Model Comparison

Rich Turner (turner@gatsby.ucl.ac.uk)

Gatsby Unit, 16/09/2005

Bayesian Model comparison (averaging)

- **Bayesian statistics makes coherrent inferences from the data based on explicit modeling assumptions**: there is no need for extra complexity control.

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)} \quad (1)$$

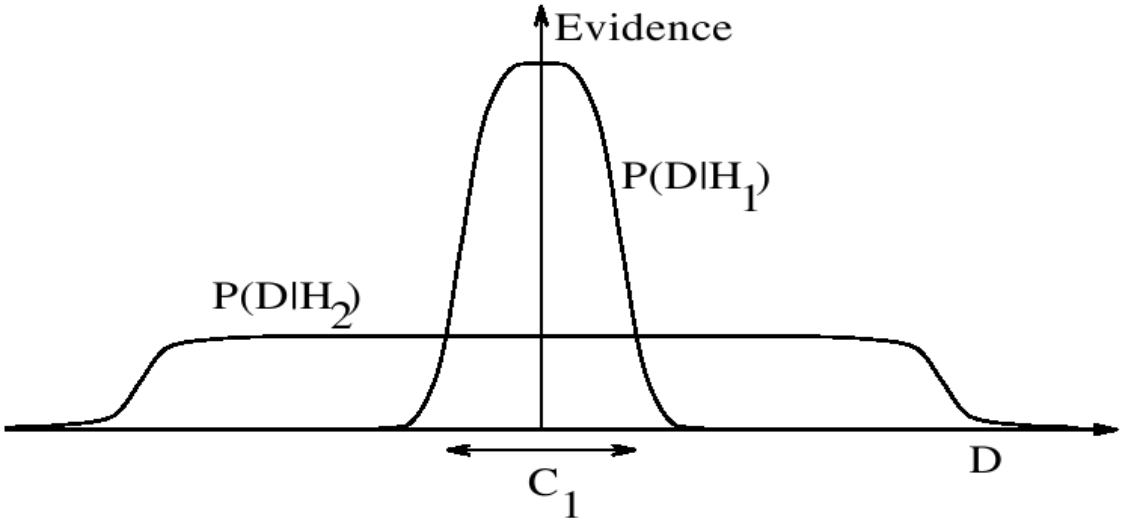
$$P(D) = \sum_i P(D|H_i)P(H_i) \quad (2)$$

$$P(D|H_i) = \sum_j P(D|H_i, \theta_j)P(\theta_j|H_i) \quad (3)$$

- **Bayesian methods incorporate Occam's razor automatically, penalising more complex models** which have more parameters.

David's intuition

"A simple model H_1 makes only a limited range of predictions; a more powerful model H_2 that has, for example, more free parameters than H_1 is able to predict a greater variety of data-sets."



Note: the ambiguity in the ordering of the data-set axis.

Two different sorts of hierarchy

The hierarchy implicit in the above picture:

- **Simple models** have **scrunched up distributions** $P(D|H)$ and data sets are therefore either very likely or very unlikely.

- More **complex model** have more knobs to twidle and **correspond to flatter marginal likelihoods**: No data set is particularly unlikely under a complex model.

But what if you could construct a hierarchy where:

- A **simple model corresponded to a very flat evidence** (all data sets equally likely under the model)

- **More complex models** have more knobs on that you can twidle to **scrunch up** the distribution (complicated data-sets very probable under the model)

