Readings:
K&F: 15.1, 15.2, 15.3, 15.4, 15.5

in BNs

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

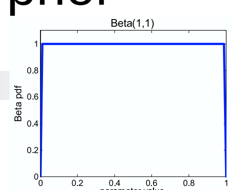
Graphical Models – 10708

Carlos Guestrin
Carnegie Mellon University

September 29th, 2006



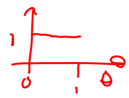
## Understanding the uniform prior over the parameters




- Likelihood: $P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$
- Prior: Uniform, Beta(1,1)
- E[θ] = $\frac{m_H+1}{m_H+m_T+2}$
- Why???

$P(\theta|D) \propto P(\theta) \cdot P(D|\theta) = \overset{1}{P(\theta)} \cdot \theta^{m_H}(1-\theta)^{m_T}$

uniform $P(\theta)=1$ $\quad = \theta^{m_H}(1-\theta)^{m_T} = Beta(m_H+1, m_T+1)$



$E[\theta] = \frac{1}{Z}\int_0^1 \theta \cdot \theta^{m_H}(1-\theta)^{m_T} = \underbrace{\frac{m_H+1}{m_H+m_T+2}}_{mean}$ mode: $\frac{m_H}{m_H+m_T}$

suppose: $m_H=1$, $m_T=0$

$\theta_{mean} = \frac{2}{3}$ $\quad \theta$ mode $= 1$

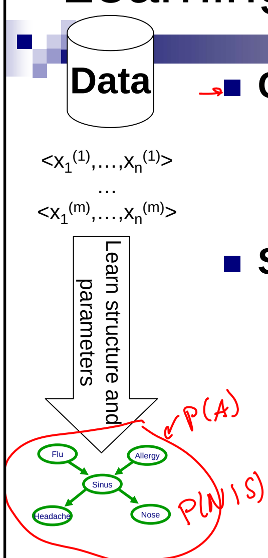
- Intuition: Prior knowledge is “don’t know anything”, so I don’t want to commit too fast

# Where are we with learning BNs?

- Given structure, estimate parameters
  - Maximum likelihood estimation MLE
  - Bayesian learning
- What about learning structure?



# Learning the structure of a BN




- **Constraint-based approach**
  - BN encodes conditional independencies
  - Test conditional independencies in data
  - Find an I-map
- **Score-based approach**
  - Finding a structure and parameters is a density estimation task
  - Evaluate model as we evaluated parameters
    - Maximum likelihood
    - Bayesian
    - etc.

# Remember: Obtaining a P-map?

- Given the independence assertions that are true for *P*
  - Obtain skeleton
  - Obtain immoralities
- From skeleton and immoralities, obtain every (and any) BN structure from the equivalence class

- **Constraint-based approach**:
  - Use Learn PDAG algorithm
  - Key question: **Independence test**



# Independence tests

$I(X_i, X_j | Z) = E_Z[I(X_i, X_j | Z=z)] = \sum_z P(z) \cdot \sum_{x_i x_j} P(x_i, x_j | z) \ldots$

- Statistically difficult task! $P(x_i, x_j) = P(x_j) \cdot P(x_i | x_j)$
- Intuitive approach: **Mutual information**

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

$\frac{P(x_i | x_j)}{P(x_i)}$ $\geq 0$

- Mutual information and independence:
  - $X_i$ and $X_j$ independent if and only if $I(X_i, X_j)=0$

Suppose $X_i \perp X_j$ $P(x_i, x_j) = P(x_i) \cdot P(x_j)$

$\frac{P(x_i) \cdot P(x_j)}{P(x_i) \cdot P(x_j)} = 1$ ; $\log 1 = 0$

- Conditional mutual information:

$(X_i \perp X_j | Z)$

$$I(X_i, X_j | Z) = \sum_{x_i, x_j, z} P(x_i, x_j, z) \cdot \log \frac{P(x_i, x_j | z)}{P(x_i | z) \cdot P(x_j | z)}$$

$I(X_i, X_j | Z) = 0 \Leftrightarrow (X_i \perp X_j | Z)$ ; since $P(x_i, x_j | z) = P(x_i | z) \cdot P(x_j | z)$

# Independence tests and the constraint based approach

- Using the data *D*
  - Empirical distribution: $\underline{\hat{P}(x_i, x_j)} \overset{\text{MLE}}{=} \frac{\text{Count}(x_i, x_j)}{m}$
  - Mutual information: $\underset{\text{MLE}}{\hat{I}(X_i, X_j)} = \sum_{x_i, x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$
  - Similarly for conditional MI

$t \to$ magic param.

