Unsupervised Learning

Bayesian Model Comparison

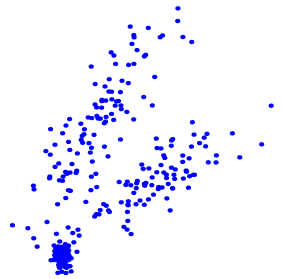
Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

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

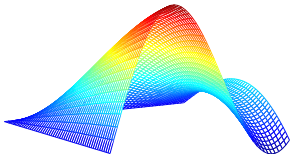
Term 1, Autumn 2006

Learning Model Structure

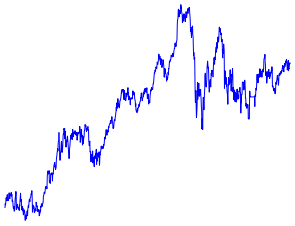
How many clusters in the data?



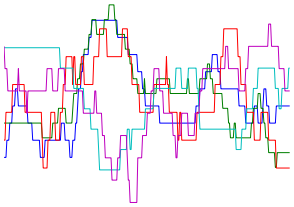
What is the intrinsic dimensionality of the data?



Is this input relevant to predicting that output?



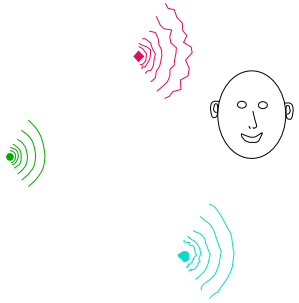
What is the order of a dynamical system?



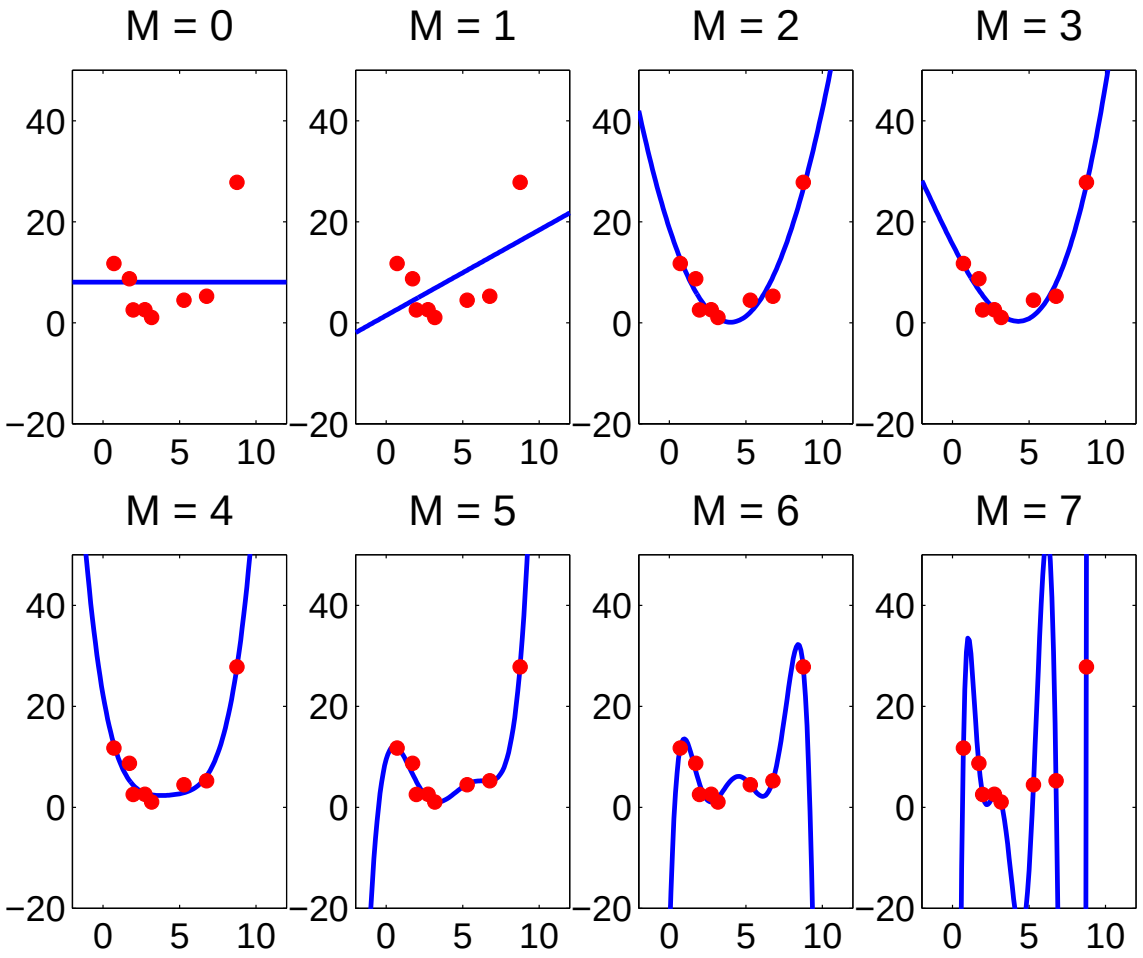
How many states in a hidden Markov model?