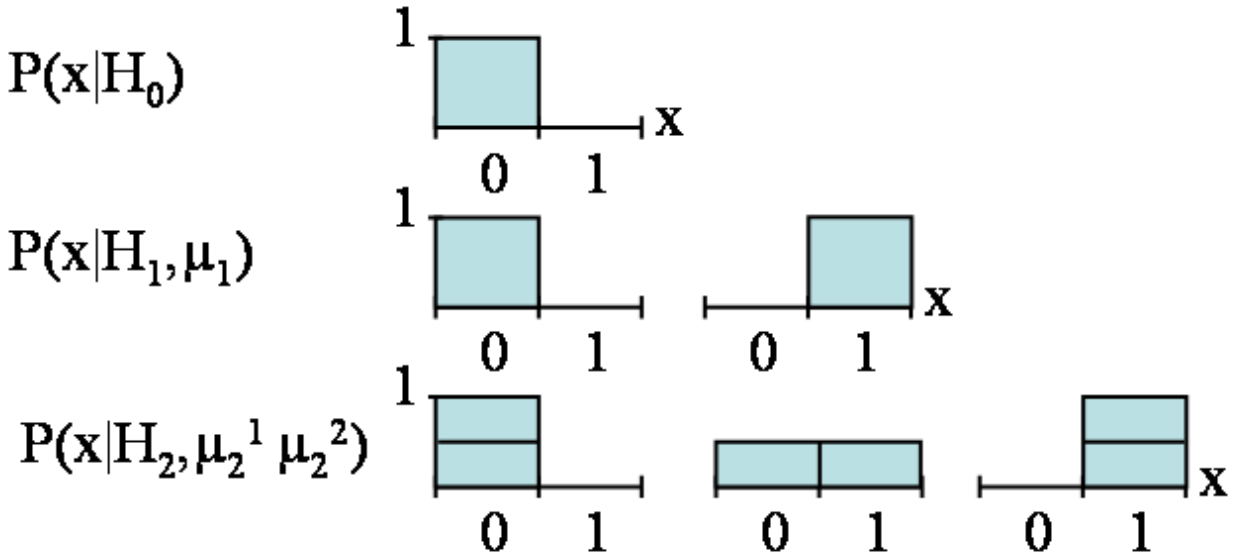
Toy Examples

Look at a couple of toy examples to lend some intuition.
The criteria these toy models should meet are:

- We want to **enumerate all possible data sets efficiently**
- We want there to be a **natural choice for the prior over parameters** and summing (integrating) them out to be easy.

Example 1

