Bayesian Analysis (2006) **1**, Number 3, pp. 625–650

Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model

Bo Wang*, and D. M. Titterington†

Abstract. In this paper we propose a generalised iterative algorithm for calculating variational Bayesian estimates for a normal mixture model and investigate its convergence properties. It is shown theoretically that the variational Bayesian estimator converges locally to the maximum likelihood estimator at the rate of $O(1/n)$ in the large sample limit.

Keywords: Mixture model, Variational Bayes, Local convergence, Laplace approximation

1 Introduction

A full Bayesian analysis of data involving missing values or based on latent structure models is almost always non-trivial; tractable closed-form expressions for Bayesian posterior or predictive distributions are rarely available. Computational tools such as Markov chain Monte Carlo methods are well established, but even in simple problems such as the analysis of mixture data, these methods are not totally straightforward (Celeux et al. (2000)). In addition, the implementation of MCMC may be impractical - because of computational explosion or analytical intractability, for instance - if the structure of the incomplete component in the data involves high dimensionality or non-trivial dependence. In addition, one has to deal with issues such as convergence and storage of the MCMC realisations.

In the face of these difficulties, deterministic variational Bayesian approximations have recently been introduced in the machine learning community (for instance by MacKay (1997) and Attias (1999, 2000)) and are widely recognised to be effective and promising in a variety of contexts, such as hidden Markov models (MacKay (1997)), graphical models (Attias (1999, 2000)), mixture models (Humphreys and Titterington (2000); Penny and Roberts (2000); Corduneanu and Bishop (2001); Ueda and Ghahramani (2003)), mixtures of factor analysers (Ghahramani and Beal (2000)) and state space models (Ghahramani and Beal (2001); Beal (2003)). Titterington (2004) gives a more extensive review and Jordan (2004) provides an overview of a general formulation of the approach in terms of convex analysis.

*School of Mathematics and Statistics, University of Newcastle upon Tyne, Newcastle, UK, mailto:b.wang@ncl.ac.uk

†Department of Statistics, University of Glasgow, Glasgow, UK, mailto:mike@stats.gla.ac.uk

Let Y denote observed data, let S denote missing data, with the letter 'S' chosen to fit in with the mixture context to be discussed in detail later, let p and Θ generically denote probability density and parameters, and let $p(S, \Theta|Y)$ denote the posterior density of (S, Θ), given Y. The variational Bayesian approximation, $q(S, \Theta)$, for $p(S, \Theta|Y)$, is defined as the minimiser of the Kullback-Leibler divergence between q and p,

$$\int q(S, \Theta) \log \frac{q(S, \Theta)}{p(S, \Theta|Y)} dS d\Theta, \tag{1}$$

with q restricted to have a special structure, usually corresponding to independence between Θ and S. If (as in the mixture problem) S is discrete, then the integral is interpreted as a summation. The variational Bayesian estimator for Θ is defined to be the mean of the corresponding approximating distribution. The minimisation of the Kullback-Leibler divergence (1) is equivalent to maximising the so-called negative free energy,

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

Clearly, $q(S, \Theta)$ depends on Y, but for simplicity we do not indicate this explicitly in the notation.

Empirically, variational Bayesian approximations have often been shown to perform well in earlier contributions, but the convergence behaviour of the algorithm has not been examined in detail, nor have the asymptotic properties of the variational Bayesian estimator of Θ been established; formal theoretical analysis of the quality of the method needs to be studied.

Hall et al. (2002) considered a likelihood-based version of the problem in which, for fixed Θ, a variational approximation $q(S)$ for $p(S|Y, \Theta)$ is chosen to maximise

$$\int q(S) \log \frac{p(S, Y|\Theta)}{q(S)} dS = F(q, \Theta). \tag{2}$$

Formula (2), with the maximising $q(S)$ substituted, provides a lower bound for the observed-data loglikelihood, evaluated at Θ. Hall et al. (2002) proved that, for certain Markov models, the parameter estimator obtained by maximising the resulting lower bound function is asymptotically consistent provided the proportion of all values that are missing tends to zero. However, their analysis is likelihood-based rather than Bayesian and, in any case, this sufficient condition is not satisfied in the case of many problems, such as state space models and mixture models.

In Wang and Titterington (2004) we investigated the consistency properties of both so-called mean field and variational Bayesian estimators in the context of linear state space models, in which the above sufficient condition obviously does not hold. The mean field estimators are obtained as follows: we assume for $q(S)$ a factorised form, with a factor for each of the individual missing values, which in this case are the state

variables; each factor involves variational parameters; alternate maximisation of $F(q, \Theta)$ is carried out with respect to the variational parameters and Θ; and the mean field estimators of Θ are the values of Θ to which this algorithm converges. The nature of the algorithm for obtaining the variational Bayesian estimators is similar but incorporates priors; details are given in the next section. We proved that the mean field approximation is asymptotically consistent when the variances of the noise variables in the system are sufficiently small, but neither the mean field estimator nor the variational Bayes estimator is always asymptotically consistent as the 'sample size' becomes large - essentially because of the unrealistic nature of the independence assumption underlying the variational approximation to the distribution of the missing states. We subsequently studied the consistency properties of variational Bayesian estimators for mixture models involving known component densities in Wang and Titterington (2003). It was shown in Wang and Titterington (2003) that, with probability 1 as the sample size increases indefinitely, the iterative algorithm for the variational Bayes approximation converges locally to the maximum likelihood estimator, in the context of that very special model.

In this paper we investigate a more general mixture model, and we consider a more general iterative algorithm. So far as the parameters Θ are concerned, we shall see in Section 2 that the iterative algorithm that leads to the variational Bayes estimators takes the form

$$\Theta^{(k)} = T(\Theta^{(k-1)}), \tag{3}$$

for $k = 1, \ldots,$ where T denotes a certain mapping and $\{\Theta^{(k)}\}$ denotes the sequence of iterates that are produced. Instead, we investigate algorithms of the form

$$\Theta^{(k)} = (1-\varepsilon)\Theta^{(k-1)} + \varepsilon T(\Theta^{(k-1)}) \stackrel{\triangle}{=} \Phi_n^{\varepsilon}(\Theta^{(k-1)}), \tag{4}$$

for $k = 1, \ldots$ and some $\varepsilon > 0$. Obviously, when $\varepsilon = 1$ algorithm (4) becomes algorithm (3). For mixture models, iterative procedures, such as the EM algorithm, for obtaining maximum likelihood estimates of the parameters, have been widely investigated; see, for example, Peters and Walker (1978), McLachlan and Peel (2000) and references therein. Salakhutdinov and Roweis (2003) studied a class of overrelaxed bound optimisation algorithms, which are generalisations of the EM algorithm, and provided theoretical analysis of the convergence properties. As we shall point out, algorithm (3) is an analogue of the EM algorithm and algorithm (4) is an analogue of an adaptation of the EM algorithm proposed by Peters and Walker (1978). Motivated by the earlier work on the EM algorithm, we investigate the version of (4) for calculating approximate Bayesian estimates, and we prove that the variational Bayesian estimator, for the parameters of mixture models of normal densities, converges locally to the maximum likelihood estimator at the rate of $O(1/n)$ in the large sample limit.

2 The mixture model and the variational approximation

We consider a mixture of m d-dimensional multivariate normal densities $p_1, \ldots, p_m$ with mean vectors $\mu_1, \ldots, \mu_m$ and precision (inverse covariance) matrices $\Gamma_1, \ldots, \Gamma_m$,

respectively. Thus the density of an observation is given by

$$p(y_i) = \sum_{s=1}^{m} p_s(y_i) p(s_i = s), \tag{5}$$

where $y_i \in \mathbb{R}^d$ denotes the ith observed data vector, and s_i indicates the hidden component that generated it. The components are labelled by $s = 1, \ldots, m$, and the component s has mixing coefficient $\pi_s = p(s_i = s)$ for any i. We write the parameters collectively as

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_m \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_m \end{pmatrix}, \quad \Theta = \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\mu} \\ \boldsymbol{\Gamma} \end{pmatrix}.$$

For each s, $1 \leq s \leq m$; π_s, μ_s and Γ_s are elements of $\mathbb{R}$, $\mathbb{R}^d$ and the set of all real, symmetric $d \times d$ matrices, respectively. We denote by $\mathcal{A}$, $\mathcal{M}$ and $\mathcal{T}$ the respective m-fold direct sums of these sets with themselves. Then $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$ and Θ are elements of $\mathcal{A}$, $\mathcal{M}$, $\mathcal{T}$ and their direct sum $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$, respectively. In fact, $\mathcal{A}$ and $\mathcal{T}$ are generously large in that the elements of $\boldsymbol{\pi}$ are probabilities and the precision matrices should be positive definite, but this does not cause a problem here - and in any case, the natures of the priors defined next automatically impose appropriate restrictions.