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY

How many auditory sources in the input?



Model complexity and overfitting: a simple example



Learning Model Structure

Models labeled by m have parameters $\boldsymbol{\theta}_m$. Which model is correct?
ML (or MAP) has no good answer: $P(\mathcal{D}|\boldsymbol{\theta}_m^{\text{ML}})$ is always larger for more complex (nested) models.

Neyman-Pearson hypothesis testing

- For nested models. Starting with simplest model ($m = 1$), compare (e.g. by likelihood ratio test) null hypothesis m to alternative $m + 1$. Continue until $m + 1$ is rejected.
- Usually only valid asympotically in data number.
- Conservative (N-P hypothesis tests are asymmetric).

Likelihood validation

- Partition data into disjoint *training* and *test* data sets $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{ts}}$. Choose model with greatest $P(\mathcal{D}_{\text{ts}}|\boldsymbol{\theta}_m^{\text{ML}})$, with $\boldsymbol{\theta}_m^{\text{ML}} = \text{argmax}\, P(\mathcal{D}_{\text{tr}}|\boldsymbol{\theta})$.
- Unbiased, but often high-variance.
- Cross-validation uses multiple partitions and averages likelihoods.

Bayesian model selection

- Choose most likely model: argmax $P(m|\mathcal{D})$.
- Principled (from a probabilistic viewpoint), but dependent on assumed priors etc.
- Can use posterior probabilities to weight models for combined predictions (no need to select at all).

The Bayesian Occam's Razor

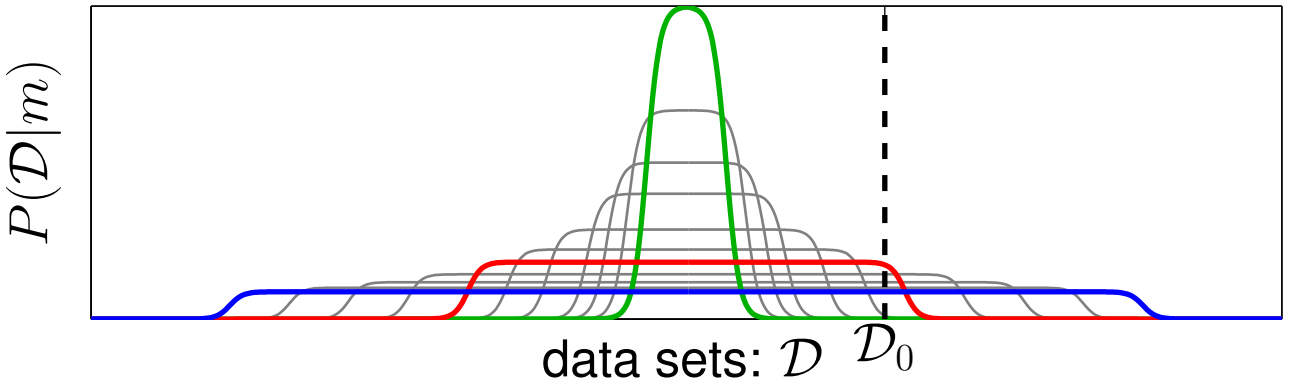
Compare model classes m using their posterior probability given the data:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}, \qquad P(\mathcal{D}|m) = \int_{\Theta_m} P(\mathcal{D}|\boldsymbol{\theta}_m, m)P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m$$

