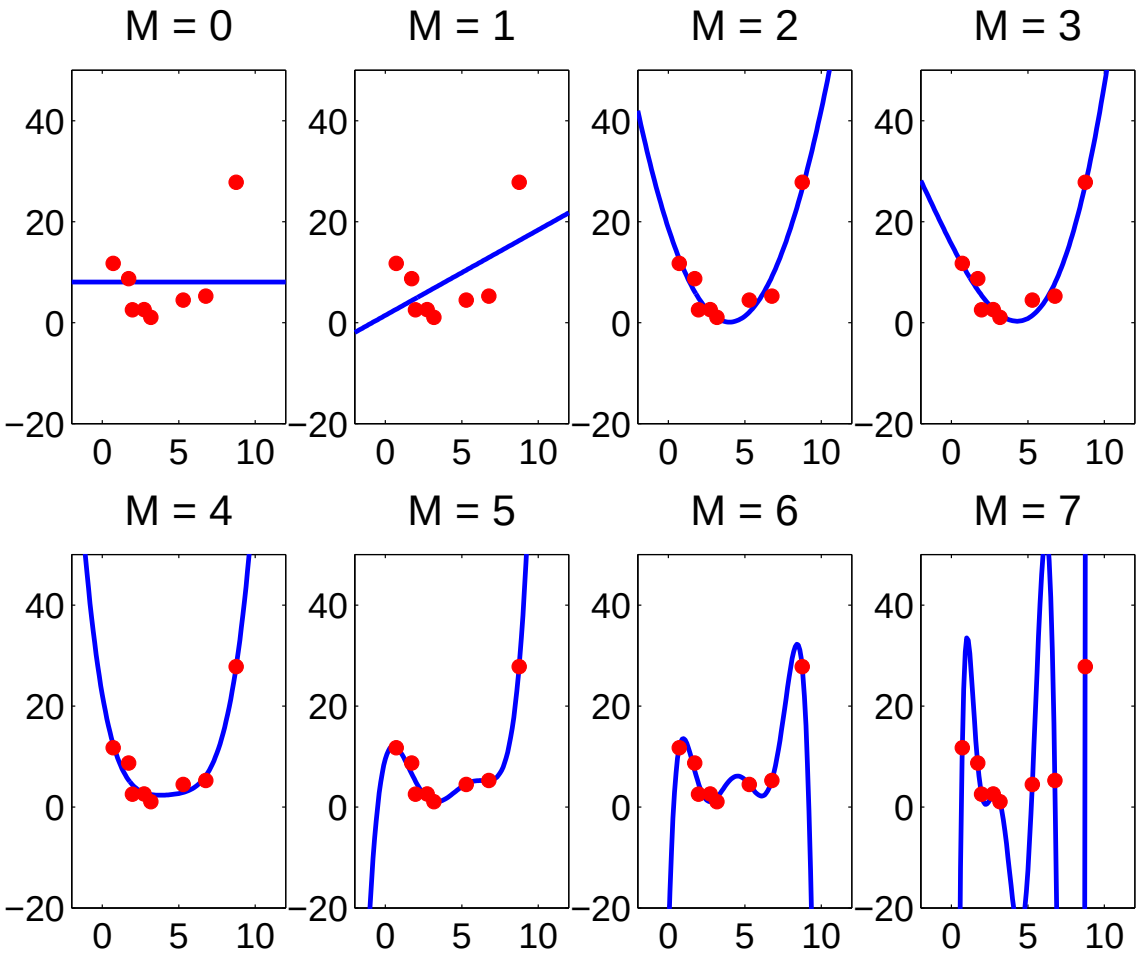
Statistical Approaches to Learning and Discovery

Bayesian Model Selection

Zoubin Ghahramani & Teddy Seidenfeld
zoubin@cs.cmu.edu & teddy@stat.cmu.edu

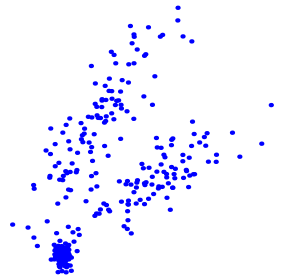
CALD / CS / Statistics / Philosophy
Carnegie Mellon University
Spring 2002

Model structure and overfitting: a simple example

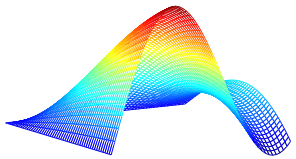


Learning Model Structure

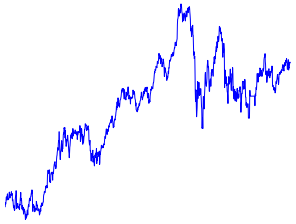
How many clusters in the data?



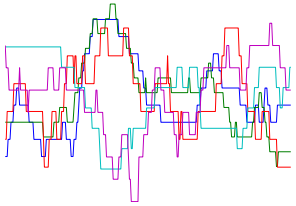
What is the intrinsic dimensionality of the data?



Is this input relevant to predicting that output?



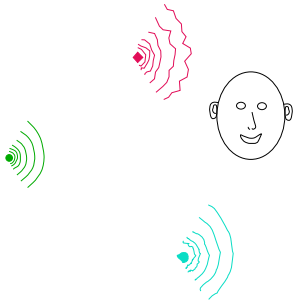
What is the order of a dynamical system?



How many states in a hidden Markov model?

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY

How many auditory sources in the input?