We use priors on the parameters Θ that would be conjugate were the data complete (i.e. were S known). The mixing coefficients $\boldsymbol{\pi}$ follow a symmetric Dirichlet distribution $\mathcal{D}(\lambda^0)$. The precisions are independently Wishart, with $\Gamma_s \sim \mathcal{W}(\nu^0, \Phi^0)$. The means, conditioned on the precisions, are independently normal, with $\mu_s | \Gamma_s \sim \mathcal{N}(\rho^0, \beta^0 \Gamma_s)$, where $\beta^0 \Gamma_s$ is the inverse covariance matrix of the normal distribution.

Suppose that we have (complete) data consisting of a random sample of size n, so that $Y = (y_1, \ldots, y_n)'$ and $S = (s_1, \ldots, s_n)'$, then the joint density of S, Θ and Y is

$$p(S, \Theta, Y) = p(\boldsymbol{\pi}) \prod_{s=1}^{m} p(\mu_s | \Gamma_s) p(\Gamma_s) \prod_{i=1}^{n} \pi_{s_i} p_{s_i}(y_i).$$

In the variational Bayesian 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 such that the factors are chosen to maximise the negative free energy

$$\int \sum_{\{S\}} q(S, \Theta) \log \frac{p(S, \Theta, Y)}{q(S, \Theta)} d\Theta. \tag{6}$$

As a result of the form of $p(S, \Theta, Y)$, it follows immediately, by a variational argument, that the optimal $q^{(S)}(S)$ and $q^{(\Theta)}(\Theta)$ must factorise as

$$q^{(S)}(S) = \prod_{i=1}^{n} q_i^{(S)}(s_i) \quad \text{and} \quad q^{(\Theta)}(\Theta) = q(\boldsymbol{\pi}) \prod_{s=1}^{m} q(\mu_s|\Gamma_s) q(\Gamma_s).$$

It also follows that, since conjugate priors are used, the factors of the variational posterior $q^{(\Theta)}(\Theta)$ are functionally identical to the priors, but with different hyperparameter values: the mixing coefficients $\boldsymbol{\pi}$ are jointly Dirichlet, with $q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi} : \lambda_1, \ldots, \lambda_m)$, say; the precisions are independently Wishart, with $q(\Gamma_s) = \mathcal{W}(\Gamma_s : \nu_s, \Phi_s)$, say; and the means conditioned on the precisions are independently normal, with $q(\mu_s|\Gamma_s) = \mathcal{N}(\mu_s : \rho_s, \beta_s\Gamma_s)$, say. Here $\mathcal{D}(\boldsymbol{\pi} : \lambda_1, \ldots, \lambda_m)$, $\mathcal{W}(\Gamma_s : \nu_s, \Phi_s)$ and $\mathcal{N}(\mu_s : \rho_s, \beta_s\Gamma_s)$ denote the relevant density functions. Note that this reveals the key simplification created by the variational approximation; the variational approximation to the posterior distribution of the parameters is a single member of the corresponding conjugate family, whereas the true posterior, based on the observed, mixture data, is a complicated mixture of a large number of such conjugate distributions.

As in Attias (1999, 2000), Humphreys and Titterington (2000), Penny and Roberts (2000), Corduneanu and Bishop (2001) and Ueda and Ghahramani (2003), the appropriate values of the hyperparameters and of $\{q_i^{(S)}(s_i), i = 1, \ldots, n\}$ are obtained by an iterative procedure. Suppose that, as we begin the kth iteration, we have current hyperparameters and values of $q^{(S)}(S)$ superscripted by $(k-1)$, for $k = 1, \ldots$. Then we perform the following two steps.

Step 1. *Optimise the hyperparameters in* $q^{(\Theta)}(\Theta)$ *for fixed* $\{q_i^{(S)(k-1)}(s_i), i = 1, \ldots, n\}$.

This gives

$$\begin{aligned}
\lambda_s^{(k)} &= \sum_{i=1}^{n} r_{is}^{(k-1)} + \lambda^0, \quad \rho_s^{(k)} = \Big(\sum_{i=1}^{n} r_{is}^{(k-1)} y_i + \beta^0\rho^0\Big) \Big/ \Big(\sum_{i=1}^{n} r_{is}^{(k-1)} + \beta^0\Big), \quad (7)\\
\beta_s^{(k)} &= \sum_{i=1}^{n} r_{is}^{(k-1)} + \beta^0, \quad \nu_s^{(k)} = \sum_{i=1}^{n} r_{is}^{(k-1)} + \nu^0,\\
\Phi_s^{(k)} &= \sum_{i=1}^{n} r_{is}^{(k-1)} (y_i - \bar{\mu}_s^{(k-1)})(y_i - \bar{\mu}_s^{(k-1)})'\\
&\quad + \Big[(\sum_{i=1}^{n} r_{is}^{(k-1)}) \beta^0 (\bar{\mu}_s^{(k-1)} - \rho^0)(\bar{\mu}_s^{(k-1)} - \rho^0)'\Big] \Big/ \Big(\sum_{i=1}^{n} r_{is}^{(k-1)} + \beta^0\Big) + \Phi^0,
\end{aligned}$$

where

$$r_{is}^{(k-1)} = q_i^{(S)(k-1)}(s_i = s), \quad \bar{\mu}_s^{(k-1)} = \Big(\sum_{i=1}^{n} r_{is}^{(k-1)} y_i\Big) \Big/ \Big(\sum_{i=1}^{n} r_{is}^{(k-1)}\Big).$$

Step 2. *Optimise* $\{q_i^{(S)}(s_i), s_i = 1, \ldots, m, i = 1, \ldots, n\}$ *for fixed* $q^{(\Theta)}(\Theta) = q^{(\Theta)(k)}(\Theta)$, *corresponding to the hyperparameters as calculated in Step* 1.

For $s = 1, \ldots, m$ and $i = 1, \ldots, n$, this results in

$$r_{is}^{(k)} = q_i^{(S)(k)}(s_i = s) \propto \tilde{\pi}_s^{(k)} \tilde{\Gamma}_s^{(k)1/2} e^{-(y_i - \rho_s^{(k)})' \bar{\Gamma}_s^{(k)} (y_i - \rho_s^{(k)})/2 - d/(2\beta_s^{(k)})} \triangleq \gamma_{is}^{(k)},$$

where

$$\tilde{\pi}_s^{(k)} = \exp\{\int q^{(k)}(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi}\}, \tag{8}$$

$$\tilde{\Gamma}_s^{(k)} = \exp\{\int q^{(k)}(\Gamma_s) \log |\Gamma_s| d\Gamma_s\}, \tag{9}$$

$$\bar{\Gamma}_s^{(k)} = \nu_s^{(k)} (\Phi_s^{(k)})^{-1}, \tag{10}$$

and

$$q^{(k)}(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi} : \lambda_1^{(k)}, \ldots, \lambda_m^{(k)}),$$
$$q^{(k)}(\Gamma_s) = \mathcal{W}(\Gamma_s : \nu_s^{(k)}, \Phi_s^{(k)}),$$
$$q^{(k)}(\mu_s | \Gamma_s) = \mathcal{N}(\mu_s : \rho_s^{(k)}, \beta_s^{(k)} \Gamma_s).$$

If we let $\gamma_i^{(k)} = \sum_{s=1}^m \gamma_{is}^{(k)}$, $i = 1, \ldots, n$, then $r_{is}^{(k)} = \gamma_{is}^{(k)} / \gamma_i^{(k)}$.

This iterative procedure can be initialised at $k = 0$ by assigning the observations in some way (either at random or with the help of a clustering algorithm) to the m components, estimating the mixing weights and component distributions based on this assignment, thereby providing point estimates $\Theta^{(0)}$ for Θ, and taking, for each i and s, $r_{is}^{(0)}$ equal to the predictive probability $p(s_i = s | y_i, \Theta^{(0)})$.

