Unsupervised Learning

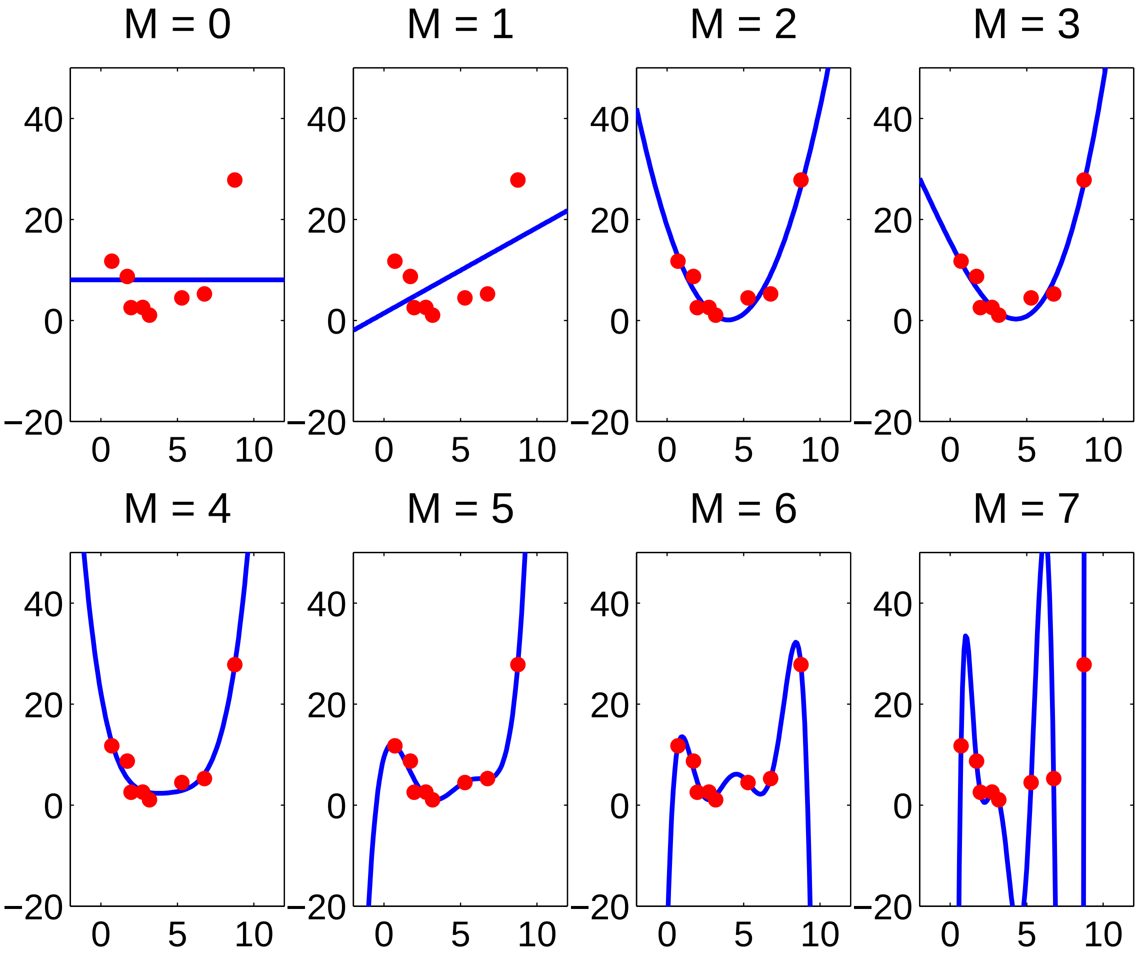
Bayesian Model Selection

Zoubin Ghahramani
zoubin@gatsby.ucl.ac.uk

**Gatsby Computational Neuroscience Unit, and
MSc in Intelligent Systems, Dept Computer Science
University College London**

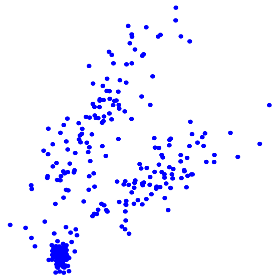
Autumn 2003

Model structure and overfitting: a simple example

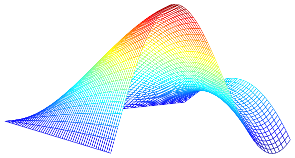


Learning Model Structure

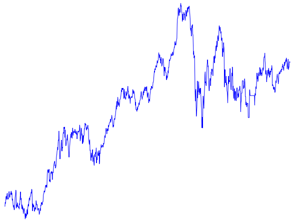
How many clusters in the data?



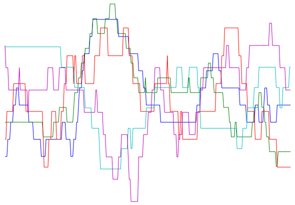
What is the intrinsic dimensionality of the data?



Is this input relevant to predicting that output?



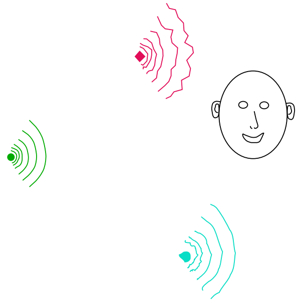
What is the order of a dynamical system?



How many states in a hidden Markov model?

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY

How many auditory sources in the input?



Using Occam's Razor to Learn Model Structure

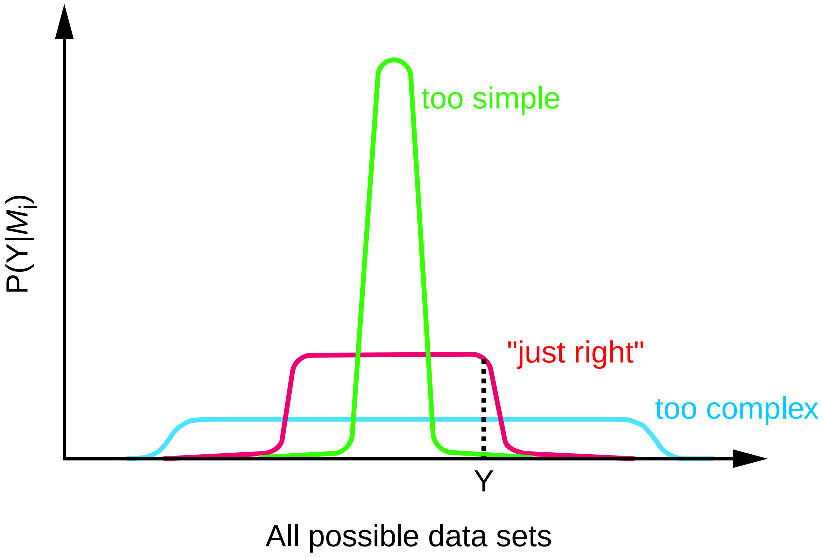
Select the model class $\mathcal{M}_i$ with the highest probability given the data:

$$P(\mathcal{M}_i|\mathbf{y}) = \frac{P(\mathbf{y}|\mathcal{M}_i)P(\mathcal{M}_i)}{P(\mathbf{y})}, \qquad P(\mathbf{y}|\mathcal{M}_i) = \int_{\Theta_i} P(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i)P(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i$$

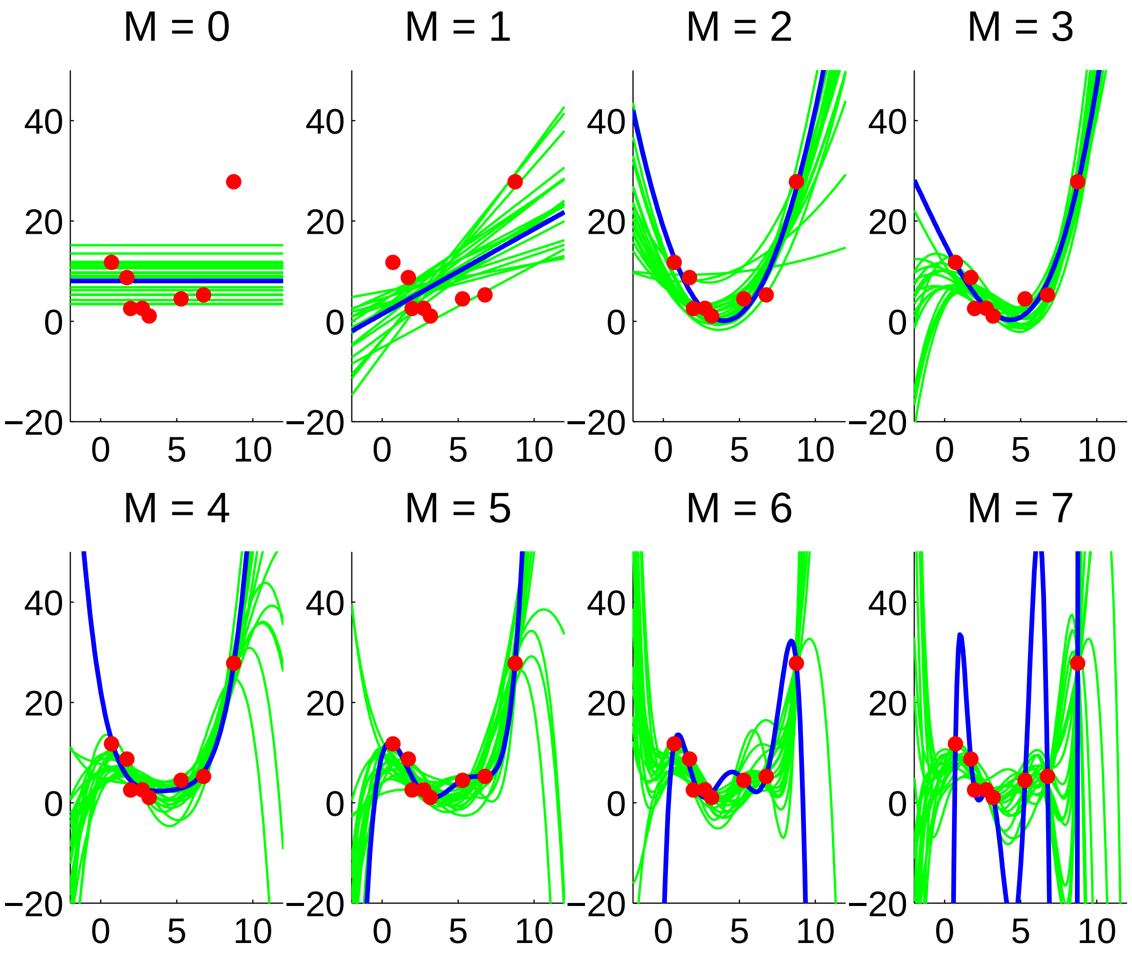
Interpretation of $P(\mathbf{y}|\mathcal{M}_i)$: The probability that *randomly selected* parameter values from the model class would generate data set $\mathbf{y}$.

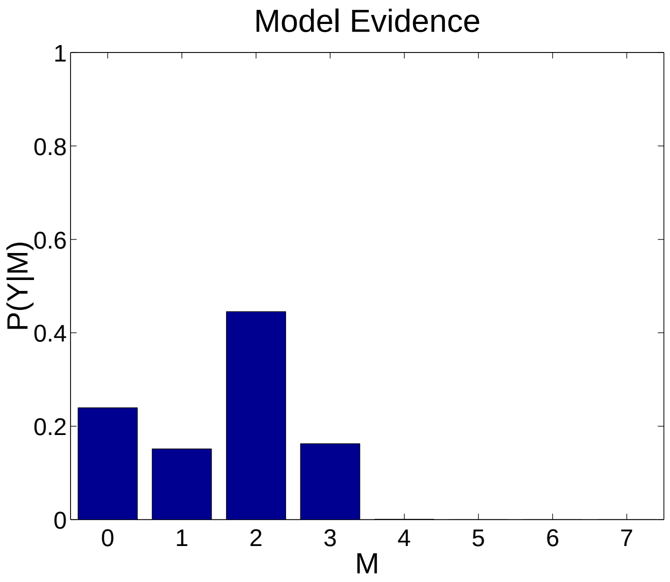
Model classes that are too simple are unlikely to generate the data set.

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: Occam's Razor at Work





e.g. for quadratic (M=2): $y = a_0 + a_1 x + a_2 x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$ and $\boldsymbol{\theta}_2 = [a_0\ a_1\ a_2\ \tau]$

`demo:  polybayes`

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.
  - Sometimes hard to assess convergence.
- Variational approximations

Note: other deterministic approximations are also available now: e.g. Bethe approximations and Expectation Propagation

Laplace Approximation

data set $\mathbf{y}$, models $\mathcal{M}_1 \ldots, \mathcal{M}_n$, parameter sets $\boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_n$

Model Selection: $P(\mathcal{M}_i|\mathbf{y}) \propto P(\mathcal{M}_i)P(\mathbf{y}|\mathcal{M}_i)$

For large amounts of data (relative to number of parameters, d) the parameter posterior is approximately Gaussian around the MAP estimate $\hat{\boldsymbol{\theta}}_i$:

$$P(\boldsymbol{\theta}_i|\mathbf{y}, \mathcal{M}_i) \approx (2\pi)^{\frac{-d}{2}}|A|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^\top A(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)\right\}$$

$$P(\mathbf{y}|\mathcal{M}_i) = \frac{P(\boldsymbol{\theta}_i, \mathbf{y}|\mathcal{M}_i)}{P(\boldsymbol{\theta}_i|\mathbf{y}, \mathcal{M}_i)}$$

Evaluating the above expression for $\ln P(\mathbf{y}|\mathcal{M}_i)$ at $\hat{\boldsymbol{\theta}}_i$:

$$\ln P(\mathbf{y}|\mathcal{M}_i) \approx \ln P(\hat{\boldsymbol{\theta}}_i|\mathcal{M}_i) + \ln P(\mathbf{y}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i) + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|A|$$

where A is the $d \times d$ negative Hessian matrix of the log posterior, $A_{\ell m} = -\frac{\partial^2}{\partial\theta_{i\ell}\partial\theta_{im}} \ln P(\boldsymbol{\theta}_i|\mathbf{y}, \mathcal{M}_i)|_{\hat{\boldsymbol{\theta}}_i}$.

This can be used for model selection.

Bayesian Information Criterion (BIC)

BIC can be obtained from the Laplace approximation:

$$\ln P(\mathbf{y}|\mathcal{M}_i) \approx \ln P(\hat{\boldsymbol{\theta}}_i|\mathcal{M}_i) + \ln P(\mathbf{y}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i) + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |A|$$

in the large sample limit ($N \to \infty$) where N is the number of data points, A grows as NA_0 for some fixed matrix A_0, so $\ln |A| \to \ln |NA_0| = \ln(N^d|A_0|) = d\ln N + \ln |A_0|$. Retaining only terms that grow in N we get:

$$\ln P(\mathbf{y}|\mathcal{M}_i) \approx \ln P(\mathbf{y}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i) - \frac{d}{2}\ln N$$

Properties:

- Quick and easy to compute
- It does not depend on the prior
- We can use the ML estimate of θ instead of the MAP estimate
- It is equivalent to the MDL criterion
- It assumes that in the large sample limit, all the parameters are well-determined (i.e. the model is identifiable; otherwise, d should be the number of well-determined parameters)
- **Danger:** counting parameters can be deceiving! (c.f. sinusoid, infinite models)

MCMC Approximations

Let's consider a non-Markov chain method, **Importance Sampling:**

$$
\begin{aligned}
\ln P(\mathbf{y}|\mathcal{M}_i) &= \ln \int_{\Theta_i} P(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i) P(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i \\
&= \ln \int_{\Theta_i} P(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i) \frac{P(\boldsymbol{\theta}_i|\mathcal{M}_i)}{Q(\boldsymbol{\theta}_i)} Q(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i \\
&\approx \ln \sum_k P(\mathbf{y}|\boldsymbol{\theta}_i^{(k)}, \mathcal{M}_i) \frac{P(\boldsymbol{\theta}_i^{(k)}|\mathcal{M}_i)}{Q(\boldsymbol{\theta}_i^{(k)})}
\end{aligned}
$$

where $\boldsymbol{\theta}_i^{(k)}$ are i.i.d. draws from $Q(\boldsymbol{\theta}_i)$. Assumes we can **sample from** and **evaluate** $Q(\boldsymbol{\theta}_i)$ (incl. normalization!) and we can **compute the likelihood** $P(\mathbf{y}|\boldsymbol{\theta}_i^{(k)}, \mathcal{M}_i)$.

Although importance sampling does not work well in high dimensions, it inspires the following approach: Create a **Markov chain,** $Q_k \to Q_{k+1} \ldots$ for which:

- $Q_k(\boldsymbol{\theta})$ can be evaluated including normalization
- $\lim_{k\to\infty} Q_k(\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_i)$

Variational Bayesian Learning

Lower Bounding the Evidence

Let the hidden latent variables 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 Learning . . .

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

$$\begin{aligned} Q^*_{\mathbf{x}}(\mathbf{x}) &\propto \exp \langle \ln P(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} && E\!-\!like\;\; step \\ Q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) &\propto P(\boldsymbol{\theta}) \exp \langle \ln P(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) \rangle_{Q_{\mathbf{x}}(\mathbf{x})} && M\!-\!like\;\; step \end{aligned}$$

Maximizing $\mathcal{F}$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $Q(\boldsymbol{\theta})Q(\mathbf{x})$ and the *true posterior*, $P(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$.

Conjugate-Exponential models

Let's focus on *conjugate-exponential* (**CE**) models, which satisfy **(1)** and **(2)**:
Condition (1). The joint probability over *variables* is in the exponential family:

$$P(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{x},\mathbf{y})\; g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x},\mathbf{y})\right\}$$

where $\phi(\boldsymbol{\theta})$ is the vector of *natural parameters*, $\mathbf{u}$ are *sufficient statistics*
Condition (2). The prior over *parameters* is conjugate to this joint probability:

$$P(\boldsymbol{\theta}|\eta,\boldsymbol{\nu}) = h(\eta,\boldsymbol{\nu})\; g(\boldsymbol{\theta})^\eta \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}\right\}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior.
Conjugate priors are computationally convenient and have an intuitive interpretation:

- η: number of pseudo-observations
- $\boldsymbol{\nu}$: values of pseudo-observations

Conjugate-Exponential examples

In the **CE** family:

- Gaussian mixtures
- factor analysis, probabilistic PCA
- hidden Markov models and factorial HMMs
- linear dynamical systems and switching models
- discrete-variable belief networks

Other as yet undreamt-of models can combine Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Not in the **CE** family:

- Boltzmann machines, MRFs (no simple conjugacy)
- logistic regression (no simple conjugacy)
- sigmoid belief networks (not exponential)
- independent components analysis (not exponential)

Note: one can often approximate these models with models in the **CE** family.

A Useful Result

Theorem Given an iid data set $\mathbf{y} = (\mathbf{y}_1, \ldots \mathbf{y}_n)$, if the model is **CE** then:

(a) $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is also conjugate, *i.e.*

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}\right\}$$

where $\tilde{\eta} = \eta + n$ and $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_i \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)$.

(b) $Q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n Q_{\mathbf{x}_i}(\mathbf{x}_i)$ is of the same form as in the E step of regular EM, but using pseudo parameters computed by averaging over $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

$$Q_{\mathbf{x}_i}(\mathbf{x}_i) \propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} = P(\mathbf{x}_i | \mathbf{y}_i, \overline{\boldsymbol{\phi}}(\boldsymbol{\theta}))$$

KEY points:

(a) the approximate parameter posterior is of the same form as the prior, so it is easily summarized in terms of two sets of hyperparameters, $\tilde{\eta}$ and $\tilde{\boldsymbol{\nu}}$;

(b) the approximate hidden variable posterior, *averaging over all parameters*, is of the same form as the hidden variable posterior for a *single setting of the parameters*, so again, it is easily computed using the usual methods.

The Variational Bayesian EM algorithm

EM for MAP estimation

Goal: maximize $p(\boldsymbol{\theta}|\mathbf{y}, m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})\, d\mathbf{x}$$

Variational Bayesian EM

Goal: lower bound $p(\mathbf{y}|m)$

VB-E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \exp\left[\int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})\, d\mathbf{x}\right]$$

Properties:

- Reduces to the EM algorithm if $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- $\mathcal{F}_m$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\bar{\phi}$.

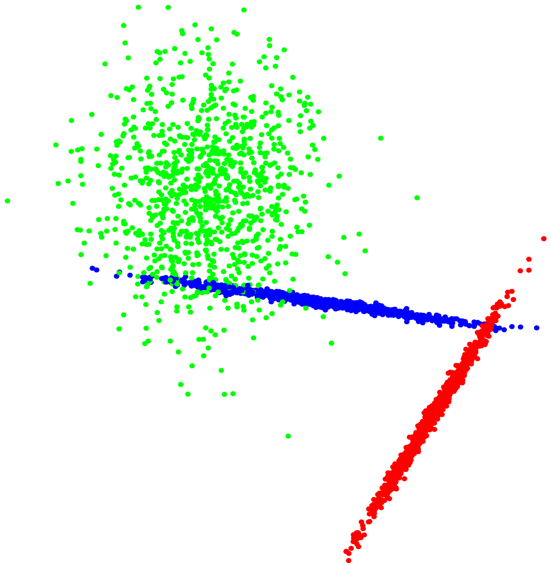
Variational Bayes: History of Models Treated

- multilayer perceptrons (Hinton & van Camp, 1993)
- mixture of experts (Waterhouse, MacKay & Robinson, 1996)
- hidden Markov models (MacKay, 1995)
- other work by Jaakkola, Jordan, Barber, Bishop, Tipping, etc

Examples of Variational Learning of Model Structure

- mixtures of factor analysers (Ghahramani & Beal, 1999)
- mixtures of Gaussians (Attias, 1999)
- independent components analysis (Attias, 1999; Miskin & MacKay, 2000; Valpola 2000)
- principal components analysis (Bishop, 1999)
- linear dynamical systems (Ghahramani & Beal, 2000)
- mixture of experts (Ueda & Ghahramani, 2000)
- discrete graphical models (Beal & Ghahramani, 2002)
- VIBES software for conjugate-exponential graphs (Winn, 2003)

Mixture of Factor Analysers



Goal:

- Infer number of clusters

- Infer intrinsic dimensionality of each cluster

Under the assumption that each cluster is Gaussian

```
embed_demo
```

Mixture of Factor Analysers

True data: 6 Gaussian clusters with dimensions: (1 7 4 3 2 2) embedded in 10-D

Inferred structure:

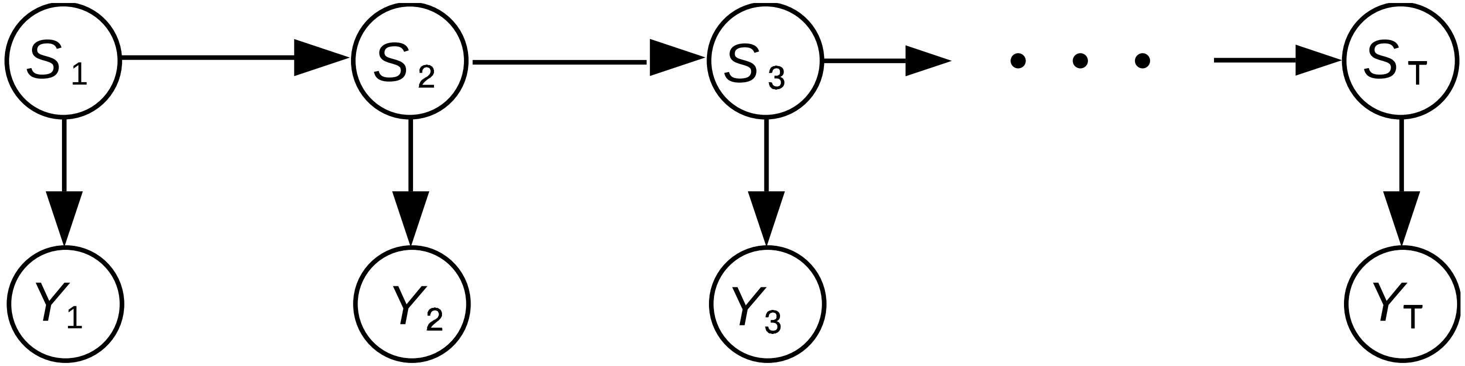
<table>
<tr><th rowspan="2">number of points per cluster</th><th colspan="6">intrinsic dimensionalities</th></tr>
<tr><th>1</th><th>7</th><th>4</th><th>3</th><th>2</th><th>2</th></tr>
<tr><td>8</td><td colspan="4">2</td><td colspan="2">1</td></tr>
<tr><td>8</td><td colspan="2">1</td><td colspan="4">2</td></tr>
<tr><td>16</td><td>1</td><td colspan="4">4</td><td>2</td></tr>
<tr><td>32</td><td>1</td><td>6</td><td>3</td><td>3</td><td>2</td><td>2</td></tr>
<tr><td>64</td><td>1</td><td>7</td><td>4</td><td>3</td><td>2</td><td>2</td></tr>
<tr><td>128</td><td>1</td><td>7</td><td>4</td><td>3</td><td>2</td><td>2</td></tr>
</table>