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

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 Comparison: Terminology

- A model class m is a set of models parameterised by $\boldsymbol{\theta}_m$, e.g. the set of all possible mixtures of m Gaussians.
- The marginal likelihood of model class m:

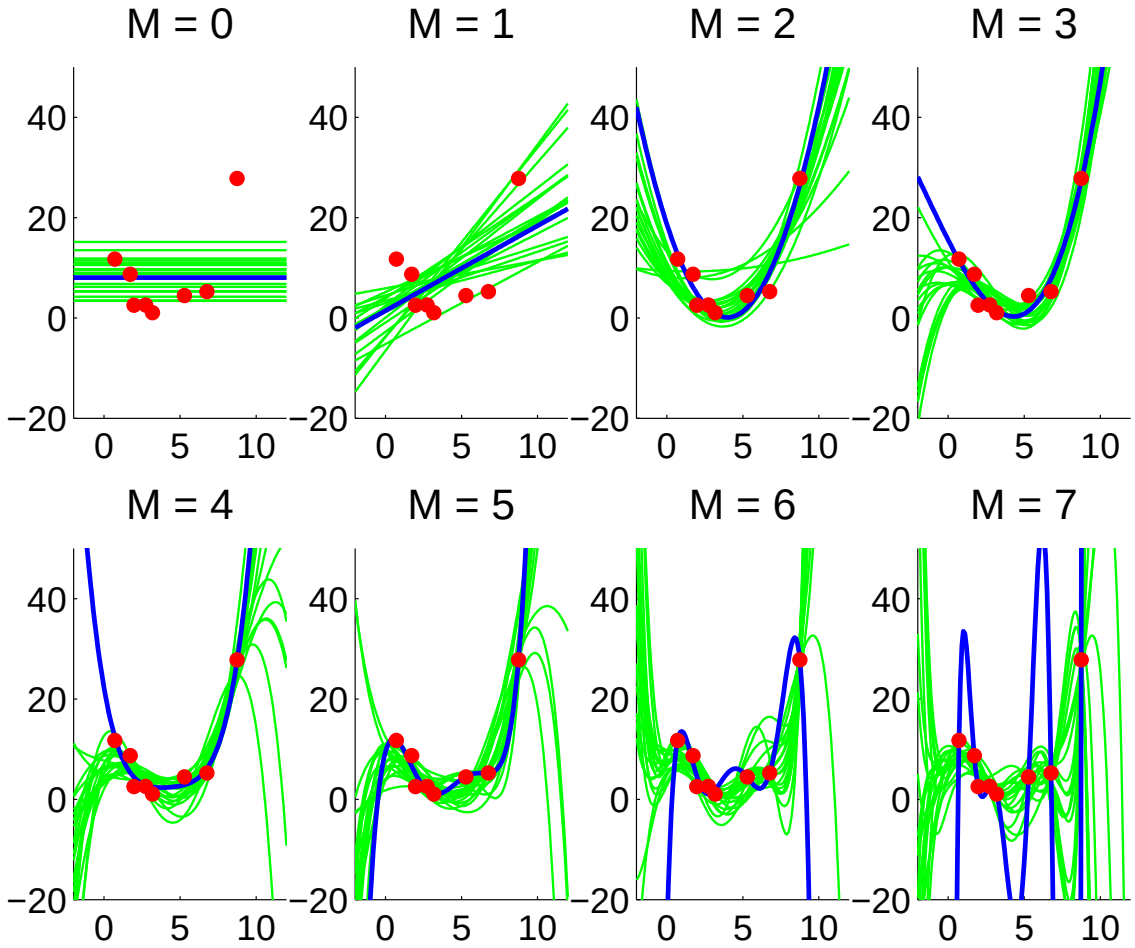
$$P(\mathcal{D}|m) = \int_{\Theta_m} P(\mathcal{D}|\boldsymbol{\theta}_m, m) P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m$$

