\Gamma(M[0]+\alpha_0)} p_1^{\;(M[1]+\alpha_1-1)} p_0^{\;(M[0]+\alpha_0-1)} d\boldsymbol{\theta}_\mathcal{G} = 1$$

$$\Rightarrow \int_{\Theta_\mathcal{G}} p_1^{\;(M[1]+\alpha_1-1)} p_0^{\;(M[0]+\alpha_0-1)} d\boldsymbol{\theta}_\mathcal{G} = \frac{\Gamma(\alpha+M)}{\Gamma(M[1]+\alpha_1)\Gamma(M[0]+\alpha_0)}$$

Marginal Likelihood (Single Variable case)

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \frac{\Gamma(\alpha+M)}{\Gamma(M[1]+\alpha_1)\Gamma(M[0]+\alpha_0)}$$

$$P(\mathcal{D}|\mathcal{G}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+M)} \cdot \frac{\Gamma(\alpha_1+M)}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_0+M)}{\Gamma(\alpha_0)}$$

We can easily generalize to a multinomial distribution over the space of values $x^1, \dots, x^k$ with a prior $Dirichlet(\alpha_1, \dots, \alpha_k)$:

$$P(\mathcal{D}|\mathcal{G}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+M)} \cdot \prod_i^k \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)}$$

Bayesian Scoring

Global parameter independence:

Let $\mathcal{G}$ be a Bayesian network structure with parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_{X_1|Pa(X_1)}, \dots, \boldsymbol{\theta}_{X_n|Pa(X_n)})$.

The distribution $P(\boldsymbol{\theta})$ satisfies global parameter independence if it has the form:

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i|Pa(X_i)})$$

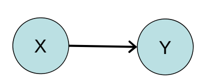
Bayesian Scoring

Local parameter independence:

Let X be a variable with parents $\boldsymbol{U}$. We say that distribution $P(\boldsymbol{\theta}_{X|\boldsymbol{U}})$ satisfies local parameter independence if:

$$P(\boldsymbol{\theta}_{X|\boldsymbol{U}}) = \prod_{\boldsymbol{u}} P(\boldsymbol{\theta}_{X|\boldsymbol{u}})$$

Example:



X	Y	P(Y|X)
0	0	$\theta_{Y\|x^0}$
0	1	
1	0	$\theta_{Y\|x^1}$
1	1	

Only one of the θs applies, depending on the value of x. In other words, the θs don't affect each other

Bayesian Scoring

Now suppose there are two binary random variables X and Y. Let $\mathcal{G}_0$ be a graph with X and Y and no edges

$$P(\mathcal{D}|\mathcal{G}_0) = \int_{\Theta_X \times \Theta_Y} P(D|\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y, \mathcal{G}_0) P(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y|\mathcal{G}_0) d[\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y]$$

1. Decompose likelihood in terms of each variable
$$P(D|\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y, \mathcal{G}_0) = \prod_{i=1}^{M} P(x[m]|\theta_X, \mathcal{G}_0)\, P(y[m]|\theta_Y, \mathcal{G}_0)$$
2. Global Parameter Independence: $P(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y|\mathcal{G}_0) = \mathrm{P}(\boldsymbol{\theta}_X|\mathcal{G}_0) P(\boldsymbol{\theta}_Y|\mathcal{G}_0)$

Bayesian Scoring

$$P(\mathcal{D}|\mathcal{G}_0) = \int_{\Theta_X \times \Theta_Y} P(D|\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y, \mathcal{G}_0) P(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y|\mathcal{G}_0) d[\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y]$$
$$= \int_{\Theta_X} \int_{\Theta_Y} \prod_{m=1}^{M} P(x[m]|\boldsymbol{\theta}_X, \mathcal{G}_0)\, P(y[m]|\boldsymbol{\theta}_Y, \mathcal{G}_0) \mathrm{P}(\boldsymbol{\theta}_X|\mathcal{G}_0) P(\boldsymbol{\theta}_Y|\mathcal{G}_0)\; d[\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y]$$
$$= \int_{\Theta_X} \int_{\Theta_Y} \left(\prod_{m=1}^{M} P(x[m]|\boldsymbol{\theta}_X, \mathcal{G}_0) \mathrm{P}(\boldsymbol{\theta}_X|\mathcal{G}_0) \right) \left(\prod_{m=1}^{M} P(y[m]|\boldsymbol{\theta}_Y, \mathcal{G}_0)\, P(\boldsymbol{\theta}_Y|\mathcal{G}_0) \right) d[\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y]$$

