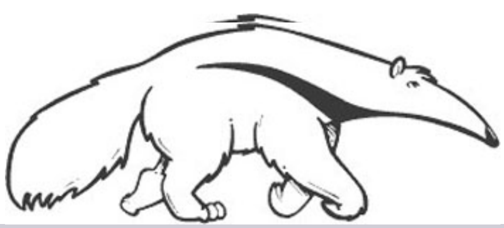
Structure Learning in Bayes Nets

Learning in Graphical Models

Prof. Alexander Ihler



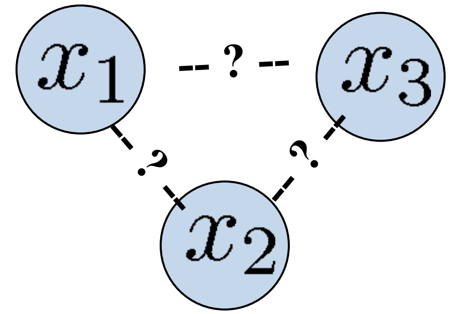
BREN:ICS
INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

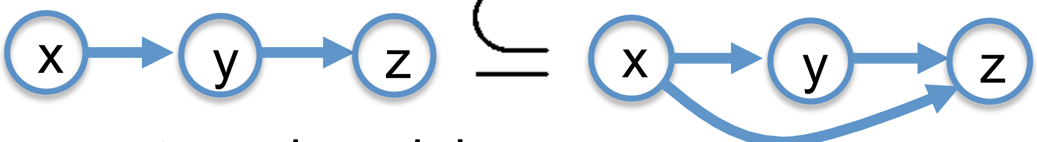
Structure learning

- Unknown structure: Select by ML also?

$$\max_G \max_{\theta_G} \log p(\{x^{(i)}\}\,;\,G,\theta_G)$$

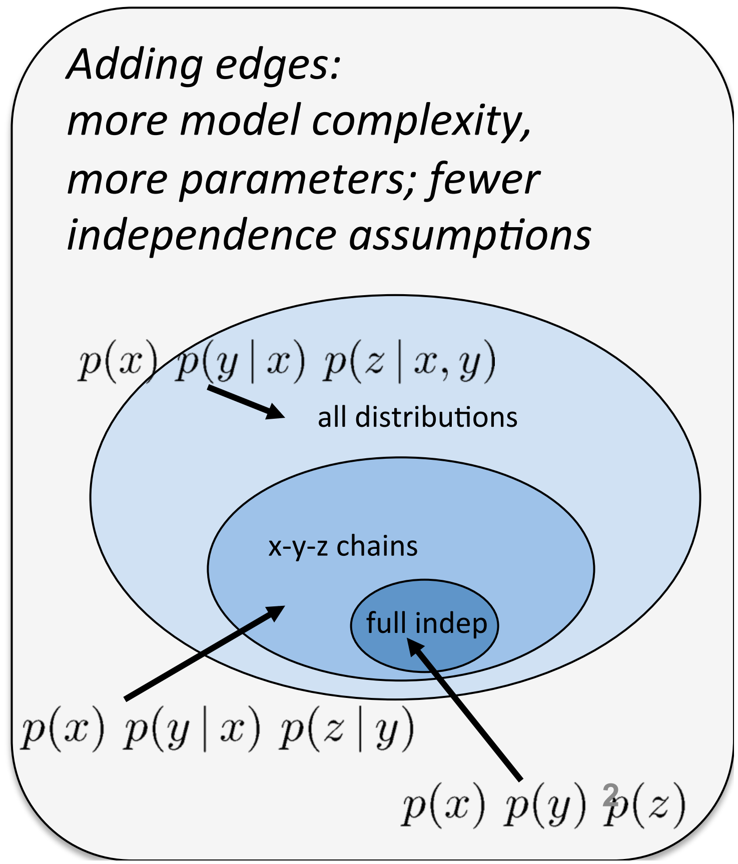


$p(x)\ p(y\,|\,x)\ p(z\,|\,y) \quad \subseteq \quad p(x)\ p(y\,|\,x)\ p(z\,|\,x,y)$



- Nested models
- ML structure is the complete graph
  - # parameters? Overfitting?