Using Occam's Razor to Learn Model Structure

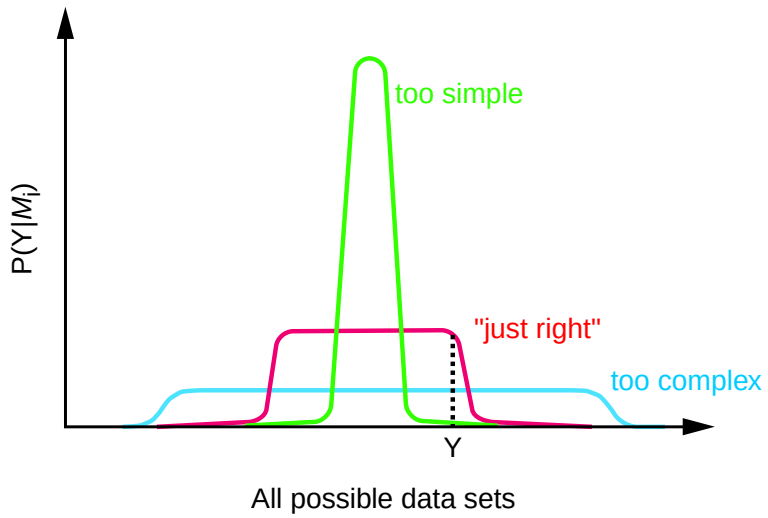
Select the model class $\mathcal{M}_i$ with the highest probability given the data:

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

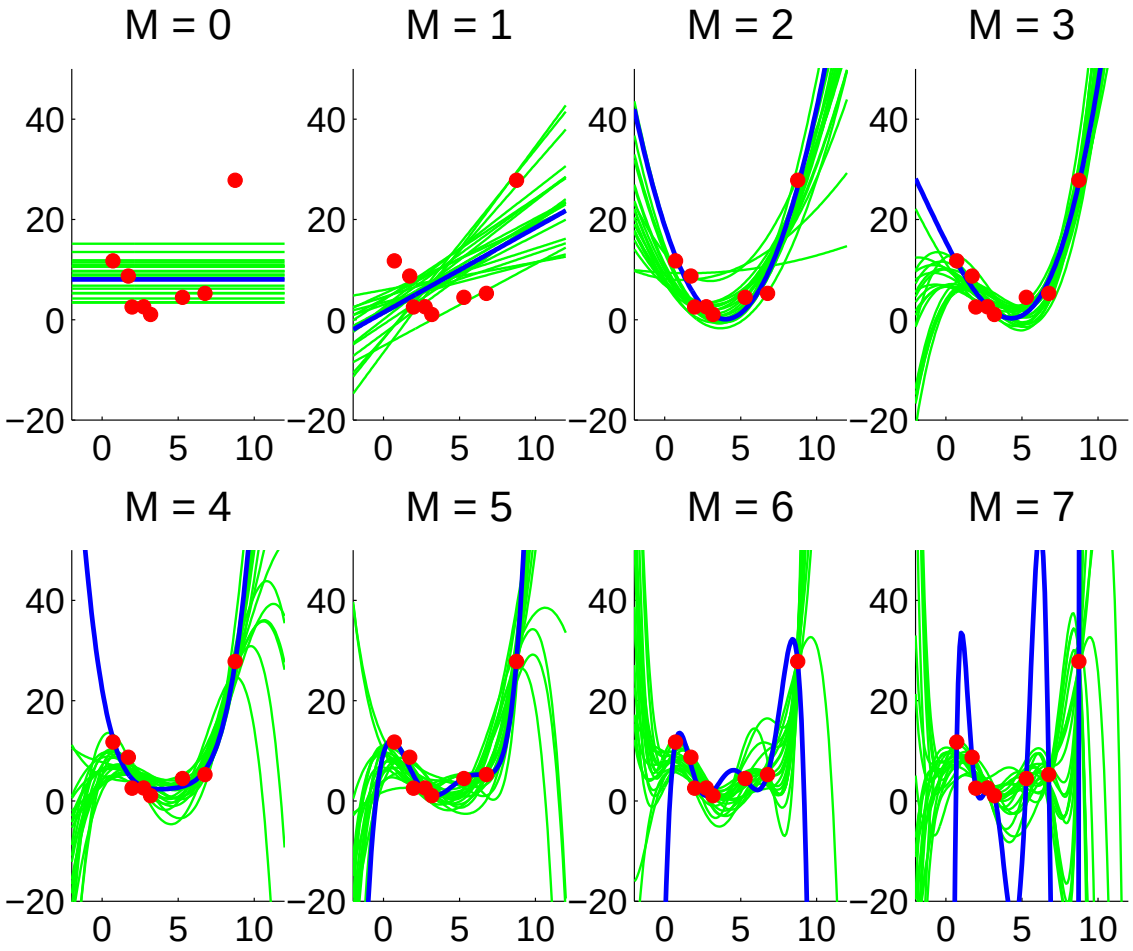
Interpretation: The probability that *randomly selected* parameter values from the model class would generate data set $\mathbf{y}$.

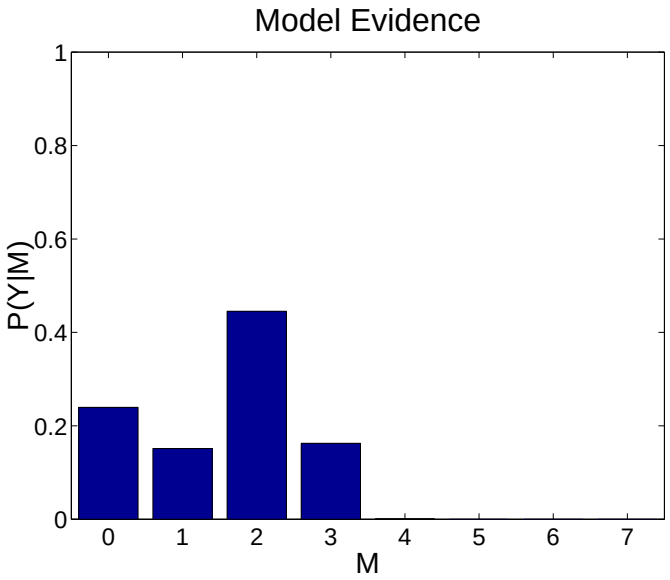
Model classes that are too simple are unlikely to generate the data set.