Based on a quadratic loss function, the Bayesian estimator of a parameter is the posterior mean; we therefore define the variational Bayesian estimators of parameters as the means of the variational approximation to the posterior distribution. Therefore, at stage k of the iteration the corresponding approximations to the variational Bayesian estimates are given by

$$\pi_s^{(k)} = \Big(\sum_{i=1}^n r_{is}^{(k-1)} + \lambda^0 \Big) / (n + m\lambda^0), \tag{11}$$

$$\mu_s^{(k)} = \Big(\sum_{i=1}^n r_{is}^{(k-1)} y_i + \beta^0 \rho^0 \Big) \Big/ \Big(\sum_{i=1}^n r_{is}^{(k-1)} + \beta^0 \Big), \tag{12}$$

$$\Gamma_s^{(k)} = \Big(\sum_{i=1}^n r_{is}^{(k-1)} + \nu^0 \Big) \Big\{ \sum_{i=1}^n r_{is}^{(k-1)} (y_i - \mu_s^{(k)})(y_i - \mu_s^{(k)})' + \Big[(\sum_{i=1}^n r_{is}^{(k-1)}) \beta^0 (\mu_s^{(k)} - \rho^0)(\mu_s^{(k)} - \rho^0)' \Big] \Big/ \Big(\sum_{i=1}^n r_{is}^{(k-1)} + \beta^0 \Big) + \Phi^0 \Big\}^{-1}. \tag{13}$$

From this it follows that the hyperparameters in the variational posterior distributions can be expressed in terms of these estimates as

$$\lambda_s^{(k)} = n\pi_s^{(k)} + \lambda^0, \quad \rho_s^{(k)} = (n\mu_s^{(k)}\pi_s^{(k)} + \beta^0\rho^0)/(n\pi_s^{(k)} + \beta^0),$$
$$\beta_s^{(k)} = n\pi_s^{(k)} + \beta^0, \quad \nu_s^{(k)} = n\pi_s^{(k)} + \nu^0,$$
$$\Phi_s^{(k)} = n\pi_s^{(k)}(\Gamma_s^{(k)})^{-1} + n\pi_s^{(k)}\beta^0(\mu_s^{(k)} - \rho^0)(\mu_s^{(k)} - \rho^0)'(n\pi_s^{(k)} + \beta^0)^{-1} + \Phi^0.$$

It is clear from the version of Step 2 corresponding to stage $(k-1)$ of the iteration that, for $k = 2, \ldots$, the $\{r_{is}^{(k-1)}\}$ are functions of the elements of $\Theta^{(k-1)}$; therefore, equations (11)-(13) encapsulate the mapping T that corresponds to iteration (3) and contributes to iteration (4).

3 Convergence of the generalised iterative algorithm for calculating variational Bayesian estimates

Our theoretical analysis will be somewhat simpler if we deal with a slight modification of the iterations (11)-(13) in the previous section, corresponding to omission of the hyperparameters associated with the priors. The theoretical results that we obtain for the modified iteration will apply also to the original version because, asymptotically, the choice of prior hyperparameters will have negligible effect on posterior distributions, whether they be exact or variational approximations; information from the sample will dominate the prior. Therefore, instead of (11)-(13) we analyse the iteration defined by

$$\pi_s^{(k)} = \frac{1}{n}\sum_{i=1}^{n} r_{is}^{(k-1)} \triangleq \Pi_s(\Theta^{(k-1)}), \tag{14}$$

$$\mu_s^{(k)} = \Big(\sum_{i=1}^{n} r_{is}^{(k-1)} y_i\Big) \Big/ \Big(\sum_{i=1}^{n} r_{is}^{(k-1)}\Big) \triangleq M_s(\Theta^{(k-1)}), \tag{15}$$

$$\Gamma_s^{(k)} = \Big(\sum_{i=1}^{n} r_{is}^{(k-1)}\Big)\Big(\sum_{i=1}^{n} r_{is}^{(k-1)}(y_i - \mu_s^{(k-1)})(y_i - \mu_s^{(k-1)})'\Big)^{-1} \triangleq S_s(\Theta^{(k)}), \tag{16}$$

where $r_{is}^{(k-1)}$ is as defined in Step 2 in Section 2, but for $k-1$ instead of k.

Let

$$\Pi(\Theta) = \begin{pmatrix} \Pi_1(\Theta) \\ \vdots \\ \Pi_m(\Theta) \end{pmatrix}, \quad M(\Theta) = \begin{pmatrix} M_1(\Theta) \\ \vdots \\ M_m(\Theta) \end{pmatrix}, \quad S(\Theta) = \begin{pmatrix} S_1(\Theta) \\ \vdots \\ S_m(\Theta) \end{pmatrix}.$$

Then Π, M and S are operators from $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$ to itself, and the iterative procedure (14)-(16) can be rewritten in the form of (3) as

$$\Theta^{(k)} = T(\Theta^{(k-1)}) = \begin{pmatrix} \Pi(\Theta^{(k-1)}) \\ M(\Theta^{(k-1)}) \\ S(\Theta^{(k-1)}) \end{pmatrix}. \tag{17}$$

Similarly, the iterative stage in the generalised algorithm corresponding to (4) is

$$\Theta^{(k)} = (1-\varepsilon)\Theta^{(k-1)} + \varepsilon \begin{pmatrix} \Pi(\Theta^{(k-1)}) \\ M(\Theta^{(k-1)}) \\ S(\Theta^{(k-1)}) \end{pmatrix} \triangleq \Phi_n^{\varepsilon}(\Theta^{(k-1)}), \tag{18}$$

for $k = 0, 1, \ldots$ and some $\varepsilon > 0$. As remarked in Section 1, when $\varepsilon = 1$ algorithm (18) becomes (17).

Suppose that the true value of the parameter Θ is Θ^*; we may then establish the following theorem.

Theorem 1 *With probability 1 as n approaches infinity, the iterative procedure (18) converges locally to the true value Θ^* whenever $0 < \varepsilon < 2$; that is, the iterative procedure (18) converges to the true value Θ^* whenever $0 < \varepsilon < 2$ and the starting values are sufficiently near to Θ^*.*

For mixture models with unknown parameters in the components, the negative free energy (6) may be multimodal (see for example Duda and Hart (1973)), so that the variational Bayes algorithm may converge to different local maxima if different starting values (or hyperparameters) are chosen. Therefore, only the local convergence property is proved here. The existence of similar multimodality is well known in maximum likelihood estimation of mixture parameters.

The details of the proof of Theorem 1 are provided in the Appendix; only a skeletal account is given here. Although the details are complicated, the strategy is a familiar one for proving convergence of iterative algorithms. The key is to show that the mapping $\Phi_n^{\varepsilon}(\Theta)$ is locally contractive at Θ^*. For a deterministic mapping, Ostrowski's Theorem provides a sufficient condition for this, namely that the matrix Fréchet derivative of the mapping should have a norm that is less than 1. Here we show that, if $\nabla\Phi_n^{\varepsilon}(\Theta^*)$ denotes the Fréchet derivative of $\Phi_n^{\varepsilon}(\Theta)$ evaluated at Θ^*, then, with probability 1, $\nabla\Phi_n^{\varepsilon}(\Theta^*)$ converges to an operator of which the sup-norm is less than 1. The main part of the proof is given in Appendix B, which follows a brief Appendix A in which appropriate norms are defined. Appendix B provides the detailed calculation of the Fréchet derivatives of Π, M and S. It then derives the almost sure limits of these derivatives, which requires the calculation of the limits of r_{is} and their derivatives. To do this, we first give two necessary lemmas in Appendix C, we then derive the limits of the Fréchet derivatives of the functions corresponding to the right-hand sides of (8)-(10) in Appendix D, and we further study the limits of the variational probabilities r_{is} of the labels and their derivatives in Appendix E. Once this has been achieved, the rest of Appendix B can go on to show that the operator corresponding to the limits of the Fréchet derivatives, evaluated at the true Θ^*, has a norm that is less than 1. This requires the final appendix, Appendix F, which contains an argument similar to that used by Peters and Walker (1978) in their maximum-likelihood work.

4 The convergence rate of the variational Bayesian estimator

It is known that in general the (non-variational) Bayesian estimator and the MLE approach each other at rate $O(1/n)$. In this section we estimate the rate at which the variational Bayesian estimator converges to the maximum likelihood estimator (MLE). Suppose that the sample size n is large, and let $\tilde{\Theta}^n$ be the strongly consistent MLE of the parameter Θ; that is, it is the solution of the following likelihood equations (see, for example, Redner and Walker (1984)). For $s = 1, \ldots, m$,

$$\begin{aligned}
L_s^n(\Theta) &\triangleq \pi_s - \frac{1}{n}\sum_{i=1}^{n}\frac{p_s(y_i)\pi_s}{p(y_i)} = 0,\\
\bar{L}_s^n(\Theta) &\triangleq \mu_s - \Big\{\frac{1}{n}\sum_{i=1}^{n} y_i\frac{p_s(y_i)}{p(y_i)}\Big\}\Big/\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{p_s(y_i)}{p(y_i)}\Big\} = 0,\\
\tilde{L}_s^n(\Theta) &\triangleq \Gamma_s - \Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{p_s(y_i)}{p(y_i)}\Big\}\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{p_s(y_i)}{p(y_i)}(y_i-\mu_s)(y_i-\mu_s)'\Big\}^{-1} = 0.
\end{aligned}$$