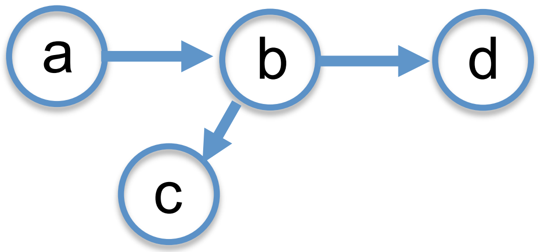
- Options:
  - Compare equal complexity (best tree…)
  - Use hold-out data
  - Use complexity penalty (BIC, …)
  - Use prior & MAP parameters

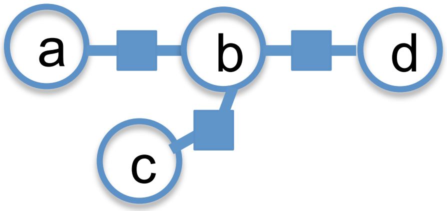


Tree-structured Bayes Nets

- Trees
  - No undirected cycles; single root node

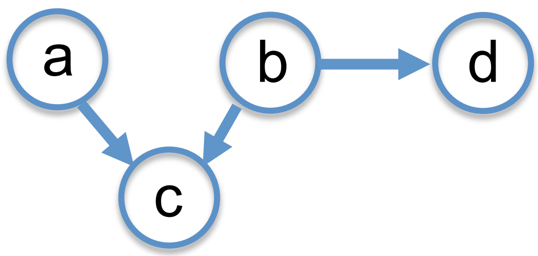
    $\Rightarrow$ Only pairwise interactions

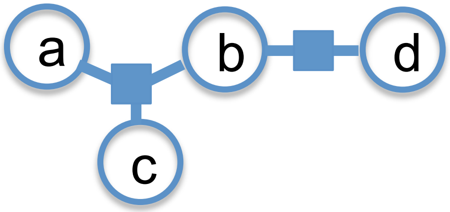




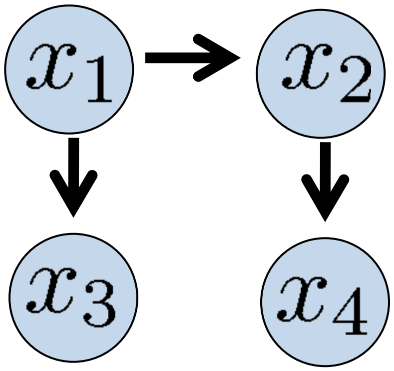
- Poly-trees
  - No undirected cycles; multiple roots

    $\Rightarrow$ Non-pairwise interactions





Generalizing to trees



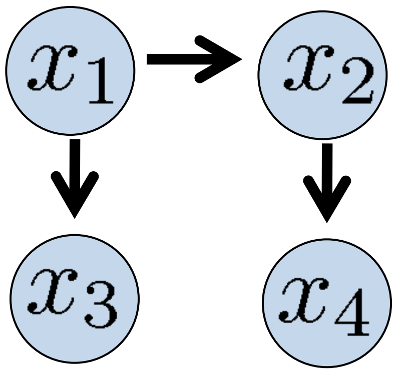
$$\mathcal{L} = \sum_i \log \left[p(x_1^i)\, p(x_2^i|x_1^i)\, p(x_3^i|x_1^i)\, p(x_4^i|x_2^i) \right]$$

- Suppose
  - Known structure, exp family
  - Fully observed data
- Then,
  - ML estimate given as before (fit each term)
  - Conditional probabilities equal their empirical estimates

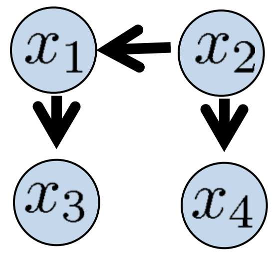
$$\max_\theta \mathcal{L} = \sum_i \log \left[\hat{p}(x_1^i)\, \hat{p}(x_2^i|x_1^i)\, \hat{p}(x_3^i|x_1^i)\, \hat{p}(x_4^i|x_2^i) \right]$$

Score different structures

Why x_1 centric view?