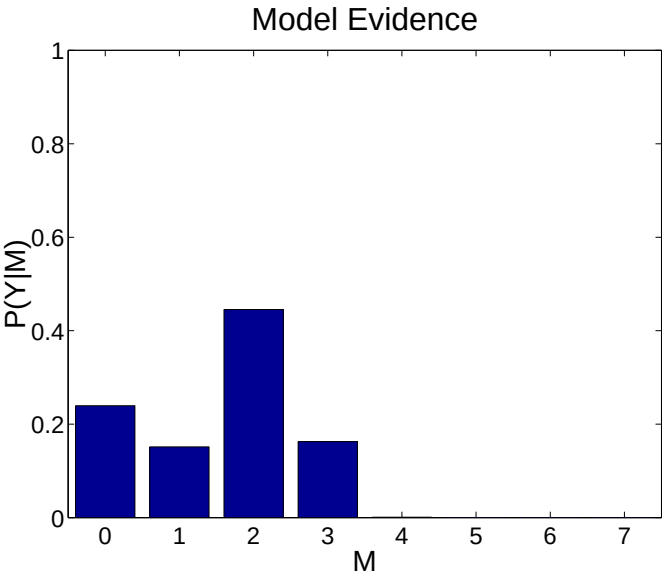
  is also known as the Bayesian evidence for model m.
- The ratio of two marginal likelihoods (or sometimes its log) is known as the Bayes factor:

$$\frac{P(\mathcal{D}|m)}{P(\mathcal{D}|m')}$$

- The Occam's Razor principle is, roughly speaking, that one should prefer simpler explanations than more complex explanations.
- Bayesian inference formalises and *automatically* implements the Occam's Razor principle, generaly preferring the simplest of nested model that can account for the data.

Bayesian Model Comparison: 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 $\boldsymbol{\theta}_2 = [a_0\ a_1\ a_2\ \tau]$

Integrals again

Can we compute $P(\mathcal{D}|m)$? Sometimes.

Suppose $P(\mathcal{D}|\boldsymbol{\theta}_m, m)$ is a member of the exponential family:

$$P(\mathbf{x}|\theta_m, m) = e^{\mathbf{s}(\mathbf{x})^\mathsf{T}\boldsymbol{\theta}_m - A(\boldsymbol{\theta}_m)}.$$

If our prior on $\boldsymbol{\theta}_m$ is conjugate:

$$P(\boldsymbol{\theta}_m|m) = e^{\mathbf{s}_p^\mathsf{T}\boldsymbol{\theta}_m - n_p A(\boldsymbol{\theta}_m)} / Z(\mathbf{s}_p, n_p)$$

then the joint is in the same family:

$$P(\mathcal{D}, \boldsymbol{\theta}_m|m) = e^{\left(\sum_i \mathbf{s}(\mathbf{x}_i) + \mathbf{s}_p\right)^\mathsf{T}\boldsymbol{\theta}_m - (N+n_p)A(\boldsymbol{\theta}_m)} / Z(\mathbf{s}_p, p)$$

and so:

$$P(\mathcal{D}|m) = \int d\boldsymbol{\theta}_m \; P(\mathcal{D}, \boldsymbol{\theta}_m|m) = Z\left(\sum_i \mathbf{s}(\mathbf{x}_i) + \mathbf{s}_p, N + n_p\right) / Z(\mathbf{s}_p, p)$$

But this is a special case. In general, we need to approximate ...

Practical Bayesian approaches

- Laplace approximations:
  - Makes a Gaussian approximation about the maximum *a posteriori* parameter estimate.
- Bayesian Information Criterion (BIC)
  - an asymptotic approximation.
- 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

This list is not exhaustive. There are a number of other deterministic approximations, including those based on, e.g. Bethe approximations and Expectation Propagation.

Laplace Approximation

We want to find $P(\mathcal{D}|m) = \int d\boldsymbol{\theta}_m \; P(\mathcal{D}, \boldsymbol{\theta}_m|m)$.