Denote by $\hat{\Theta}^n$ the variational Bayes estimator, which is the stationary point of iteration (18) in the neighbourhood of the true value; that is, $\hat{\Theta}^n$ satisfies

$$\hat{\Theta}^n - \begin{pmatrix} \Pi(\hat{\Theta}^n) \\ M(\hat{\Theta}^n) \\ S(\hat{\Theta}^n) \end{pmatrix} = 0, \tag{19}$$

and the hyperparameters in the variational posterior distributions $q(\boldsymbol{\pi})$, $q(\Gamma_s)$ and $q(\mu_s|\Gamma_s)$ are correspondingly given by

$$\begin{gathered}
\hat{\lambda}_s^n = n\hat{\pi}_s^n + \lambda^0, \quad \hat{\rho}_s^n = (n\hat{\mu}_s^n\hat{\pi}_s^n + \beta^0\rho^0)/(n\hat{\pi}_s^n + \beta^0),\\
\hat{\beta}_s^n = n\hat{\pi}_s^n + \beta^0, \quad \hat{\nu}_s^n = n\hat{\pi}_s^n + \nu^0,\\
\hat{\Phi}_s^n = n\hat{\pi}_s^n(\hat{\Gamma}_s^n)^{-1} + n\hat{\pi}_s^n\beta^0(\hat{\mu}_s^n - \rho^0)(\hat{\mu}_s^n - \rho^0)'(n\hat{\pi}_s^n + \beta^0)^{-1} + \Phi^0.
\end{gathered}$$

Denote by $\hat{\gamma}_{is}^n$ and $\hat{\gamma}_i^n$ the converged values of $\gamma_{is}^{(k)}$, $\gamma_i^{(k)}$, as $k \to \infty$, and by $\hat{p}_{is}^n$ and $\hat{p}_i^n$ the evaluations of $p_s(y_i)$ and $p(y_i)$ at $\hat{\Theta}^n$, respectively. It then follows from the first equation of (19) that

$$\begin{aligned}
0 &= \hat{\pi}_s^n - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\gamma}_{is}^n}{\hat{\gamma}_i^n}\\
&= L_s^n(\hat{\Theta}^n) + \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\hat{\pi}_s^n\hat{p}_{is}^n}{\hat{p}_i^n} - \frac{\hat{\gamma}_{is}^n}{\hat{\gamma}_i^n}\Big\}\\
&= L_s^n(\hat{\Theta}^n) + \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\pi}_s^n\hat{p}_{is}^n\hat{\gamma}_i^n - \hat{p}_i^n\hat{\gamma}_{is}^n}{\hat{p}_i^n\hat{\gamma}_i^n}.
\end{aligned} \tag{20}$$

From (31) in Appendix E we obtain

$$\hat{\gamma}_{is}^n = \hat{\pi}_s^n |\hat{\Gamma}_s^n|^{1/2} e^{-(y_i-\hat{\mu}_s^n)'\hat{\Gamma}_s^n(y_i-\hat{\mu}_s^n)/2} + O(\frac{1}{n}) = \hat{\pi}_s^n \hat{p}_{is}^n + O(\frac{1}{n})$$

and thus $\hat{\gamma}_i^n = \hat{p}_i^n + O(1/n)$, so that the second term of (20) is of order $O(1/n)$. From Taylor's expansion the first term can be rewritten as

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

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

$$0 = \nabla L_s^n\big(\tilde{\Theta}^n + \lambda(\hat{\Theta}^n - \tilde{\Theta}^n)\big)(\hat{\Theta}^n - \tilde{\Theta}^n) + O(\frac{1}{n}).$$

Along the same lines as above, from the second and the third equations of (19), we have

$$\begin{aligned} 0 &= \nabla \bar{L}_s^n\big(\tilde{\Theta}^n + \lambda(\hat{\Theta}^n - \tilde{\Theta}^n)\big)(\hat{\Theta}^n - \tilde{\Theta}^n) + O(\frac{1}{n}), \\ 0 &= \nabla \tilde{L}_s^n\big(\tilde{\Theta}^n + \lambda(\hat{\Theta}^n - \tilde{\Theta}^n)\big)(\hat{\Theta}^n - \tilde{\Theta}^n) + O(\frac{1}{n}). \end{aligned}$$

If we let

$$\mathcal{L}^n = \begin{pmatrix} L_1^n \\ \vdots \\ L_m^n \\ \bar{L}_1^n \\ \vdots \\ \bar{L}_m^n \\ \tilde{L}_1^n \\ \vdots \\ \tilde{L}_m^n \end{pmatrix},$$

the last three equations give

$$\nabla \mathcal{L}^n\big(\tilde{\Theta}^n + \lambda(\hat{\Theta}^n - \tilde{\Theta}^n)\big)(\hat{\Theta}^n - \tilde{\Theta}^n) + O(\frac{1}{n}) = 0.$$

We have proved that $\hat{\Theta}^n$ converges to the true value Θ^*, and it is known that the MLE $\tilde{\Theta}^n$ tends to Θ^*, so a derivation similar to the proof of Theorem 1 gives that, for any $B \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$, $\nabla \mathcal{L}^n\big(\tilde{\Theta}^n + \lambda(\hat{\Theta}^n - \tilde{\Theta}^n)\big)(B)$ converges to $\Psi\mathbb{E}(HR(B))$, which is positive definite, where Ψ, H and $R(\cdot)$ are as defined in the Appendix B. Therefore, it follows that $\hat{\Theta}^n = \tilde{\Theta}^n + O(1/n)$.

5 Conclusion

Exact theoretical analysis of the quality of variational Bayesian approximations is an important issue. In this paper we have investigated iterative algorithms for estimating parameters in normal mixture models. Its results are twofold. First we proposed a generalised algorithm, involving a step-size parameter ε, for obtaining the variational Bayesian estimates. Small-scale numerical experiments (not reported here) showed that, for appropriate step sizes, the generalised algorithm provides accelerated convergence relative to the basic algorithm, which corresponds to $\varepsilon = 1$; if the components in the mixture model are widely separated, the optimal ε appears to be only slightly greater than 1, whereas, if the components are nearly identical, the optimal ε is close to 2. This coincides with the theoretical analysis of Peters and Walker (1978), who discussed the optimal ε for obtaining maximum likelihood estimates.

Secondly, we proved theoretically that the variational Bayesian estimators for mixture models of normal densities converge locally to the maximum likelihood estimators at the rate of $O(1/n)$ in the large sample limit, which had not been justified in the previous literature. This implies that in mixture models, and so far as point estimation is concerned, the factorised form of the posterior distribution does not cause bias for large samples, so the variational Bayesian estimator is very effective and asymptotically consistent for mixture models. However, this property may not hold for other models; for example, we proved in Wang and Titterington (2004) that the variational Bayes estimators for linear state space models are not always asymptotically consistent as the 'sample size' becomes large, essentially because the factorised form of $Q^{(S)}(S)$ destroys the intrinsic correlations between the hidden states in the models.

Of course, proving that the means of the variational posterior distribution converge to the maximum likelihood estimators is a rather limited achievement, and it is appropriate to investigate more features of the distributions. As remarked in Section 2, in scenarios like mixtures, where the complete-data Bayesian analysis can be based on conjugate priors, the variational posterior typically takes the conjugate form whereas the correct posterior based on the observed data certainly does not. It is of particular interest to ask if the variances (covariances) of the variational posteriors have the same properties as the means; that is, do the variances (covariances) associated with variational Bayesian approximations converge to those of the true posterior distributions in some sense? In Wang and Titterington (2005) we examine this problem and investigate the performance of variational Bayesian approximations in this context for interval estimation. It turns out that the covariance matrices corresponding to the variational Bayesian approximation are normally 'too small' compared with those for the MLE, and therefore the variational Bayes approximations to interval estimates are unrealistically narrow.

Acknowledgement. This work was supported by a grant from the UK Science and Engineering Research Council and by the IST Programme of the European Community,

under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. The authors are grateful for the very helpful comments provided by the reviewers. A part of this work was presented at the International Society for Bayesian Analysis (ISBA) 2004 World Meeting, May 23-27, Viña del Mar, Chile.

Appendix: Proof of Theorem 1

Appendix A *The establishment of suitable norms.*

In order that the derivative in the vector space $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$ makes sense, we endow the various spaces with norms. We define the norm of $\mu \in \mathbb{R}^d$ as $\|\mu\| = (\mu'\mu)^{1/2}$, and the norm of a real, symmetric $d \times d$ matrix Γ as

$$\|\Gamma\| = \sup_{\mu \in \mathbb{R}^d,\ \|\mu\|=1} \|\Gamma\mu\|.$$

The norms on the direct sums $\mathcal{A}$, $\mathcal{M}$, $\mathcal{T}$ and $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$ are defined naturally as

$$\|\boldsymbol{\pi}\| = \textstyle\sum_{s=1}^m |\pi_s|, \text{ for } \boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_m \end{pmatrix} \in \mathcal{A},$$

$$\|\boldsymbol{\mu}\| = \textstyle\sum_{s=1}^m \|\mu_s\|, \text{ for } \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} \in \mathcal{M},$$

$$\|\boldsymbol{\Gamma}\| = \textstyle\sum_{s=1}^m \|\Gamma_s\|, \text{ for } \boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_m \end{pmatrix} \in \mathcal{T},$$

$$\|\Theta\| = \|\boldsymbol{\pi}\| + \|\boldsymbol{\mu}\| + \|\boldsymbol{\Gamma}\|, \text{ for } \Theta = \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\mu} \\ \boldsymbol{\Gamma} \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}.$$

For any operator Φ on the vector space $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$, its norm is defined as

$$\|\Phi\| = \sup_{\|B\|=1} \|\Phi(B)\|. \tag{21}$$

Also, $\nabla\Phi$ denotes the Fréchet derivative of Φ. When ambiguity exists, the specific vector variable of differentiation appears as a subscript of the symbol ∇. $\nabla\Phi(\Theta)$ denotes the Fréchet derivative evaluated at Θ, and is a linear operator on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$; see Chapter X of Bhatia (1997).

