Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values

Bo Wang and D. M. Titterington
Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK.

Abstract

We study the properties of variational Bayes approximations for exponential family models with missing values. It is shown that the iterative algorithm for obtaining the variational Bayesian estimator converges locally to the true value with probability 1 as the sample size becomes indefinitely large. Moreover, the variational posterior distribution is proved to be asymptotically normal.

Key words: Exponential family model, Variational Bayes, Local convergence, Asymptotic normality, Laplace approximation

1 Introduction

Variational Bayes approximations have recently been applied to complex models involving incomplete-data for which computational difficulties arise with the ideal Bayesian approach. Such models include hidden Markov models and mixture models; see for example [1, 2, 3, 6, 9, 10, 12, 13, 17]. In these earlier contributions, the approximations were shown empirically to be convergent and effective. However little has been done to investigate their theoretical properties.

Hall, Humphreys and Titterington [7] initiated a discussion of these issues and proved that, for certain Markov models, the parameter estimator obtained by maximising the variational lower bound function is asymptotically consistent provided the proportion of all values that are missing tends to zero. Later we proved in [16] that it is not always the case that a fully factorised form of variational posterior, which includes the factorisation of the joint probability function for the hidden states, provides an asymptotically

consistent estimator as the 'sample size' becomes large. We demonstrated this in particular in the context of linear state space models, in which the above sufficient condition obviously does not hold. On the other hand we showed in [15] that variational Bayes estimators for certain mixture models are asymptotically efficient for large sample sizes.

In this paper we study the properties of variational approximation algorithms for more general models, namely exponential family models with missing values. Exponential families include distributions such as Gaussian, gamma, Poisson, Dirichlet and Wishart, and exponential family models with missing values contain many models of practical interest as particular cases, such as Gaussian mixtures, hidden Markov models and linear state space models. Beal [3] and Ghahramani and Beal [6] applied the variational Bayesian method to these models and derived the iterative algorithm for learning the approximate posterior distributions of the latent states and the model parameters. The numerical expriments therein show empirically that this algorithm is convergent and efficient. In this paper we derive the iterative procedure for obtaining the variational Bayesian estimator, we provide analytical proofs of local convergence of the procedure as the sample size tends to infinity, and we show that the variational posterior distribution for the parameters is asymptotically normal.

2 Exponential family models with missing values and variational approximations

We consider the following exponential family models with missing values. Suppose that $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}^m\}$ is a family of probability distributions on a measurable space $(\Omega, \mathcal{F})$, and that x and y are sampled from the natural exponential family:

$$p(x, y|\theta) = f(x, y) \exp\{\theta^\top u(x, y) - \psi(\theta)\} \tag{1}$$

with x taking values in $\mathbb{R}^d$ and y in $\mathbb{R}^p$, where $\theta \in \mathbb{R}^m$ is the unknown parameter, and $\psi(\cdot) : \mathbb{R}^m \mapsto \mathbb{R}$ is continuously differentiable and strictly convex. The parameter θ has a conjugate prior to the complete-data likelihood (1), with density

$$p(\theta|\alpha_0, \beta_0) = h(\alpha_0, \beta_0) \exp\{\theta^\top \beta_0 - \alpha_0 \psi(\theta)\}, \tag{2}$$

where h is a normalising constant satisfying

$$h(\alpha, \beta)^{-1} = \int \exp\{\theta^\top \beta - \alpha\psi(\theta)\} d\theta, \tag{3}$$

and $\alpha_0 \in \mathbb{R}, \beta_0 \in \mathbb{R}^m$ are the hyperparameters of the prior.

Suppose that only y is observable whereas x is latent. We have a dataset consisting of a random sample of size n, with $Y = (y_1, y_2, \dots, y_n)$ and $X = (x_1, x_2, \dots, x_n)$. In the Bayesian framework we want to infer the posteriors over both the parameters and the hidden states. Unfortunately exact Bayesian inference is generally time-consuming, if not impossible, especially for large dimensionality m. Therefore approximation is usually necessary in these cases. In the variational approach, the true posterior $p(X, \theta|Y)$ is approximated by the variational distribution $q(X, \theta)$, which factorises as $q(X, \theta) = q_X(X)q_\theta(\theta)$, and is chosen to maximise the functional

$$\int q(X, \theta) \log \frac{p(\theta, X, Y)}{q(X, \theta)} d\theta dX, \tag{4}$$

equivalent to minimising the Kullback-Leibler divergence between the exact and approximate distributions of θ and X, given Y.

The functional (4) can be maximised using the following iterative procedure (see [3, 6]). In turn, the following two stages are performed.