- Use learning PDAG algorithm: $\hat{I}(X, Y \mid U) \overset{>t}{\underset{\leq t}{\gtrless}} 0$ no / yes!
  - When algorithm asks: (X⊥Y|**U**)?

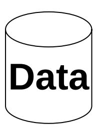
picking $t$ is about statistical significance

- Many other types of independence tests
  - See reading…



# Score-based approach

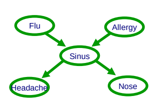
what score




<x_1^(1),…,x_n^(1)>
…
<x_1^(m),…,x_n^(m)>

**Possible structures**

DAG




F → A, S, H, N

S F A H N

S F A H N

many many

Learn parameters

**Score structure**

log space

−10 000

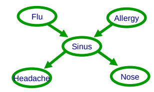
−12 000

−20 000

−10 500

# Information-theoretic interpretation of maximum likelihood

- Given structure, log likelihood of data:




$$\log P(\mathcal{D}\mid\theta_{\mathcal{G}},\mathcal{G}) = \sum_{j=1}^{m}\sum_{i=1}^{n}\log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)}\left[\mathbf{Pa}_{X_i}\right]\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}\log P\left(X_i = x_i^{(j)} \mid Pa_{X_i} = x^{(j)}[Pa_{X_i}]\right)$$

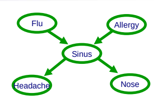
$$= m\sum_i \sum_{x_i,u} \underbrace{\frac{Count(X_i = x_i, Pa_{X_i} = u)}{m}}_{\hat{P}(X_i = x_i,\, Pa_{X_i} = u)} \log P(X_i = x_i \mid Pa_{X_i} = u)$$



# Information-theoretic interpretation of maximum likelihood 2

$\hat{H}$ estimated from counts

- Given structure, log likelihood of data:




$$\log \hat{P}(\mathcal{D}\mid\theta,\mathcal{G}) = m\sum_i \sum_{x_i,\mathbf{Pa}_{x_i,\mathcal{G}}} \hat{P}(x_i,\mathbf{Pa}_{x_i,\mathcal{G}})\log \hat{P}(x_i \mid \underbrace{\mathbf{Pa}_{x_i,\mathcal{G}}}_{=u})$$

$$= m\sum_i \left(-\hat{H}(X_i \mid Pa_{X_i})\right)$$

$$= -m\sum_i \hat{H}(X_i \mid Pa_{X_i})$$

$$\downarrow \sum_i \hat{H}(X_i \mid Pa_{X_i})$$

learn structure

$\Rightarrow$ for $X_i$ find very informative parents

Conditional Entropy:

$H(X \mid U)$

$= -\sum_{x,u} p(x,u)\log P(x\mid u)$

information of U about X
or
uncertainty of X after observing U

# Decomposable score

$I(X_i ; Pa_{X_i,G}) = H(X_i) - H(X_i | Pa_{X_i,G})$

- Log data likelihood $\neg(X_i \perp Pa_{X_i})$ [very not indep.]