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



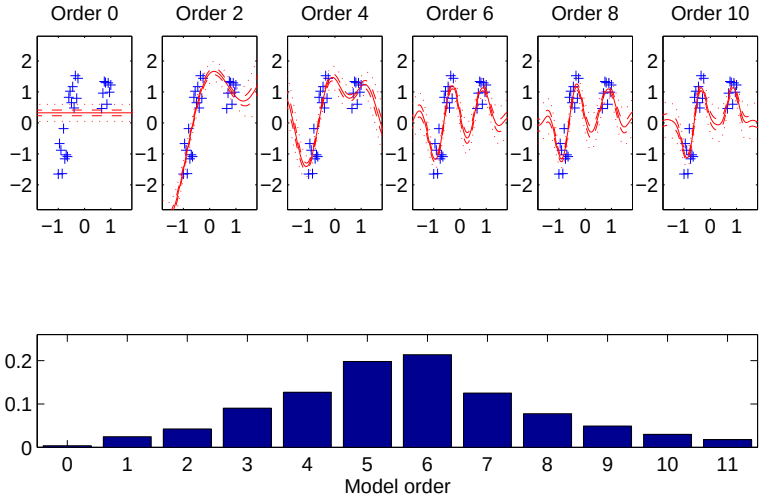


demo: polybayes

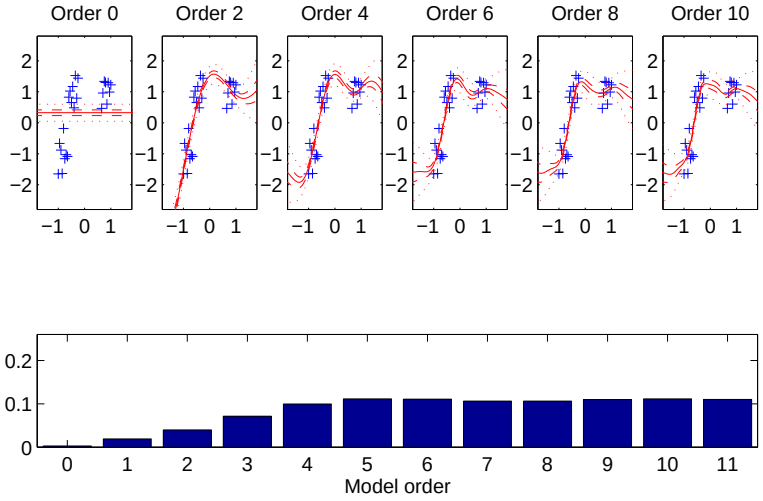
Scaling the parameter priors

It is important to consider how the parameter prior changes as the number of parameters increases — this determines whether an Occam's hill is present or not.

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.
  - 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.

This can be used for model selection.

Bayesian Information Criterion (BIC)

BIC can be obtained from the Laplace approximation:

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

by taking the large sample limit $(N \to \infty)$ where N is the number of data points:

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

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 $\boldsymbol{\phi}(\boldsymbol{\theta})$ 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 EM algorithm

VE Step: Compute the expected sufficient statistics $\sum_i \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)$ under the hidden variable distributions $Q_{\mathbf{x}_i}(\mathbf{x}_i)$.

VM Step: Compute expected natural parameters $\overline{\boldsymbol{\phi}}(\boldsymbol{\theta})$ under the parameter distribution given by $\tilde{\eta}$ and $\tilde{\boldsymbol{\nu}}$.

Properties:

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

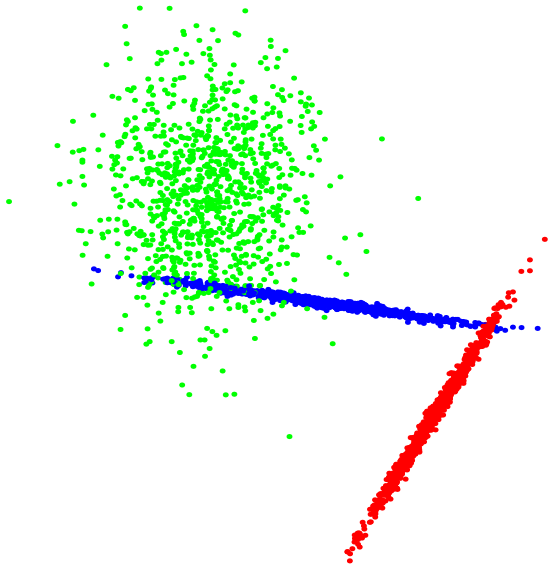
Variational Bayes: History of Models Treated

- 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 (Ghahramani & Beal, in prep)

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

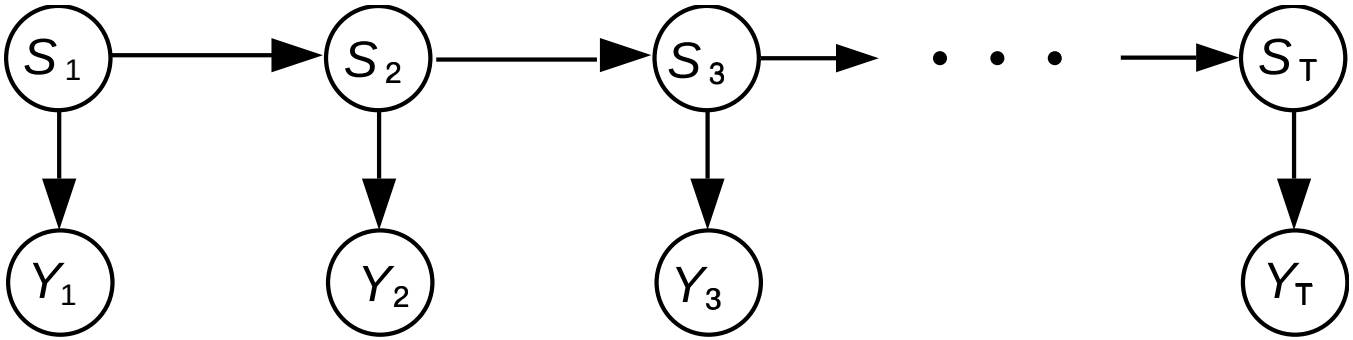
Inferred structure:

number of points per cluster	intrinsic dimensionalities					
	1	7	4	3	2	2
8	2				1	
8	1	2				
16	1	4				2
32	1	6	3	3	2	2
64	1	7	4	3	2	2
128	1	7	4	3	2	2