(i) Optimise $q_\theta(\theta)$ for fixed $\{q_{x_i}(x_i), i = 1, \dots, n\}$. This step results in

$$q_\theta(\theta) = h(\alpha, \beta) \exp\{\theta^\top \beta - \alpha\psi(\theta)\}, \tag{5}$$

where α and β are the hyperparameters of the variational posterior and are updated by

$$\alpha = n + \alpha_0, \quad \beta = \sum_{i=1}^{n} r_i + \beta_0, \quad \text{and } r_i = \langle u(x_i, y_i) \rangle_{x_i}. \tag{6}$$

Here $\langle \cdot \rangle_{x_i}$ denotes the expectation under $q_{x_i}(x_i)$.

(ii) Optimise $q_X(X)$ for fixed $q_\theta(\theta)$. This leads to the factorised form $q_X(X) = \prod_{i=1}^{n} q_{x_i}(x_i)$, where

$$q_{x_i}(x_i) = f(x_i, y_i) g(\theta, y_i) \exp\{\langle \theta \rangle_\theta^\top u(x_i, y_i) - \psi(\langle \theta \rangle_\theta)\}, \tag{7}$$

in which $g(\theta, y_i)$ is a normalising constant satisfying

$$g(\theta, y_i)^{-1} = \int f(x_i, y_i) \exp\{\langle \theta \rangle_\theta^\top u(x_i, y_i) - \psi(\langle \theta \rangle_\theta)\} dx_i, \tag{8}$$

and $\langle \cdot \rangle_\theta$ denotes the expectation under $q_\theta(\theta)$.

3 The iterative algorithm and its convergence

We define the variational Bayesian estimator $\hat{\theta}$ of the parameter θ as

$$\hat{\theta} = \int \theta q_{pos}(\theta) d\theta,$$

where q_{pos} is the variational posterior density of θ, given by the limiting form of $q_\theta(\theta)$ that results from the above iterative procedure. For the exponential family distribution (5) the corresponding variational Bayesian estimator is

$$\hat{\theta} = \int \theta q_\theta(\theta) d\theta = -\frac{D_\beta h(\alpha, \beta)}{h(\alpha, \beta)}.$$

(Throughout the paper, $D\Psi$ and $D^2\Psi$ denote the gradient and the Hessian of Ψ. When ambiguity exists, the specific variable of differentiation appears as a subscript of the symbol D and D^2.)

Thus, the procedure in the previous section can be used to derive the following algorithm for obtaining the variational Bayesian estimate of θ: starting with some initial value $\theta^{(0)}$, successive iterates are defined inductively by

$$\theta^{(k+1)} \triangleq \Phi_n(\theta^{(k)}) = -\frac{D_\beta h(\alpha, \beta)}{h(\alpha, \beta)}, \tag{9}$$

where α and β are given as in (6), and

$$q_{x_i}(x_i) = f(x_i, y_i) g(\theta^{(k)}, y_i) \exp\{(\theta^{(k)})^\top u(x_i, y_i) - \psi(\theta^{(k)})\},$$
$$g(\theta^{(k)}, y_i)^{-1} = \int f(x_i, y_i) \exp\{(\theta^{(k)})^\top u(x_i, y_i) - \psi(\theta^{(k)})\} dx_i.$$

It is of interest to investigate the questions of whether or not the algorithm (9) is convergent and, if so, what properties are possessed by the limiting value. The following theorem gives a partial answer.

Theorem 1. *With probability 1 as n approaches infinity, the iterative procedure (9) converges locally to the true value θ^*, i.e. (9) converges to θ^* whenever the starting value is sufficiently near to θ^*.*

Proof. Denote by $D\Phi_n(\theta^*)$ the gradient of $\Phi_n(\theta)$ evaluated at θ^* and write β and r_i as $\beta(\theta)$ and $r_i(\theta)$ to indicate explicitly their dependence on θ. From (9) one has

$$D\Phi_n(\theta^*) = \frac{D_\beta h(\alpha, \beta) D_\beta^\top h(\alpha, \beta) - h(\alpha, \beta) D_\beta^2 h(\alpha, \beta)}{h^2(\alpha, \beta)} D\beta(\theta^*).$$

Here h and its derivatives are evaluated at θ^*.

