# Bayesian Model Selection

Chris Williams

School of Informatics, University of Edinburgh

November 2008



## Overview

- Bayesian Learning of CPTs
- Dealing with Multiple Models
- Other Scores for Model Comparison
- Searching over Belief Network structures
- Readings: Bishop §3.4, Heckerman tutorial sections 1, 2, 3, 4, 5, 7, 8.1, 11



## Learning in Belief Networks

| | Known Structure | Unknown Structure |
|---|---|---|
| Complete Data | Statistical parameter estimation | Discrete search over structures |
| Incomplete Data | EM, stochastic sampling methods | Combined search over structures and parameters |

(Friedman and Goldszmidt, 1998)

- Data + prior/expert beliefs ⇒ Belief networks



## Bayesian Learning with Complete Data

- Belief network with $m$ nodes, $x_1, \ldots, x_m$, parameters $\boldsymbol{\theta}$
- Log likelihood

$$L(\boldsymbol{\theta}; D) = \sum_{i=1}^{n} \log p(x_1^i, \ldots, x_m^i | \boldsymbol{\theta})$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log p(x_j^i | pa_j^i, \theta_j)$$

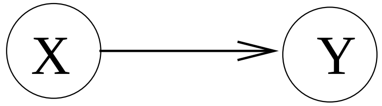
- The likelihood decomposes according to the structure of the network
- ⇒ independent estimation problems for MLE

- If priors for each CPT are independent, so are posteriors
- Posterior for each multinomial CPT $P(X_j|Pa_j)$ is Dirichlet with parameters

$$\alpha(X_j = 1|pa_j) + n(X_j = 1|pa_j), \ldots, \\ , \alpha(X_j = r|pa_j) + n(X_j = r|pa_j)$$

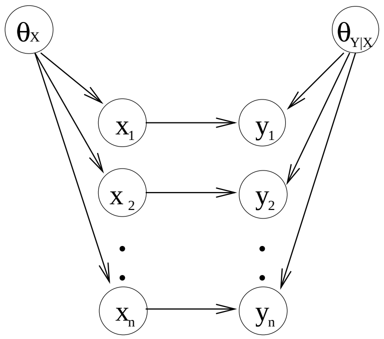


## Example: $X \rightarrow Y$




- Parameters $\theta_X$, $\theta_{Y|X}$






- Read off from network: complete data $\implies$ posteriors for $\theta_X$ and $\theta_{Y|X}$ are independent
- Reduces to 3 separate thumbtack-learning problems



## Dealing with Multiple Models

- Let $M$ index possible model structures, with associated parameters $\boldsymbol{\theta}_M$

$$p(M|D) \propto p(D|M)p(M)$$

- For complete data (plus some other assumptions) the marginal likelihood $p(D|M)$ can be computed in closed form
- Making predictions

$$\begin{aligned} p(\mathbf{x}_{n+1}|D) &= \sum_M p(M|D) p(\mathbf{x}_{n+1}|M, D) \\ &= \sum_M p(M|D) \int p(\mathbf{x}_{n+1}|\boldsymbol{\theta}_M, M) p(\boldsymbol{\theta}_M|D, M)\, d\boldsymbol{\theta}_M \end{aligned}$$

- Can approximate $\sum_M$ by keeping the best or the top few models

## Comparing models

$$\text{Bayes factor} = \frac{P(D|M_1)}{P(D|M_2)}$$

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(M_1)}{P(M_2)} \cdot \frac{P(D|M_1)}{P(D|M_2)}$$
$$\text{Posterior ratio} = \text{Prior ratio} \times \text{Bayes factor}$$

Strength of evidence from Bayes factor (Kass, 1995; after Jeffreys, 1961)

| | |
|---|---|
| 1 to 3 | Not worth more than a bare mention |
| 3 to 20 | Positive |
| 20 to 150 | Strong |
| $>$ 150 | Very strong |



## Computing $P(D|M)$

- For the thumbtack example

$$p(D|M) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{i=1}^{r} \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)}$$

- The graph $X \to Y$ corresponds to 3 separate thumbtack problems for $X$, $Y|X = \textit{heads}$ and $Y|X = \textit{tails}$



General form of $P(D|M)$ for a discrete belief network

$$p(D|M) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}$$

where

- $n_{ijk}$ is the number of cases where $X_i = x_i^k$ and $Pa_i = pa_i^j$
- $r_i$ is the number of states of $X_i$
- $q_i$ is the number of configurations of the parents of $X_i$

$$\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk} \qquad n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$$

- Formula due to Cooper and Herskovits (1992)
- Simply the product of the thumbtack result over all nodes and states of the parents



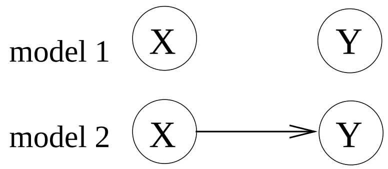
## Computation of Marginal Likelihood

Efficient closed form if

- No missing data or hidden variables
- Parameters are independent in prior
- Local distributions are in the exponential family (e.g. multinomial, Gaussian, Poisson, ...)
- Conjugate priors are used