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - \underbrace{m \sum_i \hat{H}(X_i)}_{\text{doesn't graph structure } G}$$

- Decomposable score:
  - Decomposes over families in BN (node and its parents)
  - Will lead to significant computational efficiency!!!
  - Score($G$ : $D$) = $\sum_i$ FamScore($X_i$|$\mathbf{Pa}_{X_i}$ : $D$)

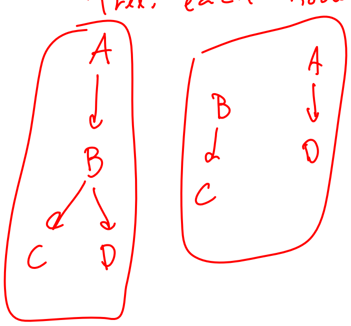
For MLE: $\text{FamScore}(X_i \mid Pa_{X_i} : D) = m\left[\hat{I}(X_i, Pa_{X_i}) - \hat{H}(X_i)\right]$



# How many trees are there?

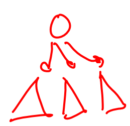
**Nonetheless – Efficient optimal algorithm finds best tree**

Tree: each node has zero or one parent




trees, not forests

- one root pick n
  - pick children
    - for each child another tree



$\sim 2^{\Theta(n \lg n)}$ → really large

# Scoring a tree 1: I-equivalent trees

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = M \sum_i \hat{I}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) - M \sum_i \hat{H}(X_i)$$

I-equiv, same score

Score(Tree) $= M \sum_{ij \in edges} I(X_i, X_j) - C$

$X \to Z \to Y$

$\cancel{I(X,\emptyset)} - H(X) + I(X,Z) - H(Z) + I(Z,Y) - H(Y)$

$I(X,Z) + I(Z,Y) + \underbrace{H(X) + H(Y) + H(Z)}_{\text{constant } C}$

$X \to Z \leftarrow Y$ not a tree

$X \leftarrow Z \to Y$

$I(X,Z) - H(X) + \cancel{I(Z;\emptyset)} - H(Z) + I(Y,Z) - H(Y)$

$X \leftarrow Z \leftarrow Y$

$I(X,Z) - H(X) + I(Z,Y) - H(Z) + \cancel{I(Y;\emptyset)} - H(Y)$



# Scoring a tree 2: similar trees

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = M \sum_i \hat{I}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) - M \sum_i \hat{H}(X_i)$$

$X \to Z \to Y$

$I(X,Z) + I(Z,Y) - C$

$Z \to Y \to X$

$I(Z,Y) + I(Y,X) - C$

Score(Tree) $= \sum_{ij \text{ edges}} I(X_i, X_j)$

best tree heaviest edges $\Rightarrow$ maximum spanning tree

difference

$I(X,Z) - I(Y,X)$

only change a few families $\to$ score only changed for these families

## Chow-Liu tree learning algorithm 1

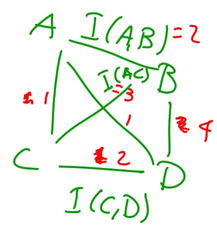
- For each pair of variables $X_i, X_j$
  - Compute empirical distribution:

$$\widehat{P}(x_i, x_j) \overset{MLE}{=} \frac{\text{Count}(x_i, x_j)}{m}$$

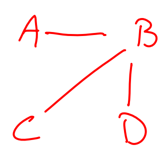
  - Compute mutual information:

$$\widehat{I}(X_i, X_j) = \sum_{x_i, x_j} \widehat{P}(x_i, x_j) \log \frac{\widehat{P}(x_i, x_j)}{\widehat{P}(x_i)\widehat{P}(x_j)}$$

- Define a graph
  - Nodes $X_1, \ldots, X_n$
  - Edge (i,j) gets weight $\widehat{I}(X_i, X_j)$




MST:

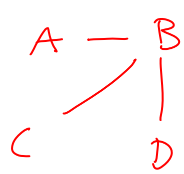




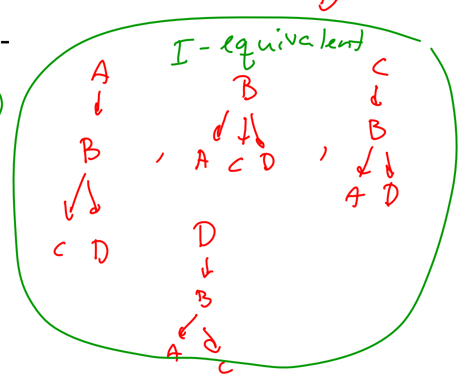

## Chow-Liu tree learning algorithm 2

$$\log \widehat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = M \sum_i \widehat{I}(x_i, \mathbf{Pa}_{x_i, \mathcal{G}}) - M \sum_i \widehat{H}(X_i)$$

- Optimal tree BN
  - Compute maximum weight spanning tree
  - Directions in BN: pick any node as root, breadth-first-search defines directions




Even though $2^{\theta(n \log n)}$ trees, found best in poly time.