Appendix B *Proof of Theorem 1.*

We first prove that, with probability 1 as n approaches infinity, the operator Φ_n^ε on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$ is *locally contractive* in the norm defined above; that is, there exists a number λ, $0 \leq \lambda < 1$, such that

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

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

Since $\bar{\Theta}$ is near Θ^*, it follows from Taylor's theorem on Banach spaces (see p.315 of Bhatia (1997)) that

$$\|\Phi_n^\varepsilon(\bar{\Theta}) - \Phi_n^\varepsilon(\Theta^*)\| \leq \|\nabla\Phi_n^\varepsilon(\Theta^*)\|\|\bar{\Theta} - \Theta^*\| + O(\|\bar{\Theta} - \Theta^*\|^2).$$

Consequently, it is sufficient to show that $\nabla\Phi_n^\varepsilon(\Theta^*)$ converges with probability 1 to an operator which has norm less than 1.

For

$$B = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{W} \end{pmatrix} = \begin{pmatrix} u_1 \\ \vdots \\ u_m \\ v_1 \\ \vdots \\ v_m \\ W_1 \\ \vdots \\ W_m \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T},$$

from the definition of the operator Φ_n^ε, we have

$$\nabla\Phi_n^\varepsilon(\Theta)(B) = (1-\varepsilon) I_{m(1+2d)} B + \varepsilon \begin{pmatrix} \nabla_{\boldsymbol{\pi}}\Pi & \nabla_{\boldsymbol{\mu}}\Pi & \nabla_{\boldsymbol{\Gamma}}\Pi \\ \nabla_{\boldsymbol{\pi}}M & \nabla_{\boldsymbol{\mu}}M & \nabla_{\boldsymbol{\Gamma}}M \\ \nabla_{\boldsymbol{\pi}}S & \nabla_{\boldsymbol{\mu}}S & \nabla_{\boldsymbol{\Gamma}}S \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \boldsymbol{W} \end{pmatrix},$$

where $I_{m(1+2d)}$ denotes the $m(1+2d) \times m(1+2d)$ identity matrix. Also, the entries of the above matrix can themselves be represented as matrices of Fréchet derivatives.

In the sequel, we drop the superscript $(k-1)$ from r_{is}, γ_{is} and γ_i to indicate that the operator Φ_n^ε is being evaluated at some given Θ, and not at a member of the iterative sequence obtained by the algorithm. After a straightforward calculation we obtain, at

the true value Θ^*,

$$\begin{aligned}
\nabla_{\pi_j}\Pi_s(\Theta^*) &= \frac{1}{n}\sum_{i=1}^{n}\nabla_{\pi_j}r_{is},\\
\nabla_{\mu_j}\Pi_s(\Theta^*) &= \frac{1}{n}\sum_{i=1}^{n}\nabla_{\mu_j}r_{is},\\
\nabla_{\Gamma_j}\Pi_s(\Theta^*) &= \frac{1}{n}\sum_{i=1}^{n}\nabla_{\Gamma_j}r_{is},
\end{aligned}$$

$$\begin{aligned}
\nabla_{\pi_j}M_s(\Theta^*) &= \Big[(\sum_{i=1}^{n}r_{is})(\sum_{i=1}^{n}y_i\nabla_{\pi_j}r_{is}) - (\sum_{i=1}^{n}r_{is}y_i)(\sum_{i=1}^{n}\nabla_{\pi_j}r_{is})\Big]\Big/(\sum_{i=1}^{n}r_{is})^2,\\
\nabla_{\mu_j}M_s(\Theta^*) &= \Big[(\sum_{i=1}^{n}r_{is})(\sum_{i=1}^{n}y_i\nabla_{\mu_j}r_{is}) - (\sum_{i=1}^{n}r_{is}y_i)(\sum_{i=1}^{n}\nabla_{\mu_j}r_{is})\Big]\Big/(\sum_{i=1}^{n}r_{is})^2,\\
\nabla_{\Gamma_j}M_s(\Theta^*) &= \Big[(\sum_{i=1}^{n}r_{is})(\sum_{i=1}^{n}y_i\nabla_{\Gamma_j}r_{is}) - (\sum_{i=1}^{n}r_{is}y_i)(\sum_{i=1}^{n}\nabla_{\Gamma_j}r_{is})\Big]\Big/(\sum_{i=1}^{n}r_{is})^2,
\end{aligned}$$

$$\begin{aligned}
\nabla_{\pi_j}S_s(\Theta^*) &= (\sum_{i=1}^{n}\nabla_{\pi_j}r_{is})\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1}\\
&-(\sum_{i=1}^{n}r_{is})\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1}\Big(\sum_{i=1}^{n}(y_i-\mu_s^*)(y_i-\mu_s^*)'\nabla_{\pi_j}r_{is}\Big)\\
&\times\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1},
\end{aligned}$$

$$\begin{aligned}
\nabla_{\mu_j}S_s(\Theta^*)v_j &= \Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1}(\sum_{i=1}^{n}\nabla_{\mu_j}r_{is}v_j)\\
&-(\sum_{i=1}^{n}r_{is})\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1}\Big(\sum_{i=1}^{n}\big[(y_i-\mu_s^*)(y_i-\mu_s^*)'\nabla_{\mu_j}r_{is}v_j\\
&-r_{is}v_j(y_i-\mu_s)'\delta_{sj}-r_{is}(y_i-\mu_s)v_j'\delta_{sj}\big]\Big)\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1},
\end{aligned}$$

$$\begin{aligned}
\nabla_{\Gamma_j}S_s(\Theta^*)W_j &= \Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1}(\sum_{i=1}^{n}\nabla_{\Gamma_j}r_{is}W_j)\\
&-(\sum_{i=1}^{n}r_{is})\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1}\Big(\sum_{i=1}^{n}(y_i-\mu_s^*)(y_i-\mu_s^*)'\nabla_{\Gamma_j}r_{is}W_j\Big)\\
&\times\Big(\sum_{i=1}^{n}r_{is}(y_i-\mu_s^*)(y_i-\mu_s^*)'\Big)^{-1},
\end{aligned}$$

where δ_{sj} is the Kronecker delta function; $\delta_{sj} = 1$ if $s = j$ and $\delta_{sj} = 0$ otherwise.

To obtain the limits of these derivatives as n tends to infinity, we need the limits of r_{is} and all its derivatives with respect to $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, evaluated at $\boldsymbol{\pi}^*$, $\boldsymbol{\mu}^*$ and $\boldsymbol{\Gamma}^*$. These limits are established in Appendix E, which is based on Appendices C and D. For $s = 1, \ldots, m$, we denote by ϕ_s^* and ϕ^* the evaluations of $p_s(y)$ and $p(y)$ at Θ^* for any random vector y distributed according to the probability density of the form (5), and introduce the notation

$$\alpha_s^1 = \frac{\phi_s^*}{\phi^*}, \quad \alpha_s^2 = \pi_s^* \frac{\phi_s^*}{\phi^*} \Gamma_s^*(y - \mu_s^*),$$
$$\alpha_s^3 = \tfrac{1}{2}\pi_s^* \frac{\phi_s^*}{\phi^*}\big[\Gamma_s^*(y - \mu_s^*)(y - \mu_s^*)'\Gamma_s^* - \Gamma_s^*\big],$$
$$\Lambda = \operatorname{diag}(\pi_s^*), \quad \Omega = \operatorname{diag}(\pi_s^{*-1}\Gamma_s^{*-1}), \quad \Sigma = \operatorname{diag}(2\pi_s^{*-1}\Gamma_s^{*-2}).$$