- Finds the clusters and dimensionalities efficiently.

- The model complexity reduces in line with the lack of data support.

demos: `run_simple` and `ueda_demo`

Hidden Markov Models



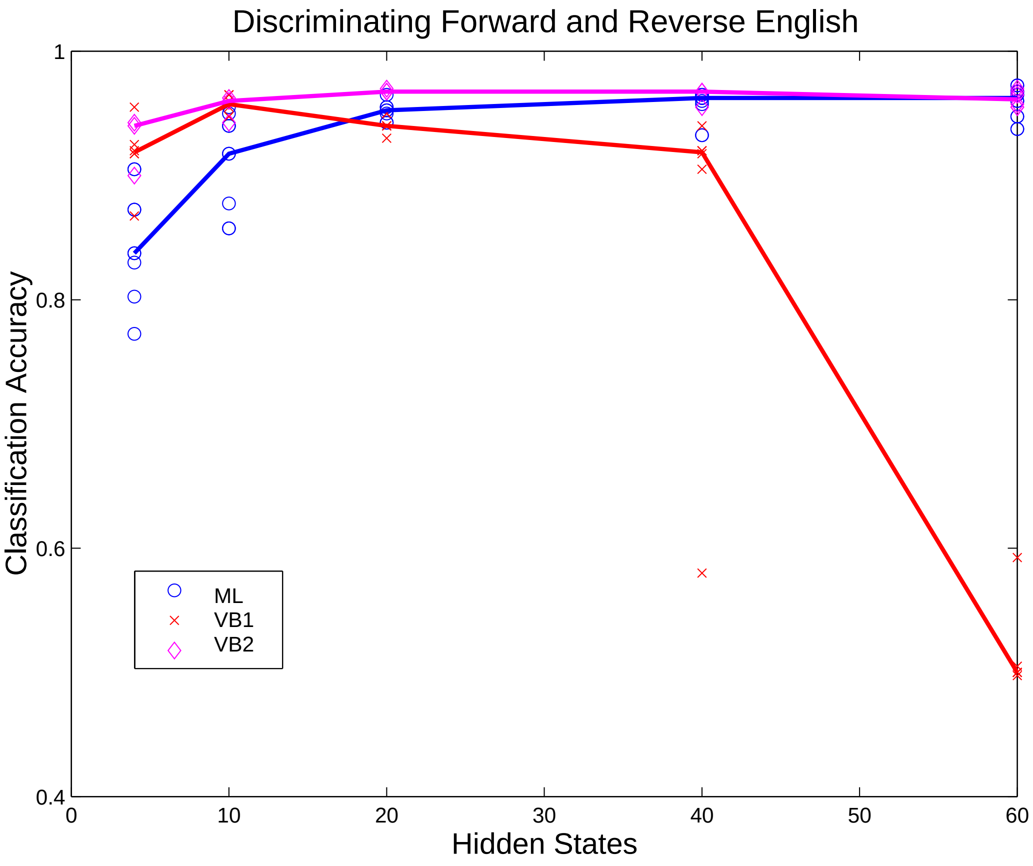
Discrete hidden states, $\mathbf{s}_t$.
Observations $\mathbf{y}_t$.