As data size N grows (relative to # of parameter, d), $\boldsymbol{\theta}_m$ becomes more constrained $\Rightarrow$ $P(\mathcal{D}, \boldsymbol{\theta}_m|m) \propto P(\boldsymbol{\theta}_m|\mathcal{D}, m)$ becomes concentrated on mode $\boldsymbol{\theta}^*$

Idea: approximate **log**$P(\mathcal{D}, \boldsymbol{\theta}_m|m)$ to second-order around $\boldsymbol{\theta}^*$.

$$
\begin{aligned}
\int d\boldsymbol{\theta}_m \; P(\mathcal{D}, \boldsymbol{\theta}_m|m) &= \int d\boldsymbol{\theta}_m \; e^{\log P(\mathcal{D}, \boldsymbol{\theta}_m|m)} \\
&= \int d\boldsymbol{\theta}_m \; e^{\log P(\mathcal{D}, \boldsymbol{\theta}^*|m) + \nabla \log P(\mathcal{D}, \boldsymbol{\theta}^*|m) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathsf{T}} \nabla\nabla \log P(\mathcal{D}, \boldsymbol{\theta}^*|m)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)} \\
&= \int d\boldsymbol{\theta}_m \; P(\mathcal{D}, \boldsymbol{\theta}^*|m) e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathsf{T}} A (\boldsymbol{\theta} - \boldsymbol{\theta}^*)} \\
&= P(\mathcal{D}|\boldsymbol{\theta}^*, m) P(\boldsymbol{\theta}^*|m) (2\pi)^{\frac{d}{2}} |A|^{-\frac{1}{2}}
\end{aligned}
$$

with A the negative of the Hessian matrix of $\log P(\mathcal{D}, \boldsymbol{\theta}|m)$ evaluated at $\boldsymbol{\theta}^*$. Note that we use the fact that the gradient at the mode vanishes.

This is equivalent to approximating the posterior by a Gaussian: an approximation which is asymptotically correct.

Bayesian Information Criterion (BIC)

BIC can be obtained from the Laplace approximation:

$$\log P(\mathcal{D}|m) \approx \log P(\boldsymbol{\theta}^*_m|m) + \log P(\mathcal{D}|\boldsymbol{\theta}^*_m, m) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|A|$$

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 $\log|A| \to \log|NA_0| = \log(N^d|A_0|) = d\log N + \log|A_0|$. Retaining only terms that grow in N we get:

$$\log P(\mathcal{D}|m) \approx \log P(\mathcal{D}|\boldsymbol{\theta}^*_m, m) - \frac{d}{2}\log N$$

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 related to the "Minimum Description Length" (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!

Sampling Approximations

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

$$\begin{aligned}
\log P(\mathcal{D}|m) &= \log \int_{\Theta_m} P(\mathcal{D}|\boldsymbol{\theta}_m, m) P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m \\
&= \log \int_{\Theta_m} P(\mathcal{D}|\boldsymbol{\theta}_m, m) \frac{P(\boldsymbol{\theta}_m|m)}{Q(\boldsymbol{\theta}_m)} Q(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m \\
&\approx \log \frac{1}{K} \sum_k P(\mathcal{D}|\boldsymbol{\theta}_m^{(k)}, m) \frac{P(\boldsymbol{\theta}_m^{(k)}|m)}{Q(\boldsymbol{\theta}_m^{(k)})}
\end{aligned}$$

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

In general, importance sampling does not work well in high dimensions. However, we can use MCMC techniques: Create a **Markov chain,** $Q_k \to Q_{k+1} \ldots$ for which:

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

Variational Bayesian Learning

Lower Bounding the Marginal Likelihood

Let the hidden latent variables be $\mathcal{Y}$, data $\mathcal{X}$ and the parameters $\boldsymbol{\theta}$.