It can be verified that

$$\mathbb{E}\Big(\frac{\phi_s^*}{\phi^*}\Big) = \int \frac{\phi_s^*}{\phi^*} p(y|\Theta^*)dy = 1, \tag{22}$$

$$\mathbb{E}\Big(\frac{\phi_s^*}{\phi^*}(y - \mu_s^*)\Big) = \int \frac{\phi_s^*(y - \mu_s^*)}{\phi^*} p(y|\Theta^*)dy = 0, \tag{23}$$

$$\mathbb{E}\Big(\frac{\phi_s^*}{\phi^*}(y - \mu_s^*)(y - \mu_s^*)'\Big) = \int \phi_s^*(y - \mu_s^*)(y - \mu_s^*)'dy = \Gamma_s^{*-1}, \tag{24}$$

$$\mathbb{E}\Big(\frac{\phi_s^*}{\phi^*}(y - \mu_s^*)(y - \mu_s^*)'(y - \mu_s^*)\Big) = 0, \tag{25}$$

$$\mathbb{E}\Big(\frac{\phi_s^*}{\phi^*}\big[(y - \mu_s^*)(y - \mu_s^*)'\big]^2\Big) = 3\Gamma_s^{*-2}. \tag{26}$$

Thus, by (32)-(35) and (22)-(26) we can study the limits of the above Fréchet derivatives. As a demonstration of how the calculations go, we consider $\nabla_{\mu_j} M_s(\Theta^*)$ for $s \neq j$. For this we have

$$\begin{aligned}
\nabla_{\mu_j} M_s(\Theta^*) &= \big(\frac{1}{n}\sum_{i=1}^n r_{is}\big)^{-2}\Big[\big(\frac{1}{n}\sum_{i=1}^n r_{is}\big)\big(\frac{1}{n}\sum_{i=1}^n y_i \nabla_{\mu_j} r_{is}\big) \\
&\quad -\big(\frac{1}{n}\sum_{i=1}^n r_{is} y_i\big)\big(\frac{1}{n}\sum_{i=1}^n \nabla_{\mu_j} r_{is}\big)\Big] \\
&\to -\mathbb{E}[y(y - \mu_j^*)'\Gamma_j^*(\phi_s^*\pi_j^*\phi_j^*/\phi^{*2})] \\
&\quad +\mathbb{E}[y\phi_s^*/\phi^*]\mathbb{E}[(y - \mu_j^*)'\Gamma_j^*(\phi_s^*\pi_j^*\phi_j^*/\phi^{*2})] \\
&= -\pi_s^{*-1}\Gamma_s^{*-1}\mathbb{E}(\alpha_s^2\alpha_j^2).
\end{aligned}$$

Using similar manipulations and after very careful calculations, we obtain

$$\nabla_{\boldsymbol{\pi}}\Pi(\Theta^*)\boldsymbol{u} \to I_m\boldsymbol{u} - \Lambda\mathbb{E}\begin{pmatrix} \alpha_1^1 \\ \vdots \\ \alpha_m^1 \end{pmatrix}\Big\{\sum_{s=1}^m \alpha_s^1 u_s\Big\},$$

$$\nabla_{\boldsymbol{\mu}}\Pi(\Theta^*)\boldsymbol{v} \to -\Lambda\mathbb{E}\begin{pmatrix}\alpha_1^1\\ \vdots\\ \alpha_m^1\end{pmatrix}\Big\{\sum_{s=1}^m(\alpha_s^2)'v_s\Big\},$$

$$\nabla_{\boldsymbol{\Gamma}}\Pi(\Theta^*)\boldsymbol{W} \to -\Lambda\mathbb{E}\begin{pmatrix}\alpha_1^1\\ \vdots\\ \alpha_m^1\end{pmatrix}\Big\{\sum_{s=1}^m\mathrm{tr}\{\alpha_s^3W_s\}\Big\},$$

$$\nabla_{\boldsymbol{\pi}}M(\Theta^*)\boldsymbol{u} \to -\Omega\mathbb{E}\begin{pmatrix}\alpha_1^2\\ \vdots\\ \alpha_m^2\end{pmatrix}\Big\{\sum_{s=1}^m\alpha_s^1u_s\Big\},$$

$$\nabla_{\boldsymbol{\mu}}M(\Theta^*)\boldsymbol{v} \to I_{md}\boldsymbol{v} - \Omega\mathbb{E}\begin{pmatrix}\alpha_1^2\\ \vdots\\ \alpha_m^2\end{pmatrix}\Big\{\sum_{s=1}^m(\alpha_s^2)'v_s\Big\},$$

$$\nabla_{\boldsymbol{\Gamma}}M(\Theta^*)\boldsymbol{W} \to -\Omega\mathbb{E}\begin{pmatrix}\alpha_1^2\\ \vdots\\ \alpha_m^2\end{pmatrix}\Big\{\sum_{s=1}^m\mathrm{tr}\{\alpha_s^3W_s\}\Big\},$$

$$\nabla_{\boldsymbol{\pi}}S(\Theta^*)\boldsymbol{u} \to -\Sigma\mathbb{E}\begin{pmatrix}\alpha_1^3\\ \vdots\\ \alpha_m^3\end{pmatrix}\Big\{\sum_{s=1}^m\alpha_s^1u_s\Big\},$$

$$\nabla_{\boldsymbol{\mu}}S(\Theta^*)\boldsymbol{v} \to -\Sigma\mathbb{E}\begin{pmatrix}\alpha_1^3\\ \vdots\\ \alpha_m^3\end{pmatrix}\Big\{\sum_{s=1}^m(\alpha_s^2)'v_s\Big\},$$

$$\nabla_{\boldsymbol{\Gamma}}S(\Theta^*)\boldsymbol{W} \to I_{md}\boldsymbol{W} - \Sigma\mathbb{E}\begin{pmatrix}\alpha_1^3\\ \vdots\\ \alpha_m^3\end{pmatrix}\Big\{\sum_{s=1}^m\mathrm{tr}\{\alpha_s^3W_s\}\Big\}.$$

Set

$$R(B) = \sum_{s=1}^m\alpha_s^1u_s + \sum_{s=1}^m(\alpha_s^2)'v_s + \sum_{s=1}^m\mathrm{tr}\{\alpha_s^3W_s\},$$

$$\Psi = \begin{pmatrix}\Lambda & 0 & 0\\ 0 & \Omega & 0\\ 0 & 0 & \Sigma\end{pmatrix}, \qquad H = \begin{pmatrix}H_1\\ H_2\\ H_3\end{pmatrix} = \begin{pmatrix}\alpha_1^1\\ \vdots\\ \alpha_m^1\\ \alpha_1^2\\ \vdots\\ \alpha_m^2\\ \alpha_1^3\\ \vdots\\ \alpha_m^3\end{pmatrix}.$$

Accordingly, we have that, as n tends to infinity, $\nabla\Phi_n^\varepsilon(\Theta^*)(B)$ converges to

$$I_{m(1+2d)}B - \varepsilon\Psi\mathbb{E}(HR(B)).$$

We define the inner product on $\mathbb{R}$ as scalar multiplication, the inner product on $\mathbb{R}^d$ as $\langle\mu,\nu\rangle = \mu'\nu$ and the inner product on the set of real, symmetric $d\times d$ matrices as $\langle A,B\rangle = \mathrm{tr}\{AB\}$. Naturally, the inner products on the direct sums $\mathcal{A}$, $\mathcal{M}$, $\mathcal{T}$ and $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{T}$ are the corresponding direct sum inner products. For instance, $\langle\Theta_1,\Theta_2\rangle = \langle\boldsymbol{\pi}_1,\boldsymbol{\pi}_2\rangle + \langle\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\rangle + \langle\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_2\rangle$, for

$$\Theta_1 = \begin{pmatrix}\boldsymbol{\pi}_1\\ \boldsymbol{\mu}_1\\ \boldsymbol{\Gamma}_1\end{pmatrix} \in \mathcal{A}\oplus\mathcal{M}\oplus\mathcal{T}, \quad \Theta_2 = \begin{pmatrix}\boldsymbol{\pi}_2\\ \boldsymbol{\mu}_2\\ \boldsymbol{\Gamma}_2\end{pmatrix} \in \mathcal{A}\oplus\mathcal{M}\oplus\mathcal{T}.$$

For any operator Φ on the vector space $\mathcal{A}\oplus\mathcal{M}\oplus\mathcal{T}$, its norm as defined in (21) is equal to $\sup_{\|B\|=1}\langle B,\Phi(B)\rangle$.

It is obvious that Ψ and $\mathbb{E}(HR(\cdot))$ are positive definite and symmetric with respect to the inner product which we have defined. Therefore, as n tends to infinity, $\nabla\Phi_n^\varepsilon(\Theta^*)(\cdot) < I_{m(1+2d)}$ whenever $\varepsilon > 0$.

Furthermore, Peters and Walker (1978) proved that, when $0<\varepsilon<2$, the operator of the form $I_{m(1+2d)} - \varepsilon\Psi\mathbb{E}(HR(\cdot))$ is greater than $-I_{m(1+2d)}$ with respect to the inner product. For completeness, we have included a brief proof in Appendix F.

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

Moreover, along lines similar to the above argument it is easy to deduce that $\Phi_n^\varepsilon(\Theta^*)$ tends to Θ^* as n approaches infinity. Therefore, since

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

the iterative procedure (18) converges locally to the true value Θ^* as n approaches infinity.