- **Most complex model** = abstraction of a **mixture of two Gaussians**, with the **two means as the free parameters**.



Mathematically

- The 'means' (μ) are binary variables
- Put a **uniform prior over parameter settings**.

$$P(x|H_0) = \delta_{1,x} \tag{4}$$

$$P(x|H_1, \mu_1) = \delta_{\mu_0,x} \tag{5}$$

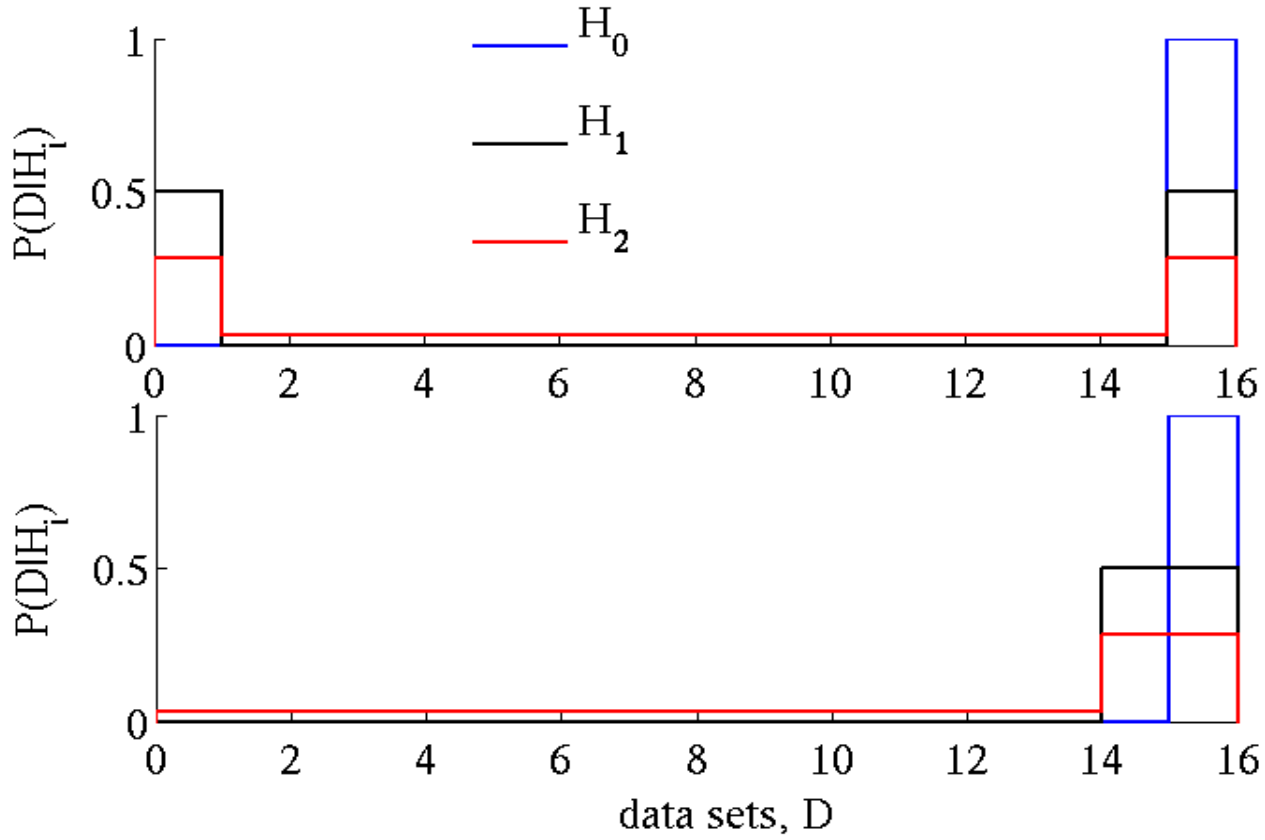
$$P(x|H_2, \mu_2^1, \mu_2^2) = \frac{1}{2}\left[\delta_{\mu_2^1,x} + \delta_{\mu_2^1,x}\right] \tag{6}$$

