Probabilistic & Unsupervised Learning

Variational EM
and Variational Bayesian Learning

Yee Whye Teh
ywteh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London

Term 1, Autumn 2011
Integrals in Statistical Modelling

- **Parameter estimation**

\[ \hat{\theta} = \arg\max_{\theta} \int dY \ P(Y|\theta)P(X,Y,\theta) \]

(or using EM)

\[ \theta^{\text{new}} = \arg\max_{\theta} \int dY \ P(Y|X,\theta^{\text{old}}) \log P(X,Y|\theta) \]

- **Prediction**

\[ p(x|D, m) = \int d\theta \ p(\theta|D, m)p(x|\theta, D, m) \]

- **Model selection or weighting** (by marginal likelihood)

\[ p(D|m) = \int d\theta \ p(\theta|m)p(D|\theta, m) \]

These integrals are often intractable:

- **Analytic intractability**: integrals may not have closed form in non-linear, non-Gaussian models ⇒ numerical integration.

- **Computational intractability**: Numerical integral (or sum if \( Y \) or \( \theta \) are discrete) may be exponential in data or model size.
Examples of Intractability

- Bayesian marginal likelihood/model evidence for Mixture of Gaussians: exact computations are exponential in number of data points

\[ p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \int d\theta \left( \prod_{i=1}^{N} \sum_{s_i} p(\mathbf{x}_i|s_i, \theta) p(s_i|\theta) \right) \]

\[ = \sum_{s_1} \sum_{s_2} \ldots \sum_{s_N} \int d\theta \left( \prod_{i=1}^{N} p(\mathbf{x}_i|s_i, \theta) p(s_i|\theta) \right) \]

- Computing the conditional probability of a variable in a very large multiply connected directed graphical model:

\[ p(x_i|X_j = a) = \sum_{\text{all settings of } y_i \neq j} p(x_i, y, X_j = a) / p(X_j = a) \]

- Computing the hidden state distribution in a general nonlinear dynamical system

\[ p(y_t|\mathbf{x}_1, \ldots, \mathbf{x}_T) \propto \int p(y_t|y_{t-1}) p(\mathbf{x}_t|y_t) p(\mathbf{y}_{t-1}|\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_t|y_t) d\mathbf{y}_{t-1} \]
In the FHMM, moralisation puts simultaneous states $s_t^{(1)}, s_t^{(2)}, s_t^{(3)}$ into a single clique.

- $M$ state variables, $K$ values $\Rightarrow$ sums over $K^{2M}$ terms.
- Factorial prior $\nRightarrow$ Factorial posterior (explaining away).

Variational methods approximate the posterior, often in a factored form. To see how they work, we need to review the free-energy interpretation of EM.
The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta$.

**Goal:** Maximize the log likelihood (i.e. ML learning) wrt $\theta$:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen’s inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} d\mathcal{Y} \overset{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} d\mathcal{Y} = \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) d\mathcal{Y}$$

$$= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y} + H[q],$$

where $H[q]$ is the entropy of $q(\mathcal{Y})$.

So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + H[q]$$
The E and M steps of EM

The log likelihood is bounded below (Jensen) by:

\[ F(q, \theta) = \langle \log P(Y, X|\theta) \rangle_{q(Y)} + H[q], \]

EM alternates between:

**E step:** optimise \( F(q, \theta) \) wrt distribution over hidden variables holding parameters fixed:

\[ q^{(k)}(Y) := \arg\max_{q(Y)} F(q(Y), \theta^{(k-1)}) = P(Y|X, \theta^{(k-1)}) \]

**M step:** maximise \( F(q, \theta) \) wrt parameters holding hidden distribution fixed:

\[ \theta^{(k)} := \arg\max_{\theta} F(q^{(k)}(Y), \theta) = \arg\max_{\theta} \langle \log P(Y, X|\theta) \rangle_{q^{(k)}(Y)} \]
EM as Coordinate Ascent in $\mathcal{F}$
EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

\[
    \ell(\theta^{(k-1)}) = \mathcal{E}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)}),
\]

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt $\theta$.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL.

If the M-step is executed so that $\theta^{(k)} \neq \theta^{(k-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of $\theta$ iff the likelihood increases.
Variational Approximations to the EM algorithm

What if finding expected sufficient stats under $P(Y|X, \theta)$ is computationally intractable?