Appendix C *Two necessary lemmas.*

Lemma 1 is a variant of the Laplace approximation; see Chapter 4 of Evans and Swartz (2000). Its role is twofold: on one hand, it proves that the mean of a function of a random vector converges to the function evaluated at the mean, under prescribed conditions, which will be used for studying the limits of r_{is}; on the other hand, it gives the order of the difference between them, which serves to estimate the convergence rate in section 4.

Lemma 1 *Suppose that $p_n(x)$ is the probability density function of the $\mathbb{R}^m$-valued random vector $X_n = (x_n^1, \ldots, x_n^m)'$, that $\mathbb{E}(X_n) = \mu_n \to \mu$ and that $Cov_{ij}(X_n) = O(1/n)$ as $n \to \infty$. Then, for any function $f(\cdot)$ with continuous second-order derivative near μ, it holds that*

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

Proof. From Taylor expansion we have that

$$\begin{aligned} f(X_n) = \quad & f(\mu_n) + \textstyle\sum_{i=1}^m \frac{\partial f(\mu_n)}{\partial x_n^i}(x_n^i - \mu_n^i) + \frac{1}{2}\sum_{i,j=1}^m \frac{\partial^2 f(\mu_n)}{\partial x_n^i \partial x_n^j}(x_n^i - \mu_n^i)(x_n^j - \mu_n^j) \\ & + o(\|X_n - \mu_n\|^2), \end{aligned}$$

and thus

$$\begin{aligned} \mathbb{E}(f(X_n)) = \quad & f(\mu_n) + \textstyle\sum_{i=1}^m \frac{\partial f(\mu_n)}{\partial x_n^i}\mathbb{E}(x_n^i - \mu_n^i) \\ & + \frac{1}{2}\sum_{i,j=1}^m \frac{\partial^2 f(\mu_n)}{\partial x_n^i \partial x_n^j}\mathbb{E}((x_n^i - \mu_n^i)(x_n^j - \mu_n^j)) + o(\mathbb{E}(\|X_n - \mu_n\|^2)). \end{aligned}$$

Since $f(\cdot)$ has continuous second-order derivative near μ, $\frac{\partial^2 f(\mu_n)}{\partial x_n^i \partial x_n^j}$ is bounded. Noting that $\mathbb{E}(x_n^i) = \mu_n^i$ and $\mathbb{E}((x_n^i - \mu_n^i)(x_n^j - \mu_n^j)) = O(1/n)$, we have $\mathbb{E}(f(X_n)) = f(\mu_n) + O(1/n)$.

The following lemma is a slight generalisation of the strong law of large numbers. It establishes a law of large numbers result for functions of a random variable when the functions involved have a limit.

Lemma 2 *If $\{X_n\}$ is a sequence of independent and identically distributed random variables and $F_n(\cdot) \to F_0(\cdot)$ uniformly, then, with probability 1,*

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

Proof. In fact, we have that

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

By the strong law of large numbers, the second term tends to zero, as does the first term because of the uniform convergence.

Appendix D *Limits of certain Fréchet derivatives.*

Here we consider the limits of the Fréchet derivatives, with respect to $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, of the quantities

$$I_1 = \int q(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi}, \quad I_2 = \int q(\Gamma_s) \log |\Gamma_s| d\Gamma_s, \quad I_3 = \nu_s (\Phi_s)^{-1},$$

with

$$\begin{aligned} q(\boldsymbol{\pi}) &= \mathcal{D}(\boldsymbol{\pi}; \lambda_1, \ldots, \lambda_m), && (27) \\ q(\Gamma_s) &= \mathcal{W}(\Gamma_s; \nu_s, \Phi_s), && (28) \\ q(\mu_s|\Gamma_s) &= \mathcal{N}(\mu_s; \rho_s, \beta_s \Gamma_s), && (29) \end{aligned}$$

and

$$\begin{gathered} \lambda_s = n\pi_s + \lambda^0, \quad \rho_s = (n\mu_s\pi_s + \beta^0\rho^0)/(n\pi_s + \beta_0), \\ \beta_s = n\pi_s + \beta^0, \quad \nu_s = n\pi_s + \nu^0, \\ \Phi_s = n\pi_s(\Gamma_s)^{-1} + n\pi_s\beta^0(\mu_s - \rho^0)(\mu_s - \rho^0)'(n\pi_s + \beta_0)^{-1} + \Phi^0. \end{gathered}$$

If we write

$$\psi(x) = \frac{d}{dx} \log \Gamma(x) = \Gamma^{-1}(x) \int_0^\infty z^{x-1} e^{-z} \log z dz,$$

where $\Gamma(x)$ is the gamma function, then we have

$$\int q(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi} = \psi(\lambda_s) - \psi(\sum_{s=1}^m \lambda_s).$$

Since $\sum_{s=1}^m \lambda_s = n + m\lambda^0$, it follows that

$$\begin{aligned} & \nabla_{\pi_s} \int q(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi} = \nabla_{\pi_s} \Big[\Gamma^{-1}(\lambda_s) \int_0^\infty z^{\lambda_s - 1} e^{-z} \log z dz \Big] \\ = \; & n\Gamma^{-2}(\lambda_s) \Big[\int_0^\infty z^{\lambda_s - 1} e^{-z} \log^2 z dz \int_0^\infty z^{\lambda_s - 1} e^{-z} dz \\ & - \Big(\int_0^\infty z^{\lambda_s - 1} e^{-z} \log z dz \Big)^2 \Big]. \qquad (30) \end{aligned}$$

We consider the integral of the form

$$\int_0^\infty z^{\lambda_s - 1} e^{-z} f(z) dz$$

for some function $f(\cdot)$ with continuous second-order derivative.

If we make the change of variable $z = u(\lambda_s - 1)$ and denote $f(u(\lambda_s - 1))$ by $h(u)$, we obtain

$$\int_0^\infty z^{\lambda_s - 1} e^{-z} f(z) dz = (\lambda_s - 1)^{\lambda_s} \int_0^\infty e^{-(\lambda_s - 1)(u - \log u)} h(u) du.$$

Obviously, $k(u) \triangleq u - \log u$ attains its global minimum at $\hat{u} = 1$, and therefore an application of the Laplace approximation yields (see for example Chapter 4 of Evans and Swartz (2000))

$$\int_0^\infty e^{-(\lambda_s-1)(u-\log u)}h(u)du = (2\pi)^{1/2}e^{-(\lambda_s-1)}\Big\{h(\hat{u})(\lambda_s-1)^{-1/2} + (\lambda_s-1)^{-3/2}[a_1h(\hat{u}) - a_2h'(\hat{u}) + a_3h''(\hat{u})] + o((\lambda_s-1)^{-3/2})\Big\},$$

where

$$a_1 = -\frac{3k^{(4)}(\hat{u})}{4!} + \frac{1}{2}\Big(\frac{k^{(3)}(\hat{u})}{3!}\Big)^2 15, \quad a_2 = \frac{3k^{(3)}(\hat{u})}{3!}, \quad a_3 = \frac{1}{2}.$$

Letting $f(z)$ be 1, $\log z$ and $\log^2 z$, respectively, we obtain

$$\int_0^\infty z^{\lambda_s-1}e^{-z}dz = (2\pi)^{1/2}e^{-(\lambda_s-1)}(\lambda_s-1)^{\lambda_s} \Big\{(\lambda_s-1)^{-1/2} + a_1(\lambda_s-1)^{-3/2} + o((\lambda_s-1)^{-3/2})\Big\},$$

$$\int_0^\infty z^{\lambda_s-1}e^{-z}\log z dz = (2\pi)^{1/2}e^{-(\lambda_s-1)}(\lambda_s-1)^{\lambda_s} \Big\{(\lambda_s-1)^{-1/2}\log(\lambda_s-1) + (\lambda_s-1)^{-3/2}[a_1\log(\lambda_s-1) - a_2 - a_3] + o((\lambda_s-1)^{-3/2})\Big\},$$

$$\int_0^\infty z^{\lambda_s-1}e^{-z}\log^2 z dz = (2\pi)^{1/2}e^{-(\lambda_s-1)}(\lambda_s-1)^{\lambda_s} \Big\{(\lambda_s-1)^{-1/2}\log^2(\lambda_s-1) + (\lambda_s-1)^{-3/2}[a_1\log^2(\lambda_s-1) - 2a_2\log(\lambda_s-1) + 2a_3(1-\log(\lambda_s-1))] + o((\lambda_s-1)^{-3/2})\Big\}.$$

Hence, after a straightforward calculation we obtain, as $n \to \infty$,

$$\nabla_{\pi_s}\int q(\boldsymbol{\pi})\log\pi_s d\boldsymbol{\pi} \sim \frac{2a_3n(\lambda_s-1)^{-2}+o((\lambda_s-1)^{-1})}{(\lambda_s-1)^{-1}+2a_1(\lambda_s-1)^{-2}+o((\lambda_s-1)^{-2})} \to \frac{1}{\pi_s}.$$

It is obvious that the derivatives of $\int q(\boldsymbol{\pi})\log\pi_s d\boldsymbol{\pi}$ with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$ and π_j $(j \neq s)$ are zero.