$$\tag{7}$$

The marginal likelihoods

- Under the simplest model, only one data-set has high probability.

- **Entropy of the marginal likelihood grows with complexity of the models**.



Example 2

- You get **two vectors** (both length N) of **binary data**.
- Does the data come from:
  - Two bent coins (H_2)
  - one bent coin (H_1),
  - one normal coin (H_0) ?

$$P(x_i|H_0) = \frac{1}{2} \tag{8}$$

$$P(x_i|H_1, \beta_1) = \beta_1^{x_i}(1-\beta_1)^{1-x_i} \tag{9}$$

$$P(x_i|H_2, \beta_2^i) = \beta_i^{x_i}(1-\beta_i)^{1-x_i} \tag{10}$$

Computing the marginal likelihoods

- Use a uniform distribution for the prior on β.
- Define the **sufficient statistics**: r_i = number of 1s in the i^{th} vector, $R = \sum r_i$ number of 1s in the entire data set

$$P(\{\mathbf{x}\}_{n=1}^{N}|H_2) = \prod_{i=1}^{2} \int \beta_i^{r_i}(1-\beta_i)^{N-r_i} d\beta_i \tag{11}$$

$$= \prod_{i=1}^{2} \frac{r_i!(N-r_i)!}{(N+1)!} \tag{12}$$

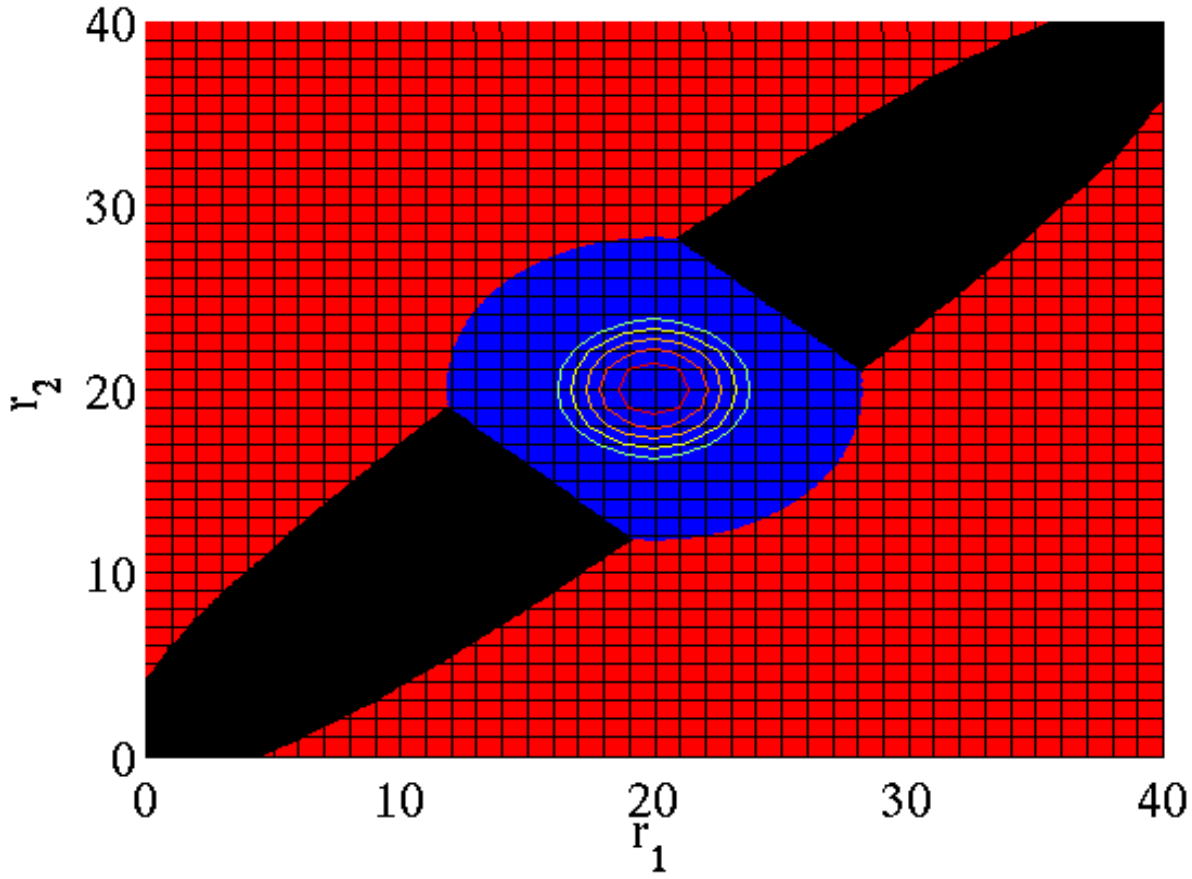
$$= \prod_{i=1}^{2} \frac{1}{N+1} \frac{1}{{}^{N}C_{r_i}} \tag{13}$$

$$P(\{\mathbf{x}\}_{n=1}^{N}|H_1) = \frac{1}{2N+1} \frac{1}{{}^{2N}C_{R}} \tag{14}$$