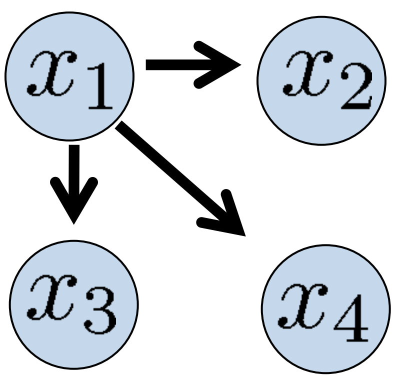


$$\max_{\theta_{G_1}} \mathcal{L} = \sum_i \log \left[\hat{p}(x_1^i)\, \hat{p}(x_2^i|x_1^i)\, \hat{p}(x_3^i|x_1^i)\, \hat{p}(x_4^i|x_2^i) \right]$$



$$\max_{\theta} \mathcal{L} = \sum_i \log \left[\hat{p}(x_2^i)\, \hat{p}(x_1^i|x_2^i) \right) \ldots$$

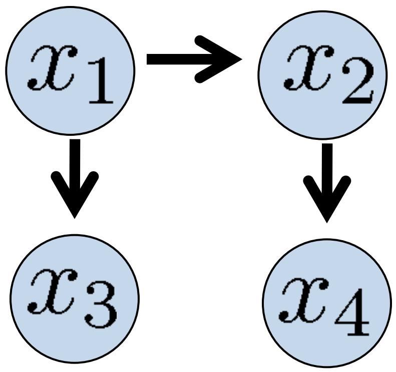
Exactly the same



$$\max_{\theta_{G_2}} \mathcal{L} = \sum_i \log \left[\hat{p}(x_1^i)\, \hat{p}(x_2^i|x_1^i)\, \hat{p}(x_3^i|x_1^i)\, \hat{p}(x_4^i|x_1^i) \right]$$

Choose structure G with highest likelihood

A more symmetric view



$$p(x_1^i)\, p(x_2^i|x_1^i) = p(x_2^i)\, p(x_1^i|x_2^i) = p(x_1^i)\, p(x_2^i)\, \frac{p(x_1^i, x_2^i)}{p(x_1^i)\, p(x_2^i)}$$

Then,

$$\mathcal{L}^* = \sum_i \log \left[\hat{p}(x_1^i)\hat{p}(x_2^i)\hat{p}(x_3^i)\hat{p}(x_4^i)\right] + \sum_i \log \frac{\hat{p}(x_1^i, x_2^i)}{\hat{p}(x_1^i)\hat{p}(x_2^i)} + \sum_i \ldots$$

Present in all models

Present in models with an edge (1,2)