- 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



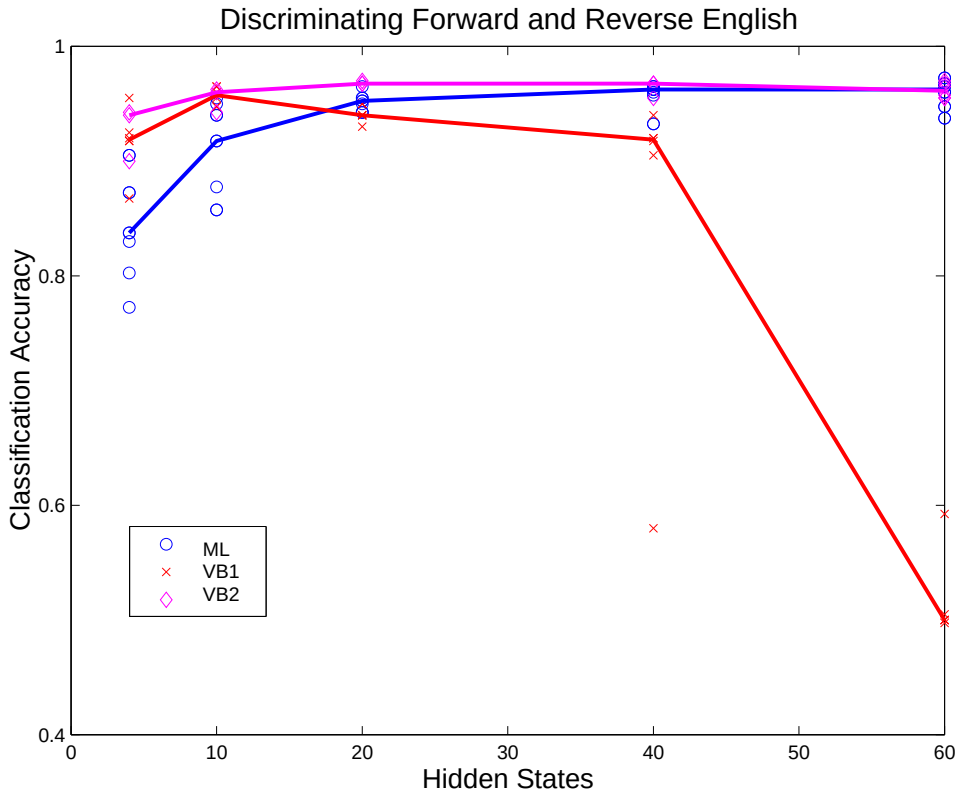
Discrete hidden states, $\mathbf{s}_t$.
Observations $\mathbf{y}_t$.

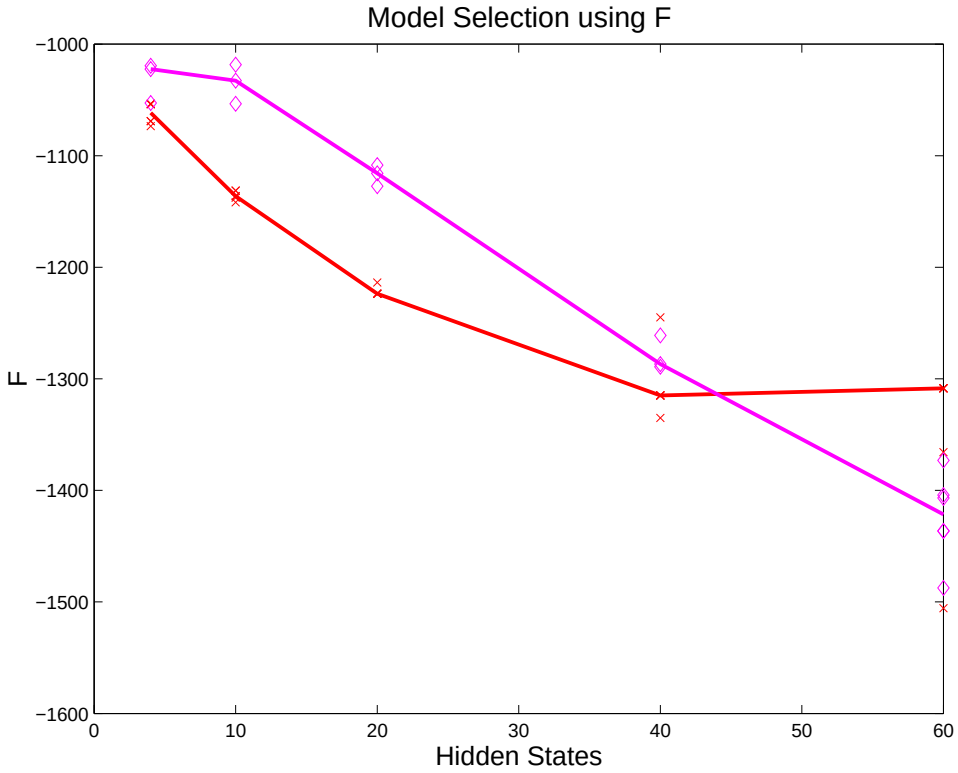
How many hidden states?
What structure state-transition matrix?

demo: `vbhmm_demo`

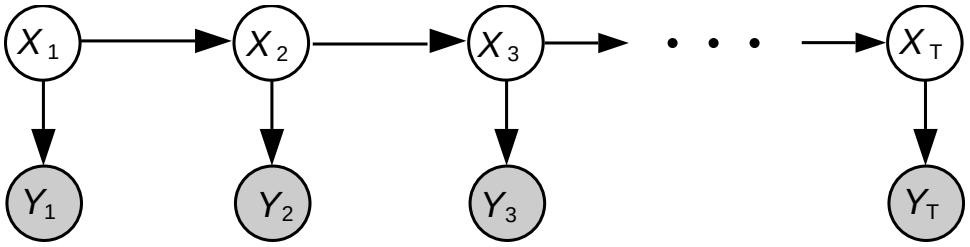
Hidden Markov Models: Discriminating Forward from Reverse English

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** $\rightarrow$ hidden state;