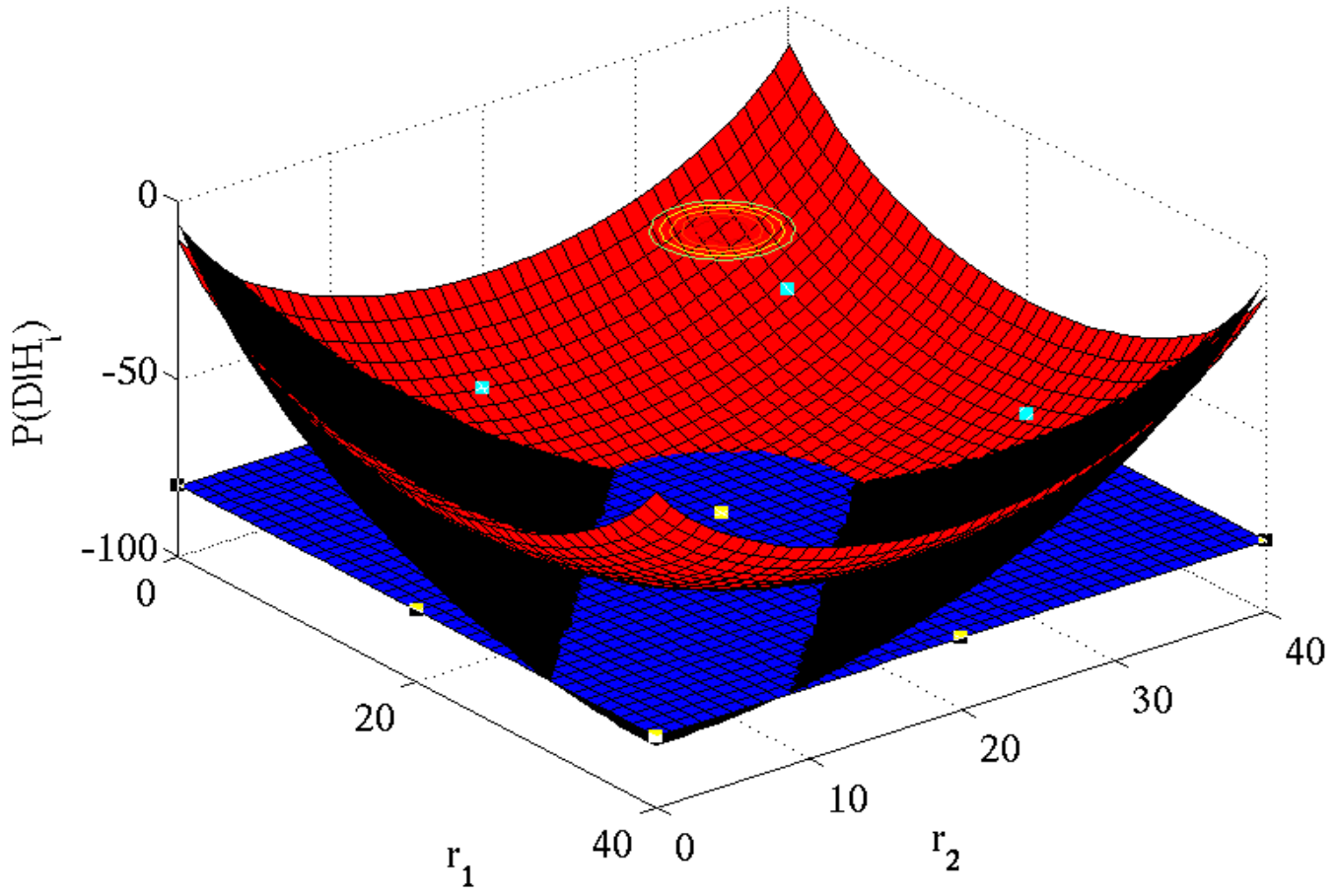
$$P(\{\mathbf{x}\}_{n=1}^{N}|H_0) \quad = \quad \left(\frac{1}{2}\right)^{2N} \tag{15}$$

- **Easy to completely enumerate large sets of data-sets** by enumerating the sufficient statistics (counting in base N up to $N^2 - 1$)
- Number of strings with sufficient statistics R is ${}^{N}C_{R}$
- Note how the **probability of an un-ordered string is independent of the sufficient statistics** (number unordered strings $= 2N + 1$).

The marginal likelihoods 1



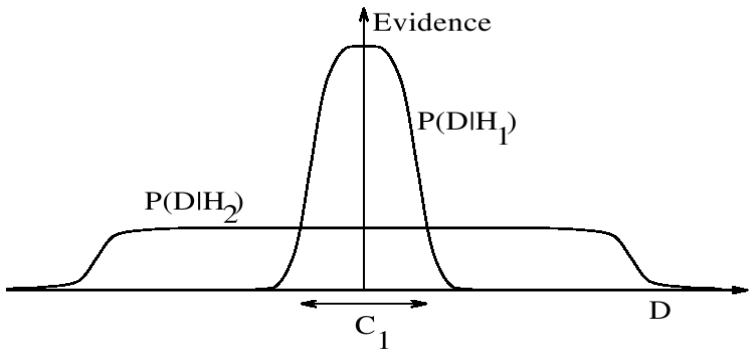
The marginal likelihoods 2



Summary

- In the first example the **entropy of the marginal likelihood increased with the complexity of the model**.

- In the second example the **entropy decreases with the complexity of the model**.



Moral: Be careful in interpreting David's picture -it's just a simple case. Maybe it's best not to talk about complex and simple models, perhaps characterisation in terms of their parameter numbers would be best. When we draw David's figure, we should draw two general distributions as well.