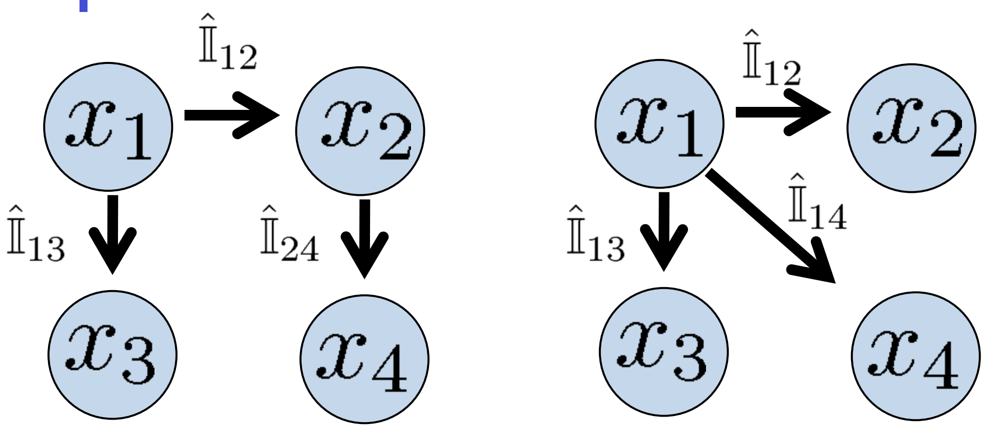
Now, reorganize sum over data samples by their value:

$$\sum_i \log \left[\hat{p}(x_1^i)\right] = m \sum_{x_1} \hat{p}(x_1) \log \hat{p}(x_1) = m\hat{\mathbb{H}}(x_1)$$

$$\sum_i \log \frac{\hat{p}(x_1^i, x_2^i)}{\hat{p}(x_1^i)\hat{p}(x_2^i)} = m \sum_{x_1, x_2} \hat{p}(x_1, x_2) \log \frac{\hat{p}(x_1, x_2)}{\hat{p}(x_1)\hat{p}(x_2)} = m\hat{\mathbb{I}}(x_1, x_2)$$

Score different structures

Chow & Liu, 1968



…

- Compute scores I_{ij} for all pairs (ij)
- Maximize the sum of terms in the tree
- Max-weight spanning tree problem
  - Find largest weight that connects two disconnected components

$$\hat{\mathbb{I}}(x_1, x_2) = \mathbb{E}_D\left[\log \frac{\hat{p}(x_1, x_2)}{\hat{p}(x_1)\hat{p}(x_2)}\right]$$

- I_{ij} is the mutual information of the empirical model \hat p
  - KL-divergence from the independent model

BIC-penalized scores

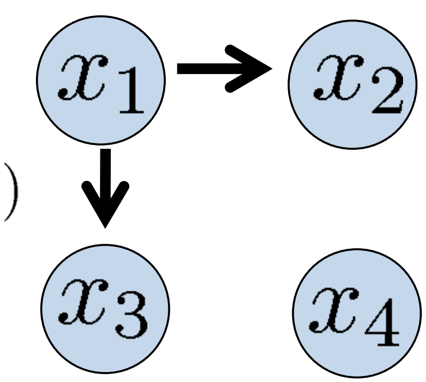
- BIC: Bayesian Information Criterion
- Penalize log-likelihood score by complexity, k:

$$\mathrm{BIC} = \mathcal{L}^* - \frac{k}{2}\log m = \Big(\max_\theta \log p(\{x^{(j)}\};\theta)\Big) - \frac{k}{2}\log m$$

- AIC: Aikike Information Criterion

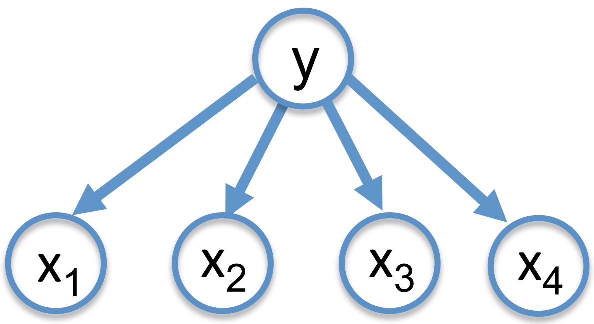
$$\mathrm{AIC} = \mathcal{L}^* - k \qquad\qquad \mathrm{AICc} = \mathcal{L}^* - k - \frac{k(k+1)}{m-k-1}$$

- Ex: BIC-penalized Chow-Liu
  - Score by $\hat{\mathbb{I}}(x_1, x_2) - \frac{\log m}{2m}(d_1 d_2 - d_1 - d_2 + 1)$
  - Note: score can be negative $\Rightarrow$ select forest