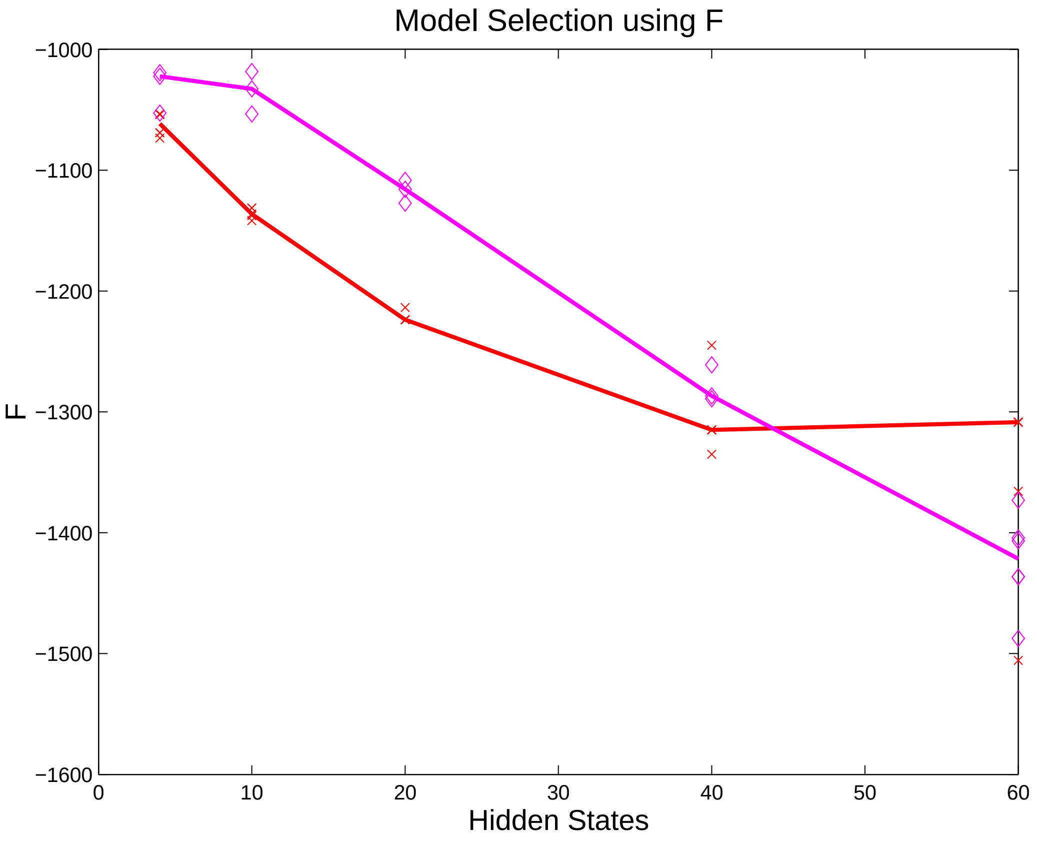
How many hidden states?
What structure state-transition matrix?

`demo:   vbhmm_demo`

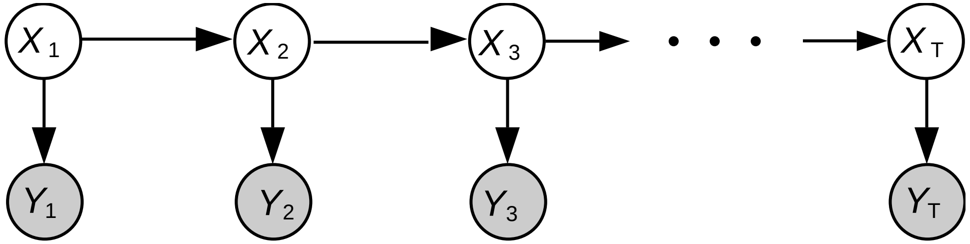
Hidden Markov Models:
Discriminating Forward from Reverse English

First 8 sentences from Alice in Wonderland.
Compare VB-HMM with ML-HMM.





Linear Dynamical Systems



- Assumes $\mathbf{y}_t$ generated from a sequence of Markov *hidden* state variables $\mathbf{x}_t$
- If transition and output functions are linear, time-invariant, and noise distributions are Gaussian, this is a linear-Gaussian state-space model:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t$$

- Three levels of inference:

I Given data, structure and parameters, **Kalman smoothing** → hidden state;

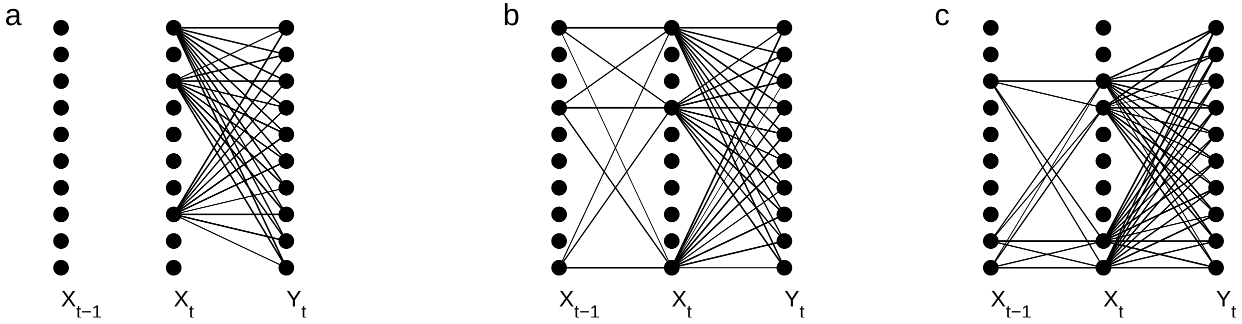
II Given data and structure, **EM** → hidden state and parameter point estimates;

III Given data only, **VEM** → **model structure** *and* **distributions over parameters** *and* **hidden state**.

