1

The EM algorithm, variational approximations and expectation propagation for mixtures

D. Michael Titterington

1.1 Preamble

The material in this chapter is largely tutorial in nature. The main goal is to review two types of deterministic approximation, variational approximations and the expectation propagation approach, which have been developed mainly in the computer science literature, but with some statistical antecedents, to assist approximate Bayesian inference. However, we believe that it is helpful to preface discussion of these methods with an elementary reminder of the EM algorithm as a way of computing posterior modes. All three approaches have now been applied to many model types, but we shall just mention them in the context of mixtures, and only a very small number of types of mixture at that.

Mixtures: Estimation and Applications, First Edition. Edited by Kerrie L. Mengersen, Christian P. Robert and D. Michael Titterington.

1.2 The EM algorithm

1.2.1 Introduction to the algorithm

Parameter estimation in mixture models often goes hand-in-hand with a discussion of the EM algorithm. This is especially so if the objective is maximum likelihood estimation, but the algorithm is also relevant in the Bayesian approach if maximum a posteriori estimates are required. If we have a set of data D from a parametric model, with parameter θ, probably multidimensional, and with likelihood function $p(D|\theta)$ and prior density $p(\theta)$, then the posterior density for θ is

$$p(\theta|D) \propto p(D|\theta)p(\theta),$$

and therefore the posterior mode $\hat{\theta}_{MAP}$ is the maximiser of $p(D|\theta)p(\theta) = p(D, \theta)$. Of course, if the prior density is uniform then the posterior mode is the same as the maximum likelihood estimate. If explicit formulae for the posterior mode do not exist then recourse has to be made to numerical methods, and the EM algorithm is a popular general method in contexts that involve incomplete data, either explicitly or by construct. Mixture data fall into this category, with the component membership indicators $\mathbf{z}$ regarded as missing values.

The EM algorithm is as follows. With data D and initial guess $\theta^{(0)}$ for $\hat{\theta}_{MAP}$, a sequence $\{\theta^{(m)}\}$ of values are generated from the following double-step that creates $\theta^{(m+1)}$ from $\theta^{(m)}$.

- **E-step**: Evaluate

$$\begin{aligned} Q(\theta; \theta^{(m)}) &= E\{\log[p(D, \mathbf{z}, \theta)]|D, \theta^{(m)}\} \\ &= \sum_{\mathbf{z}} \log[p(D, \mathbf{z}, \theta)]\ p(\mathbf{z}|D, \theta^{(m)}) \\ &= \sum_{\mathbf{z}} \log[p(D, \mathbf{z}|\theta)]\ p(\mathbf{z}|D, \theta^{(m)}) + p(\theta). \end{aligned}$$

- **M-step**: Find $\theta = \theta^{(m+1)}$ to maximise $Q(\theta; \theta^{(m)})$ with respect to θ.

Remarks

1. In many other incomplete data problems the missing values $\mathbf{z}$ are continuous, in which case the summation is replaced by an integration in the above.
2. Not surprisingly, $Q(\theta; \theta^{(m)})$ is usually very like a complete-data log-posterior, apart from a constant that is independent of θ, so that the M-step is easy or difficult according as calculation of the complete-data posterior mode is easy or difficult.
3. The usual monotonicity proof of the EM algorithm in the maximum-likelihood context can be used, with minimal adaptation, to show that

$$\log[p(D, \theta^{(m+1)})] \geq \log[p(D, \theta^{(m)})].$$

Thus, the EM algorithm 'improves' $p(D, \theta)$ at each stage and, provided the posterior density for θ is locally bounded, the values of $p(D, \theta^{(m)})$ should converge to a local maximum of $p(D, \theta)$. The corresponding sequence $\{\theta^{(m)}\}$ will also often converge, one hopes to $\hat{\theta}_{MAP}$, but convergence may not occur if, for instance, $p(D, \theta)$ contains a ridge. The niceties of convergence properties are discussed in detail, for maximum likelihood, in Chapter 3 of McLachlan and Krishnan (1997).

1.2.2 The E-step and the M-step for the mixing weights

Suppose now that the data are a random sample $D = \{y_1, \ldots, y_n\}$ from a distribution with probability density function

$$p(y|\theta) = \sum_{j=1}^{k} \lambda_j f_j(y|\phi_j),$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)$ are the mixing weights, the f_j are the component densities, each corresponding to a subpopulation, and k is finite. The density f_j is parameterised by ϕ_j and the set of all these is to be called ϕ. Often we shall assume that the component densities are of the same type, in which case we shall omit the subscript j from f_j. The complete set of parameters is $\theta = (\lambda, \phi)$.

The complete-data joint distribution can be conveniently written as

$$p(D, \mathbf{z}, \theta) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} [\lambda_j f_j(y_i|\phi_j)]^{z_{ij}} \right\} p(\theta),$$

with the help of the indicator notation, where $z_{ij} = 1$ if the ith observation comes from component j and is zero otherwise. Thus

$$\log[p(D, \mathbf{z}, \theta)] = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log[\lambda_j f_j(y_i|\phi_j)] + \log\, p(\theta).$$

For the E-step of the EM algorithm all that we have to compute are the expectations of the indicator variables. Given $\theta^{(m)} = (\lambda^{(m)}, \phi^{(m)})$, we obtain

$$\begin{aligned} z_{ij}^{(m)} &= P(z_{ij} = 1|y_i, \theta^{(m)}) \\ &= P(i\text{th observation belongs to component } j|y_i, \theta^{(m)}) \\ &= \frac{\lambda_j^{(m)} f_j(y_i|\phi_j^{(m)})}{\sum_l \lambda_l^{(m)} f_l(y_i|\phi_l^{(m)})}, \end{aligned}$$

for each i, j. We now have

$$\begin{aligned} Q(\theta;\theta^{(m)}) &= \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}^{(m)} \log[\lambda_j f_j(y_i|\phi_j)] + \log\, p(\theta) \\ &= \sum_{j=1}^{k} n_j^{(m)} \log \lambda_j + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}^{(m)} \log\, f_j(y_i|\phi_j) + \log\, p(\theta) \\ &= Q_1(\lambda) + Q_2(\phi) + \log\, p(\theta), \end{aligned}$$

say, where $n_j^{(m)} = \sum_{i=1}^{n} z_{ij}^{(m)}$, a 'pseudo' sample size associated with subpopulation j. (In the case of complete data, $n_j = \sum_{i=1}^{n} z_{ij}$ is precisely the sample size for subpopulation j.)

We now consider the M-step for the mixing weights λ. Before this we make some assumptions about the prior distributions. In particular, we assume that λ and ϕ are a priori independent and that the prior for λ takes a convenient form, namely that which would be conjugate were the data complete. Thus, we assume that, a priori, $\lambda \sim \mathrm{Dir}(a^{(0)})$; that is,

$$p(\lambda) \propto \prod_{j=1}^{k} \lambda_j^{a_j^{(0)}-1},$$

for prescribed hyperparameters $a^{(0)} = \{a_j^{(0)}\}$. Clearly, $\lambda^{(m+1)}$ must maximise

$$\sum_{j=1}^{k} (n_j^{(m)} + a_j^{(0)} - 1) \log \lambda_j,$$

which is essentially a log-posterior associated with multinomial data. Thus

$$\lambda_j^{(m+1)} = (n_j^{(m)} + a_j^{(0)} - 1)/(n + a_{\cdot}^{(0)} - k),$$

where $a_{\cdot}^{(0)} = \sum_j a_j^{(0)}$.

Example 1: Mixture of two known densities

In this simplest case with $k = 2$, write $\lambda_1 = \lambda = 1 - \lambda_2$. Then the iteration is

$$\lambda^{(r+1)} = (n_1^{(m)} + a_1^{(0)} - 1)/(n + a_{\cdot}^{(0)} - 2). \tag{1.1}$$

In examples involving mixtures of *known* densities, this completes the analysis. Otherwise the nature of the M-step for ϕ depends on the model used for the component densities.

1.2.3 The M-step for mixtures of univariate Gaussian distributions

This is by far the most commonly used finite mixture in practice. Here, for $j = 1, \ldots, k$, and using the natural notation for means and variances of

Gaussian distributions,

$$\begin{aligned} f_j(y|\phi_j) &= f(y|\mu_j, \sigma_j^2) \\ &= (2\pi\sigma_j^2)^{-1/2}\exp[-(y-\mu_j)^2/(2\sigma_j^2)]. \end{aligned}$$

Thus

$$Q_2(\phi) = \text{const.} - \frac{1}{2}\sum_{j=1}^{k}\left(n_j^{(m)}\log\sigma_j^2 + \frac{1}{\sigma_j^2}\sum_{i=1}^{n} z_{ij}^{(m)}(y_i-\mu_j)^2\right),$$

which is very like a sum of k Gaussian log-likelihoods, with the $z_{ij}^{(m)}$ included.

For the prior $p(\phi)$ we make the convenient assumptions that the parameters corresponding to the different components are a priori independent, with densities that belong to the appropriate conjugate families. Thus

$$p(\phi) = \prod_{j=1}^{k}[p(\mu_j|\tau_j)p(\tau_j)],$$

in which $\tau_j = 1/\sigma_j^2$ denotes the jth component's precision,

$$\begin{aligned} p(\mu_j|\tau_j) &= N[\mu_j; \rho_j^{(0)}, (b_j^{(0)}\tau_j)^{-1}], \\ p(\tau_j) &= \text{Ga}(\tau_j; c_j^{(0)}, d_j^{(0)}), \end{aligned}$$

where $N[\mu_j; \rho_j^{(0)}, (b_j^{(0)}\tau_j)^{-1}]$ denotes the density of the $N[\rho_j^{(0)}, (b_j^{(0)}\tau_j)^{-1}]$ distribution and $\text{Ga}(\tau_j; c_j^{(0)}, d_j^{(0)})$ denotes the density of the $\text{Ga}(c_j^{(0)}, d_j^{(0)})$ distribution. Thus, for a given j, $\mu_j^{(m+1)}$ and $\sigma_j^{2(m+1)}$ are the joint maximisers of

$$-\frac{1}{2}n_j^{(m)}\log\sigma_j^2 - \frac{1}{2\sigma_j^2}\sum_{i=1}^{n} z_{ij}^{(m)}(y_i-\mu_j)^2 - (c_j^{(0)}-1)\log\sigma_j^2 - d_j^{(0)}/\sigma_j^2$$

$$-\frac{1}{2}\log\sigma_j^2 - \frac{b_j^{(0)}}{2\sigma_j^2}(\rho_j^{(0)}-\mu_j)^2.$$

Straightforward calculus gives

$$\begin{aligned} \mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} z_{ij}^{(m)} y_i + b_j^{(0)}\rho_j^{(0)}}{n_j^{(m)} + b_j^{(0)}}, \\ \sigma_j^{2(m+1)} &= \frac{\sum_{i=1}^{n} z_{ij}^{(m)}(y_i-\mu_j^{(m+1)})^2 + 2d_j^{(0)} + b_j^{(0)}(\rho_j^{(0)}-\mu_j^{(m+1)})^2}{n_j^{(m)} + 2c_j^{(0)} - 1}. \end{aligned}$$

As usual, note the similarity in structure with formulae for posterior modes for Gaussian parameters given complete data.

1.2.4 M-step for mixtures of regular exponential family distributions formulated in terms of the natural parameters

Here $\log f_j(y|\phi_j) = \text{const.} + t(y)^\top \phi_j - C(\phi_j)$ and therefore

$$\begin{aligned} Q_2(\phi) &= \text{const.} + \sum_i \sum_j z_{ij}^{(m)} [t(y_i)^\top \phi_j - C(\phi_j)] \\ &= \text{const.} + \sum_j \left\{ \left[\sum_i z_{ij}^{(m)} t(y_i) \right]^\top \phi_j - n_j^{(m)} C(\phi_j) \right\}. \end{aligned}$$

Then an appropriate prior for ϕ_j has the form

$$\log p(\phi_j) = d_j^{(0)\top} \phi_j - c_j^{(0)} C(\phi_j) + \text{const},$$

and therefore $\phi^{(m+1)}$ is the maximiser of

$$\left[\sum_i z_{ij}^{(m)} t(y_i) + d_j^{(0)} \right]^\top \phi_j - (n_j^{(m)} + c_j^{(0)}) C(\phi_j).$$

Differentiation with respect to ϕ leads to the following equation satisfied by $\phi_j = \phi_j^{(m+1)}$:

$$\sum_i z_{ij}^{(m)} t(y_i) + d_j^{(0)} = (n_j^{(m)} + c_j^{(0)}) \frac{\partial C(\phi_j)}{\partial \phi_j}.$$

However, if we parameterise by

$$\psi_j := E[t(Y)],$$

in which the expectation is over Y assumed to belong to the jth subpopulation, then

$$\psi_j = \frac{\partial C(\phi_j)}{\partial \phi_j},$$

so the M-step is just

$$\psi_j^{(m+1)} = \frac{\sum_{i=1}^n z_{ij}^{(m)} t(y_i) + d_j^{(0)}}{n_j^{(m)} + c_j^{(0)}}.$$

Example 2: Mixture of Poisson distributions

In this example the jth component density can be written as

$$f_j(y|\phi_j) \propto e^{-\phi_j} \phi_j^y,$$

so that an appropriate conjugate prior for ϕ_j is a Ga($c_j^{(0)}$, $d_j^{(0)}$) distribution. Then it is easy to see that the M-step of the EM algorithm leads to

$$\phi_j^{(m+1)} = \frac{\sum_{i=1}^n z_{ij}^{(m)} y_i + c_j^{(0)} - 1}{n_j^{(m)} + d_j^{(0)}}.$$

The terms in $Q(\theta; \theta^{(m)})$ that involve ϕ_j constitute, up to an additive constant, the logarithm of a Ga($c_j^{(m+1)}$, $d_j^{(m+1)}$) density function, with

$$c_j^{(m+1)} = \sum_{i=1}^n z_{ij}^{(m)} y_i + c_j^{(0)},$$
$$d_j^{(m+1)} = n_j^{(m)} + d_j^{(0)}.$$

Example 3: Mixture of exponential distributions

In the 'natural' parameterisation,

$$f_j(y|\phi_j) = \phi_j \exp(-\phi_j y),$$

so that $t(y) = -y$, $C(\phi_j) = -\log \phi_j$ and $\psi_j = -1/\phi_j$. To fit in with the previous notation, we assume that the prior for ϕ_j is a Ga($c_j^{(0)}$, $d_j^{(0)}$) distribution. Thus the M-step is

$$\psi_j^{(m+1)} = -\frac{\sum_{i=1}^n z_{ij}^{(m)} y_i + d_j^{(0)}}{n_j^{(m)} + c_j^{(0)} - 1},$$

from which $\phi_j^{(m+1)} = -1/\psi_j^{(m+1)}$ can be obtained.

1.2.5 Application to other mixtures

Application of EM to various other specific mixture models is discussed in monographs such as Titterington *et al.* (1985) and McLachlan and Peel (2000), not to mention many individual papers. Included in these special cases is that of hidden Markov models, more precisely called hidden Markov chain models. In a mixture sample the complete data corresponding to the ith observation consist of observed data y_i together with the component-membership indicators $z_i = \{z_{ij}, j = 1, \ldots, k\}$. In a mixture sample the $\{z_i\}$ are missing or 'hidden', it is assumed that the ys for different observations are independent, given the zs, and also that the zs are themselves independent. In the hidden Markov model, this second assumption is modified; instead, it is assumed that the $\{z_i, i = 1, \ldots, n\}$ come from a homogeneous, first-order Markov chain with states corresponding to the component subpopulations. Thus, dependence is assumed, but is of the simplest, one-dimensional kind. This model is very popular, in areas such as ecology, speech modelling and DNA sequencing, and associated methodology has been developed based on both maximum likelihood (Rabiner, 1989) and the Bayesian approach

(Robert *et al.*, 1993). The dependence among the hidden variables leads to additional complications, in comparison to the case of mixture data, but typically not to a severe degree. More precisely, the E-step in the EM algorithm for finding a posterior mode is not explicit, but requires a (terminating) so-called forwards–backwards algorithm.

1.2.6 EM as a double expectation

There is an appealing interpretation of EM as a double maximisation rather than as an expectation followed by a maximisation. The idea goes back at least as far as Csiszár and Tusnády (1984) but has more recently been set out clearly, in the context of maximum likelihood estimation, by Neal and Hinton (1999). The version corresponding to calculation of a Bayesian posterior mode with our notation goes as follows. Define

$$F(q, \theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{p(D, \mathbf{z}, \theta)}{q(\mathbf{z})} \right],$$

where q is a density function, and suppose that we are at stage m in the algorithm.

- The first step is to choose $q = q^{(m)}$ to maximise $F(q, \theta^{(m)})$. Since we can write

$$F(q, \theta^{(m)}) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{p(\mathbf{z}|D, \theta^{(m)})}{q(\mathbf{z})} \right] + \log\, p(D, \theta^{(m)}),$$

  the solution is to take $q^{(m)}(\mathbf{z}) = p(\mathbf{z}|D, \theta^{(m)})$.
- The second step is to choose $\theta = \theta^{(m+1)}$ to maximise $F(q^{(m)}, \theta)$. This amounts to maximising $Q(\theta; \theta^{(m)})$, which is just the EM algorithm.

It is easy to see that

$$F(q^{(m)}, \theta^{(m)}) = \log\, p(D, \theta^{(m)}),$$

and therefore the above *alternating maximisation* technique leads to monotonicity:

$$\begin{aligned} \log\, p(D, \theta^{(m)}) &= F(q^{(m)}, \theta^{(m)}) \leq F(q^{(m)}, \theta^{(m+1)}) \leq F(q^{(m+1)}, \theta^{(m+1)}) \\ &= \log\, p(D, \theta^{(m+1)}). \end{aligned} \tag{1.2}$$

1.3 Variational approximations

1.3.1 Preamble

Exact Bayesian analysis of mixture data is complicated, from a computational point of view, because the likelihood function, in expanded form, consists of the sum of a large number of terms. In practice, the use of some form of approximation is

inevitable, and much space has been devoted to the idea of approximating the posterior density or predictive density of interest stochastically, in the form of a set of realisations from the distribution, typically created by an MCMC procedure. In principle the resulting inferences will be asymptotically 'correct', in that the empirical distribution associated with a very large set of realisations should be very similar to the target distribution. However, practical difficulties may arise, especially for problems involving many observations and many parameters, because of storage costs, computation time, the need to confirm convergence and so on. These considerations have led to the development of deterministic approximations to complicated distributions. In the next two sections we describe two such approaches, based respectively on *variational approximations* and on *expectation propagation*. We shall see in this section that, unlike with MCMC methods, the resulting variational approximations are not exact, even asymptotically as $n \to \infty$, but they are less unwieldy and this may be sufficiently attractive to outweigh technical considerations. In fact some attractive asymptotic properties do hold, as is discussed in Section 1.3.6.

1.3.2 Introduction to variational approximations

The fundamental idea is very natural, namely to identify a best approximating density q to a target density p, where 'best' means the minimisation of some 'distance' between q and p. There are many measures of distance between density functions, but the one generally used in this context is the Kullback–Leibler directed divergence $\mathrm{KL}(q, p)$, defined by

$$\mathrm{KL}(q, p) = \int q \log(q/p). \tag{1.3}$$

Of course, $\mathrm{KL}(q, p)$ is not symmetric in its arguments and is therefore not a metric, but it is nonnegative, and zero only if q and p are the same except on a set of measure zero. Without further constraints, the optimal q is of course p, but in practice the whole point is that the p in question is very complicated and the optimisation is carried out subject to approximating but facilitating constraints being imposed on q. In many applications the target, p, is $p_D = p(\cdot|D)$, the conditional density of a set of 'unobservables', $\mathbf{u}$, conditional on a set of observed data, D, and q is chosen to minimise

$$\mathrm{KL}(q, p_D) = \int q(\mathbf{u}) \log \left[q(\mathbf{u})/p(\mathbf{u}|D)\right] \mathrm{d}\mathbf{u},$$

subject to simplifying constraints on q.

The same solution yields a lower bound on $p(D)$, the (marginal) probability of the data. This follows because

$$\log p(D) = \int q(\mathbf{u}) \log[p(D, \mathbf{u})/q(\mathbf{u})] \mathrm{d}\mathbf{u} + \mathrm{KL}(q, p_D),$$

as can be seen by combining the two terms on the right-hand side. The properties of the Kullback–Leibler divergence then both provide the desired lower bound, in that then

$$\log\, p(D) \geq \int q(\mathbf{u}) \log[p(D, \mathbf{u})/q(\mathbf{u})] \mathrm{d}\mathbf{u} = \mathcal{F}(q), \tag{1.4}$$

say, and demonstrate that a q that minimises $\mathrm{KL}(q, p_D)$ provides the best lower bound for $\log\, p(D)$. The right-hand side of (1.4), $\mathcal{F}(q)$, is known in statistical physics as the *free energy* associated with q. The optimum q can be interpreted as the maximiser of the free energy, and it is this interpretation that stimulates the methodology described in this chapter.

In a non-Bayesian context these results permit the approximation of a likelihood function corresponding to a missing-data problem. Here $\mathbf{u}$ represents the missing data and $(D, \mathbf{u})$ the complete data. Thus, the target for q is the conditional density of the missing data, given the observed data, evaluated at a specified value θ for the parameters, and (1.4) provides a lower bound for the observed-data loglikelihood, again for the specified value of θ. More discussion of this use of variational approximations, with references, is given in Section 3.1 of Titterington (2004).

In the Bayesian context, the unobservables include the parameters themselves, as well as any missing values. Thus $\mathbf{u} = (\theta, \mathbf{z})$, where $\mathbf{z}$ denotes the missing data, which for us are the mixture component indicators. The target for q is therefore $p(\theta, \mathbf{z}|D)$, the posterior density of the parameters and the missing values, and (1.4) provides a lower bound for the marginal loglikelihood for the observed data. This marginal likelihood is called the 'evidence' by MacKay (1992) or the Type-II likelihood, and it is a key component of Bayes factors in model-comparison contexts. As we have said, constraints have to be imposed on q in order to create a workable procedure, and the standard approximation is to assume a factorised form,

$$q(\theta, \mathbf{z}) = q^{(\theta)}(\theta) q^{(\mathbf{z})}(\mathbf{z}),$$

for q. We are therefore imposing a (posterior) assumption of independence between the parameters and the missing values. This is clearly a substantive concession, and it is crucial to assess the degree to which the extra computational feasibility counteracts the loss of accuracy. One consequence of the assumption is that the factor $q^{(\theta)}(\theta)$ represents the variational approximation to the 'true' posterior, $p(\theta|D)$. The lower bound in (1.4) is

$$\log\, p(D) \geq \int \sum_{\mathbf{z}} q^{(\theta)}(\theta) q^{(\mathbf{z})}(\mathbf{z}) \log[p(D, \mathbf{z}, \theta)/q^{(\theta)}(\theta) q^{(\mathbf{z})}(\mathbf{z})] \mathrm{d}\theta = \mathcal{F}(q^{(\theta)}, q^{(\mathbf{z})}), \tag{1.5}$$

say, and can also be written

$$\log\, p(D) \geq \int q^{(\theta)}(\theta) \left\{ \sum_{\mathbf{z}} q^{(\mathbf{z})}(\mathbf{z}) \log[p(D, \mathbf{z}|\theta)/q^{(\mathbf{z})}(\mathbf{z})] + \log[p(\theta)/q^{(\theta)}(\theta)] \right\} \mathrm{d}\theta.$$

The optimal $q^{(\theta)}$ and $q^{(\mathbf{z})}$ maximise the right-hand side of (1.5). Thus, the two factors respectively maximise the coupled formulae

$$\int q^{(\theta)}(\theta) \log \left(\frac{\exp\{\sum q^{(\mathbf{z})}(\mathbf{z}) \log[p(D, \mathbf{z}, \theta)]\}}{q^{(\theta)}(\theta)} \right) \mathrm{d}\theta, \tag{1.6}$$

$$\sum q^{(\mathbf{z})}(\mathbf{z}) \log \left(\frac{\exp\{\int q^{(\theta)}(\theta) \log[p(D, \mathbf{z}, \theta)] \mathrm{d}\theta\}}{q^{(\mathbf{z})}(\mathbf{z})} \right). \tag{1.7}$$

It follows that the optimum $q^{(\theta)}$ and $q^{(\mathbf{z})}$ satisfy

$$q^{(\theta)}(\theta) \propto \exp\left\{ \sum q^{(\mathbf{z})}(\mathbf{z}) \log[p(D, \mathbf{z}, \theta)] \right\} \tag{1.8}$$

and

$$q^{(\mathbf{z})}(\mathbf{z}) \propto \exp\left\{ \int q^{(\theta)}(\theta) \log[p(D, \mathbf{z}, \theta)] \mathrm{d}\theta \right\}. \tag{1.9}$$

Explicit solution of these equations will not be possible. However, writing the equations in the shorthand form $q^{(\theta)} = T^{(\theta)}(q^{(\mathbf{z})})$, $q^{(\mathbf{z})} = T^{(\mathbf{z})}(q^{(\theta)})$ suggests the following iterative algorithm for computing $q^{(\theta)}$ and $q^{(\mathbf{z})}$ from an initial $q^{(\mathbf{z})} = q^{(\mathbf{z})(0)}$: for $m = 0, 1, \ldots$, calculate

$$\begin{aligned} q^{(\theta)(m+1)} &= T^{(\theta)}(q^{(\mathbf{z})(m)}), \\ q^{(\mathbf{z})(m+1)} &= T^{(\mathbf{z})}(q^{(\theta)(m+1)}). \end{aligned} \tag{1.10}$$

The construction of the algorithm is such that, so far as the 'free energy' $\mathcal{F}(q^{(\theta)}, q^{(\mathbf{z})})$ in (1.5) is concerned,

$$\mathcal{F}(q^{(\theta)(m)}, q^{(\mathbf{z})(m)}) \leq \mathcal{F}(q^{(\theta)(m+1)}, q^{(\mathbf{z})(m)}) \leq \mathcal{F}(q^{(\theta)(m+1)}, q^{(\mathbf{z})(m+1)}), \tag{1.11}$$

for $m = 1, 2, \ldots$, so that the free energy is monotonically increasing and bounded above by $\log\ p(D)$, and therefore the sequence of free energies converges. This behaviour is a direct parallel of that of the EM algorithm; compare (1.11) with (1.2). Whether or not there are multiple local maxima, and so on, is a different issue; in fact the existence of multiple local maxima, or multiple fixed points of the algorithm, is a very common phenomenon.

Again the reader is referred to Titterington (2004), and to Bishop (2006), for general discussions of variational Bayes approaches to missing-data problems and a substantial body of references.

1.3.3 Application of variational Bayes to mixture problems

We shall consider some of the same examples as those used in section 1.2 about the EM algorithm, with the intention of revealing strong methodological similarities between the two approaches.

Example 4: Mixture of k known densities

Recall that here we assume that the observed data are a random sample $D = \{y_1, \ldots, y_n\}$ from a distribution with probability density function

$$p(y|\theta) = \sum_{j=1}^{k} \lambda_j f_j(y),$$

where the f_js are known so that θ just consists of the λ_js. In this case,

$$p(D, \mathbf{z}, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} (\lambda^{z_{ij}} f_{ij}^{z_{ij}}) p(\theta),$$

where $z_{ij} = 1$ if the ith observation comes from component j and is zero otherwise, $f_{ij} = f_j(y_i)$ and $p(\theta)$ is the prior density for θ. If we assume a Dir($a^{(0)}$) prior for θ, then

$$p(D, \mathbf{z}, \theta) \propto \prod_{j=1}^{k} \lambda_j^{\sum_i z_{ij} + a_j^{(0)} - 1},$$

so that

$$\sum_{\mathbf{z}} q^{(\mathbf{z})}(\mathbf{z}) \log[p(D, \mathbf{z}, \theta)] = \sum_{j=1}^{k} \left(\sum_i q_{ij}^{(\mathbf{z})} + a_j^{(0)} - 1 \right) \log \lambda_j + \text{const},$$

where 'const.' does not depend on θ and $q_{ij}^{(\mathbf{z})}$ is the marginal probability that $z_{ij} = 1$, according to the distribution $q^{(\mathbf{z})}(\mathbf{z})$. From (1.8), therefore, the optimal $q^{(\theta)}(\theta)$ density is that of the Dir(a) distribution, where $a_j = \sum_i q_{ij}^{(\mathbf{z})} + a_j^{(0)}$, for $j = 1, \ldots, k$.

Next we identify the optimal $q^{(\mathbf{z})}(\mathbf{z})$ as a function of $q^{(\theta)}(\theta)$. We have

$$\int q^{(\theta)}(\theta) \log[p(D, \mathbf{z}, \theta)] \mathrm{d}\theta = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} [\log(\phi_j^{(q)}) + \log f_{ij}] + \text{const},$$

where $\phi_j^{(q)} = \exp[E_{q^{(\theta)}} \log(\lambda_j)]$ and now the 'const.' is independent of the $\{z_{ij}\}$. If we substitute this on the right-hand side of (1.9) we see that the optimal $q_{\mathbf{z}}$ takes a factorised form, with one factor for each observation, and that the optimal factors are defined by

$$q_{ij}^{(\mathbf{z})} \propto f_{ij} \phi_j^{(q)},$$

for $j = 1, \ldots k$, subject to $\sum_j q_{ij}^{(\mathbf{z})} = 1$, for each i. Properties of the Dir(a) distribution imply that $E_{q^{(\theta)}} \log(\lambda_j) = \Psi(a_j) - \Psi(a_\cdot)$, where Ψ denotes the digamma

function and $a_\cdot = \sum_j a_j$. Thus, the equations for $q^{(\mathbf{z})}$ are

$$q_{ij}^{(\mathbf{z})} = f_{ij}\exp[\Psi(a_j) - \Psi(a_\cdot)] / \sum_{r=1}^{k} f_{ir}\exp[\Psi(a_r) - \Psi(a_\cdot)],$$

for each i and j. As predicted earlier, clearly the equations for the $\{a_j\}$, which define the Dirichlet distribution that represents the variational approximation to the posterior distribution of θ, and the $\{q_{ij}^{(\mathbf{z})}\}$ cannot be solved explicitly, but the version of the iterative algorithm (1.10) works as follows. Initialise, for example by choosing $q_{ij}^{(\mathbf{z})(0)} = f_{ij}a_j^{(0)}/(\sum_{r=1}^{k} f_{ir}a_r^{(0)})$, for each i and j, and carry out the following steps, for $m = 0, 1, \ldots$.

- *Step 1*: For $j = 1, \ldots, k$, calculate

$$a_j^{(m+1)} = \sum_i q_{ij}^{(\mathbf{z})(m)} + a_j^{(0)}. \tag{1.12}$$

- *Step 2*: For $i = 1, \ldots, n$ and $j = 1, 2$, calculate

$$q_{ij}^{(\mathbf{z})(m+1)} = f_{ij}\exp\left[\Psi(a_j^{(m+1)}) - \Psi(a_\cdot^{(m+1)})\right] / \sum_{r=1}^{k} f_{ir}\exp\left[\Psi(a_r^{(m+1)}) - \Psi(a_\cdot^{(m+1)})\right]. \tag{1.13}$$

Note the strong similarity between Step 1 and an EM algorithm M-step and between Step 2 and an EM-algorithm E-step, the major difference being the replacement of $z_{ij}^{(m)}$ by $q_{ij}^{(\mathbf{z})(m)}$. Indeed, in some of the literature these algorithms are called 'variational EM'.

Example 5: Mixture of k univariate Gaussian densities

In this case

$$p(y|\theta) = \sum_{j=1}^{k} \lambda_j N(y; \mu_j, \sigma_j^2),$$

where $\lambda = \{\lambda_j\}$ are the mixing weights, $\mu = \{\mu_j\}$ are the component means and $\sigma = \{\sigma_j^2\}$ are the component variances, so that $\theta = (\lambda, \mu, \sigma)$. Given data $D = \{y_i, i = 1, \ldots, n\}$, we have

$$p(D, \mathbf{z}, \theta) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[\lambda_j N(y_i; \mu_j, \sigma_j^2)\right]^{z_{ij}} p(\theta).$$

At this point it is revealing to return to Example 4 for a couple of remarks.

1. The variational posterior for θ there took the form of a Dirichlet distribution, as a result of the factorisation assumption about $q(\theta, \mathbf{z})$ and of the choice of a Dirichlet prior. In other words, having chosen the traditional complete-data conjugate family of priors, conjugacy was obtained for the variational method.
2. The joint variational approximation for the distribution of $\mathbf{z}$ took a factorised form, essentially because $p(D, \mathbf{z}, \theta)$ took the form of a product over i; in other words the (y_i, z_i) are independent over i, where $z_i = \{z_{ij}, j = 1, \ldots .k\}$.

The second remark holds again here, so that we shall have

$$q_{ij}^{(\mathbf{z})} \propto \exp\left\{\int q^{(\theta)}(\theta)\log[p(y_i, z_{ij} = 1, \theta)]\mathrm{d}\theta\right\}, \tag{1.14}$$

for each i and j, normalised so that $\sum_j q_{ij}^{(\mathbf{z})} = 1$, for each i.

Also, if we choose a complete-data conjugate prior for θ then the optimal $q^{(\theta)}$ will be a member of that family. The appropriate hyperparameters will satisfy equations interlinked with those for the $\{q_{ij}^{(\mathbf{z})}\}$ that can be solved iteratively, by alternately updating the hyperparameters and the $\{q_{ij}^{(\mathbf{z})}\}$. This structure will clearly obtain for a wide range of other examples, but we concentrate on the details for the univariate Gaussian mixture, for which the appropriate prior density takes the form mentioned before, namely

$$p(\theta) = p(\lambda, \mu, \sigma) = p(\lambda)\prod_{j=1}^{k}[p(\mu_j|\tau_j)p(\tau_j)],$$

in which $\tau_j = 1/\sigma_j^2$ denotes the jth component precision,

$$\begin{aligned} p(\lambda) &= \mathrm{Dir}(\lambda; a^{(0)}), \\ p(\mu_j|\tau_j) &= N[\mu_j; \rho_j^{(0)}, (b_j^{(0)}\tau_j)^{-1}], \\ p(\tau_j) &= \mathrm{Ga}(\tau_j; c_j^{(0)}, d_j^{(0)}), \end{aligned}$$

where $\mathrm{Dir}(\lambda; a^{(0)})$ denotes the density of the Dirichlet distribution with parameters $a^{(0)} = \{a_j^{(0)}, j = 1, \ldots, k\}$, and the other notation has been defined already.

Often the priors will be exchangeable, so that all the $c_j^{(0)}$s will be the same, and so on, but we shall work through the more general case. The optimal variational approximation $q^{(\theta)}(\theta)$ then takes the form

$$q^{(\theta)}(\theta) = q^{(\lambda)}(\lambda)\prod_{j}[q^{(\mu_j|\tau_j)}(\mu_j|\tau_j)q^{(\tau_j)}(\tau_j)],$$

within which the factors are given by

$$\begin{aligned} q^{(\lambda)}(\lambda) &= \mathrm{Dir}(\lambda; a), \\ q^{(\mu_j|\tau_j)}(\mu_j|\tau_j) &= N[\mu_j; \rho_j, (b_j\tau_j)^{-1}], \\ q^{(\tau_j)}(\tau_j) &= \mathrm{Ga}(\tau_j; c_j, d_j), \end{aligned}$$

where the hyperparameters satisfy

$$a_j = \sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} + a_j^{(0)},$$

$$\rho_j = \left(\sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} y_i + b_j^{(0)} \rho_j^{(0)} \right) / \left(\sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} + b_j^{(0)} \right),$$

$$b_j = \sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} + b_j^{(0)},$$

$$c_j = \frac{1}{2} \sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} + c_j^{(0)},$$

$$d_j = \frac{1}{2} \left[\sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} (y_i - \bar{\mu}_j)^2 + b_j^{(0)} \left(\sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} \right) (\bar{\mu}_j - \rho_j^{(0)})^2 / \left(\sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} + b_j^{(0)} \right) \right] + d_j^{(0)},$$

in which $\bar{\mu}_j = \sum_{i=1}^{n} q_{ij}^{(\mathbf{z})} y_i / \sum_{i=1}^{n} q_{ij}^{(\mathbf{z})}$, for each j, representing a pseudo sample mean for component j in the same way that $\sum_i q_{ij}^{(\mathbf{z})}$ represents a pseudo sample size.

If we write the total set of hyperparameters as h, then the above equations take the form

$$h = G_1(q^{(\mathbf{z})}). \tag{1.15}$$

We now have to identify the optimal $q^{(\mathbf{z})}(\mathbf{z})$ as a function of $q^{(\theta)}(\theta)$ or, to be more specific, as a function of the associated hyperparameters h. As we have seen, for each i we have

$$\begin{aligned} q_{ij}^{(\mathbf{z})} &\propto \exp \left\{ \int q^{(\theta)}(\theta) \log[p(y_i, z_{ij} = 1, \theta)] \mathrm{d}\theta \right\} \\ &= \exp \left\{ \int q^{(\theta)}(\theta) \sum_j z_{ij} [\log(\lambda_j) + \frac{1}{2} \log(\tau_j) - \frac{1}{2} \tau_j (y_i - \mu_j)^2] \mathrm{d}\theta \right\} + \text{const} \\ &= \exp \left(\sum_j z_{ij} \{ E_{q^{(\theta)}} \log(\lambda_j) + \frac{1}{2} E_{q^{(\theta)}} \log(\tau_j) - \frac{1}{2} E_{q^{(\theta)}} [\tau_j (y_i - \mu_j)^2] \} \right) + \text{const} \end{aligned}$$

From Example 1 we know that

$$E_{q^{(\theta)}} \log(\lambda_j) = \Psi(a_j) - \Psi(a_\cdot).$$

Also, from properties of the gamma distribution, we have

$$E_{q^{(\theta)}} \log(\tau_j) = \Psi(c_j) - \log(d_j),$$

and, averaging out first conditionally on τ_j and then over τ_j, we obtain

$$E_{q^{(\theta)}} [\tau_j (y_i - \mu_j)^2] = c_j d_j^{-1} (y_i - \rho_j)^2 + b_j^{-1}.$$

Thus

$$q_{ij}^{(\mathbf{z})} \propto \exp\left(\Psi(a_j) + \frac{1}{2}[\Psi(c_j) - \log(d_j)] - \frac{1}{2}[c_j d_j^{-1}(y_i - \rho_j)^2 + b_j^{-1}]\right),$$

with the normalisation carried out over j for each i. The set of relationships can be represented concisely in the form

$$q^{(\mathbf{z})} = G_2(h). \tag{1.16}$$

The obvious algorithm for obtaining the variational approximations involves initialising, by setting $q^{(\mathbf{z})}$ to some $q^{(\mathbf{z})(0)}$, and then calculating $h^{(m+1)} = G_1(q^{(\mathbf{z})(m)})$ and $q^{(\mathbf{z})(m+1)} = G_2(h^{(m+1)})$, for $m = 0, 1, \ldots$. Again, this is the form of (1.10) corresponding to this example. Since the component densities are unknown, initialisation is not so straightforward, especially if exchangeable priors are assumed for the parameters in the component densities. An ad hoc approach that seemed to be adequate in practice is to perform a cluster analysis of the data into k clusters and to base the initial $q^{(\mathbf{z})}$ on that. For univariate data the crude method of using sample quantiles to determine the clusters can be adequate, and for multivariate data a k-means clustering can be effective enough.

1.3.4 Application to other mixture problems

As already mentioned, Titterington (2004) and Bishop (2006) list many references to particular cases of variational Bayes. The case of mixtures of multivariate Gaussian distributions has been treated in a number of places. The natural analogue of the univariate case as discussed in Example 5 is considered by McGrory (2005) and McGrory and Titterington (2007). The usual conjugate prior is used for the parameters of the component distributions, namely independent Wishart distributions for the inverses of the covariance matrices and independent Gaussian priors for the mean vectors, conditionally on the inverse covariance matrices. The version of the variational posterior distribution, $q^{(\theta)}(\theta)$, naturally has the same structure.

The same mixture model was investigated by Corduneanu and Bishop (2001), but with a different structure for the prior. The prior distributions for the component means were independent Gaussians with zero means, but with covariance matrices that were βI, where I is the identity matrix and β is a positive scalar; this is clearly different from the usual conjugate prior structure. Furthermore, they did not assume any prior distribution for the mixing weights. Instead, they kept the mixing weights λ as fixed parameters and chose $q^{(\theta)}(\theta)q^{(\mathbf{z})}(\mathbf{z})$, where θ consists of the component mean vectors and precision matrices, so as to maximise

$$\sum_{\mathbf{z}} \int q^{(\theta)}(\theta) \log[(D, \mathbf{z}, \theta|\lambda)/q^{(\theta)}(\theta)q^{(\mathbf{z})}(\mathbf{z})] \mathrm{d}\theta,$$

which is a lower bound $\mathcal{F}_\lambda(q)$ for the 'marginal loglikelihood' $\log\, p(D|\lambda)$. Corduneanu and Bishop (2001) alternately calculate the optimal q for the current value of λ and then obtain λ so as to maximise $\mathcal{F}_\lambda(q)$ for that q. In fact they only perform

one iteration of the algorithm for the calculation of the optimal q before moving on to obtain a new λ. As the algorithm progresses the values of $\mathcal{F}_\lambda$ increase, reflecting the fact that the algorithm is similar to the generalised EM algorithm (Dempster *et al.*, 1977). Hyperparameters such as β are pre-set, although explicit procedures are not stated in the paper.

The method of Corduneanu and Bishop (2001) is unconventional in statistical terms in not using the usual structure for the prior distributions for the component mean vectors and in not treating the mixing weights in a fully Bayesian way, as was done in McGrory (2005) and McGrory and Titterington (2007). The more traditional structure is also followed by Ueda and Ghahramani (2002), who consider a more complicated mixture model known as the mixture-of-experts model (Jordan and Jacobs, 1994). This model assumes the existence of covariates, the mixing weights depend on the covariates and the component distributions correspond to regression models, usually Gaussian, again depending on the covariates. The calculations for the variational approach involve much careful detail but are again evocative of complete-data conjugate Bayesian analysis.

Ueda and Ghahramani (2002) emphasise other issues beyond the calculation of approximate posterior distributions, and in particular they cover prediction and model choice. So far as prediction is concerned, in the case of 'ordinary' Gaussian mixtures, the calculation of the predictive density of a new observation is straightforward, being a mixture of the corresponding component predictive densities, with mixing weights given by the means of the variational posterior distribution for the mixing weights. In the univariate case this gives a mixture of Student's t distributions. In the case of mixture-of-experts models the complexity of the mixing weights requires a mild amount of approximation in order to obtain a closed-form predictive density. Ueda and Ghahramani deal with model choice by assuming a specified finite class of models, with associated prior probabilities, incorporating a factor $q^{(M)}(m)$ in the formula to represent the variational posterior distribution on the class of models, and basing model choice on the resulting values of $q^{(M)}(m)$. In their empirical work on Gaussian mixtures, Corduneanu and Bishop (2001) observed automatic model selection taking place, when analysing data that were actually simulated from mixture distributions. If a model were fitted that contained more components than the true model, then, as the 'marginal likelihood' maximisation progressed, at least one of the estimates of the mixing weights would become smaller and smaller. At that stage that component was dropped from the model, the algorithm proceeded and this automatic pruning stopped only when the estimated mixture had the same number of components as the true model. If the fitted model did not have as many components as the true model then no pruning occurred. The same phenonemon occurred in the fully Bayesian approach of McGrory and Titterington (2007) in that, if the fitted model was too rich, then, for some j, $\sum_i q_{ij}^{(\mathbf{z})}$ became very small, corresponding to a component for which the other parameters were very similar to those of a second component. As we have said, $\sum_i q_{ij}^{(\mathbf{z})}$ is a pseudo sample size of those observations nominally assigned to the jth component, and if that number fell much below 1 the component was dropped from the fitted model. There does not seem to be any formal explanation for this intriguing yet helpful behaviour.

Application to hidden Markov models is described in MacKay (1997) and Humphreys and Titterington (2001). Software for variational Bayes developed by J. Winn is available at http://vibes.sourceforge.net/.

1.3.5 Recursive variational approximations

Example 1 (Revisited): Mixture of two known densities

We introduce the idea of recursive methods by returning to the very simple case of a mixture of two known densities, with λ denoting the mixing weight, which is the sole unknown parameter. The variational posterior turned out to be a beta distribution, provided that a beta prior was proposed. Alternative ways of deriving beta approximations to the true posterior are given by recursive approximations, in which the observations are processed one by one and the posterior is updated each time, within the beta class. If, after the ith observation has been dealt with, the approximate posterior is the $\mathrm{Be}(a_1^{(i)}, a_2^{(i)})$ distribution, then the recursion computes hyperparameters $(a_1^{(i+1)}, a_2^{(i+1)})$ on the basis of $(a_1^{(i)}, a_2^{(i)})$ and y_{i+1}. Nonvariational versions of these recursions are motivated by the exact Bayesian step for incorporating observation $i+1$, given a 'current' $\mathrm{Be}(a_1^{(i)}, a_2^{(i)})$ prior. The correct posterior is the mixture

$$p(\lambda|y_{i+1}) = w_{i+1,1}\mathrm{Be}(\lambda; a_1^{(i)}+1, a_2^{(i)}) + w_{i+1,2}\mathrm{Be}(\lambda; a_1^{(i)}, a_2^{(i)}+1),$$

where, for $j = 1, 2$, $w_{i+1,j} = f_{i+1,j}a_j^{(i)}/(f_{i+1,1}a_1^{(i)} + f_{i+1,2}a_2^{(i)})$ and $\mathrm{Be}(\lambda; a_1, a_2)$ denotes the $\mathrm{Be}(a_1, a_2)$ density.

Recursions such as these were discussed in detail in Chapter 6 of Titterington *et al.* (1985). For this simple, one-parameter problem, two particular versions are the quasi-Bayes method (Makov and Smith, 1977; Smith and Makov, 1978, 1981), in which $a_j^{(i+1)} = a_j^{(i)} + w_{i+1,j}$, for $j = 1, 2$, and the probabilistic editor (Athans *et al.*, 1977; Makov, 1983), in which $a_1^{(i+1)}$ and $a_2^{(i+1)}$ are calculated to ensure that the first two moments of the beta approximation match the 'correct' moments corresponding to the mixture. This moment-matching is possible beause there are two hyperparameters.

It is easy to define a sequence of beta variational approximations $\{\mathrm{Be}(a_1^{(i)}, a_2^{(i)})\}$. Given an approximating $\mathrm{Be}(a_1^{(i)}, a_2^{(i)})$ at stage i, choose $q_{i+1,j}^{(\mathbf{z})(0)} = f_{i+1,j}a_j^{(i)}/\{\sum_{r=1}^{2}(f_{i+1,r}a_r^{(i)})\}$, for each $j = 1, 2$, and carry out the following steps, for $m = 0, 1, \ldots$.

- *Step 1*: For $j = 1, 2$, calculate

$$a_j^{(i+1)(m+1)} = q_{i+1,j}^{(\mathbf{z})(m)} + a_j^{(i)}.$$

- *Step 2*: For $j = 1, 2$, calculate

$$q_{i+1,j}^{(\mathbf{z})(m+1)} = \frac{f_{i+1,j}\exp[\Psi(a_j^{(i+1)(m+1)}) - \Psi(a_\cdot^{(i+1)(m+1)})]}{\sum_{r=1}^{2} f_{i+1,r}\exp[\Psi(a_r^{(i+1)(m+1)}) - \Psi(a_\cdot^{(i+1)(m+1)})]}.$$

Note the obvious similarity to Equations (1.12) and (1.13).

Clearly, at each recursive stage, the iterative procedure may take some time, but for large i very few iterations should be necessary, and even terminating at $m = 1$ may eventually be adequate. One important factor is that the results obtained will depend on the order in which the observations are incorporated, as tends to be the case with all recursive methods.

Humphreys and Titterington (2000) report an experiment based on a sample of size 50 from a mixture of the $N(3, 1)$ and $N(5, 1)$ densities with mixing weight $\lambda = 0.65$, starting from a Un(0,1) prior, so that $a_1^{(0)} = a_2^{(0)} = 1$. So far as the posterior variances are concerned, the values that were obtained empirically from the nonrecursive variational method, the quasi-Bayes and the recursive variational methods were 0.0043, 0.0044 and 0.0043, respectively. In contrast, the values obtained from Gibbs sampling and the Probabilistic Editor (PE) were 0.0088 and 0.0091, respectively.

Of particular note is the fact that only the probabilistic editor leads to a posterior variance for λ that is very similar to the 'correct' value provided by the Gibbs sampling result; all the other methods, including the nonrecursive variational method, 'underestimate' the posterior variability, although they produce a reasonable mean. Further detailed elucidation of this behaviour is provided in Section 1.4.4.

As remarked above, the recursive results are influenced by the ordering of the data. In the case of a hidden Markov chain, a natural ordering does exist, and indeed recursive (online) analysis could be of genuine practical importance. Humphreys and Titterington (2001) develop versions of the quasi-Bayes, probabilistic editor and recursive variational methods for this problem.

1.3.6 Asymptotic results

Variational Bayes approximations have been developed for many particular scenarios, including a number involving mixture models, and have led to useful empirical results in applications. However, it is important to try to reinforce these pragmatic advantages with consideration of the underlying theoretical properties. What aspects, if any, of the variational posteriors match those of the true, potentially highly complex, posteriors, at least asymptotically? Attias (1999, 2000) and Penny and Roberts (2000) claim that, in certain contexts, the variational Bayes estimator, given by the variational posterior mean, approaches the maximum likelihood estimator in the large sample limit, but no detailed proof was given.

More recently, some more detailed investigations have been carried out. In Wang and Titterington (2004a) the case of a mixture of k known densities was considered, for which the variational approximation is a Dirichlet distribution whereas the true posterior corresponds to a mixture of Dirichlets. However, it was proved that the variational Bayes estimator (the variational posterior mean) of the mixing weights converges locally to the maximum likelihood estimator. Wang and Titterington (2005a) extended the treatment of this mixture example by proving asymptotic normality of the variational posterior distribution of the parameters. In terms of asymptotic mean and distribution type, therefore, the variational approximation behaved satisfactorily. However, the paper went on to examine the asymptotic covariance matrix and showed that it was 'too small' compared with

that of the maximum likelihood estimators. As a result, interval estimates based on the variational approximation will be unrealistically narrow.

This confirms, through theory, intuition inspired by the fact that the variational approximation is a 'pure' Dirichlet rather than a complicated mixture of Dirichlets, as well as reinforcing the empirical results described in the previous section. For the case of $k = 2$, the asymptotic variational posterior variance of the mixing weight λ is $\lambda(1-\lambda)/n$, in other words the same as the variance of the complete-data maximum likelihood estimator. For the example in Humphreys and Titterington (2000), mentioned in the previous section, in which $\lambda = 0.65$ and $n = 50$, this variance is 0.00455, which is rather similar to some of the estimates reported in Section 1.3.5. As already mentioned, we return to this in Section 1.4.4.

In Wang and Titterington (2005b) the demonstration of this unsatisfactory nature of posterior second moments was extended to the case of mixtures of multivariate normal distributions with all parameters unknown. Also in the context of this type of mixture, Wang and Titterington (2006) allowed more flexibility in the algorithm for computing the hyperparameters in $q^{(\theta)}$ by introducing a variable step-length ϵ reminiscent of the extension of the EM algorithm in Peters and Walker (1978); the standard algorithm corresponds to $\epsilon = 1$. It was proved that, for $0 < \epsilon < 2$, the algorithm converges and that the variational Bayes estimators of the parameters converge to the maximum likelihood estimators. Wang and Titterington (2004b) considered the more general context of data from exponential family models with missing values, establishing local convergence of the iterative algorithm for calculating the variational approximation and proving asymptotic normality of the estimator.

In the context of hidden Markov chains, the use of a fully factorised version of $q^{(\mathbf{z})}$ does not reflect the correct solution of the variational problem and this shows up in unsatisfactory empirical results. Wang and Titterington (2004c) reinforced this more formally by showing that, in the context of a simple state-space model, which is a continuous-time analogue of the hidden Markov chain model, the variational Bayes estimators could be asymptotically biased; the factorisation constraints imposed on $q^{(\mathbf{z})}$ are inconsistent with the essential correlations between the hidden states that constitute $\mathbf{z}$.

1.4 Expectation–propagation

1.4.1 Introduction

In this section we describe, in the mixtures context, another deterministic approximation to Bayesian analysis, as an alternative to the variational approximation of the previous section and as before stimulated by the fact that exact calculation of posterior and predictive distributions is not feasible when there are latent or missing variables.

Suppose that we have data of the form $D := \{y_1, \ldots, y_n\}$. The posterior distribution of interest is given formally by

$$p(\theta|D) \propto p(D|\theta)\, p(\theta),$$

where $p(\theta)$ on the right-hand side is the prior density for θ. In this account we shall assume that the data are independently distributed, so that we can write

$$p(\theta|D) \propto t_0(\theta) \prod_{i=1}^{n} t_i(\theta),$$

in which $t_0(\theta) = p(\theta)$ and $t_i(\theta) = p(y_i|\theta)$, the probability density or mass function for the ith observation, for $i = 1, \ldots, n$.

The alternative class of deterministic approximations is that provided by the method of expectation propagation (Minka, 2001a, 2001b). In this approach it is assumed that the posterior distribution for θ is approximated by a product of terms,

$$q_\theta(\theta) = \prod_{i=0}^{n} \tilde{t}_i(\theta),$$

where the $i = 0$ term corresponds to the prior density and, typically, if the prior comes from a conjugate family then, as functions of θ, all the other terms take the same form so that $q_\theta(\theta)$ does also. The explicit form of the approximation is calculated iteratively.

- *Step 1*: From an initial or current proposal for $q_\theta(\theta)$ the ith factor $\tilde{t}_i$ is discarded and the resulting form is renormalised, giving $q_{\theta,\setminus i}(\theta)$.
- *Step 2*: This $q_{\theta,\setminus i}(\theta)$ is then combined with the 'correct' ith factor $t_i(\theta)$ (implying that the correct posterior does consist of a product) and a new $q_\theta(\theta)$ of the conjugate form is selected that gives an 'optimal' fit with the new product, $p_i(\theta)$. In Minka (2001b) optimality is determined by Kullback–Leibler divergence, in that, with

  $$p_i(\theta) \propto q_{\theta,\setminus i}(\theta) t_i(\theta),$$

  then the new $q_\theta(\theta)$ minimises

  $$\text{KL}(p_i, q) = \int p_i \log (p_i/q). \tag{1.17}$$

  However, in some cases, simpler solutions are obtained if moment-matching is used instead; if the conjugate family is Gaussian then the two approaches are equivalent.
- *Step 3*: Finally, from $q_{\theta,\setminus i}(\theta)$ and the new $q_\theta(\theta)$, a new factor $\tilde{t}_i(\theta)$ can be obtained such that

  $$q_\theta(\theta) \propto q_{\theta,\setminus i}(\theta) \tilde{t}_i(\theta).$$

This procedure is iterated over choices of i repeatedly until convergence is attained.

We note in passing that the position of q among the arguments of the Kullback–Leibler divergence in (1.17) is different from that in (1.3).

As with the variational Bayes approach, an approximate posterior is obtained that is a member of the complete-data conjugate family. The important question is to what extent the approximation improves on the variational approximation. Much empirical evidence suggests that it is indeed better, but it is important to investigate this issue at a deeper level. Does expectation propagation achieve what variational Bayes achieves in terms of Gaussianity and asymptotically correct mean and does it in addition manage to behave appropriately, or at least better, in terms of asymptotic second posterior moments?

This section takes a few first steps by discussing some very simple scenarios and treating them without complete rigour. In some cases positive results are obtained, but the method is shown not to be uniformly successful in the above terms.

1.4.2 Overview of the recursive approach to be adopted

As Minka (2001b) explains, expectation propagation (EP) was motivated by a recursive version of the approach known as assumed density filtering (ADF) (Maybeck, 1982). In ADF an approximation to the posterior density is created by incorporating the data one by one, carrying forward a conjugate-family approximation to the true posterior and updating it observation-by-observation using the sort of Kullback–Leibler-based strategy described above in Step 2 of the EP method; the same approach was studied by Bernardo and Girón (1988) and Stephens (1997, Chapter 5) and there are some similarities with Titterington (1976). The difference from EP is that the data are run through only once, in a particular order, and therefore the final result is order-dependent. However, recursive procedures such as these often have desirable asymptotic properties, sometimes going under the name of stochastic approximations. There is also a close relationship with the probabilistic editor. Indeed, our analysis concentrates on the recursive step defined by Step 2 in the EP approach, which is essentially the probabilistic editor if moment-matching is used to update.

We shall summarise results for two one-parameter scenarios, namely normal mixtures with an unknown mean parameter, for which the conjugate prior family is Gaussian, and mixtures with an unknown mixing weight, for which the conjugate prior family is beta. Full details will appear in a paper co-authored with N. L. Hjort.

We shall suppose that the conjugate family takes the form $q(\theta|a)$, where a represents a set of hyperparameters and θ is a scalar. For simplicity we shall omit the subscript θ attached to q. Of interest will be the relationship between consecutive sets of hyperparameters, $a^{(n-1)}$ and $a^{(n)}$, corresponding to the situation before and after the nth observation, y_n, is incorporated. Thus, $q(\theta|a^{(n)})$ is the member of the conjugate family that is closest, in terms of Kullback–Leibler divergence, or perhaps of moment-matching, to the density that is given by

$$q(\theta|a^{(n-1)})t_n(\theta) = q(\theta|a^{(n-1)})f(y_n|\theta),$$

suitably normalised. If E_n and V_n are the functions of $a^{(n)}$ that represent the mean and the variance corresponding to $q(\theta|a^{(n)})$, then we would want E_n and V_n asymptotically to be indistinguishable from the corresponding values for the correct posterior.

We would also expect asymptotic Gaussianity. So far as E_n is concerned, it is a matter of showing that it converges in some sense to the true value of θ. The correct asymptotic variance is essentially given by the asymptotic variance of the maximum likelihood estimator, with the prior density having negligible effect, asymptotically. Equivalently, the increase in precision associated with the addition of y_n, $V_n^{-1} - V_{n-1}^{-1}$, should be asymptotically the same, in some sense, as the negative of the second derivative with respect to θ of $\log f(y_n|\theta)$, again by analogy with maximum likelihood theory.

1.4.3 Finite Gaussian mixtures with an unknown mean parameter

The assumption is that the observed data are a random sample from a univariate mixture of J Gaussian distributions, with means and variances $\{c_j\mu, \sigma_j^2; j = 1, \ldots, J\}$ and with mixing weights $\{v_j; j = 1, \ldots, J\}$. The $\{c_j, \sigma_j^2, v_j\}$ are assumed known, so that μ is the only unknown parameter. For observation y, therefore, we have

$$p(y|\mu) \propto \sum_{j=1}^{J} \frac{v_j}{\sigma_j} \exp\left[-\frac{1}{2\sigma_j^2}(y - c_j\mu)^2\right].$$

In the case of a Gaussian distribution the obvious hyperparameters are the mean and the variance themselves. For comparative simplicity of notation we shall write those hyperparameters before treatment of the nth observation as (a, b), the hyperparameters afterwards as (A, B) and the nth observation itself as y.

In the recursive step the hyperparameters A and B are chosen to match the moments of the density for μ that is proportional to

$$b^{-1/2} \exp\left[-\frac{1}{2b}(\mu - a)^2\right] \sum_{j=1}^{J} \frac{v_j}{\sigma_j} \exp\left[-\frac{1}{2\sigma_j^2}(y - c_j\mu)^2\right].$$

Detailed calculation, described in Hjort and Titterington (2010), shows that the changes in mean and precision satisfy, respectively,

$$A - a = b \sum_j R_j S_j T_j / \left[\sum_{j'} T_{j'} + o(b)\right] \tag{1.18}$$

and

$$B^{-1} - b^{-1} = \frac{\sum_j R_j^2 T_j}{\sum_j T_j} - \frac{\sum_j T_j R_j^2 S_j^2}{\sum_j T_j} + \frac{(\sum_j T_j R_j S_j)^2}{(\sum_j T_j)^2} + o(1), \tag{1.19}$$

where $R_j = c_j/\sigma_j$, $S_j = (x - c_j a)/\sigma_j$ and

$$T_j = \frac{v_j}{\sigma_j} \exp\left[-\frac{(y - ac_j)^2}{2\sigma_j^2}\right].$$

As explained earlier, the 'correct' asymptotic variance is given by the inverse of the Fisher information, so that the expected change in the inverse of the variance is the Fisher information corresponding to one observation, i.e. the negative of the expected second derivative of

$$\begin{aligned}\log\, p(y|\mu) &= \text{const.} + \log\, \left\{ \sum_{j=1}^{J} \frac{v_j}{\sigma_j} \exp\left(-\frac{1}{2\sigma_j^2}(y - c_j\mu)^2 \right) \right\} \\ &= \text{const.} + \log \sum_j T_j',\end{aligned}$$

where T_j' is the same as T_j except that a is replaced by μ. In what follows, R_j' and S_j' are similarly related to R_j and S_j. Then it is straightforward to show that the observed information is

$$-\frac{\partial^2}{\partial\mu^2} \log\, p(y|\mu) = \frac{\sum_j R_j'^2 T_j'}{\sum_j T_j'} - \frac{\sum_j T_j' R_j'^2 S_j'^2}{\sum_j T_j'} + \frac{(\sum_j T_j' R_j' S_j')^2}{(\sum_j T_j')^2}. \qquad (1.20)$$

The right-hand sides of (1.19) and (1.20) differ only in that (1.19) involves a whereas (1.20) involves μ, and (1.19) is correct just to $O(b)$ whereas (1.20) is exact. However, because of the nature of (1.18), stochastic approximation theory as applied by Smith and Makov (1981) will confirm that asymptotically a will converge to μ and terms of $O(b)$ will be negligible.

Thus, the approximate posterior distribution derived in this section behaves as we would wish, in terms of its mean and variance; by construction it is also Gaussian.

Hjort and Titterington (2010) show that, for this problem, the change in precision achieved by the variational approximation in incorporating an extra observation is unrealistically large and in fact equal, to first order, to the change corresponding to the complete-data scenario. They also report details for particular normal mixtures, including Minka's (2001b) 'clutter' problem.

1.4.4 Mixture of two known distributions

Recall that in this case

$$t_i(\lambda) = p(y_i|\lambda) = \lambda f_1(y_i) + (1 - \lambda) f_2(y_i),$$

in which λ is an unknown mixing weight between zero and one and f_1 and f_2 are known densities. The prior density for λ is assumed to be that of $\text{Be}(a^{(0)}, b^{(0)})$, and the beta approximation for the posterior based on D is assumed to be the $\text{Be}(a^{(n)}, b^{(n)})$ distribution, for hyperparameters $a^{(n)}$ and $b^{(n)}$. The expectation and variance of the beta approximation are respectively

$$\begin{aligned}E_n &= a^{(n)}/(a^{(n)} + b^{(n)}), \\ V_n &= (a^{(n)} b^{(n)})/[(a^{(n)} + b^{(n)})^2 (a^{(n)} + b^{(n)} + 1)] \\ &= E_n(1 - E_n)/(L_n + 1),\end{aligned}$$

where $L_n = a^{(n)} + b^{(n)}$. The limiting behaviour of E_n and V_n is of key interest. We would want E_n to tend to the true λ in some sense and V_n to tend to the variance of the correct posterior distribution of λ. Asymptotic normality of the approximating distribution is also desired. If E_n behaves as desired then $E_n(1 - E_n)$ tends to $\lambda(1 - \lambda)$ and, for V_n, the behaviour of $a^{(n)} + b^{(n)} + 1 = L_n + 1$, and therefore of L_n, is then crucial.

Hjort and Titterington (2010) compile the following results.

- For the case of *complete data*, E_n is, to first order, $\hat{\lambda}_{\mathrm{CO}}$, the proportion of the n observations that belong to the first component, it therefore does tend to λ, by the law of large numbers, and the limiting version of V_n is

$$V_{\mathrm{CO}} = \lambda(1 - \lambda)/n,$$

  for large n. Asymptotic normality of the posterior distribution of λ follows from the Bernstein–von Mises mirror result corresponding to the central limit result for $\hat{\lambda}_{\mathrm{CO}}$.
- The behaviour of the *correct* posterior distribution will be dictated by the behaviour of the maximum likelihood estimator $\hat{\lambda}_{\mathrm{ML}}$; for large n, approximately,

$$\begin{aligned} E(\hat{\lambda}_{\mathrm{ML}}) &= \lambda, \\ \mathrm{var}(\hat{\lambda}_{\mathrm{ML}}) &= \frac{1}{n \int \frac{[f_1(x) - f_2(x)]^2}{f(x)} \mathrm{d}x} \\ &= V_{\mathrm{ML}}, \end{aligned}$$

  say. Again, these properties can be transferred to the posterior distribution of λ, by a Bernstein–von Mises argument. This transference will apply as a general rule.
- For the *variational approximation*, which is of course a beta distribution, the limiting version of V_n is

$$V_{\mathrm{VA}} = \lambda(1 - \lambda)/n,$$

  as already mentioned in Section 1.3.6. This is the same as V_{CO}. It is therefore 'smaller than it should be' and would lead to unrealistically narrow interval estimates for λ.
- For the recursive *quasi-Bayes* approach of Smith and Makov (1978), the posterior mean is consistent for large n, and

$$V_n \doteq V_{\mathrm{QB}} = \lambda(1 - \lambda)/n;$$

  this is the same as for the confirmed-data case and the variational approximation, thereby falling foul of the same criticism as the latter as being 'too small'. Similar remarks apply to a recursive version of the variational approximation, implemented in Humphreys and Titterington (2000).
- For the *probabilistic editor*, the sequence $\{E_n\}$ of posterior means will converge (to the true λ value), and the posterior variance of the PE-based beta

approximation based on n observations is approximately, for large n,

$$V_{\mathrm{PE}} = n^{-1}\lambda(1-\lambda)\left[1 - \int \frac{f_1 f_2}{\lambda f_1 + (1-\lambda) f_2}\right]^{-1},$$

which Hjort and Titterington (2010) then show is equal to V_{ML}. Thus, asymptotically, the probabilistic editor, and by implication the moment-matching version of expectation propagation, get the variance right. The same is shown to be true for the standard EP and ADF approaches based on minimisation of the Kullback–Leibler divergence rather than on moment-matching.

1.4.5 Discussion

The simulation exercise of Humphreys and Titterington (2000), mentioned in Section 1.3.5, compared empirically the nonrecursive variational approximation, its recursive variant, the recursive quasi-Bayes and probabilistic editor, and the Gibbs sampler, the last of which can be regarded as providing a reliable estimate of the true posterior. As can be expected on the basis of the above analysis, the approximation to the posterior density provided by the probabilistic editor is very similar to that obtained from the Gibbs sampler, whereas the other approximations are 'too narrow'. Furthermore, the variances associated with the various approximations are numerically very close to the 'asymptotic' values derived above. The implication is that this is also the case for the corresponding version of the EP algorithm based on moment-matching, mentioned in Section 3.3.3 of Minka (2001a), of which the PE represents an online version. Of course EP updates using KL divergence rather than (always) matching moments, but the two versions perform very similarly in Minka's empirical experiments, and Section 1.4.4 reflects this asymptotically. Recursive versions of the algorithm with KL update, i.e. versions of ADF, are outlined and illustrated by simulation in Chapter 5 of Stephens (1997) for mixtures of known densities, extending earlier work by Bernardo and Girón (1988), and for mixtures of Gaussian densities with all parameters unknown, including the mixing weights. For mixtures of two known densities, Stephens notes that, empirically, the KL update appears to produce an estimate of the posterior density that is indistinguishable from the MCMC estimate, and is much superior to the quasi-Bayes estimate, which is too narrow. For a mixture of four known densities, for which the conjugate prior distributions are four-cell Dirichlet distributions, Stephens shows the KL update clearly to be better than the quasi-Bayes update, but somewhat more 'peaked' than it should be. This is because, in terms of the approach of the present paper, there are insufficient hyperparameters in the conjugate family to match all first and second moments. For a J-cell Dirichlet, with J hyperparameters, there are $J-1$ independent first moments and $J(J-1)/2$ second moments, so that full moment-matching is not possible for $J > 2$, that is for any case but mixtures of $J = 2$ known densities.

This is implicit in the work of Cowell *et al.* (1996) on recursive updating, following on from Spiegelhalter and Lauritzen (1990) and Spiegelhalter and Cowell (1992), and referred to in Section 3.3.3 of Minka (2001a) and Section 9.7.4 of

Cowell *et al.* (1999). They chose Dirichlet hyperparameters to match first moments and the average variance of the parameters. Alternatively, one could match the average of the variances and covariances. However, the upshot is that there is no hope that a pure Dirichlet approximation will produce a totally satisfactory approximation to the 'correct' posterior for $J > 2$, whether through EP or the recursive alternatives, based on KL updating or moment-matching. Nevertheless, these versions should be a distinct improvement on the quasi-Bayes and variational approximations in terms of variance. A possible way forward for small J is to approximate the posterior by a mixture of a small number of Dirichlets. To match all first- and second-order moments of a posterior distribution of a set of J-cell multinomial probabilities one would need a mixture of K pure J-cell Dirichlets, where

$$KJ + (K - 1) = (J - 1) + (J - 1) + (J - 1)(J - 2)/2,$$

i.e. Dirichlet hyperparameters + mixing weights = first moments + variances + covariances. This gives $K = J/2$. Thus, for even J, the match can be exact, but for odd J there would be some redundancy. In fact, even for J as small as 4, the algebraic details of the moment-matching become formidable.

Acknowledgements

This work has benefited from contact with Philip Dawid, Steffen Lauritzen, Yee Whye Teh, Jinghao Xue and especially Nils Lid Hjort. The chapter was written, in part, while the author was in residence at the Isaac Newton Institute in Cambridge, taking part in the Research Programme on Statistical Theory and Methods for Complex High-Dimensional Data.

References

Athans, M., Whiting, R. and Gruber, M. (1977) A suboptimal estimation algorithm with probabilistic editing for false measurements with application to target tracking with wake phenomena. *IEEE Transactions on Automatic Control*, **AC-22**, 273–384.

Attias, H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 21–30. Morgan Kaufman.

Attias, H. (2000) A variational Bayesian framework for graphical models. In *Advances in Neural Information Processes and Systems* **12** (eds S. A. Solla, T. K. Leen, and K. L. Müller), pp. 209–215. MIT Press.

Bernardo, J. M. and Girón, F. J. (1988) A Bayesian analysis of simple mixture problems (with discussion). In *Bayesian Statistics* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 67–78.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning.* Springer.

Corduneanu, A. and Bishop, C. M. (2001) Variational Bayesian model selection for mixture distributions. In *Proceedings of the 8th Conference on Artificial Intelligence and Statistics* (eds T. Richardson and T. Jaakkola), pp. 27–34. Morgan Kaufman.

Cowell, R. G., Dawid, A. P. and Sebastiani, P. (1996) A comparison of sequential learning methods for incomplete data. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 533–541. Clarendon Press.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems.* Springer.

Csiszár, I. and Tusnády, G. (1984) Information geometry and alternating minimization procedures. *Statistics & Decisions, Supplement*, **1**, 205–237.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Hjort, N. L. and Titterington, D. M. (2010) On Expectation Propagation and the Probabilistic Editor in some simple mixture problems. Preprint.

Humphreys, K. and Titterington, D. M. (2000) Approximate Bayesian inference for simple mixtures. In *COMPSTAT 2000* (eds J. G. Bethlehem and P. G. M. van der Heijden), pp. 331–336. Physica-Verlag.

Humphreys, K. and Titterington, D. M. (2001) Some examples of recursive variational approximations. In *Advanced Mean Field Methods: Theory and Practice* (eds M. Opper and D. Saad), pp. 179–195. MIT Press.

Jordan, M. I. and Jacobs, R. A. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181–214.

McGrory, C. A. (2005) Variational approximations in Bayesian model selection. PhD Thesis, University of Glasgow.

McGrory, C. A. and Titterington, D. M. (2007) Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statististics and Data Analysis*, **51**, 5352–5367.

MacKay, D. J. C. (1992) Bayesian interpolation. *Neural Computation*, **4**, 415–447.

MacKay, D. J. C. (1997) Ensemble learning for hidden Markov models. Technical Report, Cavendish Laboratory, University of Cambridge.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions.* John Wiley & Sons, Ltd.

McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. John Wiley & Sons, Ltd.

Makov, U. E. (1983) Approximate Bayesian procedures for dynamic linear models in the presence of jumps. *The Statistician*, **32**, 207–213.

Makov, U. E. and Smith, A. F. M. (1977) A quasi-Bayes unsupervised learning procedure for priors. *IEEE Transactions on Information Theory*, **IT-23**, 761–764.

Maybeck, P. S. (1982) *Stochastic Models, Estimation and Control.* Academic Press.

Minka, T. (2001a) A family of algorithms for approximate Bayesian inference. PhD Dissertation, Massachusetts Institute of Technology.

Minka, T. (2001b) Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artifical Intelligence*.

Neal, R. M. and Hinton, G. E. (1999) A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* (ed. M. Jordan), pp. 355–368. MIT Press.

Penny, W. D. and Roberts, S. J. (2000) Variational Bayes for 1-dimensional mixture models. Technical Report PARG-2000-01, Oxford University.

Peters, B. C. and Walker, H. F. (1978) An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics*, **35**, 362–378.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recogntion. *Proceedings of the IEEE*, **77**, 257–286.

Robert, C. P., Celeux, G. and Diebolt, J. (1993) Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics and Probability Letters*, **16**, 77–83.

Smith, A. F. M. and Makov, U. E. (1978) A quasi-Bayes sequential procedure for mixtures. *Journal of the Royal Statistical Society, Series B*, **40**, 106–112.

Smith, A. F. M. and Makov, U. E. (1981) Unsupervised learning for signal versus noise. *IEEE Transactions on Information Theory*, **IT-27**, 498–500.

Spiegelhalter, D. J. and Cowell, R. G. (1992) Learning in probabilistic expert systems. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 447–465. Clarendon Press.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990) Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.

Stephens, M. (1997) Bayesian methods for mixtures of normal distributions. DPhil Dissertation, University of Oxford.

Titterington, D. M. (1976) Updating a diagnostic system using unconfirmed cases. *Applied Statistics*, **25**, 238–247.

Titterington, D. M. (2004) Bayesian methods for neural networks and related models. *Statistical Science*, **19**, 128–139.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, Ltd.

Ueda, N. and Ghahramani, Z. (2002) Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, **15**, 1223–1241.

Wang, B. and Titterington, D. M. (2004a) Local convergence of variational Bayes estimators for mixing coefficients. In *2004 JSM Proceedings*, published on CD Rom. American Statistical Association.

Wang, B. and Titterington, D. M. (2004b) Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (eds M. Chickering and J. Halperin), pp. 577–584. AUAI Press.

Wang, B. and Titterington, D. M. (2004c) Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, **20**, 151–170.

Wang, B. and Titterington, D. M. (2005a) Variational Bayes estimation of mixing coefficients. In *Deterministic and Statistical Methods in Machine Learning*, Lecture Notes in Artificial Intelligence, Vol. 3635 (eds J. Winkler, M. Niranjan and N. Lawrence), pp. 281–295. Springer-Verlag.

Wang, B. and Titterington, D. M. (2005b) Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pp. 373–380.

Wang, B. and Titterington, D. M. (2006) Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, **1**, 625–650.