In the **generalised EM** algorithm, we argued that intractable maximisations could be replaced by gradient M-steps. For the E-step we could:

- **Parameterise** $q = q_\rho(Y)$ and take a gradient step in $\rho$.
- **Assume** some simplified form for $q$, usually factored: $q = \prod_i q_i(Y_i)$ where $Y_i$ partition $Y$, and maximise within this form.

In both cases, we assume $q \in Q$, and optimise within this class:

**VE step**: maximise $\mathcal{F}(q, \theta)$ wrt restricted latent distribution given parameters:

$$q^{(k)}(Y) := \arg\max_{q(Y) \in Q} \mathcal{F}(q(Y), \theta^{(k-1)}) \, .$$

**M step**: unchanged

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(Y), \theta) = \arg\max_{\theta} \int q^{(k)}(Y) \log p(Y, X|\theta) dY \, .$$

This maximises a lower bound on the log likelihood.
What do we lose?

What does restricting $q$ to $Q$ cost us?

- Recall that the free-energy is bounded above by Jensen:

$$
\mathcal{F}(q, \theta) \leq \ell(\theta^{\text{ML}})
$$

Thus, as long as every step increases $\mathcal{F}$, convergence is still guaranteed.

- But, since $P(\mathcal{Y}|\mathcal{X}, \theta^{(k)})$ may not lie in $Q$, we no longer saturate the bound after the E-step. Thus, the likelihood may not increase on each full EM step.

$$
\ell(\theta^{(k-1)}) \geq \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)}),
$$

- Thus, we may not converge to a maximum of $\ell$.

The hope is that by increasing a lower bound on $\ell$ we will find a decent solution. [Note that if $P(\mathcal{Y}|\mathcal{X}, \theta^{\text{ML}}) \in Q$, then $\theta^{\text{ML}}$ is a fixed point of the variational algorithm.]
Recall that

\[ F(q, \theta) = \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{q(\mathcal{Y})} + H[q] \]

\[ = \langle \log P(\mathcal{X}|\theta) + \log P(\mathcal{Y}|\mathcal{X}, \theta) \rangle_{q(\mathcal{Y})} - \langle \log q(\mathcal{Y}) \rangle_{q(\mathcal{Y})} \]

\[ = \langle \log P(\mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} - \text{KL}[q || P(\mathcal{Y}|\mathcal{X}, \theta)]. \]

Thus,

**E step maximise** \( F(q, \theta) \) wrt the distribution over latents, given parameters:

\[ q^{(k)}(\mathcal{Y}) := \text{argmax}_{q(\mathcal{Y})\in Q} F(q(\mathcal{Y}), \theta^{(k-1)}). \]

is equivalent to:

**E step minimise** \( \text{KL}[q || p(\mathcal{Y}|\mathcal{X}, \theta)] \) wrt distribution over latents, given parameters:

\[ q^{(k)}(\mathcal{Y}) := \text{argmin}_{q(\mathcal{Y})\in Q} \int q(\mathcal{Y}) \log \frac{q(\mathcal{Y})}{p(\mathcal{Y}|\mathcal{X}, \theta^{(k-1)})} d\mathcal{Y} \]

So, in each E step, the algorithm is trying to find the best approximation to \( P(\mathcal{Y}|\mathcal{X}) \) in \( Q \).

This is related to ideas in *information geometry*. 
Factored Variational E-step

The most common form of variational approximation partitions $\mathcal{Y}$ into disjoint sets $\mathcal{Y}_i$ with

$$Q = \{q \mid q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i)\}.$$ 

In this case the E-step is itself iterative:

(Factored VE step)$_i$: maximise $\mathcal{F}(q, \theta)$ wrt $q_i(\mathcal{Y}_i)$ given other $q_j$ and parameters:

$$q_i^{(k)}(\mathcal{Y}_i) := \arg\max_{q_i(\mathcal{Y}_i)} \mathcal{F}(q_i(\mathcal{Y}_i) \prod_{j \neq i} q_j(\mathcal{Y}_j), \theta^{(k-1)})$$

The $q_i$s can be updated iteratively until convergence before moving on to the M-step. Alternatively, we can make a single pass over all $q_i$ (starting from values at the last step) and then perform an M-step. Each VE step increases $\mathcal{F}$, so convergence is still guaranteed.
Factored Variational E-step

The Factored Variational E-step has a general form.

