Structure Learning 2

Structure Scores

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

1. Search space
   • Start with initial state (eg. disconnected graph or randomly generated one)

![Initial State Diagram]

- Edge Addition
- Edge Deletion
- Edge Reversal

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

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
Structure Scores

Likelihood Scores

$$\max_{\mathcal{G}, \mathbf{\theta}_g} L((\mathcal{G}, \mathbf{\theta}_g): \mathcal{D})$$
$$= \max_{\mathcal{G}} \left[ \max_{\mathbf{\theta}_g} L((\mathcal{G}, \mathbf{\theta}_g): \mathcal{D}) \right]$$
$$= \max_{\mathcal{G}} \left[ L((\mathcal{G}, \hat{\mathbf{\theta}}_g): \mathcal{D}) \right]$$

Graph structure that maximizes the likelihood

Maximum likelihood estimates of parameters

$$\text{score}_L(\mathcal{G}: \mathcal{D}) = l(\hat{\mathbf{\theta}}_g; \mathcal{D})$$

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

Mutual Information

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

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

Claim:

$$score_L(G: D) = M \sum_{i=1}^{n} I_p(X_i; Parent s(X_i, G)) - M \sum_{i=1}^{n} H_p(X_i)$$

Likelihood Scores

Proof:

$$l(\hat{\theta}_G; D) = \sum_{i=1}^{n} \left[ \sum_{u_i \in Val(Parent s(X_i, G))} \sum_{x_i} M[x_i, u_i] log \hat{\theta}_{x_i|u_i} \right]$$

Take one of these terms and let $U_i = Parent s(X_i, G)$

$$\frac{1}{M} \sum_{u_i} \sum_{x_i} M[x_i, u_i] log \hat{\theta}_{x_i|u_i}$$
Likelihood Scores

\[
\frac{1}{M} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \theta_{x_i|u_i}
\]

\[
= \sum_{u_i} \sum_{x_i} \hat{p}(x_i, u_i) \log \hat{p}(x_i|u_i)
\]

\[
= \sum_{u_i} \sum_{x_i} \hat{p}(x_i, u_i) \log \left( \frac{\hat{p}(x_i, u_i)}{\hat{p}(u_i)} \cdot \hat{p}(x_i) \right)
\]

\[
= \sum_{u_i} \sum_{x_i} \hat{p}(x_i, u_i) \log \left( \frac{\hat{p}(x_i, u_i)}{\hat{p}(u_i)} \cdot \hat{p}(x_i) \right) + \sum_{x_i} \left( \sum_{u_i} \hat{p}(x_i, u_i) \right) \log \hat{p}(x_i)
\]

\[
= I_{\hat{p}}(X_i; U_i) - \sum_{x_i} \hat{p}(x_i) \log \frac{1}{\hat{p}(x_i)}
\]

\[
= I_{\hat{p}}(X_i; U_i) - H_{\hat{p}}(X_i)
\]

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

Likelihood Scores

What are the implications of

\[
I_{\hat{p}}(X_i; U_i) - H_{\hat{p}}(X_i)
\]

\[\text{\textcolor{red}{Depends on network structure (because}} \ U_i = Parents(X_i, G)\text{\textcolor{red}{)}}. \]

\[\text{\textcolor{blue}{Only need to maximize this.}}\]

\[\text{\textcolor{blue}{Does not depend on network structure}}.\]

The likelihood of a network measures how informative \( Parents(X_i) \) are about \( X_i \)
Likelihood Scores

An alternate representation:

\[
\frac{1}{M} \text{score}_L(G; \mathcal{D}) = H_p(X_1, \ldots, X_n) - \sum_{i=1}^{n} I_p(X_i; \{X_1, \ldots, X_{i-1}\} \mid \text{Parents}(X_i, G))
\]

**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 \(\Rightarrow\) larger score)

Problems with Likelihood Score

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

**\(G_1**

\[
\begin{align*}
G_1 & : X \rightarrow Y \\
G_0 & : X \rightarrow Y
\end{align*}
\]

**\(score_L(G_1; \mathcal{D}) \geq score_L(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 $P(G)$
  – Parameter prior $P(\theta_g | G)$

Recall:

$$P(G | D) = \frac{P(D|G)P(G)}{P(D)} = \alpha P(D|G)P(G)$$

$$score_B(G; D) = \log P(D | G) + \log P(G)$$

Marginal Likelihood (dominates the score)

Structure prior

$$P(D | G) = \int_{\theta_g} P(D | \theta_g, G)P(\theta_g | G) d\theta_g$$

“Averages” out $P(D | \theta_g, G)$ over the distribution of $\theta_g$. Contrast this with maximum likelihood which finds the $\theta_g$ 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 \( \hat{\theta} \) value
- Bayesian score works with a distribution of \( \theta_g \) and averages \( P(D|\theta_g, G) \) 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 \( \text{Dirichlet}(\alpha_1, \alpha_0) \)
- Let the data \( D = \{x[1], \ldots, x[M]\} \) have \( M[1] \) heads and \( M[0] \) tails
- Maximum likelihood value given \( D \)
  
  \[
  P(D|\hat{\theta}) = \left( \frac{M[1]}{M} \right)^{M[1]} \cdot \left( \frac{M[0]}{M} \right)^{M[0]}
  \]