## Example

Given data $D$, compare the two models




Counts: $hh = 6$, $ht = 2$, $th = 8$, $tt = 4$, from marginal probabilities $P(X = h) = 0.4$ and $P(Y = h) = 0.7$

Bayes factor $= \frac{P(D|M_1)}{P(D|M_2)} = 1.97$ in favour of model 1

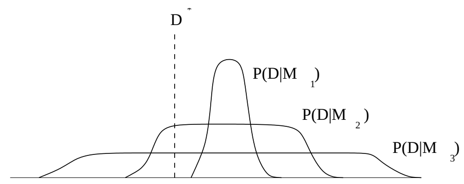
Log Likelihood criterion favours model 2

$\log L(M1) - \log L(M2) = -0.08$



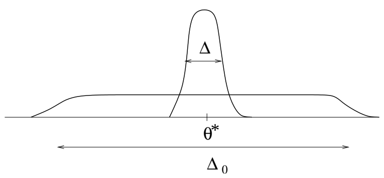
## How Bayesian model comparison works

- Consider three models $M_1$, $M_2$ and $M_3$ which are under complex, just right and over complex for a particular dataset $D^*$
- Note that $P(D|M_i)$ must be *normalized*




- Warning: it can make sense to use a model with an infinite number of parameters (but in a way that the prior is "nice")






- Another view (for a single parameter $\theta$)

$$\begin{aligned} P(D|M_i) &= \int p(D|\theta, M_i)p(\theta|M_i)d\theta \\ &\simeq p(D|\theta^*, M_i)p(\theta^*|M_i)\Delta \\ &\simeq p(D|\theta^*, M_i)\frac{\Delta}{\Delta_0} \end{aligned}$$

- This last term is known as an *Occam factor*
- The analysis can be extended to multidimensional $\boldsymbol{\theta}$. Pay an Occam factor on each dimension if parameters are well-determined by data; thus models with more parameters can be penalized more



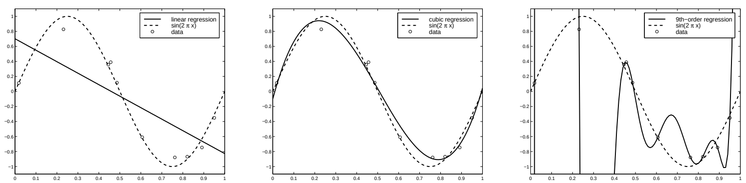
## Other scores for comparing models

Above we have used $P(D|M)$ to score models. Other ideas include

- **Maximum likelihood**

$$L(M; D) = \max_{\theta_M} L(\theta_M, M; D)$$

- Bad choice: adding arcs always helps
- Example from supervised learning



- **Penalize More Complex Models**: e.g. AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), Structural Risk Minimization (penalize hypothesis classes based on their VC dimension). BIC can be seen as large $n$ approximation ot full Bayesian method.
- **Minimum description length**: (Rissanen, Wallace) closely related to Bayesian method
- **Restrict the hypothesis space** to limit the capability for overfitting: but how much?
- **Holdout/Cross-validation**: validate generalization on data withheld during training—but this “wastes” data . . .

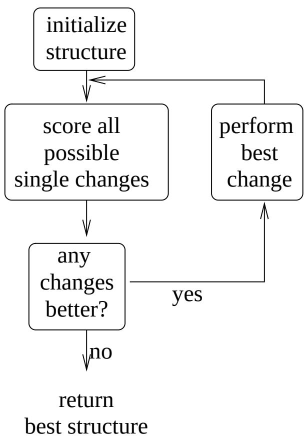


## Searching over structures

- Number of possible structures over $m$ variables is super-exponential in $m$
- Finding the BN with the highest marginal likelihood among those structures with at most $k$ parents is NP-hard if $k > 1$ (Chickering, 1995)
- Note: efficient search over trees
- Otherwise, use heuristic methods such as greedy search



## Greedy search






## Example

College plans of high-school seniors (Heckerman, 1995/6).

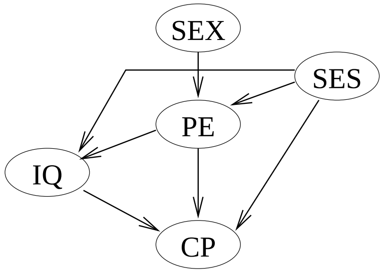
Variables are

- Sex: male, female
- Socioeconomic status: low, low mid, high mid, high
- IQ: low, low mid, high mid, high
- Parental encouragement: low, high
- College plans: yes, no

Priors

- **Structural prior**: SEX has no parents, CP has no children, otherwise uniform
- **Parameter prior**: Uniform distributions

Best network found




- Odd that SES has a direct link to IQ: suggests that a hidden variable is needed
- Searching over structures for visible variables is hard; inferring hidden structure is even harder...



Acknowledgements: this presentation has been greatly aided by the tutorials by Nir Friedman and Moises Goldszmidt
`http://www.erg.sri.com/people/moises/tutorial/index.htm`
ans David Heckerman
`http://research.microsoft.com/~heckerman/`