From (6) we have that $D\beta(\theta) = \sum_{i=1}^n Dr_i(\theta)$ and

$$\begin{aligned}
Dr_i(\theta^*) &= \int u(x_i, y_i) D_\theta^\top q_{x_i}(x_i) dx_i \\
&= \int u(x_i, y_i) f(x_i, y_i) D_\theta^\top g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i \\
&\quad + \int u(x_i, y_i) f(x_i, y_i) g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} \\
&\qquad \cdot \big[u^\top(x_i, y_i) - D^\top \psi(\theta^*)\big] dx_i \\
&= \int \big[u(x_i, y_i) - D\psi(\theta^*)\big] f(x_i, y_i) D_\theta^\top g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i \\
&\quad + \int f(x_i, y_i) g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} \\
&\qquad \cdot \big[u(x_i, y_i) - D\psi(\theta^*)\big] \big[u^\top(x_i, y_i) - D^\top \psi(\theta^*)\big] dx_i,
\end{aligned}$$

where in the last equality we used the fact that

$$\begin{aligned}
&\int \big[u(x_i, y_i) - D\psi(\theta^*)\big] f(x_i, y_i) g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i \\
&\quad + \int f(x_i, y_i) D_\theta g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i = 0, \qquad (10)
\end{aligned}$$

which is obtained by differentiating, with respect to θ^*,

$$\int f(x_i, y_i) g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i = 1.$$

Since it follows from (8) that

$$\begin{aligned}
D_\theta g(\theta^*, y_i) = -&\int f(x_i, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} \big[u(x_i, y_i) - D\psi(\theta^*)\big] dx_i \\
&\cdot \Big\{ \int f(x_i, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i \Big\}^{-2},
\end{aligned}$$

equality (10) can be rewritten as

$$\begin{aligned}
&\int \big[u(x_i, y_i) - D\psi(\theta^*)\big] f(x_i, y_i) g(\theta^*, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i \\
&\qquad \cdot \int f(x_i, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} dx_i \\
&= \int f(x_i, y_i) \exp\{\theta^{*\top} u(x_i, y_i) - \psi(\theta^*)\} \big[u(x_i, y_i) - D\psi(\theta^*)\big] dx_i. \qquad (11)
\end{aligned}$$

Differentiating both sides of (11) with respect to θ^*, we have

$$\begin{aligned}
&\Big\{ \int \big[u(x_i,y_i) - D\psi(\theta^*)\big] f(x_i,y_i) D_\theta^\top g(\theta^*,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} dx_i \\
&\qquad -D^2\psi(\theta^*) + \int f(x_i,y_i) g(\theta^*,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} \\
&\qquad\qquad \cdot \big[u(x_i,y_i) - D\psi(\theta^*)\big]\big[u^\top(x_i,y_i) - D^\top\psi(\theta^*)\big] dx_i \Big\} \\
&\qquad\qquad \cdot \int f(x_i,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} dx_i \\
&\quad + \int \big[u(x_i,y_i) - D\psi(\theta^*)\big] f(x_i,y_i) g(\theta^*,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} dx_i \\
&\qquad \cdot \int f(x_i,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta)\} \big[u^\top(x_i,y_i) - D^\top\psi(\theta^*)\big] dx_i \\
&= \int f(x_i,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} \big[u(x_i,y_i) - D\psi(\theta^*)\big]\big[u^\top(x_i,y_i) - D^\top\psi(\theta^*)\big] dx_i \\
&\qquad\qquad - \int f(x_i,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} dx_i \cdot D^2\psi(\theta^*). \qquad (12)
\end{aligned}$$

Write $\phi = u(x_i,y_i) - D\psi(\theta^*)$. The marginal distribution of y_i is $\int p(x_i,y_i|\theta^*)dx_i$, and therefore it follows from the strong law of large numbers that, with probability 1,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \nabla_\theta r_i \\
&\to \int \Big\{ \nabla_\theta r_i \int p(x_i,y_i|\theta^*) dx_i \Big\} dy_i \\
&= -\int \Big\{ \mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top] \int f(x_i,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} dx_i \Big\} dy_i + D^2\psi(\theta^*) \\
&= D^2\psi(\theta^*) - \mathbb{E}_{y_i}\big\{ \mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top] \big\},
\end{aligned}$$

where we have used equality (12) and the fact that

$$\begin{aligned}
&\int f(x_i,y_i) \exp\{\theta^{*\top} u(x_i,y_i) - \psi(\theta^*)\} \\
&\qquad \cdot \big[u(x_i,y_i) - D\psi(\theta^*)\big]\big[u^\top(x_i,y_i) - D^\top\psi(\theta^*)\big] dx_i dy_i = D^2\psi(\theta^*), \qquad (13)
\end{aligned}$$

and $\mathbb{E}_{x_i}$ denotes expectation under q_{x_i}.

Thus, we obtain

$$\frac{1}{n} D\beta(\theta^*) \to D^2\psi(\theta^*) - \mathbb{E}_{y_i}\big\{ \mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top] \big\}, \ a.s., \qquad (14)$$

where 'a.s.' means 'almost surely'.