Marginal Likelihood (Single Variable case)

What about the marginal likelihood?

\[
P(D|\mathcal{G}) = \int_{\Theta_\mathcal{G}} P(D|\theta_\mathcal{G}, \mathcal{G})P(\theta_\mathcal{G}|\mathcal{G})d\theta_\mathcal{G}
\]

\[
\left(\frac{M[1]}{M}\right)^{M[1]} \cdot \left(\frac{M[0]}{M}\right)^{M[0]} \text{ Dirichlet}(\alpha_1, \alpha_0)
\]

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

Marginal Likelihood (Single Variable case)

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

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

Note:

\[
\int_{\Theta_\mathcal{G}} \text{Beta}(M[1] + \alpha_1, M[0] + \alpha_0) d\theta_\mathcal{G} = 1
\]

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

\[
\Rightarrow \int_{\Theta_\mathcal{G}} p_1^{(M[1]+\alpha_1-1)} p_0^{(M[0]+\alpha_0-1)} d\theta_\mathcal{G} = \frac{\Gamma(M[1] + \alpha_1)\Gamma(M[0] + \alpha_0)}{\Gamma(\alpha + M)}
\]
Marginal Likelihood (Single Variable case)

\[ P(D|\theta) = \int_{\theta} p_1^{M_1} p_0^{M_0} \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} p_1^{(\alpha_1-1)} p_0^{(\alpha_0-1)} d\theta \]

\[ = \frac{\Gamma(\alpha)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \int_{\theta} p_1^{(M_1+\alpha_1-1)} p_0^{(M_0+\alpha_0-1)} d\theta \]

Note:

\[ \int_{\theta} \text{Beta}(M[1]+\alpha_1, M[0]+\alpha_0) d\theta = 1 \]

\[ \Rightarrow \int_{\theta} \frac{\Gamma(\alpha + M)}{\Gamma(M[1]+\alpha_1)\Gamma(M[0]+\alpha_0)} p_1^{(M_1+\alpha_1-1)} p_0^{(M_0+\alpha_0-1)} d\theta = 1 \]

\[ \Rightarrow \int_{\theta} p_1^{(M_1+\alpha_1-1)} p_0^{(M_0+\alpha_0-1)} d\theta = \frac{\Gamma(M[1]+\alpha_1)\Gamma(M[0]+\alpha_0)}{\Gamma(\alpha + M)} \]

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:

\[ \Gamma(n+1) = n! \]

Marginal Likelihood (Single Variable case)

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

\[ P(D|\theta) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + M)} \cdot \frac{\Gamma(\alpha_1 + M[1])}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_0 + M[0])}{\Gamma(\alpha_0)} \]

We can easily generalize to a multinomial distribution over the space of values \( x_1, ..., x_k \) with a prior \( \text{Dirichlet}(\alpha_1, ..., \alpha_k) \):

\[ P(D|\theta) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + M)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)} \]
Bayesian Scoring

Global parameter independence:
Let $\mathcal{G}$ be a Bayesian network structure with parameters $\theta = (\theta_{X_1|Pa(X_1)}, \ldots, \theta_{X_n|Pa(X_n)})$. The distribution $P(\theta)$ satisfies global parameter independence if it has the form:

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

Bayesian Scoring

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

$$P(\theta_{X|U}) = \prod_{u} P(\theta_{X|u})$$

Example:

Only one of the $\theta$ applies, depending on the value of $x$. In other words, the $\theta$s 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

1. Decompose likelihood in terms of each variable

$$P(D|\mathcal{G}_0) = \int_{\theta_X \times \theta_Y} P(D|\theta_X, \theta_Y, \mathcal{G}_0)P(\theta_X, \theta_Y|\mathcal{G}_0)d[\theta_X, \theta_Y]$$

2. Global Parameter Independence:

$$P(\theta_X, \theta_Y|\mathcal{G}_0) = P(\theta_X|\mathcal{G}_0)P(\theta_Y|\mathcal{G}_0)$$

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

$$= \left(\int_{\theta_X} \prod_{m=1}^M P(x[m]|\theta_X, \mathcal{G}_0)d\theta_X\right) \left(\int_{\theta_Y} \prod_{m=1}^M P(y[m]|\theta_Y, \mathcal{G}_0)d\theta_Y\right)$$

Note: decomposes into one term for each random variable
Bayesian Scoring

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

![Graph](image)

| X | P(X) | X | Y | P(Y|X) |
|---|------|---|---|--------|
| 0 | $\theta_X$ | 0 | 0 | $\theta_{Y|X=0}$ |
| 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | $\theta_{Y|X=1}$ |
| 1 | 1 |