Lower bound the marginal likelihood (Bayesian model evidence) using Jensen's inequality:

$$\begin{aligned}
\log P(\mathcal{X}) &= \log \int d\mathcal{Y}\, d\boldsymbol{\theta}\; P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}) && ||m \\
&= \log \int d\mathcal{Y}\, d\boldsymbol{\theta}\; Q(\mathcal{Y}, \boldsymbol{\theta}) \frac{P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{Q(\mathcal{Y}, \boldsymbol{\theta})} \\
&\geq \int d\mathcal{Y}\, d\boldsymbol{\theta}\; Q(\mathcal{Y}, \boldsymbol{\theta}) \log \frac{P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{Q(\mathcal{Y}, \boldsymbol{\theta})}.
\end{aligned}$$

The saturating $Q(\mathcal{Y}, \boldsymbol{\theta}) = P(\mathcal{Y}, \boldsymbol{\theta}|\mathcal{X})$ is almost always intractable.
Use a simpler, factorised approximation $Q(\mathcal{Y}, \boldsymbol{\theta}) = Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$\begin{aligned}
\log P(\mathcal{X}) &\geq \int d\mathcal{Y}\, d\boldsymbol{\theta}\; Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\
&= \mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathcal{X}).
\end{aligned}$$

Maximize this lower bound. The resulting value is the approximation to the evidence.

Variational Bayesian Learning . . .

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

$$Q^*_{\mathcal{Y}}(\mathcal{Y}) \;\propto\; \exp\left\langle \log P(\mathcal{Y},\mathcal{X}|\boldsymbol{\theta})\right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \qquad E-like\ step$$

$$Q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \;\propto\; P(\boldsymbol{\theta}) \exp\left\langle \log P(\mathcal{Y},\mathcal{X}|\boldsymbol{\theta})\right\rangle_{Q_{\mathcal{Y}}(\mathcal{Y})} \qquad M-like\ step$$

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

$$\begin{aligned} \log P(\mathcal{X}) - \mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathcal{X}) \;&=\\ \log P(\mathcal{X}) - \int d\mathcal{Y}\, d\boldsymbol{\theta}\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})}{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \;&=\\ \int d\mathcal{Y}\, d\boldsymbol{\theta}\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\mathcal{Y},\boldsymbol{\theta}|\mathcal{X})} \;&=\; KL(Q||P) \end{aligned}$$

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:

$$P(\mathcal{Y}, \mathcal{X}|\boldsymbol{\theta}) = f(\mathcal{Y}, \mathcal{X})\ g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathcal{Y}, \mathcal{X})\right\}$$

  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

Given an iid data set $\mathcal{X} = (\mathcal{X}_1, \ldots \mathcal{X}_n)$, 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 $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_i \overline{\mathbf{u}}(\mathcal{Y}_i, \mathcal{X}_i)$.

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

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

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 $p(\boldsymbol{\theta}|\mathcal{X}, m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) = p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})\, d\mathcal{Y}$$

Variational Bayesian EM

Goal: lower bound $p(\mathcal{X}|m)$

VB-E Step: compute

$$q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) = p(\mathcal{Y}|\mathcal{X}, \bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \exp\left[\int q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})\, d\mathcal{Y}\right]$$

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**, $\bar{\boldsymbol{\phi}}$.

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)

Hyperparameters and Evidence Optimisation

In some cases, we need to choose between a family of continuously parameterised models.

$$p(\mathcal{D}|\eta) = \int d\theta_m p(\mathcal{D}|\theta_m) p(\theta_m|\eta)$$