Similarly, one has

$$\frac{1}{n}\beta(\theta^*) \to D\psi(\theta^*),\ a.s. \tag{15}$$

For convenience we write $h(\alpha,\beta)^{-1}$ evaluated at θ^* as $\tilde{h}(\alpha,\beta)$, from which

$$D_\beta \tilde{h}(\alpha,\beta) = \int \exp\{\theta^\top \beta - \alpha\psi(\theta)\}\theta d\theta,$$

$$D^2_\beta \tilde{h}(\alpha,\beta) = \int \exp\{\theta^\top \beta - \alpha\psi(\theta)\}\theta\theta^\top d\theta.$$

Let $b(\cdot) : \mathbb{R}^m \mapsto \mathbb{R}$ be a four-times continuously differentiable function of θ and write

$$a_n(\theta) = (1+\frac{\alpha_0}{n})\psi(\theta) - \theta^\top(\frac{1}{n}\sum_{i=1}^n r_i + \frac{\beta_0}{n}), \tag{16}$$

$$h_b = \int b(\theta)\exp\{-na_n(\theta)\}d\theta. \tag{17}$$

Since $\psi(\theta)$ is continuously differentiable and strictly convex, it is obvious that $a_n(\theta)$ is also continuously differentiable and strictly convex in θ. Therefore $a_n(\theta)$ has a unique global minimiser, denoted by $\hat{\theta}_n$, which also satisfies the equation

$$D\psi(\theta) = (\frac{1}{n}\sum_{i=1}^n r_i + \frac{\beta_0}{n})/(1+\frac{\alpha_0}{n}). \tag{18}$$

Under these conditions, in the Appendix we show the validity of Laplace's method by verifying the assumptions of Kass, Tierney and Kadane [11]. Hence application of Laplace's approximation yields

$$\begin{aligned}\int b(\theta)\exp\{-na_n(\theta)\}d\theta &= (2\pi)^{m/2}[\det(nD^2a_n)]^{-1/2}\exp\{-na_n(\hat{\theta}_n)\}\\ &\cdot\Big\{b(\hat{\theta}_n) + \frac{1}{n}\Big[\frac{1}{2}\sum_{i,j=1}^m \sigma_n^{ij}b_{ij} - \frac{1}{6}\sum_{\substack{i,j=1\\k,s=1}}^m a_n^{ijk}b_s\mu^4_{ijks}\\ &\quad + \frac{1}{72}b(\hat{\theta}_n)\sum_{\substack{i,j,k=1\\q,r,s=1}}^m a_n^{ijk}h_n^{qrs}\mu^6_{ijkqrs}\\ &\quad - \frac{1}{24}b(\hat{\theta}_n)\sum_{\substack{i,j=1\\k,s=1}}^m a_n^{ijks}\mu^4_{ijks}\Big] + O(n^{-2})\Big\},\ a.s.\end{aligned} \tag{19}$$

where μ^4_{ijks} and μ^6_{ijkqrs} are the fourth and sixth central moments of a multivariate normal distribution having covariance matrix $(D^2 a_n)^{-1}$; that is,

$$\begin{aligned}
\mu^4_{ijks} =& \sigma^{ij}_n \sigma^{ks}_n + \sigma^{ik}_n \sigma^{js}_n + \sigma^{is}_n \sigma^{jk}_n, \\
\mu^6_{ijkqrs} =& \sigma^{ij}_n \sigma^{kq}_n \sigma^{rs}_n + \sigma^{ij}_n \sigma^{kr}_n \sigma^{qs}_n + \sigma^{ij}_n \sigma^{ks}_n \sigma^{qr}_n \\
&+ \sigma^{ik}_n \sigma^{jq}_n \sigma^{rs}_n + \sigma^{ik}_n \sigma^{jr}_n \sigma^{qs}_n + \sigma^{ik}_n \sigma^{js}_n \sigma^{qr}_n \\
&+ \sigma^{iq}_n \sigma^{jk}_n \sigma^{rs}_n + \sigma^{iq}_n \sigma^{jr}_n \sigma^{ks}_n + \sigma^{iq}_n \sigma^{js}_n \sigma^{kr}_n \\
&+ \sigma^{ir}_n \sigma^{jk}_n \sigma^{qs}_n + \sigma^{ir}_n \sigma^{jq}_n \sigma^{ks}_n + \sigma^{ir}_n \sigma^{js}_n \sigma^{kq}_n \\
&+ \sigma^{is}_n \sigma^{jk}_n \sigma^{qr}_n + \sigma^{is}_n \sigma^{jq}_n \sigma^{kr}_n + \sigma^{is}_n \sigma^{jr}_n \sigma^{kq}_n,
\end{aligned}$$

where $D^2 a_n$ denotes the Hessian of a_n, its (i,j)-component is written as a^{ij}_n and the components of its inverse are written as σ^{ij}_n; moreover, b_s and b_{ij} denote the components of the first- and second-order derivatives of b, respectively. All derivatives are evaluated at $\hat{\theta}_n$.