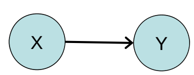
Integral of a product of independent functions is the product of integrals:

$$= \underbrace{\left(\int_{\Theta_X} \prod_{m=1}^{M} P(x[m]|\boldsymbol{\theta}_X, \mathcal{G}_0) \mathrm{P}(\boldsymbol{\theta}_X|\mathcal{G}_0)\, d\boldsymbol{\theta}_X \right)} \underbrace{\left(\int_{\Theta_Y} \prod_{m=1}^{M} P(y[m]|\boldsymbol{\theta}_Y, \mathcal{G}_0)\, P(\boldsymbol{\theta}_Y|\mathcal{G}_0)\; d\boldsymbol{\theta}_Y \right)}$$

Note: decomposes into one term for each random variable

Bayesian Scoring

Now suppose there are two binary random variables X and Y and let $\mathcal{G}_{X \to Y}$ be the graph below:

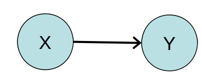


X	P(X)
0	θ_X
1	

X	Y	P(Y|X)
0	0	$\theta_{Y\|x^0}$
0	1	
1	0	$\theta_{Y\|x^1}$
1	1	

Bayesian Scoring

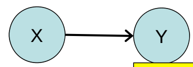
Now suppose there are two binary random variables X and Y and let $\mathcal{G}_{X \to Y}$ be the graph below:



$$P(\mathcal{D}|\mathcal{G}_{X \to Y}) = \left(\int_{\Theta_X} \prod_{m=1}^{M} P(x[m]|\boldsymbol{\theta}_X, \mathcal{G}_{X \to Y}) \mathrm{P}(\boldsymbol{\theta}_X|\mathcal{G}_{X \to Y})\, d\boldsymbol{\theta}_X \right) \cdot$$
$$\left(\int_{\Theta_{Y|x^0}} \prod_{m:x[m]=x^0}^{M} P(y[m]|\boldsymbol{\theta}_{Y|x^0}, \mathcal{G}_{X \to Y})\, P(\boldsymbol{\theta}_{Y|x^0}|\mathcal{G}_{X \to Y})\; d\boldsymbol{\theta}_{Y|x^0} \right) \cdot$$
$$\left(\int_{\Theta_{Y|x^1}} \prod_{m:x[m]=x^1}^{M} P(y[m]|\boldsymbol{\theta}_{Y|x^1}, \mathcal{G}_{X \to Y})\, P(\boldsymbol{\theta}_{Y|x^1}|\mathcal{G}_{X \to Y})\; d\boldsymbol{\theta}_{Y|x^1} \right)$$

Bayesian Scoring

Now suppose there are two binary random variables X and Y and let $\mathcal{G}_{X\rightarrow Y}$ be the graph below:



One term for each parameter family.
Each term has a closed form solution

$$P(\mathcal{D}|\mathcal{G}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + M)} \cdot \prod_{i}^{k} \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)}$$

$$P(\mathcal{D}|\mathcal{G}_{X\rightarrow Y}) = \left(\int_{\Theta_X} \prod_{m=1}^{M} P(x[m]|\boldsymbol{\theta}_X, \mathcal{G}_{X\rightarrow Y}) \right.$$

$$\left(\int_{\Theta_{Y|x^0}} \prod_{m:x[m]=x^0}^{M} P(y[m]|\boldsymbol{\theta}_{Y|x^0}, \mathcal{G}_{X\rightarrow Y})\, P(\boldsymbol{\theta}_{Y|x^0}|\mathcal{G}_{X\rightarrow} \quad {}_{Y|x^0} \right) \cdot$$

$$\left(\int_{\Theta_{Y|x^1}} \prod_{m:x[m]=x^1}^{M} P(y[m]|\boldsymbol{\theta}_{Y|x^1}, \mathcal{G}_{X\rightarrow Y})\, P(\boldsymbol{\theta}_{Y|x^1}|\mathcal{G}_{X\rightarrow Y}) \ d\boldsymbol{\theta}_{Y|x^1} \right)$$