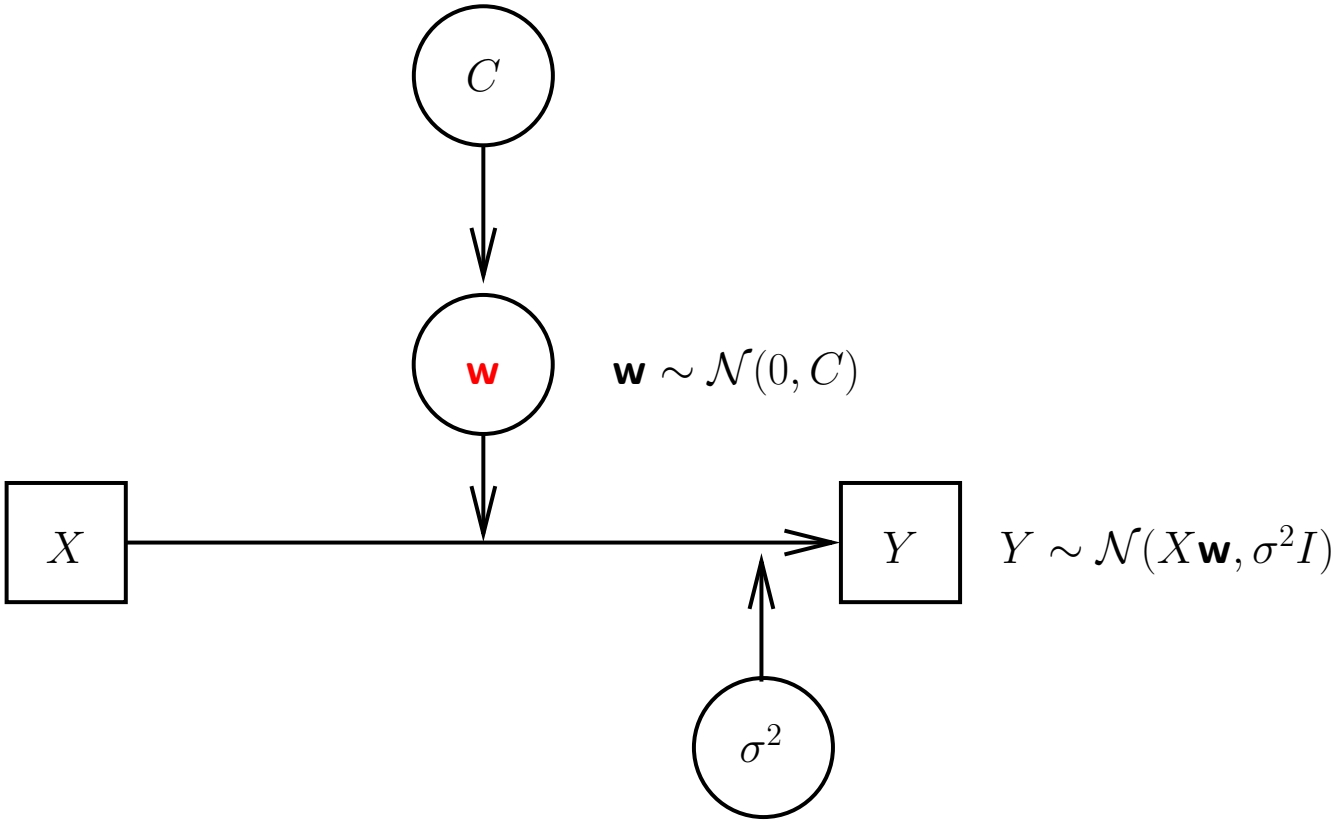
η: hyperparameters

This can often be done by gradient ascent in:

- The exact evidence (if tractable).
- Approximated evidence (Laplace, EP, Bethe, ...)
- Free-energy bound on the evidence (VB)
- Samples *with fixed random generators*

Evidence Optimisation – a supervised example

Consider simple linear regression:



- Assume a generative process for the regression weights.
- Maximize $\mathrm{P}\left(Y \mid X, C, \sigma^2\right) = \int d\mathbf{w}\ \mathrm{P}\left(Y \mid X, \mathbf{w}, \sigma^2\right) \mathrm{P}\left(\mathbf{w} \mid C\right)$ to find optimal C, σ^2.
- Estimate $\mathbf{w} = \mathrm{argmax}\ \mathrm{P}\left(Y \mid X, \mathbf{w}, \sigma^2\right) \mathrm{P}\left(\mathbf{w} \mid C\right)$ given these optimal values.

The Evidence for Linear Regression

The posterior on **w** is normal, with variance $\Sigma = (\frac{XX^\mathsf{T}}{\sigma^2} + C^{-1})^{-1}$ and mean $\mu = \Sigma\frac{XY^\mathsf{T}}{\sigma^2}$.