It is obvious that $D^2 a_n$ converges to $D^2\psi$ with probability 1 as $n \to \infty$. Letting $b(\theta)$ be 1, θ_i and $\theta_i\theta_j$ $(i,j = 1,\ldots,m)$ correspondingly in (19) and after a straightforward calculation, we obtain that, as n tends to infinity, with probability 1,

$$\frac{nD^{2,ij}_\beta \tilde{h}(\alpha,\beta)\tilde{h}(\alpha,\beta) - nD^i_\beta \tilde{h}(\alpha,\beta) D^j_\beta \tilde{h}(\alpha,\beta)}{\tilde{h}^2(\alpha,\beta)} \to \frac{1}{2}\sigma^{ij}_\infty = \frac{1}{2}[D^2\psi(\theta)]^{-1}_{ij}.$$

Therefore, combining (14) with the last limiting result we obtain that, with probability 1,

$$\begin{aligned}
D\Phi_n(\theta^*) \to& \frac{1}{2}[D^2\psi(\theta^*)]^{-1}\big[D^2\psi(\theta^*) - \mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top]\big\}\big] \\
=& \frac{1}{2}I_m - \frac{1}{2}[D^2\psi(\theta^*)]^{-1}\mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top]\big\},
\end{aligned}$$

where I_m denotes the $m \times m$ identity matrix.

Since ψ is continuously differentiable and strictly convex, $D^2\psi(\theta^*)$ is positive definite and symmetric. Obviously $\mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top]\big\}$ is positive semidefinite and symmetric. Hence, as n tends to infinity, $D\Phi_n(\theta^*) \le \frac{1}{2}I_m$; that is, $D\Phi_n(\theta^*) - \frac{1}{2}I_m$ is negative semidefinite.

Next we show that

$$[D^2\psi(\theta^*)]^{-1}\mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top]\big\} \le I_m.$$

Since $D^2\psi(\theta^*)$ is positive definite and symmetric, it is sufficient to prove that

$$\theta^\top \mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top]\big\}\theta \le \theta^\top D^2\psi(\theta^*)\theta, \tag{20}$$

for any $\theta \in \mathbb{R}^m$.

In fact, we have

$$\begin{aligned}
&\theta^\top \mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi]\mathbb{E}_{x_i}[\phi^\top]\big\}\theta \\
=&\mathbb{E}_{y_i}\Big\{\sum_{j,k=1}^{m}\theta_j\theta_k\mathbb{E}_{x_i}[\phi_j]\mathbb{E}_{x_i}[\phi_k]\Big\} \\
=&\mathbb{E}_{y_i}\Big\{\sum_{j,k=1}^{m}\theta_j\theta_k\mathbb{E}_{x_i}[\phi_j\phi_k] - \sum_{j,k=1}^{m}\theta_j\theta_k\mathbb{E}_{x_i}\big[(\phi_j-\mathbb{E}_{x_i}[\phi_j])(\phi_k-\mathbb{E}_{x_i}[\phi_k])\big]\Big\} \\
=&\mathbb{E}_{y_i}\Big\{\sum_{j,k=1}^{m}\theta_j\theta_k\mathbb{E}_{x_i}[\phi_j\phi_k] - \theta^\top\mathbb{E}_{x_i}\big[(\phi-\mathbb{E}_{x_i}[\phi])(\phi-\mathbb{E}_{x_i}[\phi])^\top\big]\theta\Big\} \\
\leq&\sum_{j,k=1}^{m}\theta_j\theta_k\mathbb{E}_{y_i}\big\{\mathbb{E}_{x_i}[\phi_j\phi_k]\big\} \\
=&\theta^\top D^2\psi(\theta^*)\theta,
\end{aligned}$$

where the last equality is a consequence of (13).

Therefore, we obtain

$$0 \leq D\Phi_n(\theta^*) \leq \frac{1}{2}I_m.$$

Moreover, if we use Laplace's approximation (19) it is easy to deduce that $\Phi_n(\theta^*) = -D_\beta h(\alpha,\beta)/h(\alpha,\beta) \to \theta^*$ with probability 1 as n tends to infinity.

