

Unsupervised Learning

Bayesian Learning of Model Structure

Zoubin Ghahramani

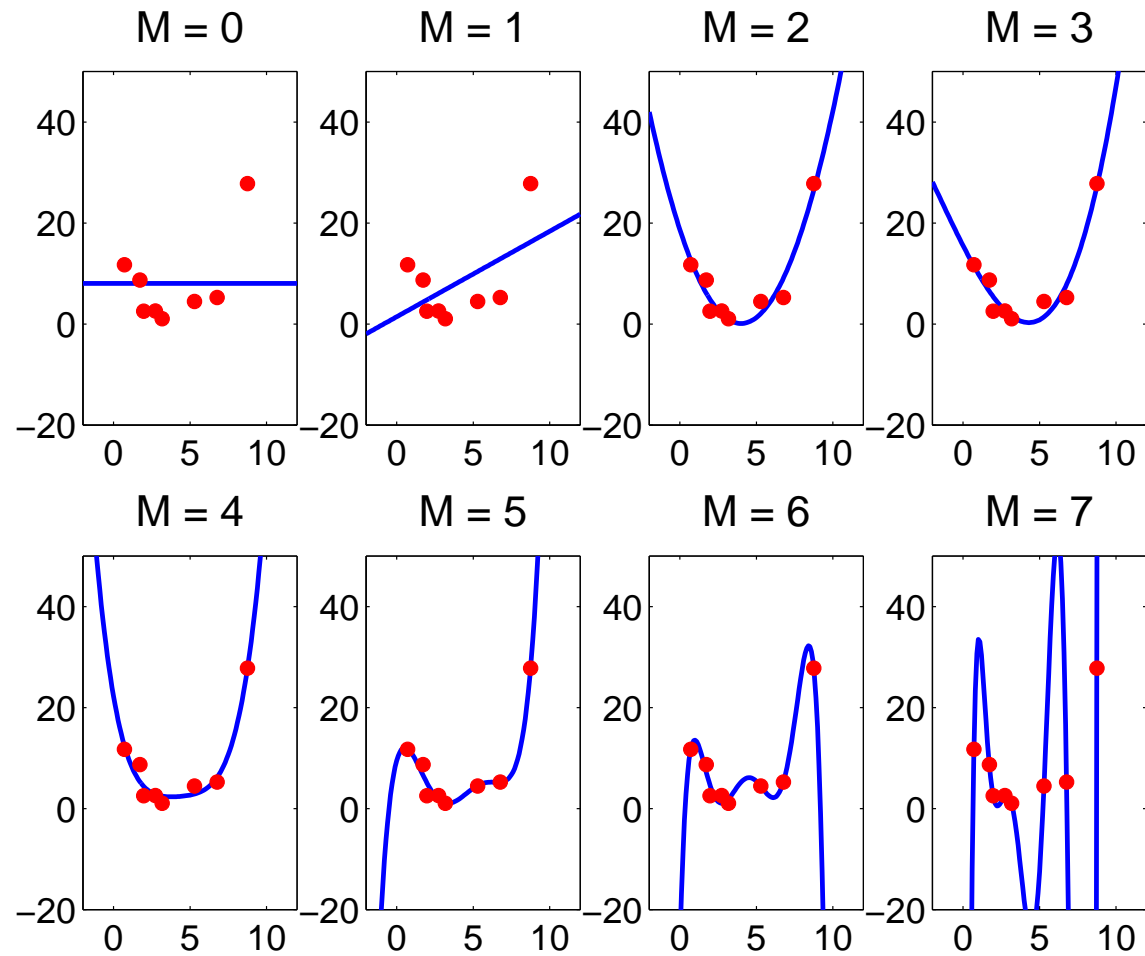
`zoubin@gatsby.ucl.ac.uk`

Carl Edward Rasmussen

`edward@gatsby.ucl.ac.uk`

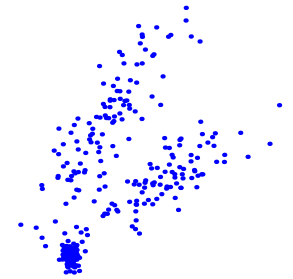
**Gatsby Computational Neuroscience Unit
MSc in Intelligent Systems, Fall 2001**

Model structure and overfitting: a simple example

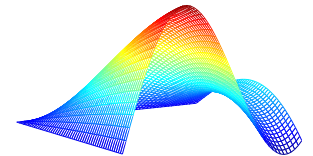


Model Selection Questions

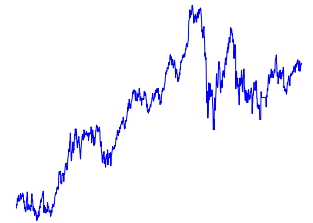
How many clusters in the data?



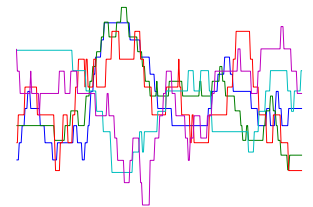
What is the intrinsic dimensionality of the data?



Is this input relevant to predicting that output?



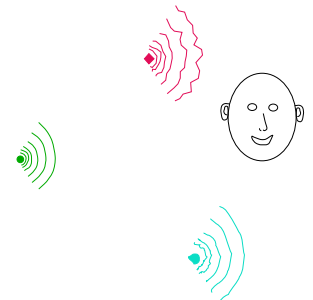
What is the order of this dynamical system?



How many states for this hidden Markov model?

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY

How many auditory sources in the input?



Bayesian Learning

data Y

models $\mathcal{M}_1 \dots, \mathcal{M}_n$

parameter sets $\theta_1 \dots, \theta_n$

(let's ignore hidden variables X for the moment, this will just introduce another level of averaging/integration)

Model Selection:

$$P(\mathcal{M}_i|Y) = \frac{P(Y|\mathcal{M}_i)P(\mathcal{M}_i)}{P(Y)}$$

Model Averaging:

$$P(y|Y) = \sum_i P(y|Y, \mathcal{M}_i)P(\mathcal{M}_i|Y)$$

Ockham's Razor

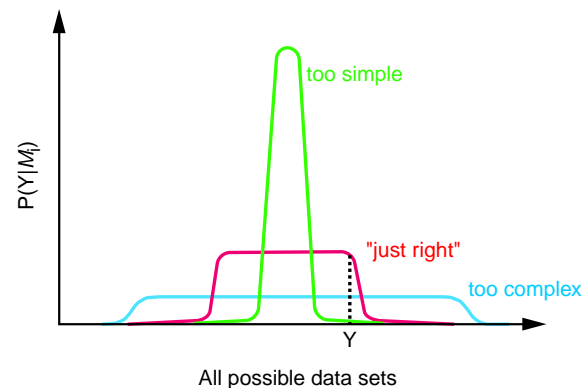
$$P(\mathcal{M}_i|Y) = \frac{P(Y|\mathcal{M}_i)P(\mathcal{M}_i)}{P(Y)}$$

$$P(Y|\mathcal{M}_i) = \int_{\theta_i} P(Y|\theta_i, \mathcal{M}_i)P(\theta_i|\mathcal{M}_i)$$

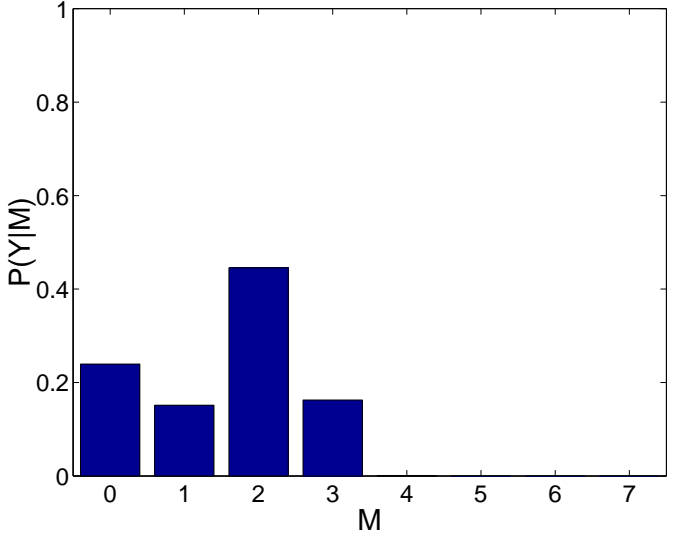
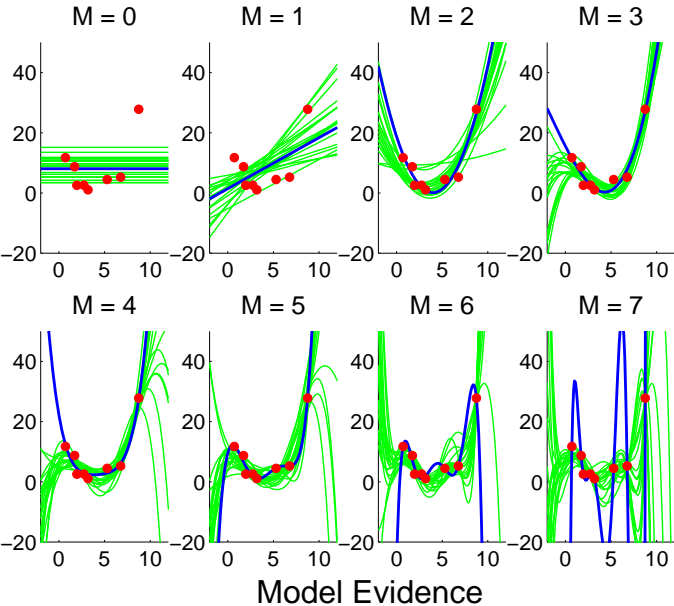
What is the probability that if you *randomly selected* parameter values from your model class you would generate data set Y ?

Model classes that are **too simple** will be very unlikely to generate that particular data set.

Model classes that are **too complex** can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.

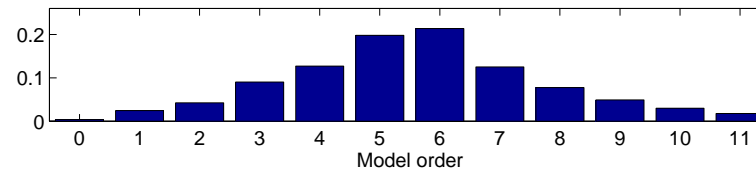
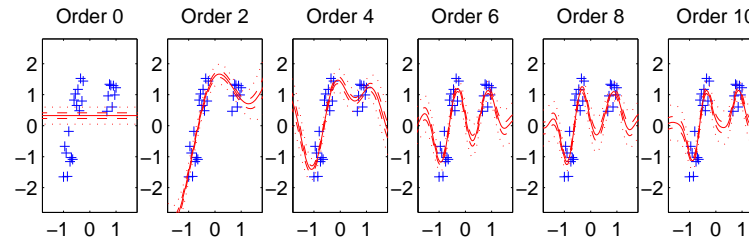


Bayesian Model Selection

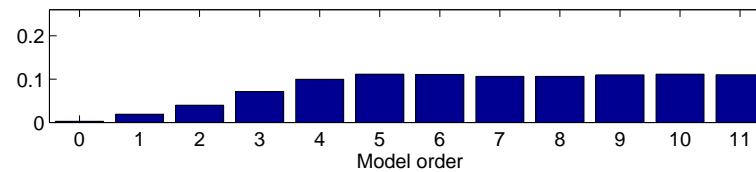
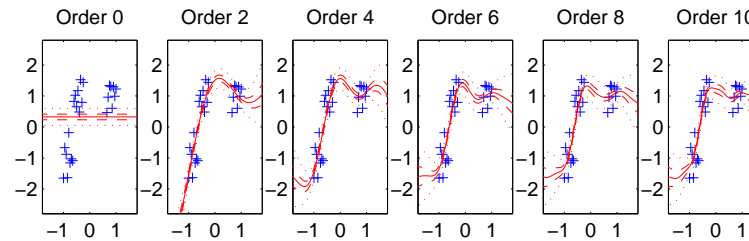


A subtle point about Ockham's Hill

Unscaled models:



Scaled models:



Practical Bayesian approaches

- Laplace approximations:
 - Appeals to Central Limit Theorem making a Gaussian approximation about maximum *a posteriori* parameter estimate.
- Large sample approximations (e.g. BIC).
- Markov chain Monte Carlo methods (MCMC):
 - In the limit are guaranteed to converge, but:
 - Many samples required to ensure accuracy.
 - Hard to assess convergence.
- Variational approximations...

Variational Bayesian Learning

Let the hidden states be \mathbf{x} , data \mathbf{y} and the parameters $\boldsymbol{\theta}$.

We can **lower bound** the **evidence** (Jensen's inequality):

$$\begin{aligned}\ln P(\mathbf{y}|\mathcal{M}) &= \ln \int d\mathbf{x} d\boldsymbol{\theta} P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|\mathcal{M}) \\ &= \ln \int d\mathbf{x} d\boldsymbol{\theta} Q(\mathbf{x}, \boldsymbol{\theta}) \frac{P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{Q(\mathbf{x}, \boldsymbol{\theta})} \\ &\geq \int d\mathbf{x} d\boldsymbol{\theta} Q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{Q(\mathbf{x}, \boldsymbol{\theta})}.\end{aligned}$$

Use a simpler, factorised approximation to $Q(\mathbf{x}, \boldsymbol{\theta})$:

$$\begin{aligned}\ln P(\mathbf{y}) &\geq \int d\mathbf{x} d\boldsymbol{\theta} Q_{\mathbf{x}}(\mathbf{x})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{Q_{\mathbf{x}}(\mathbf{x})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\ &= \mathcal{F}(Q_{\mathbf{x}}(\mathbf{x}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}).\end{aligned}$$

Variational Bayesian EM

Maximising this **lower bound**, \mathcal{F} , leads to **EM-like** updates:

$$Q_{\mathbf{x}}^*(\mathbf{x}) \propto \exp \langle \ln P(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \quad E\text{-like step}$$

$$Q_{\boldsymbol{\theta}}^*(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}) \exp \langle \ln P(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_{Q_{\mathbf{x}}(\mathbf{x})} \quad M\text{-like step}$$

Equivalent to minimizing KL-divergence between the *approximating* and *true* posteriors.

A Generative Model for Generative Models