The evidence, $\mathcal{E}(C, \sigma^2) = \int d\mathbf{w}\ \mathsf{P}\left(Y \mid X, \mathbf{w}, \sigma^2\right) \mathsf{P}\left(\mathbf{w} \mid C\right)$, is given by:

$$\mathcal{E} = \sqrt{\frac{|2\pi\Sigma|}{|2\pi\sigma^2 I|\,|2\pi C|}} \exp -\frac{1}{2} Y \left(\frac{I}{\sigma^2} - \frac{X^\mathsf{T}\Sigma X}{\sigma^4}\right) Y^\mathsf{T}$$

For optimization, general forms for the gradients are available. If θ is a parameter in C:

$$\frac{\partial}{\partial\theta}\log\mathcal{E} = \frac{1}{2}\mathsf{Tr}\left[(C - \Sigma - \mu\mu^\mathsf{T})\frac{\partial}{\partial\theta}C^{-1}\right]$$
$$\frac{\partial}{\partial\sigma^2}\log\mathcal{E} = \frac{1}{\sigma^2}\left(-T + \mathsf{Tr}\left[I - \Sigma C^{-1}\right] + \frac{1}{\sigma^2}(Y - \mu^\mathsf{T}X)(Y - \mu^\mathsf{T}X)^\mathsf{T}\right)$$

ARD

The standard form of evidence optimization for regression (due to MacKay and Neal [3]) takes $C^{-1} = \text{diag}(\boldsymbol{\alpha})$ (i.e. $w_i \sim \mathcal{N}(0, \alpha_i^{-1})$) and then optimizes the precisions $\{\alpha_i\}$. Setting the gradients to 0 and solving gives

$$\alpha_i^{\text{new}} = \frac{1 - \alpha_i \Sigma_{ii}}{\mu_i^2}$$

$$(\sigma^2)^{\text{new}} = \frac{(Y - \mu^\mathsf{T} X)(Y - X^\mathsf{T} \mu)}{T - \sum_i (1 - \Sigma_{ii} \alpha_i)}$$

During optimization the α_i's meet one of two fates

$$\begin{array}{rcll} \alpha_i \to \infty & \Rightarrow & w_i = 0 & \textit{irrelevant} \\ \alpha_i \text{ finite} & \Rightarrow w_i = \text{argmax P}\,(w_i \mid X, Y, \alpha_i) & & \textit{relevant} \end{array}$$

This procedure, Automatic Relevance Determination (ARD), yields sparse solutions that improve on ML regression.

Evidence optimisation is also called maximum marginal likelihood or ML-2.

ARD for unsupervised learning

A similar idea can be used with Variational Bayesian methods to learn the dimensionality of a latent space. Consider factor analysis:

$$\mathbf{x} \sim \mathcal{N}(\Lambda \mathbf{y}, \Psi) \qquad \mathbf{y} \sim \mathcal{N}(0, I)$$

with a prior

$$\Lambda_i \sim \mathcal{N}\left(0, \alpha_i^{-1} I\right)$$

The VB free energy is a function of the data, $Q_{\mathcal{Y}}(\mathcal{Y})$, $Q_\Lambda(\Lambda)$ and $\boldsymbol{\alpha}$:

$$\mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_\Lambda(\Lambda), \mathcal{X}, \boldsymbol{\alpha}) = \left\langle \log P(\mathcal{X}, \mathcal{Y}|\Lambda, \Psi) + \log P(\Lambda|\boldsymbol{\alpha}) + \log P(\Psi) \right\rangle_{Q_{\mathcal{Y}} Q_\Lambda} + \mathbf{H}[Q_{\mathcal{Y}}] + \mathbf{H}[Q_\Lambda]$$

Optimising this wrt the distributions and $\boldsymbol{\alpha}$ in turn (like EM) causes some α_i to diverge, restricting the effective dimensionality of $\mathbf{y}$.

Practical Bayesian approaches

- Laplace approximations:
  - Makes a Gaussian approximation about the maximum *a posteriori* parameter estimate.
- Bayesian Information Criterion (BIC)
  - an asymptotic approximation.
- 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

This list is not exhaustive. There are a number of other deterministic approximations, including those based on, e.g. Bethe approximations and Expectation Propagation.