# Can we extend Chow-Liu 1

- Tree augmented naïve Bayes (TAN) [Friedman et al. '97]
  - Naïve Bayes model overcounts, because correlation between features not considered
  - Same as Chow-Liu, but score edges with:

$$\hat{I}(X_i, X_j \mid C) = \sum_{c,x_i,x_j} \hat{P}(c, x_i, x_j) \log \frac{\hat{P}(x_i, x_j \mid c)}{\hat{P}(x_i \mid c)\hat{P}(x_j \mid c)}$$

OPTIMAL

TAN:


each $X_i$ (except for root) has two parents $C$ and some $X_j$

C → $X_1$ ... $X_n$ Naive Bayes $(X_i \perp X_j \mid C)$

what if $X_1 = X_2$ double count

Explain $X_2$ by:

C → $X_1$, C ⇢ $X_2$, $X_1 \to X_2$

decreases or avoids double counting



# Can we extend Chow-Liu 2

- (Approximately learning) models with tree-width up to *k*
  - [Narasimhan & Bilmes '04]
  - But, $O(n^{k+1})$…
    - and more subtleties

## What you need to know about learning BN structures so far

- Decomposable scores
  - Maximum likelihood Score
  - Information theoretic interpretation
- Best tree (Chow-Liu)
- Best TAN
- Nearly best k-treewidth (in $O(N^{k+1})$)



## Announcements

- Homework 2 out
  - Due Oct. 11th
- Project description out next week

## Maximum likelihood score overfits!

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = M \sum_i \hat{I}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) - M \sum_i \hat{H}(X_i)$$

- Information never hurts:
- Adding a parent always increases score!!!



## Bayesian score

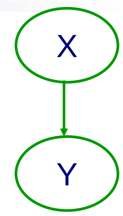
- Prior distributions:
  - Over structures
  - Over parameters of a structure
- Posterior over structures given data:

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

# Bayesian score and model complexity

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

True model:




P(Y=t|X=t) = 0.5 + $\alpha$
P(Y=t|X=f) = 0.5 - $\alpha$

- Structure 1: X and Y independent
  - Score doesn't depend on alpha
- Structure 2: X $\rightarrow$ Y
  - Data points split between P(Y=t|X=t) and P(Y=t|X=f)
  - For fixed M, only worth it for large $\alpha$
    - Because posterior over parameter will be more diffuse with less data



# Bayesian, a decomposable score

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

- As with last lecture, assume:
  - Local and global parameter independence
- Also, prior satisfies **parameter modularity**:
  - If $X_i$ has same parents in *G* and *G'*, then parameters have same prior
- Finally, structure prior P(*G*) satisfies **structure modularity**
  - Product of terms over families
  - E.g., P(*G*) $\propto$ $c^{|G|}$
- Bayesian score decomposes along families!

## BIC approximation of Bayesian score

- Bayesian has difficult integrals
- For Dirichlet prior, can use simple Bayes information criterion (BIC) approximation
  - In the limit, we can forget prior!
  - **Theorem**: for Dirichlet prior, and a BN with Dim(*G*) independent parameters, as $M \to \infty$:

$$\log P(D \mid \mathcal{G}) = \log P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) - \frac{\log M}{2}\mathrm{Dim}(\mathcal{G}) + O(1)$$



## BIC approximation, a decomposable score

- BIC: $\mathrm{Score}_{\mathrm{BIC}}(\mathcal{G} : D) = \log P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) - \frac{\log M}{2}\mathrm{Dim}(\mathcal{G})$
- Using information theoretic formulation:

$$\mathrm{Score}_{\mathrm{BIC}}(\mathcal{G} : D) = M \sum_i \hat{I}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}}) - M \sum_i \hat{H}(X_i) - \frac{\log M}{2} \sum_i \mathrm{Dim}(P(X_i \mid \mathbf{Pa}_{x_i,\mathcal{G}}))$$