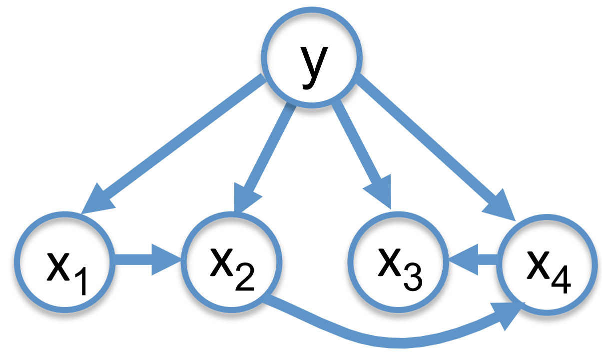
Tree-augmented Naïve Bayes

Naïve Bayes



$$p(x|y=c) = \prod_i p(x_i|y=c)$$

TAN Bayes (Friedman et al. 1997)



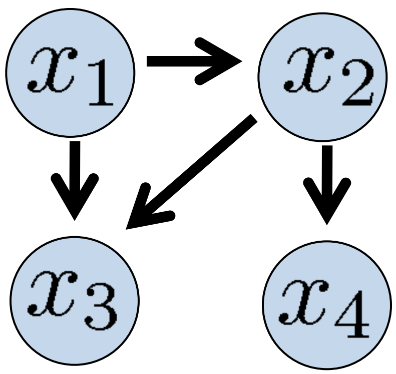
$$p(x|y=c) = \prod_i p(x_i|x_{\mathrm{pa}(i)}, y=c)$$

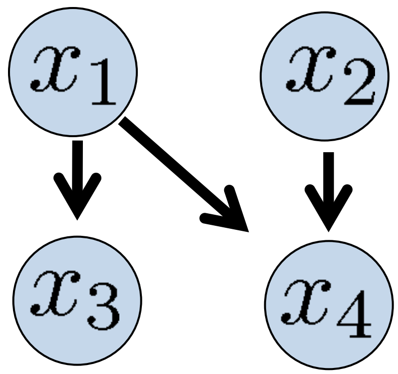
- Naïve Bayes: model features independently given class
- Correlated features: can overcount evidence (e.g., x_1=x_2)

- TAN Bayes: account for simple model over x
  - Score $\hat{\mathbb{I}}(x_1, x_2|y) = \sum_{x_1,x_2,y} \hat{p}(x_1, x_2, y) \big[\log \frac{\hat{p}(x_1, x_2|y)}{\hat{p}(x_1|y)\hat{p}(x_2|y)} \big]$

- Also easy to make graph G depend on y

Learning Bayes net structures



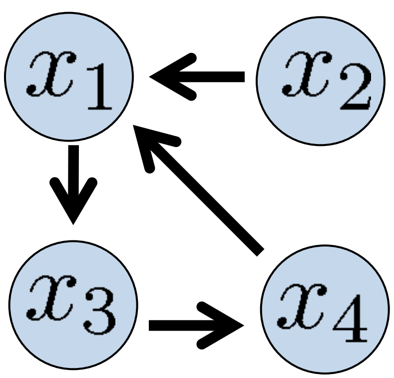


…

For any BN structure & fully observed data, still easy to

1. Compute ML estimates
2. Score a structure (e.g., penalized ML)

So?