Define the norm of $\theta \in \mathbb{R}^m$ as $\|\theta\| \triangleq (\theta^\top\theta)^{1/2}$ and the norm of the real $m \times m$ matrix A as $\|A\| \triangleq \sup_{\|\theta\|=1}\|A\theta\|$. Therefore, since the starting value is sufficiently near to θ^* we have

$$\begin{aligned}
\|\theta^{(k+1)}-\theta^*\| \leq&\|\Phi_n(\theta^{(k)})-\Phi_n(\theta^*)\| + \|\Phi_n(\theta^*)-\theta^*\| \\
\leq&\|\nabla\Phi_n(\theta^*)\|\cdot\|\theta^{(k)}-\theta^*\| + \|\Phi_n(\theta^*)-\theta^*\| \\
=&\sup_{\|\theta\|=1}|\theta^\top\nabla\Phi_n(\theta^*)\theta|\cdot\|\theta^{(k)}-\theta^*\| + \|\Phi_n(\theta^*)-\theta^*\| \\
\leq&\frac{1}{2}\|\theta^{(k)}-\theta^*\| + \|\Phi_n(\theta^*)-\theta^*\|,
\end{aligned}$$

and therefore the iterative procedure (9) converges locally to the true value θ^* with probability 1 as n approaches infinity . □

4 Asymptotic normality of the variational posterior distribution

There have been a large number of contributions about the asymptotic normality of posterior distributions including exponential families; see for instance Walker [14], Heyde and Johnstone [8], Chen [5] and Bernardo and Smith [4]. Under appropriate conditions the (true) posterior density converges in distribution to a normal density. In this section, we show that the variational posterior distribution for the parameter θ 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. We have proved that the algorithm (9) is convergent, so there exists an equilibrium point denoted by $\tilde{\theta}_n$. It follows from (5) and (7) that, at $\tilde{\theta}_n$,

$$\tilde{\alpha}_n = n + \alpha_0, \quad \tilde{\beta}_n = \sum_{i=1}^{n} r_i + \beta_0, \quad r_i = \langle u(x_i, y_i) \rangle_{x_i},$$
$$q(x_i) = f(x_i, y_i) g(\tilde{\theta}_n, y_i) \exp\{\tilde{\theta}_n^\top u(x_i, y_i) - \psi(\tilde{\theta}_n)\}.$$

Therefore, the variational posterior density of θ at the equilibrium point is

$$q_n(\theta) = h(\tilde{\alpha}_n, \tilde{\beta}_n) \exp\{\theta^\top \tilde{\beta}_n - \tilde{\alpha}_n \psi(\theta)\}. \tag{21}$$

Let $\hat{\theta}_n$ maximise $\theta^\top \tilde{\beta}_n - \tilde{\alpha}_n \psi(\theta)$. Then we have

$$D\psi(\hat{\theta}_n) = (\frac{1}{n} \sum_{i=1}^{n} r_i + \frac{\beta_0}{n}) / (1 + \frac{\alpha_0}{n}).$$

By the same arguments as used in the previous section and noting that $\tilde{\theta}_n \to \theta^*$ with probability 1 by Theorem 1, we have that $\frac{1}{n} \sum_{i=1}^{n} r_i$ converges to $D\psi(\theta^*)$ almost surely. Since $D\psi$ is strictly increasing and continuous, $\hat{\theta}_n \to \theta^*$ with probability 1 as n tends to infinity.

Define

$$L_n(\theta) \triangleq \log q_n(\theta) = \log h(\tilde{\alpha}_n, \tilde{\beta}_n) + \theta^\top \tilde{\beta}_n - \tilde{\alpha}_n \psi(\theta).$$

Then we have

$$\Sigma_n \triangleq -[D^2 L_n(\hat{\theta}_n)]^{-1} = [(n + \alpha_0) D^2 \psi(\hat{\theta}_n)]^{-1}.$$

Denote by $B(\theta, \varepsilon)$ the open ball of radius ε centred at θ. According to Chen [5], under the assumption of the consistency of $\hat{\theta}_n$ for θ^*, the posterior density q_n converges in distribution to $\mathcal{N}(\hat{\theta}_n, \Sigma_n)$ if the following basic conditions hold.

(C1) "Steepness". $\sigma_n^2 \to 0$ with P_{θ^*}-probability 1 as $n \to \infty$, where σ_n^2 is the largest eigenvalue of Σ_n.