The free energy is:

$$F\left(\prod_j q_j(Y_j), \theta^{(k-1)}\right) = \langle \log P(X, Y|\theta^{(k-1)}) \rangle_{\prod_j q_j(Y_j)} + H\left[\prod_j q_j(Y_j)\right]$$

$$= \int dY_i q_i(Y_i) \langle \log P(X, Y|\theta^{(k-1)}) \rangle_{\prod_{j\neq i} q_j(Y_j)} + H[q_i] + \sum_{j \neq i} H[q_j]$$

Now, taking the variational derivative of the Lagrangian (enforcing normalisation of $q_i$):

$$\frac{\delta}{\delta q_i} \left( F + \lambda \left( \int q_i - 1 \right) \right) = \langle \log P(X, Y|\theta^{(k-1)}) \rangle_{\prod_{j \neq i} q_j(Y_j)} - \log q_i(Y_i) - 1 + \lambda$$

$$= 0 \quad \Rightarrow \quad q_i(Y_i) \propto \exp \left\langle \log P(X, Y|\theta^{(k-1)}) \right\rangle_{\prod_{j \neq i} q_j(Y_j)}$$

In general, this depends only on the expected sufficient statistics under $q_j$. Thus, once again, we don’t actually need the entire distributions, just the relevant expectations.
Mean-field Approximations

If \( Y_i = y_i \) (i.e., \( q \) is factored over all variables) then the variational technique is often called a “mean field” approximation.

Suppose \( P(\mathcal{X}, \mathcal{Y}) \) is log-linear, e.g. the Boltzmann machine:

\[
P(\mathcal{X}, \mathcal{Y}) = \frac{1}{Z} \exp \left( \sum_{ij} W_{ij} s_i s_j + \sum_i b_i s_i \right)
\]

with some \( s_i \in \mathcal{Y} \) and others observed.

Expectations wrt a fully factored \( q \) distribute over all \( s_i \in \mathcal{Y} \)

\[
\langle \log P(\mathcal{X}, \mathcal{Y}) \rangle_{\Pi q_i} = \sum_{ij} W_{ij} \langle s_i \rangle_{q_i} \langle s_j \rangle_{q_j} + \sum_i b_i \langle s_i \rangle_{q_i}
\]

(where \( q_i \) for \( s_i \in \mathcal{X} \) is a delta function on observed value).

Thus, we can update each \( q_i \) in turn given the means of the others. Each variable is seeing the mean field imposed by its neighbours. We update these fields until they all agree.
Mean-field FHMM

The mean-field approach to the FHMM with

$$q(s_{1:T}^1) = \prod_{m,t} q^m_t(s^m_t)$$

yields a variant of the usual forward-backward algorithm. Coupling between the different chains only takes place through the joint output distribution. Each update depends only on the immediate neighbours.

$$q^m_t(s^m_t) \propto \exp \left\langle \sum_{m} \sum_{t} \log P(s^m_t | s^m_{t-1}) + \sum_{t} \log P(x_t | s_{1:T}^1) \right\rangle \prod_{\neg (m',t')} q^m_{t'}(s^m_{t'})$$

$$= \exp \left\langle \sum_{m} \sum_{t} \log P(s^m_t | s^m_{t-1}) + \sum_{t} \log P(x_t | s_{1:T}^1) \right\rangle \prod_{\neg (m',t')} q^m_{t'}$$

$$\propto \exp \left[ \left\langle \log P(s^m_{t'} | s^m_{t'-1}) \right\rangle_{q^m_{t'-1}} + \left\langle \log P(s^m_{t+1} | s^m_{t'}) \right\rangle_{q^m_{t'+1}} + \left\langle \log P(x_{t'} | s_{1:T}^1) \right\rangle_{\prod_{m} q^m_{t'}} \right]$$
Structured Variational Approximations

$q(Y)$ need not be completely factorized.

For example, suppose you can partition $Y$ into sets $Y_1$ and $Y_2$ such that computing the expected sufficient statistics under $q(Y_1)$ and $q(Y_2)$ is tractable. Then $q(Y) = q(Y_1)q(Y_2)$ is tractable.

