Koller & Friedman Chapter 13

Structure Learning 2:
the good, the bad, the ugly

Graphical Model – 10708
Carlos Guestrin
Carnegie Mellon University

October 26th, 2005

Consistency of BIC and Bayesian scores

Consistency is limiting behavior, says nothing about finite sample size!!!

- A scoring function is **consistent** if, for true model G^*, as $M \to \infty$, with probability 1
  - G^* maximizes the score
  - All structures **not I-equivalent** to G^* have strictly lower score
- **Theorem**: BIC score is consistent
- **Corollary**: the Bayesian score is consistent
- What about maximum likelihood?

G^*: X Y ; X → Y

same likelihood score

cond. 2 violated

Priors for general graphs

- For finite datasets, prior is important!
- Prior over structure satisfying prior modularity

e.g., $P(G) \propto c^{-|G|}$

- What about prior over parameters, how do we represent it?
  - *K2 prior*: fix an α, $P(\theta_{X_i|\mathbf{Pa}_{X_i}}) = \text{Dirichlet}(\alpha, \ldots, \alpha)$
  - K2 is "inconsistent"

$P(X_i | A, B) \leftarrow 4\alpha$ "samples" of X_i

$P(X_i | A) \leftarrow 2\alpha$ "samples" " "

BDe prior

- Remember that Dirichlet parameters analogous to "fictitious samples"
- Pick a fictitious sample size M'
- For each possible family, define a prior distribution $P^0(X_i,\mathbf{Pa}_{X_i})$
  - Represent with a BN
  - Usually independent (product of marginals)

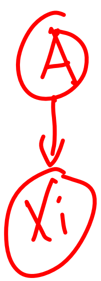
$P^0(X_i)\ \forall i \in \{1,\ldots,n\}$

$P^0(X_1,X_2,X_3) = P^0(X_1)\,P^0(X_2)\,P^0(X_3)$

- **BDe prior**:

$$\alpha_{x_i|pa_{X_i}} = M' \cdot P^0(x_i, pa_{X_i})$$

- Has "consistency property":



$P(X_i|A)$

$$\alpha_{x_i} = \sum_a \alpha_{x_i|A=a} = M' \sum_a P^0(x_i, A=a) = M'\,P^0(x_i) = \alpha_{x_i}$$

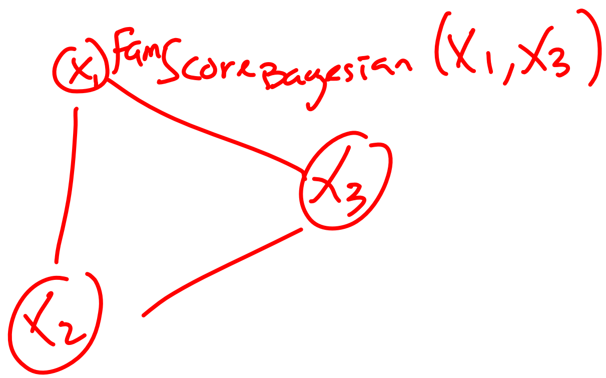
Score equivalence

- If *G* and *G'* are I-equivalent then they have same score

- **Theorem**: ~~Maximum~~ likelihood score and BIC scores satisfy score equivalence
- **Theorem**:
  - If P(*G*) assigns same prior to I-equivalent structures (e.g., edge counting)
  - and parameter prior is dirichlet
  - then Bayesian score satisfies score equivalence if and only if prior over parameters represented as a BDe prior!!!!!!

Chow-Liu for Bayesian score

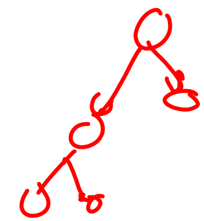
- Edge weight $w_{X_j \to X_i}$ is advantage of adding X_j as parent for X_i



- Now have a directed graph, need directed spanning forest
  - Note that adding an edge can hurt Bayesian score – choose forest not tree
  - But, if score satisfies score equivalence, then $w_{X_j \to X_i} = w_{X_j \to X_i}$!
  - Simple maximum spanning forest algorithm works