33

Bayesian Scoring

The general case: let $\mathcal{G}$ be a network structure, and let $P(\theta_{\mathcal{G}}|\mathcal{G})$ be a parameter prior satisfying global parameter independence. Then:

$$\boxed{P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{n} \int_{\Theta_{X_i|Pa(X_i)}} \prod_{m=1}^{M} P(x_i[m]|pa(X_i)[m], \boldsymbol{\theta}_{X_i|Pa(X_i)}, \mathcal{G})\, P(\boldsymbol{\theta}_{X_i|Pa(X_i)}|\mathcal{G}) d\boldsymbol{\theta}_{X_i|Pa(X_i)}}$$

If $P(\boldsymbol{\theta}_{\mathcal{G}})$ also satisfies local parameter independence, then

$$\boxed{P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{n} \prod_{\boldsymbol{u}_i \in Val(Pa^{\mathcal{G}}(X_i))} \int_{\Theta_{X_i|\boldsymbol{u}_i}} \prod_{m:\, \boldsymbol{u}_i[m]=\boldsymbol{u}_i}^{M} P(x_i[m]|\boldsymbol{u}_i)[m], \boldsymbol{\theta}_{X_i|\boldsymbol{u}_i}, \mathcal{G})\, P(\boldsymbol{\theta}_{X_i|\boldsymbol{u}_i}|\mathcal{G}) d\boldsymbol{\theta}_{X_i|\boldsymbol{u}_i}}$$

34

Bayesian Scoring

If we have a Bayesian network with Dirichlet priors where $P(\theta_{X_i|pa(X_i)}|\mathcal{G})$ has hyperparameters $\left\{\alpha^{\mathcal{G}}{}_{x^j{}_i|\boldsymbol{u}_i} : j = 1, \ldots, |X_i|\right\}$ then

$$P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{n} \prod_{u_i \in Val(Pa^{\mathcal{G}}(X_i))} \frac{\Gamma(\alpha^{\mathcal{G}}{}_{X_i|\boldsymbol{u}_i})}{\Gamma(\alpha^{\mathcal{G}}{}_{X_i|\boldsymbol{u}_i} + M[u_i])} \prod_{x_i{}^j \in Val(X_i)} \left[\frac{\Gamma\left(\alpha^{\mathcal{G}}{}_{x_i{}^j|\boldsymbol{u}_i}\right) + M[x_i{}^j, \boldsymbol{u}_i]}{\Gamma(\alpha^{\mathcal{G}}{}_{x_i{}^j|\boldsymbol{u}_i})} \right]$$

Where:

$$\alpha^{\mathcal{G}}{}_{X_i|\boldsymbol{u}_i} = \sum_j \alpha^{\mathcal{G}}{}_{x_i{}^j|\boldsymbol{u}_i}$$

35

Bayesian Scoring

If we have a Bayesian network with Dirichlet priors where $P(\theta_{X_i|pa(X_i)}|\mathcal{G})$ has hyperparameters $\left\{\alpha^{\mathcal{G}}{}_{x^j{}_i|\boldsymbol{u}_i} : j = 1, \ldots, |X_i|\right\}$ then

$$P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{n} \prod_{u_i \in Val(Pa^{\mathcal{G}}(X_i))} \frac{\Gamma(\alpha^{\mathcal{G}}{}_{X_i|\boldsymbol{u}_i})}{\Gamma(\alpha^{\mathcal{G}}{}_{X_i|\boldsymbol{u}_i} + M[u_i])} \prod_{x_i{}^j \in Val(X_i)} \left[\frac{\Gamma\left(\alpha^{\mathcal{G}}{}_{x_i{}^j|\boldsymbol{u}_i}\right) + M[x_i{}^j, \boldsymbol{u}_i]}{\Gamma(\alpha^{\mathcal{G}}{}_{x_i{}^j|\boldsymbol{u}_i})} \right]$$