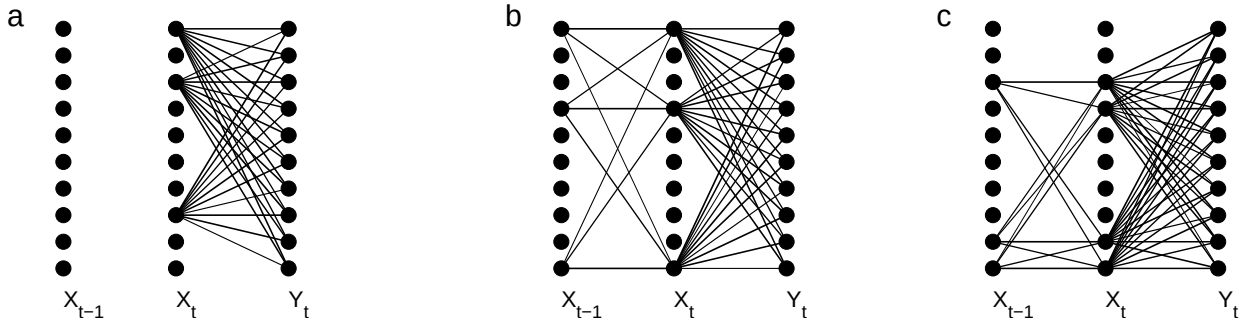
II Given data and structure, **EM** $\rightarrow$ hidden state and parameter point estimates;

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

Linear Dynamical System Results

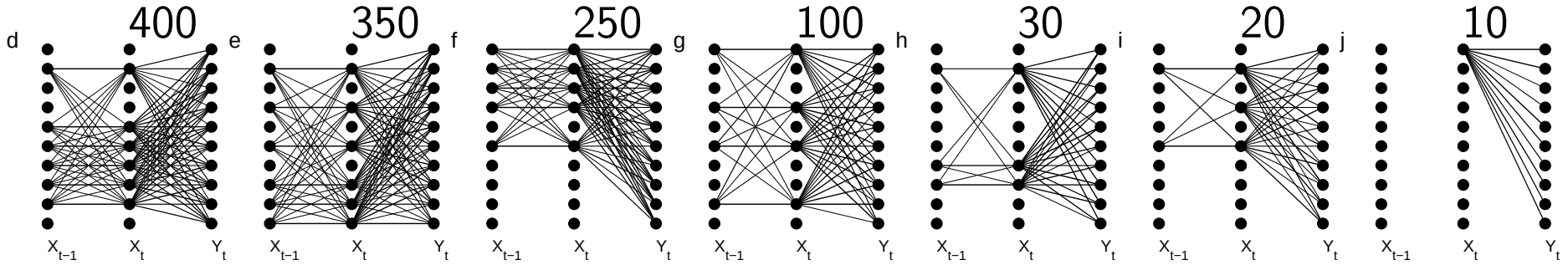
Inferring model structure:

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 $\times$ 100 msec speech and music sources linearly mixed to produce 11 signals (microphones)

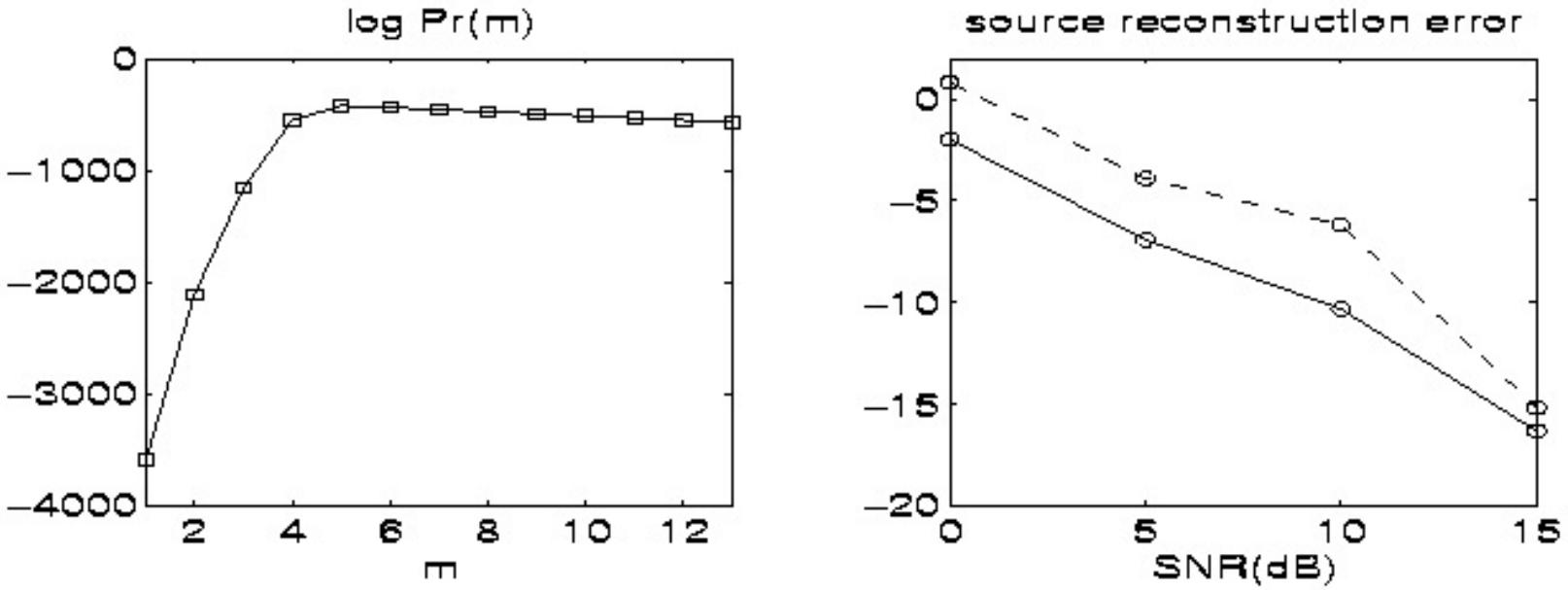


Figure 2. Application of VB to blind source separation algorithm (see text).

from Attias (2000)

Part 2: Another view of Model Selection: Nonparametric Methods and Infinite Models

We ought not to limit the complexity of our model a priori (e.g. number of hidden states, number of basis functions, number of mixture components, etc) since we don't believe that the real data was actually generated from a statistical model with a small number of parameters.

Therefore, regardless of how much training data we have, we should consider models with as many parameters as we can handle computationally.