(C2) "Smoothness". For any $\varepsilon > 0$, there exists an integer N and $\delta > 0$ such that, for any $n > N$ and $\theta \in B(\theta, \delta)$, $D^2 L_n(\theta)$ exists and satisfies

$$I_m - A(\varepsilon) \leq D^2 L_n(\theta)[D^2 L_n(\hat{\theta}_n)]^{-1} \leq I_m + A(\varepsilon), \ a.s.,$$

where $A(\varepsilon)$ is an $m \times m$ symmetric positive semidefinite matrix whose largest eigenvalue tends to zero with P_{θ^*}-probability 1 as $\varepsilon \to 0$.

(C3) "Concentration". For any $\delta > 0$, $\int_{B(\theta,\delta)} q_n(\theta) d\theta \to 1$ with P_{θ^*}-probability 1 as n tends to infinity.

In fact, since $\hat{\theta}_n \to \theta^*$, the components of $D^2\psi(\hat{\theta}_n)$ are bounded above and away from 0 almost surely if n is large enough, so the largest eigenvalue of Σ_n tends to 0.

(C2) is obvious because $D^2 L_n(\theta)[D^2 L_n(\hat{\theta}_n)]^{-1} = D^2\psi(\theta)[D^2\psi(\hat{\theta}_n)]^{-1}$ and $\psi(\cdot)$ is continuously differentiable.

From Kass, Tierney and Kadane [11], assumption (iii) in the Appendix is stronger than (C3). Therefore all the conditions are verified.

Acknowledgement. This work was supported by a grant from the UK Science and Engineering Research Council.

Appendix. We show that under our framework the Laplace approximation (19) is justified. The proof consists of verifying the *analytical assumptions for Laplace's method* in Kass, Tierney and Kadane [11], which are listed here for convenience. Since a_n defined in (16) is of random nature, some minor revisions are made to adapt our settings.

Suppose that $\{a_n : n = 1, 2, \dots\}$ is a sequence of six-times continuously differentiable real functions and that b is a four-times continuously differentiable function of θ. The pair $(\{a_n\}, b)$ is said to satisfy the *analytical assumptions for Laplace's method* if there exist positive numbers ε, M and η, and an integer n_0 such that $n > n_0$ implies the following:

(i) for all $\theta \in B(\hat{\theta}_n, \varepsilon)$ and all $1 \leq j_1, \dots, j_d \leq m$ with $0 \leq d \leq 6$, $|\partial_{j_1 \cdots j_d} a_n(\theta)| < M$ with P_{θ^*}-probability 1;

(ii) $\det(D^2 a_n(\hat{\theta}_n)) > \eta$ with P_{θ^*}-probability 1;

(iii) the integral h_b defined in equation (17) exists and is finite, and for all δ for which $0 < \delta < \varepsilon$,

$$\big[\det(nD^2 a_n(\hat{\theta}_n))\big]^{1/2} \int_{\theta \notin B(\hat{\theta}_n, \delta)} b(\theta) \exp\{-n(a_n(\hat{\theta}_n) - a_n(\theta))\} d\theta = O(n^{-2})$$

with P_{θ^*}-probability 1; or, more strongly,

(iii') for all δ for which $0 < \delta < \varepsilon$,

$$\limsup_{n\to\infty} \sup_{\theta} \{a_n(\hat{\theta}_n) - a_n(\theta) : \theta \notin B(\hat{\theta}_n, \delta)\} < 0$$

with P_{θ^*}-probability 1.

Under our assumptions, it has been shown in (15) that $\frac{1}{n}\sum_{i=1}^n r_i \to D\psi(\theta^*)$ with probability 1, so, when n large enough, $\frac{1}{n}\sum_{i=1}^n r_i$ is almost surely bounded in $B(\hat{\theta}_n, \varepsilon)$. Since ψ is continuously differentiable (i) obviously holds.

Condition (ii) is clear because ψ is strictly convex.

As n tends to infinity, for any $\theta \in \mathbb{R}^m$, $a_n(\theta)$ converges with P_{θ^*}-probability 1 to

$$a_0(\theta) = \psi(\theta) - \theta^\top D\psi(\theta^*).$$

Since $\hat{\theta}_n$ maximises a_n, we have

$$\hat{\theta}_n = (D\psi)^{-1}\big((\frac{1}{n}\sum_{i=1}^n r_i + \frac{\beta_0}{n})/(1 + \frac{\alpha_0}{n})\big),$$

so it follows that, as n tends to infinity, with probability 1,

$$\hat{\theta}_n \to (D\psi)^{-1}\big(D\psi(\theta^*)\big) = \theta^*.$$

Therefore, for all δ for which $0 < \delta < \varepsilon$ and $\theta \notin B(\hat{\theta}_n, \delta)$, we have that, $\forall \varepsilon_0$ satisfying $0 < \varepsilon_0 < \delta/2$, there exists an integer N such that, if $n > N$, it holds that, for all $\theta \in \mathbb{R}^m$,