Iterates over # of random variables

Iterates over # of instantiations of parents of X_i

Iterates over # of values of X_i

Where:

$$\alpha^{\mathcal{G}}{}_{X_i|\boldsymbol{u}_i} \quad \sum_j$$

36

Bayesian Scoring

If we use a Dirichlet parameter prior for all parameters in our network, then, because $M \to \infty$ (proof omitted), we have:

$$logP(\mathcal{D}|\mathcal{G}) = l(\hat{\theta}_{\mathcal{G}}:\mathcal{D}) - \frac{logM}{2}\underbrace{Dim[\mathcal{G}]}_{\text{\# of independent parameters in } \mathcal{G}} + O(1)$$

This is the Bayesian Information Criterion (BIC) score

37

Bayesian Scoring

This is the Bayesian Information Criterion (BIC) score:

$$score_{BIC}(\mathcal{G}:\mathcal{D}) = \underbrace{l(\hat{\theta}_{\mathcal{G}}:\mathcal{D})}_{\text{Fit to data}} - \underbrace{\frac{logM}{2}Dim[\mathcal{G}]}_{\text{Model complexity}} + O(1)$$

Can also interpret this as the # of bits to encode the model and the data given the model (minimum description length)

38

Bayesian Scoring

$$score_{BIC}(\mathcal{G}:\mathcal{D}) = M\sum_{i=1}^{n} \boldsymbol{I}_{\hat{P}}(X_i; Pa(X_i)) - M\sum_{i=1}^{n} \boldsymbol{H}_{\hat{P}}(X_i) - \frac{logM}{2}Dim[\mathcal{G}]$$

Things to note:

- Entropy term $M\sum_{i=1}^{n} \boldsymbol{H}_{\hat{P}}(X_i)$ can be ignored (doesn't depend on graph structure)
- Trades off fit to data and model complexity
  - The stronger the dependence of a variable on its parents, the higher the score (grows linearly)
  - The more complex the network, the lower the score (grows logarithmically)
- As M grows, the score pays more attention to the data fit

39

Bayesian Scoring

Assume that our data are generated by some distribution P^* for which the network $\mathcal{G}^*$ is a perfect map.

We say that a scoring function is consistent if the following properties hold as the amount of data $M \to \infty$, with probability that approaches 1 (over all possible choices of data set $\mathcal{D}$):

- The structure $\mathcal{G}^*$ will maximize the score
- All structures $\mathcal{G}$ that are not I-equivalent to $\mathcal{G}^*$ will have strictly lower score

40

Bayesian Scoring

- The BIC score (and the Bayesian score) is consistent [proof omitted]
- In practice though, the BIC score tends to have a very strong preference for simpler structures

41

Structure Priors

Recall that

$$score_B(\mathcal{G}:\mathcal{D}) = \underbrace{log P(\mathcal{D} \mid \mathcal{G})}_{\text{Grows linearly with the number of examples (dominates the score)}} + \underbrace{\log P(\mathcal{G})}_{\text{Structure prior (stays constant). Only matters for small sample sizes}}$$

42

Structure Priors

- Typically assign uniform priors over structures
- If you can provide an informed structure prior, you could penalize edges in the graph:
  - $P(\mathcal{G}) \propto c^{|\mathcal{G}|}$ (where $c < 1$ and $|\mathcal{G}|$ is the number of edges)
- Mathematically convenient to have structure prior with structure modularity:
  - $P(\mathcal{G}) \propto \prod_i \underbrace{P(Pa(X_i) = Pa^{\mathcal{G}}(X_i))}_{\text{Uses local properties not global properties of the graph}}$

43