Neal (1994) showed that MLPs with large numbers of hidden units achieved good performance on small data sets. He used MCMC to average over parameters.

Here there is no model order selection task:

- No need to evaluate evidence (which is often difficult).
- We don't need or want to use Occam's razor to limit the number of parameters in our model.

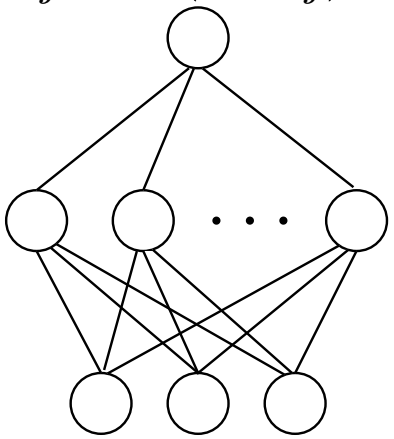
In fact, we may even want to consider doing inference in models with an infinite number of parameters...

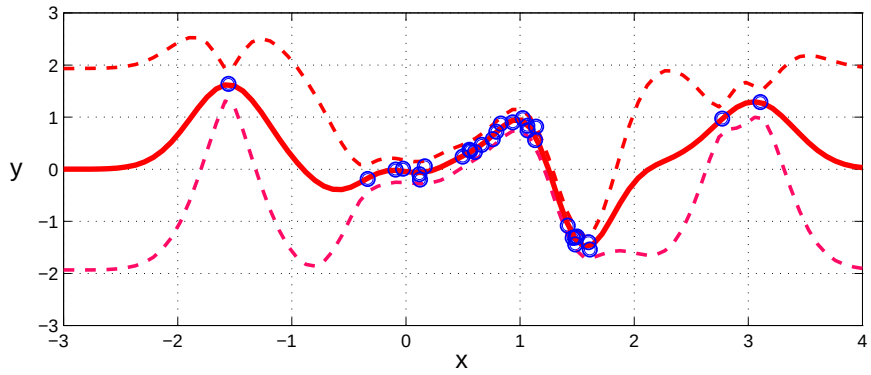
Infinite Models 1: Gaussian Processes

Neal (1994) showed that a one-hidden-layer neural network with bounded activation function and Gaussian prior over the weights and biases converges to a (nonstationary) Gaussian process prior over functions.

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(0, C(\mathbf{x}))$$

where e.g. $C_{ij} \equiv C(x_i, x_j) = g(|x_i - x_j|)$.



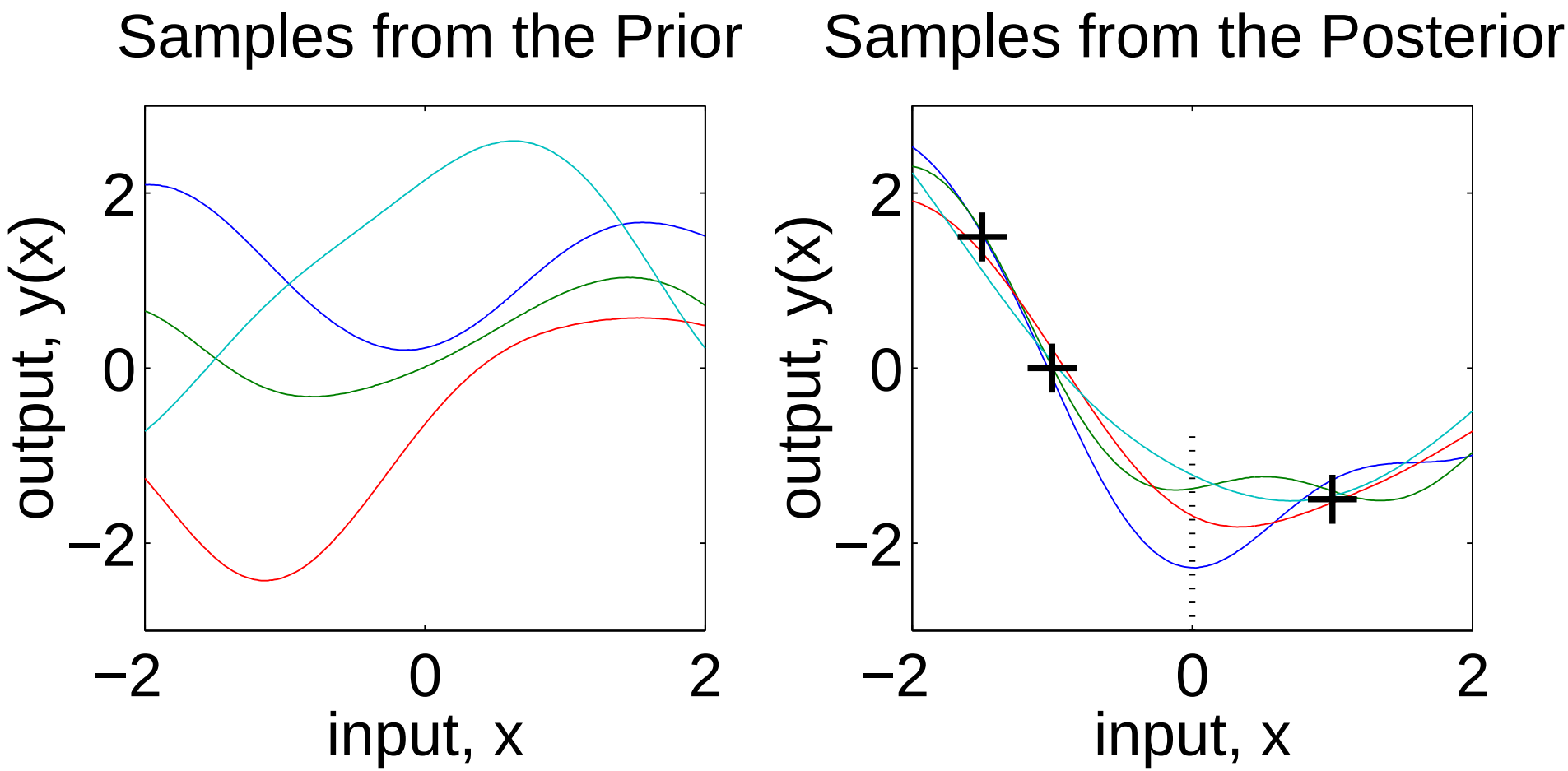


Gaussian Process with Error Bars

Bayesian inference is GPs is conceptually and algorithmically much easier than inference in large neural networks.