$$|a_n(\theta) - a_0(\theta)| < \varepsilon_0,$$

$$\|\hat{\theta}_n - \theta^*\| < \varepsilon_0, \ a.s.$$

$$|a_0(\hat{\theta}_n) - a_0(\theta^*)| < \varepsilon_0, \ a.s.$$

Thus,

$$\begin{aligned} a_n(\hat{\theta}_n) - a_n(\theta) &= a_n(\hat{\theta}_n) - a_0(\hat{\theta}_n) + a_0(\hat{\theta}_n) - a_0(\theta^*) \\ &\quad + a_0(\theta^*) - a_0(\theta) + a_0(\theta) - a_n(\theta) \\ &< a_0(\theta^*) - a_0(\theta) + 3\varepsilon_0, \ a.s. \end{aligned}$$

so that

$$\begin{aligned}
&\sup\{a_n(\hat{\theta}_n) - a_n(\theta) : \theta \notin B(\hat{\theta}_n, \delta)\} \\
\leq &\sup\{a_0(\theta^*) - a_0(\theta) : \theta \notin B(\hat{\theta}_n, \delta)\} + 3\varepsilon_0 \\
\leq &\sup\{a_0(\theta^*) - a_0(\theta) : \theta \notin B(\theta^*, \delta - \varepsilon_0)\} + 3\varepsilon_0, \ a.s. \qquad (22)
\end{aligned}$$

since $B(\theta^*, \delta - \varepsilon_0) \subset B(\hat{\theta}_n, \delta)$.

Since $a_0(\cdot)$ is strictly convex, for $\theta \notin B(\theta^*, \delta - \varepsilon_0)$, we have $a_0(\theta) - a_0(\theta^*) > c$, where $c = \inf\{a_0(\theta) - a_0(\theta^*) : \theta$ lies in the boundary of $B(\theta^*, \delta/2)\} > 0$. Consequently, we get

$$\sup\{a_0(\theta^*) - a_0(\theta) : \theta \notin B(\theta^*, \delta - \varepsilon_0)\} \leq -c.$$

Combining the last estimate with (22) we have that, $\forall \varepsilon_0$ satisfying $0 < \varepsilon_0 < \delta$, there exists an integer N such that $n > N$ implies

$$\sup\{a_n(\hat{\theta}_n) - a_n(\theta) : \theta \notin B(\hat{\theta}_n, \delta)\} \leq -c + 3\varepsilon_0, \ a.s.;$$

that is, (iii') holds.

References

[1] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, pages 21–30, 1999.

[2] H. Attias. A variational Bayesian framework for graphical models. In S. Solla, T. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, Cambridge, MA, 2000.

[3] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.

[4] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc, New York, 1994.

[5] C.-F. Chen. On asymptotic normality of limiting density functions with Bayesian implications. *J. R. Statist. Soc. B*, 47:540–546, 1985.

[6] Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing systems 12*, pages 507–513. MIT Press, Cambridge, MA, 2000.

[7] P. Hall, K. Humphreys, and D. M. Titterington. On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *Journal of the Royal Statistical Society Series B*, 64:549–564, 2002.

[8] C. C. Heyde and I. M. Johnstone. On asymptotic posterior normality for stochastic processes. *J. R. Statist. Soc. B*, 41:184–189, 1979.

[9] K. Humphreys and D. M. Titterington. Approximate Bayesian inference for simple mixtures. In J. G. Bethlehem and P. G. M. van der Heijden, editors, *COMPSTAT2000*, pages 331–336. Physica-Verlag, Heidelberg, 2000.

[10] K. Humphreys and D. M. Titterington. Some examples of recursive variational approximations for Bayesian inference. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods: Theory and Practice*, pages 179–195. MIT Press, 2001.

[11] R. E. Kass, L. Tierney, and J. B. Kadane. The validity of posterior expansions based on Laplace's method. In S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, editors, *Bayesian and Likelihood Methods in Statistics and Econometrics*, pages 473–488. Elsevier Science Publishers, North-Holland, 1990.

[12] D. J. C. MacKay. Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge, 1997.

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

[14] A. M. Walker. On the asymptotic behaviour of posterior distributions. *J. R. Statist. Soc. B*, 31:80–88, 1969.

[15] B. Wang and D. M. Titterington. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Preprint, 2003.

[16] B. Wang and D. M. Titterington. Lack of consistency of mean field and variational Bayes approximations for state space models. Technical Report 03-5, University of Glasgow, 2003.

[17] B. Wang and D. M. Titterington. Variational Bayesian inference for partially observed diffusions. Preprint, 2003.