If you have a graphical model, you may want to factorize $q(Y)$ into a product of trees, which are tractable distributions.
The most natural structured approximation in the FHMM is to factor each chain from the others

\[
q(s_{1:T}^1) = \prod_m q_m(s_{1:T}^m)
\]

Updates within each chain are then found by a forward-backward algorithm, with a modified "likelihood" term.

\[
q^{m'}(s_{1:T}^{m'}) \propto \exp \left\langle \log P(s_{1:T}^{1:M}, x_{1:T}) \right\rangle \prod_{m \neq m'} q_m(s_{1:T}^m)
\]

\[
= \exp \left\langle \sum_m \sum_t \log P(s_t^m | s_t^m) + \sum_t \log P(x_t | s_{1:T}^{1:M}) \right\rangle \prod_{m \neq m'} q_m
\]

\[
\propto \exp \left[ \sum_t \log P(s_t^{m'} | s_t^{m'}) + \sum_t \left\langle \log P(x_t | s_{1:T}^{1:M}) \right\rangle \prod_{m \neq m'} q_m s_t^{m'} \right]
\]

\[
= \prod_t P(s_t^{m'} | s_t^{m'}) \prod_t e^{\left\langle \log P(x_t | s_{1:T}^{1:M}) \right\rangle \prod_{m \neq m'} q_m s_t^{m'}}
\]
Let $q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i)$.

Variational approximation maximises $\mathcal{F}$:

$$
\mathcal{F}(q) = \int q(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}) d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) d\mathcal{Y}
$$

Focusing on one term, $q_j$, we can write this as:

$$
\mathcal{F}(q_j) = \int q_j(\mathcal{Y}_j) \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{-q_j(\mathcal{Y}_j)} d\mathcal{Y}_j + \int q_j(\mathcal{Y}_j) \log q_j(\mathcal{Y}_j) d\mathcal{Y}_j + \text{const}
$$

Where $\langle \cdot \rangle_{-q_j(\mathcal{Y}_j)}$ denotes averaging w.r.t. $q_i(\mathcal{Y}_i)$ for all $i \neq j$

Optimum occurs when:

$$
q_j^*(\mathcal{Y}_j) = \frac{1}{Z} \exp \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{-q_j(\mathcal{Y}_j)}
$$
Optimum occurs when:

\[
q_j^* (\mathcal{Y}_j) = \frac{1}{Z} \exp \left\langle \log p(\mathcal{Y}, \mathcal{X}) \right\rangle_{-q_j(\mathcal{Y}_j)}
\]

Assume graphical model: \( p(\mathcal{Y}, \mathcal{X}) = \prod_i p(X_i | \text{pa}_i) \)

\[
\log q_j^*(\mathcal{Y}_j) = \left\langle \sum_i \log p(X_i | \text{pa}_i) \right\rangle_{-q_j(\mathcal{Y}_j)} + \text{const}
\]

\[
= \left\langle \log p(\mathcal{Y}_j | \text{pa}_j) \right\rangle_{-q_j(\mathcal{Y}_j)} + \sum_{k \in \text{ch}_j} \left\langle \log p(X_k | \text{pa}_k) \right\rangle_{-q_j(\mathcal{Y}_j)} + \text{const}
\]

This defines messages that get passed between nodes in the graph. Each node receives messages from its Markov boundary: parents, children and parents of children.
Variational Approximations to Bayesian Learning

\[ \log p(\mathcal{X}) = \log \int \int p(\mathcal{X}, \mathcal{Y}|\theta)p(\theta) \, d\mathcal{Y} \, d\theta \]

\[ \geq \int \int q(\mathcal{Y}, \theta) \log \frac{p(\mathcal{X}, \mathcal{Y}, \theta)}{q(\mathcal{Y}, \theta)} \, d\mathcal{Y} \, d\theta \]

Constrain \( q \in \mathcal{Q} \) s.t. \( q(\mathcal{Y}, \theta) = q(\mathcal{Y})q(\theta) \).

This results in the **variational Bayesian EM algorithm**.
Let the hidden latent variables be $\mathcal{Y}$, data $\mathcal{X}$ and the parameters $\theta$.

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

$$
\log P(\mathcal{X}) = \log \int d\mathcal{Y} d\theta \ P(\mathcal{X}, \mathcal{Y}, \theta) \\
= \log \int d\mathcal{Y} d\theta \ Q(\mathcal{Y}, \theta) \frac{P(\mathcal{X}, \mathcal{Y}, \theta)}{Q(\mathcal{Y}, \theta)} \\
\geq \int d\mathcal{Y} d\theta \ Q(\mathcal{Y}, \theta) \log \frac{P(\mathcal{X}, \mathcal{Y}, \theta)}{Q(\mathcal{Y}, \theta)}.
$$

The saturating $Q(\mathcal{Y}, \theta) = P(\mathcal{Y}, \theta|\mathcal{X})$ is almost always intractable.