Williams (1995; 1996) and Rasmussen (1996) have evaluated GPs as regression models and shown that they are very good.

Gaussian Processes: prior over functions



Linear Regression $\Rightarrow$ Gaussian Processes

in four steps...

1. Linear Regression with inputs $\mathbf{x}_i$ and outputs y_i: $$y_i = \sum_k w_k x_{ik} + \epsilon_i$$

2. Kernel Linear Regression: $$y_i = \sum_k w_k \phi_k(\mathbf{x}_i) + \epsilon_i$$

3. Bayesian Kernel Linear Regression:

$$w_k \sim N(0, \beta_k) \quad \text{[indep. of } w_\ell\text{]}, \qquad \epsilon_i \sim N(0, \sigma^2)$$

4. Now, *integrate out* the weights, w_k:

$$\langle y_i \rangle = 0, \qquad \langle y_i y_j \rangle = \sum_k \beta_k \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) + \delta_{ij}\sigma^2 \equiv C_{ij}$$

This is a Gaussian process with covariance function:

$$C(\mathbf{x}, \mathbf{x}') = \sum_k \beta_k \phi_k(\mathbf{x})\phi_k(\mathbf{x}') + \delta_{ij}\sigma^2 \equiv C_{ij}$$

This is a Gaussian process with finite number of basis functions. Many useful GP covariance functions correspond to infinitely many kernels.

Infinite Models 2: Infinite Gaussian Mixtures

Following Neal (1991), Rasmussen (2000) showed that it is possible to do inference in countably infinite mixtures of Gaussians.

$$P(x_1, \ldots, x_N | \boldsymbol{\pi}, \mu, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j \, \mathcal{N}(x_i | \mu_j, \Sigma_j)$$

$$= \sum_{\mathbf{s}} P(\mathbf{s}, \mathbf{x} | \boldsymbol{\pi}, \mu, \boldsymbol{\Sigma}) = \sum_{\mathbf{s}} \prod_{i=1}^{N} \prod_{j=1}^{K} [\pi_j \, \mathcal{N}(x_i | \mu_j, \Sigma_j)]^{\delta(s_i, j)}$$

Joint distribution of indicators is **multinomial**

$$P(s_1, \ldots, s_N | \boldsymbol{\pi}) = \prod_{j=1}^{K} \pi_j^{n_j}, \qquad n_j = \sum_{i=1}^{N} \delta(s_i, j) \; .$$

Mixing proportions are given symmetric Dirichlet **prior**

$$P(\boldsymbol{\pi} | \beta) = \frac{\Gamma(\beta)}{\Gamma(\beta/K)^K} \prod_{j=1}^{K} \pi_j^{\beta/K - 1}$$

Infinite Gaussian Mixtures (continued)

Joint distribution of indicators is **multinomial**

$$P(s_1, \ldots, s_N|\boldsymbol{\pi}) = \prod_{j=1}^{K} \pi_j^{n_j}, \qquad n_j = \sum_{i=1}^{N} \delta(s_i, j) \ .$$

Mixing proportions are given symmetric Dirichlet **conjugate prior**

$$P(\boldsymbol{\pi}|\beta) = \frac{\Gamma(\beta)}{\Gamma(\beta/K)^K} \prod_{j=1}^{K} \pi_j^{\beta/K-1}$$

Integrating out the mixing proportions we obtain

$$P(s_1, \ldots, s_N|\beta) = \int d\boldsymbol{\pi} \ P(s_1, \ldots, s_N|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\beta) = \frac{\Gamma(\beta)}{\Gamma(n+\beta)} \prod_{j=1}^{K} \frac{\Gamma(n_j + \beta/K)}{\Gamma(\beta/K)}$$

This yields a Dirichlet Process over indicator variables.

Dirichlet Process Conditional Probabilities

Conditional Probabilities: Finite K

$$P(s_i = j | \mathbf{s}_{-i}, \beta) = \frac{n_{-i,j} + \beta/K}{N - 1 + \beta}$$

where $\mathbf{s}_{-i}$ denotes all indices except i, and $n_{-i,j}$ is total number of observations of indicator j excluding i^{th}.
DP: more populous classes are more more likely to be joined

Conditional Probabilities: Infinite K

Taking the limit as $K \to \infty$ yields the conditionals

$$P(s_i = j | \mathbf{s}_{-i}, \beta) = \begin{cases} \frac{n_{-i,j}}{N-1+\beta} & j \text{ represented} \\ \frac{\beta}{N-1+\beta} & \text{all } j \text{ not represented} \end{cases}$$

Left over mass, β, $\Rightarrow$ **countably infinite** number of indicator settings.
Gibbs sampling from posterior of indicators is easy!