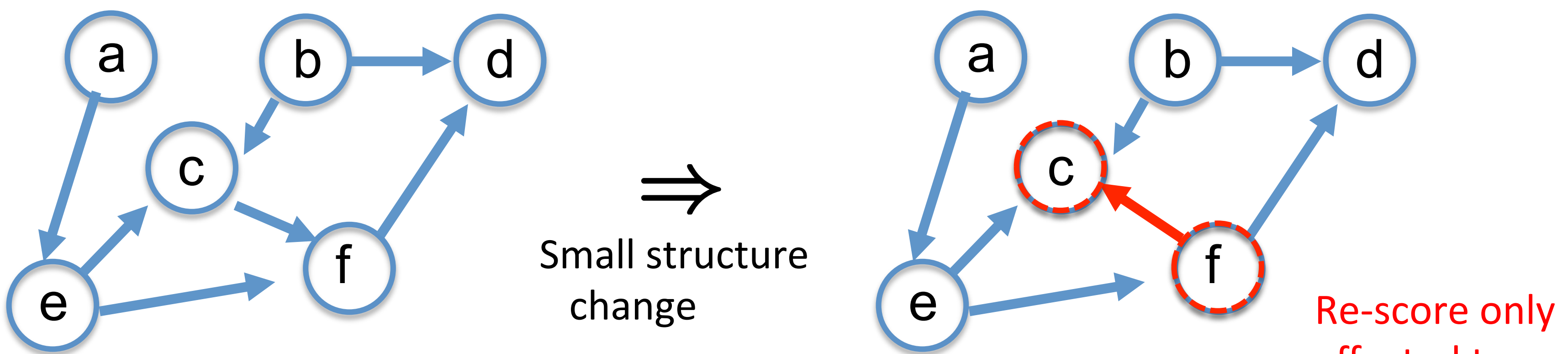
Not consistent with any variable order
(i.e., no conditional decomposition)

Ordering and parent "constraints" are hard to describe compactly, and hard to search over

Local search over structures

- Many scores (e.g. penalized likelihood) *decompose* on G

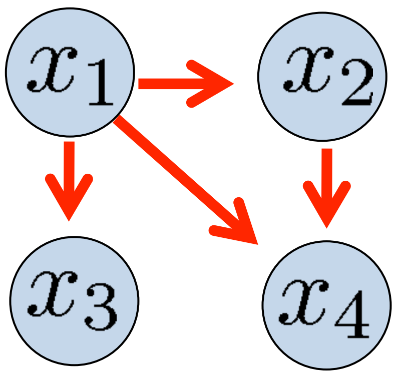
$$S(\{x^{(j)}\}\,;\,G) = \sum_i S(\{x_i^{(j)}, x_{\mathrm{pa(i)}}^{(j)}\}\,;\,G)$$



- Search locally over structures
  - Hill-climbing, stochastic search, MCMC, …
- Works even with fairly general priors on G, etc.

Exhaustive search over structures

- Suppose we have ordering 1,2,3,4



Just try all the possible parent sets

(Easy to restrict by model complexity, e.g. all parent sets of size < 3, etc.)

(1) Has no parents (no earlier variables…)
(2) Score $p(x_2)$ vs $p(x_2 \mid x_1)$ with penalized likelihood
(3) Score $p(x_3)$ vs $p(x_3 \mid x_1)$ vs $p(x_3 \mid x_2)$ vs $p(x_3 \mid x_1, x_2)$ …
(4) Score $p(x_4)$ vs …

- Now, just enumerate over all possible orders

Linear program over structures

- Score all possible (conditional probability) factors

$$\log p(x_1),\quad \log p(x_1|x_2),\quad \dots\quad \log p(x_1 \mid x_3, x_4),\quad \dots$$
$$\log p(x_2),\quad \log p(x_2|x_1),\quad \dots\quad \log p(x_2 \mid x_3, x_4),\quad \dots$$
$$\vdots$$
$$\log p(x_4),\quad \log p(x_4|x_1),\quad \dots\quad \log p(x_4 \mid x_2, x_3),\quad \dots$$

- Our model score is the sum of the terms we include
  - But some terms are incompatible with others…
- Set this up as an integer linear program
  - Maximize sum of included terms, subject to (lots of structure restrictions)
- Cutting plane methods:
  - Solve with few constraints
  - Check if any cycles exist
  - If so, add those constraints and re-solve