## 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 score?



## Priors for general graphs

- For finite datasets, prior is important!
- Prior over structure satisfying prior modularity
- What about prior over parameters, how do we represent it?
  - *K2 prior*: fix an $\alpha$, $P(\theta_{X_i|\mathbf{Pa}_{X_i}}) = \text{Dirichlet}(\alpha, \ldots, \alpha)$
  - K2 is "inconsistent"

## BDe prior

- Remember that Dirichlet parameters analogous to “fictitious samples”
- Pick a fictitious sample size m’
- For each possible family, define a prior distribution $P(X_i, \mathbf{Pa}_{X_i})$
  - Represent with a BN
  - Usually independent (product of marginals)
- **BDe prior**:
- Has “consistency property”:



## Score equivalence

- If *G* and *G’* are I-equivalent then they have same score
- **Theorem 1**: Maximum likelihood score and BIC score satisfy score equivalence
- **Theorem 2**:
  - 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 $w_{Xj \to Xi}$ is advantage of adding $X_j$ as parent for $X_i$

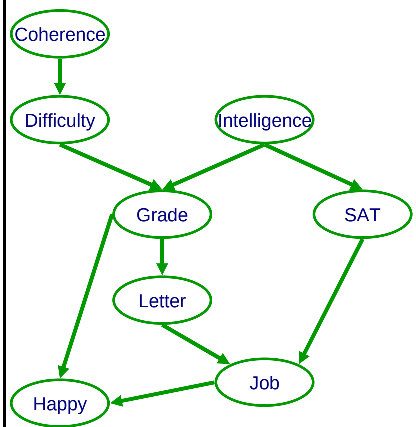
- Now have a directed graph, need directed spanning forest
  - Note that adding an edge can hurt Bayesian score – choose forest not tree
  - But, if score satisfies score equivalence, then $w_{Xj \to Xi} = w_{Xj \to Xi}$ !
  - Simple maximum spanning forest algorithm works



## 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






## Fixed variable order 1

- 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

# Fixed variable order 2

- Fix max number of parents to 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
    - $\mathbf{Pa}_{X_i}$ 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



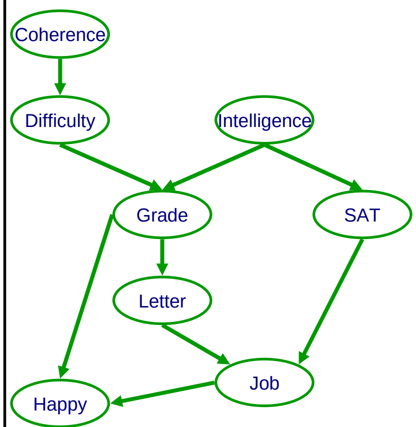
# Learn BN structure using local search

**Starting from Chow-Liu tree**

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

**Select using favorite score**

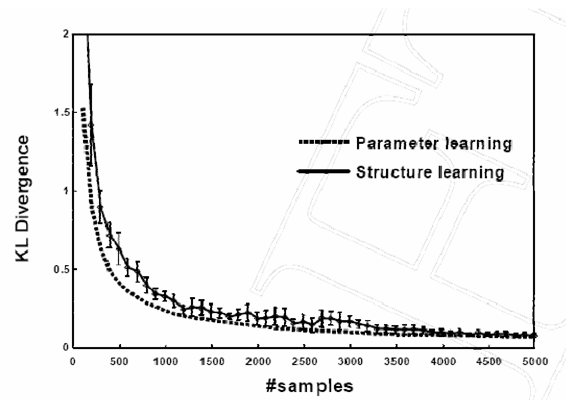
# Exploit score decomposition in local search




- Add edge and delete edge:
  - Only rescore one family!
- Reverse edge
  - Rescore only two families



# Some experiments




Alarm network

## 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
- 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
  - Data likelihood
  - 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