Structure learning for general graphs

- In a tree, a node only has one parent

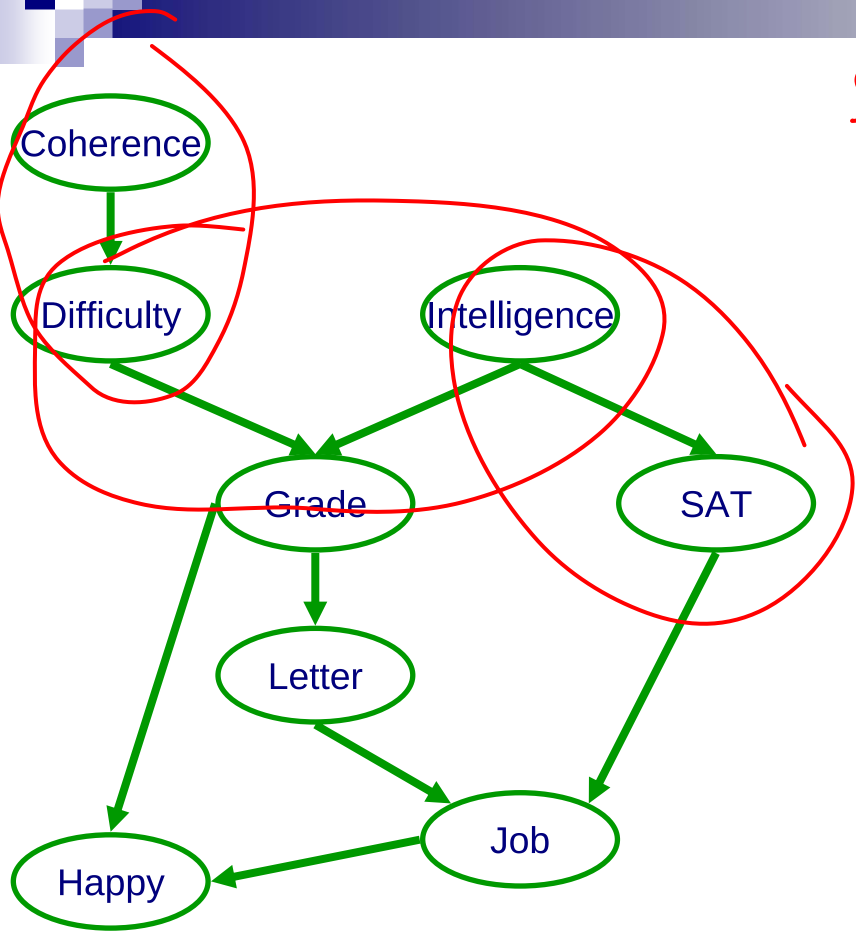


- **Theorem**:
  - The problem of learning a BN structure with at most *d* parents is NP-hard for any (fixed) $d \geq 2$

- Most structure learning approaches use heuristics
  - Exploit score decomposition
  - (Quickly) Describe two heuristics that exploit decomposition in different ways

Understanding score decomposition

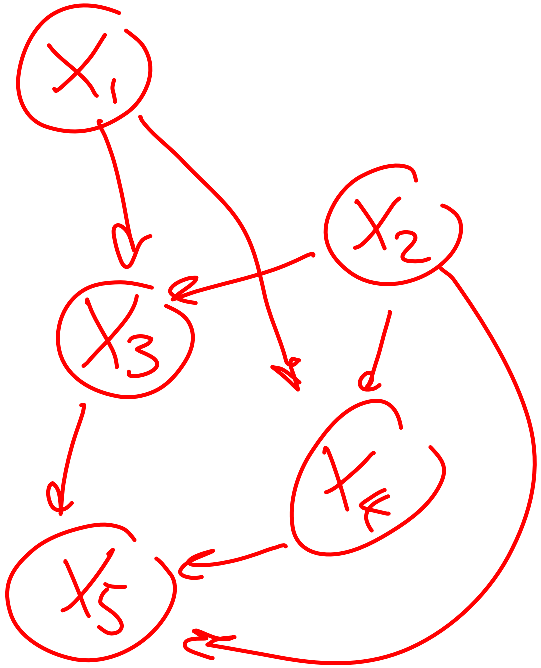
$$Score(G : D) = \sum_i FamScore(X_i \mid Pa_{X_i} : D)$$



Fixed variable order 1

$\mathrm{Score}(G : D) = \sum_i \mathrm{FamScore}(X_i \mid Pa_{X_i} : D)$

- Pick a variable order $\prec$
  - e.g., $X_1,\ldots,X_n$
- X_i can only pick parents in $\{X_1,\ldots,X_{i-1}\}$
  - Any subset
  - Acyclicity guaranteed!
- Total score = sum score of each node



Globally optimal!! :)

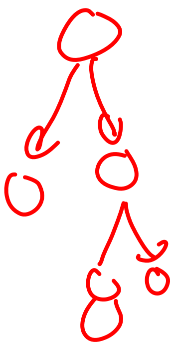
Fixed variable order 2

- Fix max number of parents to k $\binom{n}{k}$
- For each *i* in order $\prec$
  - Pick $\mathbf{Pa}_{X_i} \subseteq \{X_1,\ldots,X_{i-1}\}$
    - Exhaustively search through all possible subsets 2^k
    - $\mathbf{Pa}_{X_i}$ is maximum $\mathbf{U} \subseteq \{X_1,\ldots,X_{i-1}\}$ FamScore($X_i|\mathbf{U} : D$)