Other infinite models

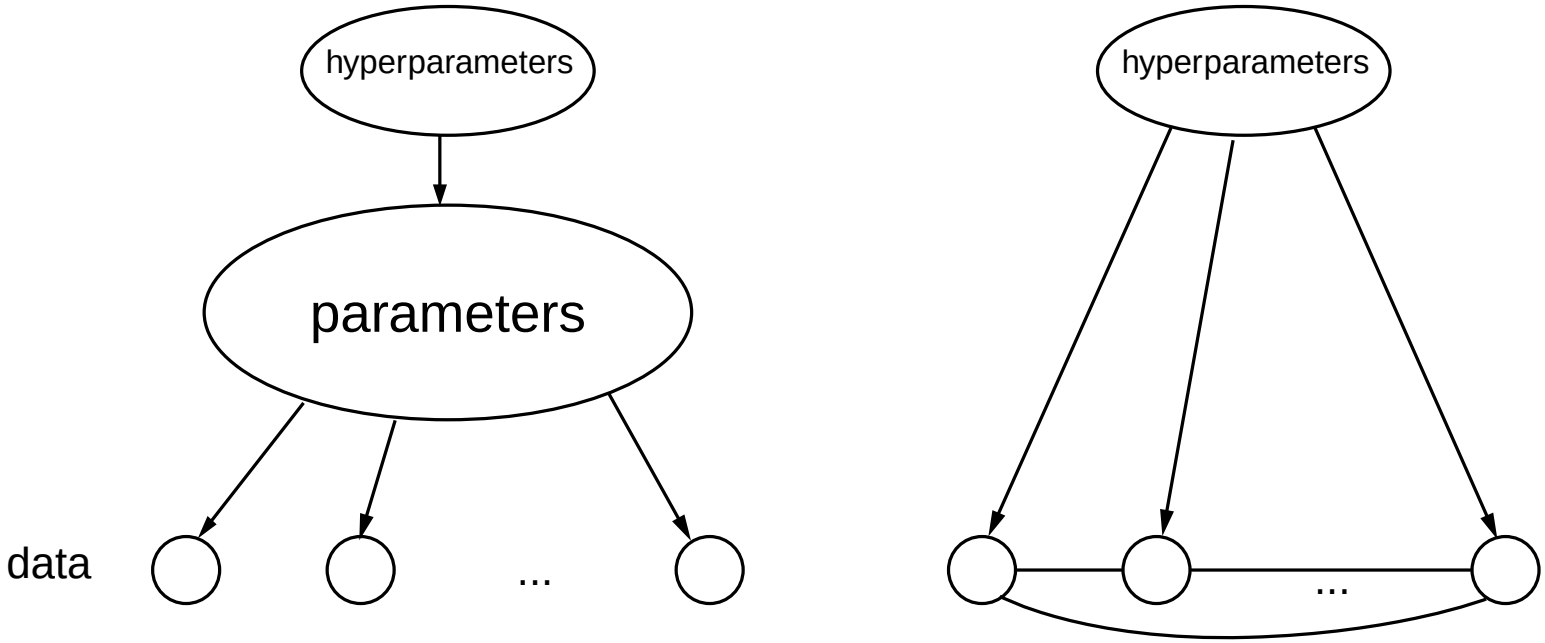
Infinite Mixtures of Experts where each each expert is a Gaussian process

Infinite hidden Markov Models

etc...

Which view: model order selection or nonparametric models?

I think, in theory, the large/infinite models view is more natural and preferable. But models become nonparametric and often require sampling or $\mathcal{O}(n^3)$ computations (e.g. GPs).



In practice, model order selection via Occam's razor is sometimes attractive, yielding smaller models and allowing deterministic (e.g. variational) approximation methods.

Summary & Conclusions

- Bayesian learning avoids overfitting and can be used to do model selection.
- Two views: model selection via Occam's Razor, versus large/infinite models.
- View 1 - a practical approach: variational approximations
  - Variational EM for CE models and propagation algorithms
- View 2 - Gaussian processes, infinite mixtures, mixture of experts & HMMs.
  - Results in non-parametric models, often requires sampling.
- In the limit of small amounts of data, we don't necessarily favour small models — rather the posterior over model orders becomes flat.
- The two views can be reconciled in the following way: Model complexity $\neq$ number of parameters, Occam's razor can still work selecting between different infinite models (e.g. rough vs smooth GPs).

Some (Biased) References

1. Attias H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. Proc. 15th Conference on Uncertainty in Artificial Intelligence.
2. Barber D., Bishop C. M., (1998) Ensemble Learning for MultiLayer Networks. Advances in Neural Information Processing Systems 10..
3. Bishop, C. M. (1999). Variational principal components. Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99 (pp. 509–514).
4. Beal, M. J., Ghahramani, Z. and Rasmussen, C. E. (2001) The Infinite Hidden Markov Model. To appear in NIPS 14.
5. Ghahramani, Z. and Beal, M.J. (1999) Variational inference for Bayesian mixtures of factor analysers. In Neural Information Processing Systems 12.
6. Ghahramani, Z. and Beal, M.J. (2000) Propagation algorithms for variational Bayesian learning. In Neural Information Processing Systems 13
7. Hinton, G. E., and van Camp, D. (1993) Keeping neural networks simple by minimizing the description length of the weights. In Proc. 6th Annu. Workshop on Comput. Learning Theory , pp. 5–13. ACM Press, New York, NY.
8. MacKay, D. J. C. (1995) Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6: 469–505.
9. Miskin J. and D. J. C. MacKay, Ensemble learning independent component analysis for blind separation and deconvolution of images, in Advances in Independent Component Analysis, M. Girolami, ed., pp. 123–141, Springer, Berlin, 2000.
10. Neal, R. M. (1991) Bayesian mixture modeling by Monte Carlo simulation, Technical Report CRG-TR-91-2, Dept. of Computer Science, University of Toronto, 23 pages.
11. Neal, R. M. (1994) Priors for infinite networks, Technical Report CRG-TR-94-1, Dept. of Computer Science, University of Toronto, 22 pages.

12. Rasmussen, C. E. (1996) Evaluation of Gaussian Processes and other Methods for Non-Linear Regression. Ph.D. thesis, Graduate Department of Computer Science, University of Toronto.
13. Rasmussen, C. E. (1999) The Infinite Gaussian Mixture Model. Advances in Neural Information Processing Systems 12, S.A. Solla, T.K. Leen and K.-R. Mller (eds.), pp. 554-560, MIT Press (2000).
14. Rasmussen, C. E and Ghahramani, Z. (2000) Occam's Razor. Advances in Neural Information Systems 13, MIT Press. (2001).
15. Rasmussen, C. E and Ghahramani, Z. (2001) Infinite Mixtures of Gaussian Process Experts. In NIPS 14.
16. Ueda, N. and Ghahramani, Z. (2000) Optimal model inference for Bayesian mixtures of experts. IEEE Neural Networks for Signal Processing. Sydney, Australia.
17. Waterhouse, S., MacKay, D.J.C. & Robinson, T. (1996). Bayesian methods for mixtures of experts. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8. Cambridge, MA: MIT Press.
18. Williams, C. K. I., and Rasmussen, C. E. (1996) Gaussian processes for regression. In Advances in Neural Information Processing Systems 8 , ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo.