

# Variational Bayes Estimation of Mixing Coefficients

Bo Wang and D. M. Titterington

Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK.

**Abstract.** We investigate theoretically some properties of variational Bayes approximations based on estimating the mixing coefficients of known densities. We show that, with probability 1 as the sample size  $n$  grows large, the iterative algorithm for the variational Bayes approximation converges locally to the maximum likelihood estimator at the rate of  $O(1/n)$ . Moreover, the variational posterior distribution for the parameters is shown to be asymptotically normal with the same mean but a different covariance matrix compared with those for the maximum likelihood estimator. Furthermore we prove that the covariance matrix from the variational Bayes approximation is ‘too small’ compared with that for the MLE, so that resulting interval estimates for the parameters will be unrealistically narrow.

## 1 Introduction

Variational Bayes approximations, introduced by Attias [1, 2], have been successfully applied to complex models involving incomplete-data where computational difficulties arise in the ideal Bayesian approaches, as for instance with hidden Markov models and mixture models. The approximations are widely recognised to be effective and promising in a series of papers, such as [3–10] and the references therein. In these earlier contributions, the approximations were shown empirically to be convergent and consistent. In Attias [1, 2] and Penny and Roberts [10] the authors claimed that the variational Bayes estimator approaches the maximum likelihood estimator in the large sample limit, but no rigorous proof was given. While experimental results are often satisfactory in practice, exact theoretical assessment of the quality of this method is an important issue.

In this paper we study some properties of variational Bayes approximations theoretically for certain mixture models. Based on estimating the mixing coefficients of known densities, we show that, with probability 1 as the sample size  $n$  grows large, the iterative algorithm for the variational Bayes approximation converges locally to the maximum likelihood estimator at the rate of  $O(1/n)$ . Moreover, we prove that the variational posterior distribution for the parameters is asymptotically normal with the same mean but a different covariance matrix compared with those for the maximum likelihood estimator. Further developments show that the covariance matrix from the variational Bayes approximation is ‘too small’ compared with that for the MLE, so that resulting interval estimates for the parameters will be too narrow. Numerical examples reinforce the theoretical analysis.

## 2 The mixture model and the variational approximation

We consider a model in which we have a mixture of  $m$  known densities  $p_1, p_2, \dots, p_m$ . The density of an observation is given by

$$p(y_i|\Theta) = \sum_{s=1}^m p_s(y_i)p(s_i = s|\Theta), \quad (1)$$

where  $y_i \in \mathbb{R}^d$  denotes the  $i$ th observed data vector, and  $s_i$  indicates the hidden component that generated it. The components are labelled by  $s = 1, 2, \dots, m$ , and the component  $s$  has mixing coefficient  $\theta_s = p(s_i = s|\Theta)$  for any  $i$ . We write the parameters collectively as  $\Theta = (\theta_1, \theta_2, \dots, \theta_m)'$ , and assign a Dirichlet prior distribution  $\mathcal{D}(a_1^{(0)}, a_2^{(0)}, \dots, a_m^{(0)})$  to  $\Theta$ .

Suppose that we have (complete) data consisting of a random sample of size  $n$ , with  $Y = (y_1, y_2, \dots, y_n)'$  and  $S = (s_1, s_2, \dots, s_n)'$ . Then the joint density of  $\Theta$ ,  $S$  and  $Y$  is

$$p(\Theta, S, Y) \propto \left\{ \prod_{s=1}^m \theta_s^{a_s^{(0)} - 1} \right\} \prod_{i=1}^n \{ \theta_{s_i} p_{s_i}(y_i) \}.$$

In the variational Bayes approach, we use an approximating density  $q(S, \Theta)$  for  $p(S, \Theta|Y)$ , which factorises as  $q(S, \Theta) = q^{(S)}(S)q^{(\Theta)}(\Theta)$ , and is chosen to maximise

$$\int \sum_{\{S\}} q(S, \Theta) \log \frac{p(\Theta, S, Y)}{q(S, \Theta)} d\Theta.$$

It follows that  $q^{(S)}(S)$  factorises as  $q^{(S)}(S) = \prod_{i=1}^n q_i^{(S)}(s_i)$  and the variational posterior can be obtained by the following iterative procedure. In turn, we perform the following two stages.

(i) Optimise  $q^{(\Theta)}(\Theta)$  for fixed  $\{q_i^{(S)}(s_i), i = 1, \dots, n\}$ . This step results in

$$q^{(\Theta)}(\Theta) \sim \mathcal{D}(\{a_s^{(0)} + \sum_{i=1}^n r_{is}\}_{s=1}^m), \quad (2)$$

where  $r_{is} = q_i^{(S)}(s_i = s)$ .

(ii) Optimise  $\{q_i^{(S)}(s_i), s_i = 1, \dots, m, i = 1, \dots, n\}$  for fixed  $q^{(\Theta)}(\Theta)$ . This results in

$$r_{is} = q_i^{(S)}(s_i = s) = \frac{p_s(y_i)\phi_s}{\sum_{s=1}^m p_s(y_i)\phi_s}, \quad s = 1, \dots, m, \quad (3)$$

where

$$\phi_s = \exp\left\{ \int q^{(\Theta)}(\alpha) \log \alpha_s d\alpha \right\}, \quad \alpha = (\alpha_1, \dots, \alpha_m)'$$

We write  $\phi = (\phi_1, \dots, \phi_m)'$ .

This iterative procedure can be initialised by taking, for each  $i$  and  $s$ ,

$$r_{is} = \frac{p_s(y_i)a_s^{(0)}}{\sum_{s=1}^m p_s(y_i)a_s^{(0)}}.$$

### 3 Local convergence of the iterative procedure

An iterative procedure is said to *converge locally* to a limit if the iterates converge to that limit whenever the starting values are sufficiently near to the limit. We suppose that the true value of the parameter  $\Theta$  is  $\Theta^*$ , with  $0 < \theta_s^* < 1$ ,  $s = 1, \dots, m$ . In this section we shall show that the algorithm presented in the previous section converges locally to  $\Theta^*$ .

In the  $k$ th iteration, we write

$$\theta_s^{(k)} = \frac{1}{n} \sum_{i=1}^n r_{is}^{(k)}, \quad \Theta^{(k)} = (\theta_1^{(k)}, \dots, \theta_m^{(k)})',$$

where the notation for  $r$  now recognises that the  $r$ -values change from iteration to iteration.

We define the variational Bayesian estimator of a parameter as its variational posterior mean. Then the procedure given by (2) and (3) suggests the following algorithm for calculating the variational Bayesian estimate of  $\Theta$ : starting with some initial value  $\Theta^{(1)}$ , successive iterates are defined inductively by

$$\Theta^{(k+1)} = \Phi_n(\Theta^{(k)}) \tag{4}$$

for  $k = 1, 2, \dots$ , where  $\Phi_n = (\Phi_n^1, \dots, \Phi_n^m)'$ ,

$$\begin{aligned} \Phi_n^s(\Theta^{(k)}) &\triangleq \frac{1}{n} \sum_{i=1}^n \frac{p_s(y_i) \phi_s^{(k)}}{\sum_{s=1}^m p_s(y_i) \phi_s^{(k)}}, \\ \phi_s^{(k)} &= \exp\left\{ \int q^{(\Theta)^{(k)}}(\alpha) \log \alpha_s d\alpha \right\}, \end{aligned} \tag{5}$$

and

$$q^{(\Theta)^{(k)}}(\alpha) \sim \mathcal{D}(\{a_s^{(0)} + n\theta_s^{(k)}\}_{s=1}^m). \tag{6}$$

We have the following theorem.

**Theorem 1.** *With probability 1 as  $n$  approaches infinity, the iterative procedure (4) converges locally to the true value  $\Theta^*$ .*

*Proof.* We first prove that, with probability 1 as  $n$  approaches infinity, the operator  $\Phi_n$  is *locally contractive*; that is, there exists a number  $\lambda$ ,  $0 \leq \lambda < 1$ , such that

$$\|\Phi_n(\bar{\Theta}) - \Phi_n(\Theta^*)\| \leq \lambda \|\bar{\Theta} - \Theta^*\|,$$

whenever  $\bar{\Theta}$  lies sufficiently near  $\Theta^*$ .

Since  $\bar{\Theta}$  is near  $\Theta^*$ , one can write

$$\Phi_n(\bar{\Theta}) - \Phi_n(\Theta^*) = \nabla \Phi_n(\Theta^*)(\bar{\Theta} - \Theta^*) + O(\|\bar{\Theta} - \Theta^*\|^2),$$

where  $\nabla \Phi_n(\Theta^*)$  denotes the gradient of  $\Phi_n(\Theta)$  evaluated at  $\Theta^*$ . Consequently, it is sufficient to show that  $\nabla \Phi_n(\Theta^*)$  converges with probability 1 to an operator which has norm less than 1.

From the definition of the operator  $\Phi_n$ , we have, for  $s, j = 1, \dots, m$ , that

$$\nabla_j \Phi_n^s(\Theta^*) = \frac{1}{n} \sum_{i=1}^n \frac{p_s(y_i) \phi_s^j(\sum_{t=1}^m p_t(y_i) \phi_t) - p_s(y_i) \phi_s(\sum_{t=1}^m p_t(y_i) \phi_t^j)}{(\sum_{t=1}^m p_t(y_i) \phi_t)^2},$$

where  $\phi_s$  is as given in (5)-(6) but with  $\Theta^{(k)}$  replaced by  $\Theta^*$  in (6), and where  $\phi_s^j$  denotes the derivative of  $\phi_s$  with respect to  $\theta_j$ . However, it is obvious that, as  $n$  tends to infinity, the mean of  $\theta_s$  corresponding to the density  $q^{(\Theta)}(\Theta)$  is

$$\frac{a_s^{(0)} + n\theta_s^*}{\sum_{s=1}^m a_s^{(0)} + n} \rightarrow \theta_s^*,$$

the covariance between  $\theta_s$  and  $\theta_t$ , for  $s \neq t$ , is

$$-\frac{(a_s^{(0)} + n\theta_s^*)(a_t^{(0)} + n\theta_t^*)}{(\sum_{s=1}^m a_s^{(0)} + n)^2 (\sum_{s=1}^m a_s^{(0)} + n + 1)} = O\left(\frac{1}{n}\right) \rightarrow 0,$$

and the variance of  $\theta_s$  is

$$\frac{(a_s^{(0)} + n\theta_s^*)(\sum_{s=1}^m a_s^{(0)} - a_s^{(0)} + n - n\theta_s^*)}{(\sum_{s=1}^m a_s^{(0)} + n)^2 (\sum_{s=1}^m a_s^{(0)} + n + 1)} = O\left(\frac{1}{n}\right) \rightarrow 0.$$

From these we show in Appendix A that, as  $n$  tends to infinity,

$$\begin{aligned} \phi_s &\rightarrow \theta_s^*, \\ \phi_s^j &\rightarrow \begin{cases} 1, & \text{if } j = s; \\ 0, & \text{if } j \neq s. \end{cases} \end{aligned}$$

Thus from Appendix B we obtain that, with probability 1,

$$\begin{aligned} \nabla_j \Phi_n^s(\Theta^*) &\rightarrow \begin{cases} \mathbb{E} \left\{ \frac{p_s(y_i) (\sum_{s=1}^m p_s(y_i) \theta_s^*) - p_s^2(y_i) \theta_s^*}{(\sum_{s=1}^m p_s(y_i) \theta_s^*)^2} \right\}, & \text{if } j = s; \\ -\mathbb{E} \left\{ \frac{p_s(y_i) p_j(y_i) \theta_s^*}{(\sum_{s=1}^m p_s(y_i) \theta_s^*)^2} \right\}, & \text{if } j \neq s, \end{cases} \\ &= \mathbb{E} \left\{ \frac{p_s(y_i)}{\sum_{s=1}^m p_s(y_i) \theta_s^*} \right\} \delta_{js} - \theta_s^* \cdot \mathbb{E} \left\{ \frac{p_s^2(y_i)}{(\sum_{s=1}^m p_s(y_i) \theta_s^*)^2} \right\}, \end{aligned}$$

where  $\delta_{js}$  is the Kronecker delta function and the expectation corresponds to the true model in which  $\Theta = \Theta^*$ .

Since

$$\mathbb{E} \left\{ \frac{p_s(y_i)}{\sum_{s=1}^m p_s(y_i) \theta_s^*} \right\} = \int \frac{p_s(y_i)}{\sum_{s=1}^m p_s(y_i) \theta_s^*} \cdot p(y_i) dy_i = 1, \quad (7)$$

where  $p(y_i) = \sum_{s=1}^m p_s(y_i) \theta_s^*$  is the (true unconditional) probability density of each observation, the last expression can be rewritten as

$$\nabla \Phi_n(\Theta^*) \rightarrow I - \Xi \Psi,$$

where  $I$  denotes the identity matrix,  $\Xi \triangleq \text{diag}(\theta_1^*, \theta_2^*, \dots, \theta_m^*)$  and

$$\Psi = \mathbb{E} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_m \end{pmatrix} (\nu_1 \nu_2 \cdots \nu_m), \quad \nu_s = \frac{p_s(y_i)}{\sum_{s=1}^m p_s(y_i) \theta_s^*}. \quad (8)$$

Obviously  $\Psi$  is a positive definite matrix. Therefore, as  $n$  tends to infinity,

$$\nabla \Phi_n(\Theta^*) < I.$$

Next we prove that  $I - \Xi\Psi \geq 0$ . Since  $\Xi$  is a positive diagonal matrix, it suffices to show that

$$\Theta' \Xi^{-1} \Theta \geq \Theta' \Xi^{-1} \Xi \Psi \Theta = \Theta' \Psi \Theta \quad (9)$$

for any  $\Theta = (\theta_1, \dots, \theta_m)'$ .

In fact, one has

$$\begin{aligned} \Theta' \Psi \Theta &= \mathbb{E}(\theta_1 \nu_1 + \theta_2 \nu_2 + \cdots + \theta_m \nu_m)^2 \\ &= \mathbb{E} \left[ \sum_{s=1}^m (\theta_s^{*-1} \theta_s \cdot \theta_s^* \nu_s) \right]^2. \end{aligned}$$

As a corollary of Schwarz's inequality we have that, if  $\eta_s \geq 0$  for  $s = 1, \dots, m$  and  $\sum_{s=1}^m \eta_s = 1$ , then

$$\left| \sum_{s=1}^m \xi_s \eta_s \right|^2 \leq \sum_{s=1}^m \xi_s^2 \eta_s \quad (10)$$

for all  $\{\xi_s\}_{s=1, \dots, m}$  (see [11]).

Applying this result and noting that  $\sum_{s=1}^m \theta_s^* \nu_s = 1$ , we obtain

$$\begin{aligned} \Theta' \Psi \Theta &\leq \mathbb{E} \left[ \sum_{s=1}^m (\theta_s^{*-1} \theta_s)^2 \theta_s^* \nu_s \right] = \mathbb{E} \left[ \sum_{s=1}^m \theta_s^{*-1} \theta_s^2 \nu_s \right] \\ &= \sum_{s=1}^m \theta_s^{*-1} \theta_s^2 \mathbb{E} \nu_s = \sum_{s=1}^m \theta_s^{*-1} \theta_s^2 = \Theta' \Xi^{-1} \Theta, \end{aligned}$$

because of (7).

Thus we have proved that  $\nabla \Phi_n(\Theta^*)$  converges with probability 1 to an operator with norm less than 1, and consequently the operator  $\Phi_n$  is *locally contractive*.

From the above proof, it is then obvious that  $\Phi_n(\Theta^*)$  tends to  $\Theta^*$  as  $n$  approaches infinity. Since

$$\begin{aligned} \|\Theta^{(k+1)} - \Theta^*\| &\leq \|\Phi_n(\Theta^{(k)}) - \Phi_n(\Theta^*)\| + \|\Phi_n(\Theta^*) - \Theta^*\| \\ &\leq \lambda \|\Theta^{(k)} - \Theta^*\| + \|\Phi_n(\Theta^*) - \Theta^*\|, \end{aligned}$$

the iterative procedure (4) converges locally to the true value  $\Theta^*$  as  $n$  approaches infinity.  $\square$

## 4 The convergence rate of the variational Bayes estimator

In this section we consider the rate at which the variational Bayes estimator converges to the maximum likelihood estimator (MLE). Suppose the sample size  $n$  is large. Let  $\hat{\Theta}^n = (\hat{\theta}_1^n, \dots, \hat{\theta}_m^n)'$  and  $\tilde{\Theta}^n = (\tilde{\theta}_1^n, \dots, \tilde{\theta}_m^n)'$  denote the fixed point of iteration (4) in the neighbourhood of the true value and the variational Bayes estimator, respectively; that is, from (4),

$$\hat{\theta}_s^n = \frac{1}{n} \sum_{i=1}^n \frac{p_s(y_i) \phi_s}{\sum_{s=1}^m p_s(y_i) \phi_s}, \quad (11)$$

$$q^{(\Theta)}(\alpha) \sim \mathcal{D}(\{a_s^{(0)} + n\hat{\theta}_s^n\}_{s=1}^m),$$

$$\phi_s = \exp\left\{\int q^{(\Theta)}(\alpha) \log \alpha_s d\alpha\right\},$$

and

$$\tilde{\theta}_s^n \triangleq \int \alpha q^{(\Theta)}(\alpha) d\alpha = \frac{a_s^{(0)} + n\hat{\theta}_s^n}{\sum_{s=1}^m a_s^{(0)} + n}.$$

Hence  $\tilde{\Theta}^n = \hat{\Theta}^n + O(1/n)$ .

Suppose that  $\bar{\Theta}^n = (\bar{\theta}_1^n, \dots, \bar{\theta}_m^n)'$  is the strongly consistent MLE of the parameter  $\Theta$ ; that is,  $\bar{\theta}_s^n$  is the solution of the following likelihood equation (see Peters and Walker [11], Redner and Walker [12]):

$$l_s^n(\Theta) \triangleq \theta_s - \frac{1}{n} \sum_{i=1}^n \frac{p_s(y_i) \theta_s}{\sum_{s=1}^m p_s(y_i) \theta_s} = 0,$$

for  $s = 1, \dots, m$ . Let  $l^n = (l_1^n, \dots, l_m^n)'$ . Then we have the following lemma.

**Lemma 1.** *With probability 1 as  $n$  tends to infinity, the gradient of the likelihood function  $l^n$ , evaluated at the true value  $\Theta^*$ , converges uniformly to  $\Xi\Psi$ , where  $\Xi$  and  $\Psi$  are defined as in the previous section.*

*Proof.* After some straightforward manipulations, the convergence is a direct corollary of the strong law of large numbers.  $\square$

Meanwhile, from (11) we have that

$$\begin{aligned} 0 &= \hat{\theta}_s^n - \frac{1}{n} \sum_{i=1}^n \frac{p_s(y_i) \phi_s}{\sum_{s=1}^m p_s(y_i) \phi_s} \\ &= l_s^n(\hat{\Theta}^n) + \frac{1}{n} \sum_{i=1}^n \left[ \frac{p_s(y_i) \hat{\theta}_s^n}{\sum_{s=1}^m p_s(y_i) \hat{\theta}_s^n} - \frac{p_s(y_i) \phi_s}{\sum_{s=1}^m p_s(y_i) \phi_s} \right] \\ &= l_s^n(\hat{\Theta}^n) + \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^m p_s(y_i) p_j(y_i) (\hat{\theta}_s^n \phi_j - \phi_s \hat{\theta}_j^n)}{[\sum_{s=1}^m p_s(y_i) \hat{\theta}_s^n] [\sum_{s=1}^m p_s(y_i) \phi_s]}. \end{aligned}$$

It follows from Appendix A that  $\phi_s = \hat{\theta}_s^n + O(1/n)$ , so  $\hat{\theta}_s^n \phi_j - \phi_s \hat{\theta}_j^n = O(1/n)$ . Thus the second term is of order  $O(1/n)$ . From Taylor's expansion the first term can be rewritten as

$$\begin{aligned} l^n(\hat{\Theta}^n) &= l^n(\bar{\Theta}^n) + \nabla l^n(\bar{\Theta}^n + \lambda(\hat{\Theta}^n - \bar{\Theta}^n))(\hat{\Theta}^n - \bar{\Theta}^n) \\ &= \nabla l^n(\bar{\Theta}^n + \lambda(\hat{\Theta}^n - \bar{\Theta}^n))(\hat{\Theta}^n - \bar{\Theta}^n), \end{aligned}$$

where  $0 \leq \lambda \leq 1$ . Thus, we obtain

$$0 = \nabla l^n(\bar{\Theta}^n + \lambda(\hat{\Theta}^n - \bar{\Theta}^n))(\hat{\Theta}^n - \bar{\Theta}^n) + O\left(\frac{1}{n}\right).$$

We have proved that  $\hat{\Theta}^n$  converges to  $\Theta^*$ , and it is well known that  $\bar{\Theta}^n$  tends to  $\Theta^*$ , so it follows from Lemma 1 that  $\nabla l^n(\bar{\Theta}^n + \lambda(\hat{\Theta}^n - \bar{\Theta}^n))$  converges to  $\Xi\Psi$ . Therefore, we have that  $\hat{\Theta}^n = \bar{\Theta}^n + O(1/n)$ , and consequently that  $\hat{\Theta}^n = \bar{\Theta}^n + O(1/n)$ .

*Remark 1.* It is known that the (non-variational) Bayes estimator and the MLE get closer to each other at rate  $O(1/n)$ , so the variational Bayes estimator is asymptotically consistent at the same rate.

## 5 Asymptotic normality of the variational posterior distribution

In this section, we show that the variational posterior distribution for the parameter  $\Theta$  obtained by the iterative procedure has also the property of asymptotic normality. This implies that the variational posterior becomes more and more concentrated around the true parameter value as the sample size grows.

Suppose the sample size  $n$  is large. Denote by  $\hat{\Theta}^n = (\hat{\theta}_1^n, \dots, \hat{\theta}_m^n)'$  the fixed point of the iteration (4). Thus the variational posterior density of  $\Theta$  at  $\hat{\Theta}^n$  is

$$q_n^{(\Theta)}(\Theta) \sim \mathcal{D}(\{\hat{a}_s^{(n)}\}_{s=1}^m), \quad (12)$$

where  $\hat{a}_s^{(n)} = a_s^{(0)} + \sum_{i=1}^n \hat{r}_{is}$  and  $\hat{r}_{is} = q_i^{(S)}(s_i = s)$ .

In the rest of the paper, we express  $\theta_m$  explicitly as  $1 - \sum_{s=1}^{m-1} \theta_s$ . Thus the density (12) can be rewritten as a density of exponential family type:

$$\begin{aligned} q_n^{(\Theta)}(\Theta) &\propto \exp\{(\hat{a}_1^{(n)} - 1) \log \theta_1 + \dots + (\hat{a}_m^{(n)} - 1) \log(1 - \sum_{s=1}^{m-1} \theta_s)\} \\ &\triangleq \exp\{h'(\Theta)\beta - \alpha\psi(\Theta)\}, \end{aligned} \quad (13)$$

where

$$\begin{aligned} \beta &\triangleq (\hat{a}_1^{(n)} - 1, \dots, \hat{a}_{m-1}^{(n)} - 1)', \\ h(\Theta) &\triangleq (\log \theta_1, \dots, \log \theta_{m-1})', \\ \alpha &\triangleq 1 - \hat{a}_m^{(n)}, \text{ and } \psi(\Theta) \triangleq \log\left(1 - \sum_{s=1}^{m-1} \theta_s\right). \end{aligned}$$





for any  $\Theta = (\theta_1, \dots, \theta_{m-1})'$ .

In fact, along the same lines as the proof of inequality (9) one can obtain

$$\Theta' \mathbb{E}(VV') \Theta \leq \sum_{s=1}^m \theta_s^2 \theta_s^{*-1},$$

where  $\theta_m \triangleq -\sum_{s=1}^{m-1} \theta_s$ .

On the other hand, by (14) it can be easily checked that

$$\Theta' \Lambda^{-1} \Theta = \sum_{s=1}^m \theta_s^2 \theta_s^{*-1}.$$

The proof is complete.  $\square$

By (10) equality in (16) holds if and only if the mixture model (1) has only one component or the supports of the component densities are disjoint. If the components are well separated or have little overlap, the mixture distribution can be regarded approximately as multinomial. In this case, at a given observation  $y_i$ , there exists one  $p_s(y_i)$  which is far larger than the others, and therefore the inverse of Fisher information matrix is close to the covariance matrix of the variational posterior distribution. Proposition 1 shows that in general the covariance matrix from the variational Bayes approximation is ‘too small’ compared with that for the MLE, so that resulting interval estimates for the parameters will be too narrow.

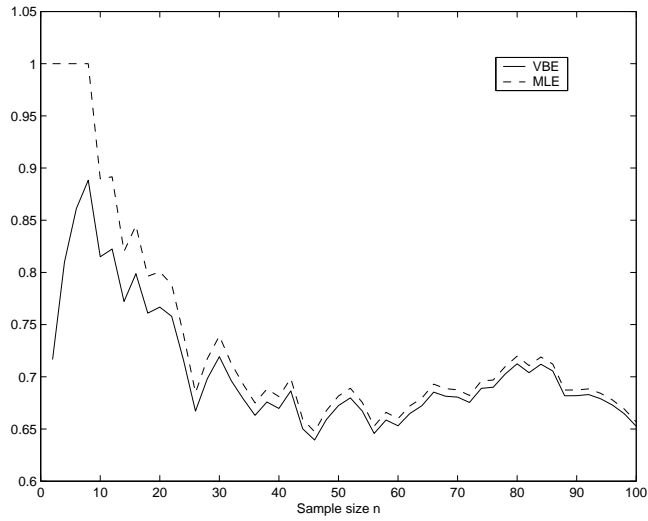
## 7 Numerical experiments

We demonstrate our results with a simple mixture of two known normal densities.

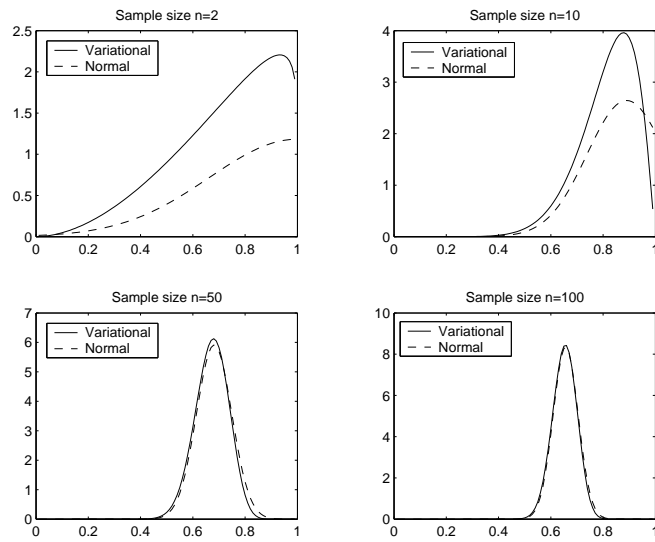
First we fix the two normal densities to have means of 2 and 4 and unit variance, and generate a total of 100 observations using  $\theta = 0.65$ . For different sample sizes up to  $n = 100$  the MLE  $\hat{\theta}^n$  and the variational Bayes estimate  $\hat{\theta}^n$  based on a Beta prior distribution for  $\theta$  with  $a_1^{(0)} = a_2^{(0)} = 1$  are computed using the first  $n$  observations, and are plotted in Figure 1. When the sample size is small, there is a gap between the two estimates, but quickly they track each other very closely as the sample size grows.

In Figure 2, we plot the corresponding variational posterior densities and the normal density  $\mathcal{N}(\hat{\theta}^n, \Lambda/n)$  for the sample sizes  $n = 2, 10, 50, 100$ , where  $\Lambda$  is defined in (14). It can be seen that the variational posterior density becomes closer and closer to the limiting normal density.

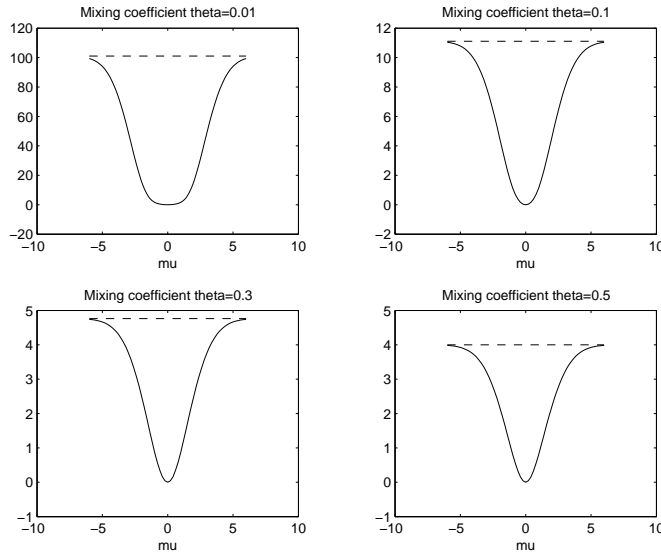
Finally, to compare the variance  $\Lambda$  associated with the variational Bayes approximation with the Fisher information, we fix one component to have mean zero and unit variance and compute the inverse of the variance  $\Lambda$  and Fisher information, allowing the other component to have varying mean  $\mu$ . The results are plotted in Figure 3 for various values of the mixing coefficient  $\theta$ . Obviously, if the components in the mixture model are widely separated, these two quantities are very similar, whereas, if the components are nearly identical, they are very different.



**Fig. 1.** Variational Bayesian estimate of a mixing weight and MLE plotted against the sample size.



**Fig. 2.** Variational posteriors of the parameter  $\theta$  and normal densities for different sample sizes.



**Fig. 3.** The inverse of the variance associated with the variational Bayes approximation and the Fisher information for different mixing coefficients. The solid lines denote the Fisher information and the dashed horizontal lines indicate the inverse of the variance for the variational Bayes approximation.

## 8 Conclusion

We have investigated some properties of variational Bayes approximations, namely consistency and asymptotic normality, and compared the true covariance matrix of the posterior distribution (the inverse of Fisher information) with the covariance matrix associated with its variational Bayes approximation. The results reveal that in mixture models the point estimate obtained by using a factorised form  $q^{(S)}(S)q^{(\Theta)}(\Theta)$  for the posterior distributions of  $\Theta$  and  $S$  does not lead to bias for large samples, but the interval estimates for the parameters will be too narrow in general.

## Acknowledgement

This work was supported by a grant from the UK Science and Engineering Research Council. This work was also supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

## Appendix A

First the following conclusion holds. Suppose that  $p_n(x)$  is the probability density function of the random variable  $X_n$ , that  $\mathbb{E}X_n = \mu_n \rightarrow \mu$ , as  $n \rightarrow \infty$ , and that  $\text{Cov}(X_n) = O(1/n)$ . Then, for any function  $f(\cdot)$  with continuous second-order derivative near  $\mu$ , we have

$$\mathbb{E}f(X_n) = f(\mu_n) + O\left(\frac{1}{n}\right).$$

This follows from the Taylor expansion

$$\begin{aligned} f(X_n) &= f(\mu_n) + \nabla f(\mu_n) \cdot (X_n - \mu_n) + \frac{1}{2}(X_n - \mu_n)' \nabla^2 f(\mu_n)(X_n - \mu_n) \\ &\quad + o(\|X_n - \mu_n\|^2), \end{aligned}$$

because then

$$\begin{aligned} \mathbb{E}f(X_n) &= f(\mu_n) + \nabla f(\mu_n) \cdot (\mathbb{E}X_n - \mu_n) \\ &\quad + \frac{1}{2}\mathbb{E}[(X_n - \mu_n)' \nabla^2 f(\mu_n)(X_n - \mu_n)] + o(\mathbb{E}\|X_n - \mu_n\|^2) \\ &= f(\mu_n) + \frac{1}{2}\mathbb{E}[(X_n - \mu_n)' \nabla^2 f(\mu_n)(X_n - \mu_n)] + o(\mathbb{E}\|X_n - \mu_n\|^2) \\ &= f(\mu_n) + O\left(\frac{1}{n}\right). \end{aligned}$$

Applying this to the case of

$$f(x) = \log x, \quad \text{and} \quad X_n : q^{(\Theta)}(\alpha) \sim \mathcal{D}(\{a_s^{(0)} + n\theta_s\}_{s=1}^m),$$

we easily obtain that

$$\mathbb{E}f(X_n) = \int q^{(\Theta)}(\alpha) \log \alpha_s d\alpha = \log(\mathbb{E}X_n) + O\left(\frac{1}{n}\right).$$

Hence, from Taylor expansion, we have

$$\begin{aligned} \phi_s &= \exp\left\{\int q^{(\Theta)}(\alpha) \log \alpha_s d\alpha\right\} = \exp\left\{\log(\mathbb{E}X_n) + O\left(\frac{1}{n}\right)\right\} \\ &= \exp\{\log(\mathbb{E}X_n)\} + \exp\{\log(\mathbb{E}X_n)\} \cdot O\left(\frac{1}{n}\right) + O\left(\frac{1}{n^2}\right) \\ &= \mathbb{E}X_n + \mathbb{E}X_n \cdot O\left(\frac{1}{n}\right) \\ &= \theta_s + O\left(\frac{1}{n}\right) \rightarrow \theta_s. \end{aligned}$$