Use a simpler, factorised approximation $Q(\mathcal{Y}, \theta) = Q_\mathcal{Y}(\mathcal{Y})Q_\theta(\theta)$:

$$
\log P(\mathcal{X}) \geq \int d\mathcal{Y} d\theta \ Q_\mathcal{Y}(\mathcal{Y})Q_\theta(\theta) \log \frac{P(\mathcal{X}, \mathcal{Y}, \theta)}{Q_\mathcal{Y}(\mathcal{Y})Q_\theta(\theta)} \\
= \mathcal{F}(Q_\mathcal{Y}(\mathcal{Y}), Q_\theta(\theta), \mathcal{X}).
$$

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:

$$Q^*_Y(Y) \propto \exp \langle \log P(Y, X|\theta) \rangle_{Q_\theta(\theta)}$$  \hspace{1cm} E-like step

$$Q^*_\theta(\theta) \propto P(\theta) \exp \langle \log P(Y, X|\theta) \rangle_{Q_Y(Y)}$$  \hspace{1cm} M-like step

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

$$\log P(X) - \mathcal{F}(Q_Y(Y), Q_\theta(\theta), X) =$$

$$\log P(X) - \int dY d\theta \ Q_Y(Y)Q_\theta(\theta) \log \frac{P(X, Y, \theta)}{Q_Y(Y)Q_\theta(\theta)} =$$

$$\int dY d\theta \ Q_Y(Y)Q_\theta(\theta) \log \frac{Q_Y(Y)Q_\theta(\theta)}{P(Y, \theta|X)} = KL(Q||P)$$
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:

  \[
  P(Y, X|\theta) = f(Y, X) g(\theta) \exp \left\{ \phi(\theta)^\top u(Y, X) \right\}
  \]

  where \( \phi(\theta) \) is the vector of *natural parameters*, \( u \) are *sufficient statistics*

- **Condition (2).** The prior over parameters is conjugate to this joint probability:

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

  where \( \eta \) and \( \nu \) are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- \( \eta \): number of pseudo-observations
- \( \nu \): 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) \( Q_\theta(\theta) \) is also conjugate, i.e.

\[
Q_\theta(\theta) = h(\tilde{\eta}, \tilde{\nu}) g(\theta)^{\tilde{\eta}} \exp \{ \phi(\theta)^\top \tilde{\nu} \}
\]

where \( \tilde{\eta} = \eta + n \) and \( \tilde{\nu} = \nu + \sum_i u(Y_i, X_i) \).

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

\[
Q_{Y_i}(Y_i) \propto f(Y_i, X_i) \exp \{ \overline{\phi}(\theta)^\top u(Y_i, X_i) \} = P(Y_i | X_i, \overline{\phi}(\theta))
\]

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 \( \tilde{\nu} \);

(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

<table>
<thead>
<tr>
<th><strong>EM for MAP estimation</strong></th>
<th><strong>Variational Bayesian EM</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal:</strong> maximize $p(\theta</td>
<td>\mathcal{X}, m)$ w.r.t. $\theta$</td>
</tr>
<tr>
<td><strong>E Step:</strong> compute</td>
<td><strong>VB-E Step:</strong> compute</td>
</tr>
<tr>
<td>$q^{(t+1)}(\mathcal{Y}) = p(\mathcal{Y}</td>
<td>\mathcal{X}, \theta^{(t)})$</td>
</tr>
<tr>
<td><strong>M Step:</strong></td>
<td><strong>VB-M Step:</strong></td>
</tr>
<tr>
<td>$\theta^{(t+1)} = \arg \max_{\theta} \int q^{(t+1)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}, \theta) d\mathcal{Y}$</td>
<td>$q^{(t+1)}(\theta) = \exp \left[ \int q^{(t+1)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}, \theta) d\mathcal{Y} \right]$</td>
</tr>
</tbody>
</table>

**Properties:**

- Reduces to the EM algorithm if $q_{\theta}(\theta) = \delta(\theta - \theta^*)$.
- $\mathcal{F}_m$ 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)
ARD for unsupervised learning

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

\[ \mathbf{x} \sim \mathcal{N}(\mathbf{\Lambda} \mathbf{y}, \Psi) \quad \mathbf{y} \sim \mathcal{N}(0, I) \]

with a prior

\[ \mathbf{\Lambda}_i \sim \mathcal{N}(0, \alpha_i^{-1} I) \]

The VB free energy is a function of the data, \( Q_Y(Y), Q_\Lambda(\Lambda) \) and \( \alpha \):

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

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