- Optimal BN for each order!!!
- Greedy search through space of orders:
  - E.g., try switching pairs of variables in order
  - If neighboring vars in order are switch, only need to recompute score for this pair
    - O(n) speed up per iteration
    - Local moves may be worse

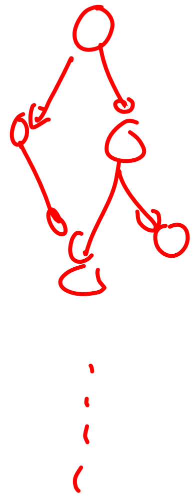
Learn BN structure using local search

Starting from Chow-Liu tree



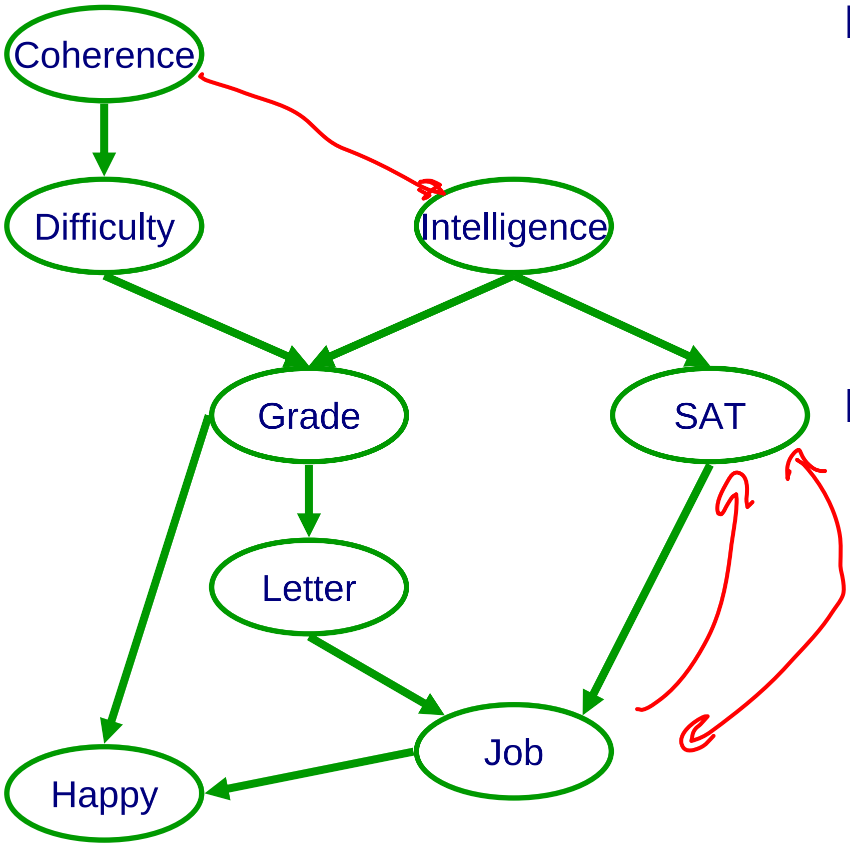
Local search,
possible moves:
Only if acyclic!!!
- Add edge
- Delete edge
- Invert edge

Select using favorite score



0.002

Exploit score decomposition in local search



- Add edge and delete edge:
  - Only rescore one family!

    $FamScore(X_i \mid Pa_{X_i} : D)$
- Reverse edge
  - Rescore only two families

    both scores change

Order search versus graph search

- Order search advantages
  - For fixed order, optimal BN – more “global” optimization
  - Space of orders much smaller than space of graphs

orders $2^{n \lg n}$ graphs $2^{(n^2)}$

- Graph search advantages
  - Not restricted to k parents
    - Especially if exploiting CPD structure, such as CSI
  - Cheaper per iteration
  - Finer moves within a graph

Bayesian model averaging

- So far, we have selected a single structure
- But, if you are really Bayesian, must average over structures
  - Similar to averaging over parameters

$$\log P(D \mid \mathcal{G}) = \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- Inference for structure averaging is very hard!!!
  - Clever tricks in reading

What you need to know about learning BN structures

- Decomposable scores
  - ~~Maximum~~ likelihood Score
  - Information theoretic interpretation
  - Bayesian
  - BIC approximation
- Priors
  - Structure and parameter assumptions
  - BDe if and only if score equivalence
- Best tree (Chow-Liu)
- Best TAN
- Nearly best k-treewidth (in $O(N^{k+1})$)
- Search techniques
  - Search through orders
  - Search through structures
- Bayesian model averaging