We have assumed that  $0 < \theta_s^* < 1$  and our conclusions are local in nature, so there is no loss of generality in restricting the discussion to  $0 < \varepsilon_0 \leq \theta_s^* \leq 1 - \varepsilon_0 < 1$  for a small positive constant  $\varepsilon_0$ . Thus the above convergences are also uniform in  $\theta_s$ . As a result we have

$$\phi_s^j \rightarrow \begin{cases} 1, & \text{if } j = s; \\ 0, & \text{if } j \neq s. \end{cases}$$

## Appendix B

We show that, if  $F_n(\cdot) \rightarrow F_0(\cdot)$  uniformly, then, with probability 1,

$$\frac{1}{n} \sum_{i=1}^n F_n(X_i) \rightarrow \mathbb{E}F_0(X_i)$$

for any sequence  $\{X_n\}$  of independent and identically distributed random variables.

This is a direct corollary of the strong law of large numbers and the following calculation:

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n F_n(X_i) - \mathbb{E}F_0(X_i) \right| \\ & \leq \left| \frac{1}{n} \sum_{i=1}^n F_n(X_i) - \frac{1}{n} \sum_{i=1}^n F_0(X_i) \right| + \left| \frac{1}{n} \sum_{i=1}^n F_0(X_i) - \mathbb{E}F_0(X_i) \right| \\ & \leq \frac{1}{n} \sum_{i=1}^n |F_n(X_i) - F_0(X_i)| + \left| \frac{1}{n} \sum_{i=1}^n F_0(X_i) - \mathbb{E}F_0(X_i) \right| \\ & \leq \sup_x |F_n(x) - F_0(x)| + \left| \frac{1}{n} \sum_{i=1}^n F_0(X_i) - \mathbb{E}F_0(X_i) \right|. \end{aligned}$$

## References

1. Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes. In Prade, H., Laskey, K., eds.: Proc. 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, Morgan Kaufmann Publishers (1999) 21–30
2. Attias, H.: A variational Bayesian framework for graphical models. In Solla, S., Leen, T., Muller, K.R., eds.: Advances in Neural Information Processing Systems 12. MIT Press, Cambridge, MA (2000) 209–215
3. Beal, M.J.: Variational Algorithms for Approximate Bayesian Inference. PhD thesis, University College London (2003)
4. Corduneanu, A., Bishop, C.M.: Variational Bayesian model selection for mixture distributions. In Richardson, T., Jaakkola, T., eds.: Proceedings Eighth International Conference on Artificial Intelligence and Statistics, Morgan Kaufmann (2001) 27–34
5. Ghahramani, Z., Beal, M.J.: Propagation algorithms for variational Bayesian learning. In Leen, T., Dietterich, T., Tresp, V., eds.: Advances in Neural Information Processing Systems 13. MIT Press, Cambridge, MA (2001) 507–513
6. Humphreys, K., Titterton, D.M.: Approximate Bayesian inference for simple mixtures. In Bethlehem, J.G., van der Heijden, P.G.M., eds.: COMPSTAT2000. Physica-Verlag, Heidelberg (2000) 331–336
7. Jaakkola, T.S., Jordan, M.I.: Bayesian logistic regression: a variational approach. *Statistics and Computing* **10** (2000) 25–37

8. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. In Jordan, M.I., ed.: *Learning in Graphical Models*. MIT Press, Cambridge (1999) 105–162
9. MacKay, D.J.C.: Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge (1997)
10. Penny, W.D., Roberts, S.J.: Variational Bayes for 1-dimensional mixture models. Technical Report PARG-2000-01, Oxford University (2000)
11. Peters, B.C., Walker, H.F.: An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35** (1978) 362–378
12. Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26** (1984) 195–239
13. Wang, B., Titterton, D.M.: Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In Chickering, M., Halpern, J., eds.: *Proceedings of the twentieth conference on Uncertainty in Artificial Intelligence*, Banff, Canada, AUAI Press (2004) 577–584
14. Walker, A.M.: On the asymptotic behaviour of posterior distributions. *J. R. Statist. Soc. B* **31** (1969) 80–88
15. Heyde, C.C., Johnstone, I.M.: On asymptotic posterior normality for stochastic processes. *J. R. Statist. Soc. B* **41** (1979) 184–189
16. Chen, C.F.: On asymptotic normality of limiting density functions with Bayesian implications. *J. R. Statist. Soc. B* **47** (1985) 540–546
17. Bernardo, J.M., Smith, A.F.M.: *Bayesian Theory*. John Wiley & Sons, Inc, New York (1994)
18. Ghosal, S., Ghosh, J.K., Samanta, T.: On convergence of posterior distributions. *The Annals of Statistics* **23** (1995) 2145–2152