Bayesian Scoring

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

![Graph](image)

$$P(D|G_{X \rightarrow Y}) = \left( \int_{\theta_X} \prod_{m=1}^{M} P(x[m]|\theta_X, G_{X \rightarrow Y}) P(\theta_X|G_{X \rightarrow Y}) \, d\theta_X \right) \cdot \left( \int_{\theta_{Y|X=0}} \prod_{m:x[m]=0}^{M} P(y[m]|\theta_{Y|X=0}, G_{X \rightarrow Y}) P(\theta_{Y|X=0}|G_{X \rightarrow Y}) \, d\theta_{Y|X=0} \right) \cdot \left( \int_{\theta_{Y|X=1}} \prod_{m:x[m]=1}^{M} P(y[m]|\theta_{Y|X=1}, G_{X \rightarrow Y}) P(\theta_{Y|X=1}|G_{X \rightarrow Y}) \, d\theta_{Y|X=1} \right).$$
Bayesian Scoring

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

\[
P(D|G_{X \rightarrow Y}) = \left( \int_{\theta_{X}} \prod_{m=1}^{M} P(X[m]|\theta_{X}, G_{X \rightarrow Y}) d\theta_{X} \right) \cdot \left( \int_{\theta_{Y|X}} \prod_{m|X[m]=x}^{M} P(Y[m]|\theta_{Y|X}, G_{X \rightarrow Y}) P(\theta_{Y|X}|G_{X \rightarrow Y}) d\theta_{Y|X} \right)
\]

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

\[
P(D|\mathcal{G}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + M)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + M[x^i])}{\Gamma(\alpha_i)}
\]

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

\[
P(D|\mathcal{G}) = \prod_{i=1}^{n} \int_{\theta_{X_i|Pa(X_i)}} \prod_{m=1}^{M} P(X_i[m]|Pa(X_i)[m], \theta_{X_i|Pa(X_i)}, G) P(\theta_{X_i|Pa(X_i)}|G) d\theta_{X_i|Pa(X_i)}
\]

If $P(\theta)$ also satisfies local parameter independence, then

\[
P(D|\mathcal{G}) = \prod_{i=1}^{n} \prod_{u_{i} \in \text{Val}(Pa^c(X_i))} \int_{\theta_{X_i|u_i}} \prod_{m|u_i[m]=u_i}^{M} P(X_i[m]|u_i[m], \theta_{X_i|u_i}, G) P(\theta_{X_i|u_i}|G) d\theta_{X_i|u_i}
\]
Bayesian Scoring

If we have a Bayesian network with Dirichlet priors where

\[ P(\theta_{X_i|pa(X_i)} | \mathcal{G}) \]

has hyperparameters \( \{ \alpha^g_{x_i|u_i} : j = 1, \ldots, |X_i| \} \)
then

\[
P(\mathcal{D} | \mathcal{G}) = \prod_{i=1}^{n} \prod_{u_i \in \text{Val}(pa(X_i))} \frac{\Gamma(\alpha^g_{x_i|u_i})}{\Gamma(\alpha^g_{x_i|u_i} + M[u_i])} \prod_{x_i' \in \text{Val}(X_i)} \frac{\Gamma\left(\alpha^g_{x_i'|u_i}\right) + M[x_i', u_i]}{\Gamma(\alpha^g_{x_i'|u_i})}
\]

Where:

\[ \alpha^g_{x_i|u_i} = \sum_j \alpha^g_{x_i'|u_i} \]
Bayesian Scoring

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

$$log P(\mathcal{D}|\mathcal{G}) = l(\hat{\theta}_G; \mathcal{D}) - \frac{\log M}{2} \text{Dim}[\mathcal{G}] + O(1)$$

# of independent parameters in $\mathcal{G}$

This is the Bayesian Information Criterion (BIC) score

Bayesian Scoring

This is the Bayesian Information Criterion (BIC) score:

$$score_{BIC}(\mathcal{G}; \mathcal{D}) = l(\hat{\theta}_G; \mathcal{D}) - \frac{\log M}{2} \text{Dim}[\mathcal{G}] + O(1)$$

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

\[ \text{score}_{\text{BIC}}(G; D) = M \sum_{i=1}^{n} I_p(X_i; Pa(X_i)) - M \sum_{i=1}^{n} H_p(X_i) - \frac{\log M}{2} \cdot \text{Dim}[G] \]

Things to note:
- Entropy term \( M \sum_{i=1}^{n} H_p(X_i) \) 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

Bayesian Scoring

Assume that our data are generated by some distribution \( P^* \) for which the network \( G^* \) 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 \( D \)):
- The structure \( G^* \) will maximize the score
- All structures \( G \) that are not I-equivalent to \( G^* \) will have strictly lower score
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

Structure Priors

Recall that

\[ \text{score}_B(G: D) = \log P(D | G) + \log P(G) \]

Grows linearly with the number of examples (dominates the score)

Structure prior (stays constant). Only matters for small sample sizes
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:
  – \( P(\mathcal{G}) \propto \prod_i P(Pa(X_i) = Pa^G(X_i)) \)

Uses local properties not global properties of the graph