Linear Dynamical System Results

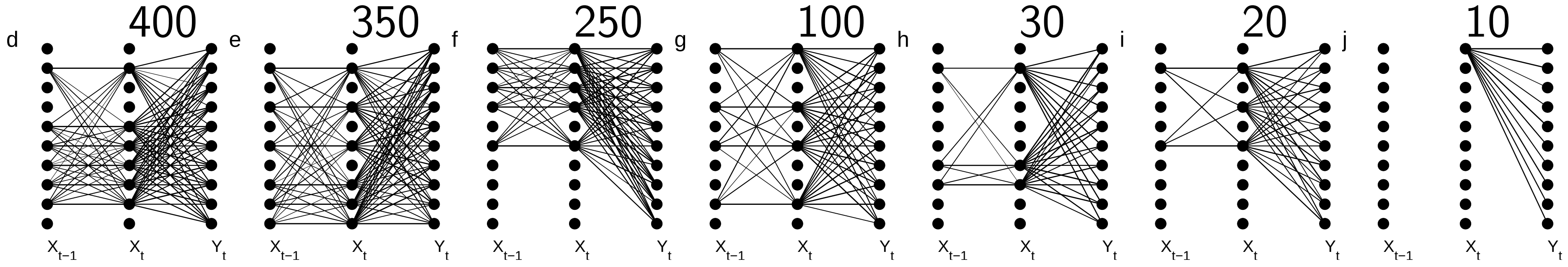
Inferring model structure:

a) SSM(0,3) i.e. FA b) SSM(3,3) c) SSM(3,4)



Inferred model complexity reduces with less data:

True model: SSM(6,6) • 10-dim observation vector.



```
demo:  bayeslds
```

Independent Components Analysis

Blind Source Separation: 5 × 100 msec speech and music sources linearly mixed to produce 11 signals (microphones)

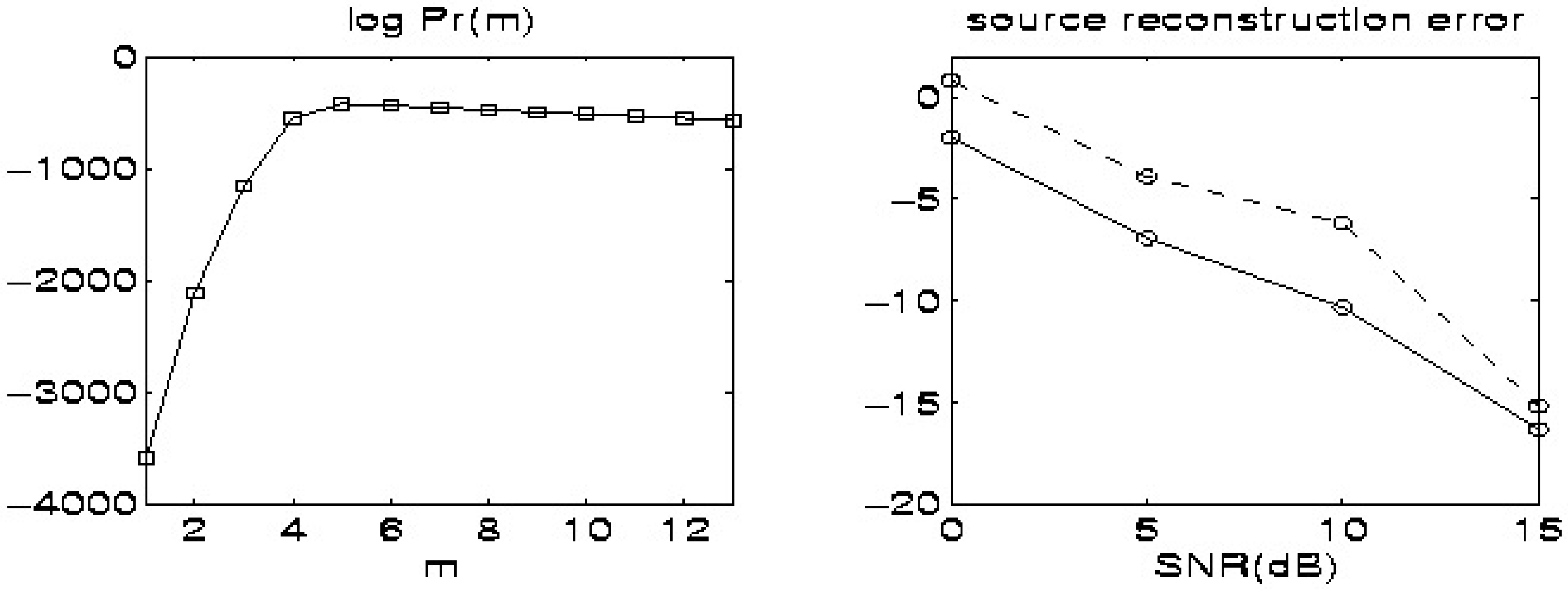


Figure 2. Application of VB to blind source separation algorithm (see text).

from Attias (2000)

Summary & Conclusions

- Bayesian learning avoids overfitting and can be used to do model selection.
- But we need approximations:
- Laplace
- BIC
- Sampling
- Variational...