The integral I_2 can be rewritten as

$$\int q(\Gamma_s)\log|\Gamma_s|d\Gamma_s = \sum_{k=1}^{d}\psi((\nu_s+1-k)/2) - \log|\Phi_s| + d\log 2,$$

and it follows that

$$\begin{aligned} & \nabla_{\pi_s} \int q(\Gamma_s) \log |\Gamma_s| d\Gamma_s \\ = \; & \nabla_{\pi_s} \Big\{ \textstyle\sum_{k=1}^{d} \Gamma^{-1}((\nu_s + 1 - k)/2) \int_0^\infty z^{(\nu_s+1-k)/2-1} e^{-z} \log z dz \Big\} \\ & -\nabla_{\pi_s} \log |\Phi_s|, \end{aligned}$$

Along the same lines as (30), we obtain

$$\nabla_{\pi_s} \Big\{ \Gamma^{-1}((\nu_s + 1 - k)/2) \int_0^\infty z^{(\nu_s+1-k)/2-1} e^{-z} \log z dz \Big\} \to \frac{d}{\pi_s},$$

and it is easy to show that

$$\nabla_{\pi_s} \log |\Phi_s| = \mathrm{tr}\Big(\Phi_s^{-1} \nabla_{\pi_s} \Phi_s \Big) \to \mathrm{tr}\Big(\big[\pi_s \Gamma_s^{-1}\big]^{-1} \Gamma_s^{-1} \Big) = \frac{d}{\pi_s}.$$

Therefore,

$$\nabla_{\pi_s} \int q(\Gamma_s) \log |\Gamma_s| d\Gamma_s \to 0.$$

Similarly, for any real, symmetric $d \times d$ matrix W,

$$\begin{aligned} & \Big\{ \nabla_{\Gamma_s} \int q(\Gamma_s) \log |\Gamma_s| d\Gamma_s \Big\} W \\ = \; & \Big\{ \nabla_{\Gamma_s} \Big[\textstyle\sum_{k=1}^{d} \Gamma^{-1}((\nu_s + 1 - k)/2) \int_0^\infty z^{(\nu_s+1-k)/2-1} e^{-z} \log z dz \\ & \quad - \log |\Phi_s| + d \log 2 \Big] \Big\} W \\ = \; & -\Big\{ \nabla_{\Gamma_s} \log |\Phi_s| \Big\} W = -\mathrm{tr}\big\{ \Phi_s^{-1} \nabla_{\Gamma_s} \Phi_s W \big\} \\ \to \; & \mathrm{tr}\big\{ \big[\pi_s \Gamma_s^{-1}\big]^{-1} \pi_s \Gamma_s^{-1} W \Gamma_s^{-1} \big\} = \mathrm{tr}\{\Gamma_s^{-1} W\}. \end{aligned}$$

The derivatives of $\int q(\Gamma_s) \log |\Gamma_s| d\Gamma_s$ in $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, Γ_j $(j \neq s)$ are zero.

It is easy to obtain that $\nabla_{\Gamma_s} I_3$ converges to I_d and the other derivatives converge to zero.

Appendix E *Now we study the limits of r_{is} and their derivatives.*

Since $\boldsymbol{\pi}$, Γ_s and μ_s have the variational posterior densities as (27)-(29), it is obvious that, as n tends to infinity, the mean of π_s corresponding to the density $q(\boldsymbol{\pi})$ is

$$\lambda_s \Big/ \sum_{s=1}^{m} \lambda_s = (n\pi_s + \lambda^0) \Big/ \sum_{s=1}^{m} (n\pi_s + \lambda^0) \to \pi_s,$$

the covariance between π_s and π_t, for $s \neq t$, is

$$\begin{aligned} & -\lambda_s \lambda_t \Big/ \Big[(\textstyle\sum_{s=1}^{m} \lambda_s)^2 (\sum_{s=1}^{m} \lambda_s + 1) \Big] \\ = \; & -(n\pi_s + \lambda^0)(n\pi_t + \lambda^0) / [(n + m\lambda^0)^2 (n + m\lambda^0 + 1)] \\ = \; & O(\tfrac{1}{n}) \to 0, \end{aligned}$$

and the variance of π_s is

$$\lambda_s(\sum_{s=1}^{m}\lambda_s-\lambda_s)\Big/\big[(\sum_{s=1}^{m}\lambda_s)^2(\sum_{s=1}^{m}\lambda_s+1)\big]=O(\frac{1}{n})\to 0;$$

similarly, the mean of Γ_s corresponding to the density $q(\Gamma_s)$ is

$$\nu_s(\Phi_s)^{-1}\to\Gamma_s,$$

and its covariance matrix is $2\nu_s(\Phi_s)^{-1}\otimes(\Phi_s)^{-1}=2\nu_s(\Phi_s\otimes\Phi_s)^{-1}$, whose components obviously tend to 0 at the rate of $O(1/n)$, where $\otimes$ denotes the Kronecker product.

Thus from Lemma 1 of Appendix C and Taylor's expansion we have

$$\tilde{\pi}_s=\exp\{\int q(\boldsymbol{\pi})\log\pi_s d\boldsymbol{\pi}\}=\exp\{\log(\pi_s)+O(\tfrac{1}{n})\}=\pi_s+O(\tfrac{1}{n}),$$
$$\tilde{\Gamma}_s=\exp\{\int q(\Gamma_s)\log|\Gamma_s|d\Gamma_s\}=|\Gamma_s|+O(\tfrac{1}{n}).$$

It is also obvious that $\bar{\Gamma}_s=\nu_s(\Phi_s)^{-1}=\Gamma_s+O(1/n)$, $\rho_s=\mu_s+O(1/n)$ and $1/\beta_s=O(1/n)$.

Therefore, noting the definition of γ_{is} we obtain that

$$\gamma_{is}=\pi_s|\Gamma_s|^{1/2}e^{-(y_i-\mu_s)'\Gamma_s(y_i-\mu_s)/2}+O(\frac{1}{n}). \tag{31}$$

Furthermore, we note that the r_{is} and their derivatives with respect to $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ are functions of y_i, which converge uniformly in y_i as n tends to infinity, according to (31) and Appendix D. Thus, if $\{y_i\}$ is a sequence of samples of the random vector y distributed according to the probability density of the form (5) and $f(\cdot)$ is any continuous function on $\mathbb{R}^d$, it follows from Lemma 2 of Appendix C that, with probability 1,

$$\frac{1}{n}\sum_{i=1}^{n}f(y_i)r_{is}\to\mathbb{E}[f(y)\pi_s\phi_s/\phi], \tag{32}$$

$$\frac{1}{n}\sum_{i=1}^{n}f(y_i)\nabla_{\pi_j}r_{is}\to\delta_{sj}\mathbb{E}[f(y)\phi_s/\phi]-\mathbb{E}[f(y)\pi_s\phi_s\phi_j/\phi^2], \tag{33}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}f(y_i)\nabla_{\mu_j}r_{is}&\to\delta_{sj}\mathbb{E}[f(y)\phi_s\pi_s(y-\mu_s)'\Gamma_s/\phi]\\&\quad-\mathbb{E}[f(y)\pi_s\phi_s(y-\mu_j)'\Gamma_j\pi_j\phi_j/\phi^2],\end{aligned} \tag{34}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}f(y_i)\nabla_{\Gamma_j}r_{is}W_j&\to\delta_{sj}\mathbb{E}\Big[f(y)\pi_s\phi_s/(2\phi)\mathrm{tr}\big\{\big[\Gamma_s^{-1}-(y-\mu_s)(y-\mu_s)'\big]W_j\big\}\Big]\\&\quad-\mathbb{E}\Big[f(y)\pi_s\phi_s\pi_j\phi_j/(2\phi^2)\mathrm{tr}\big\{\big[\Gamma_j^{-1}-(y-\mu_j)(y-\mu_j)'\big]W_j\big\}\Big],\end{aligned} \tag{35}$$

where δ_{sj} is the Kronecker delta function. Note that ϕ_s and ϕ denote the evaluations of $p_s(y)$ and $p(y)$ at $\boldsymbol{\pi}$, Γ_s and μ_s, $s=1,\ldots,m$.

Appendix F *Here we prove that* $I_{m(1+2d)} - \varepsilon\Psi\mathbb{E}(HR(\cdot)) > -I_{m(1+2d)}$.

Since $0 < \varepsilon < 2$ and Ψ is positive definite diagonal matrix, it suffices to show that

$$\langle B, \mathbb{E}(HR(B))\rangle \leq \langle B, \Psi^{-1}B\rangle.$$

In fact, we have

$$\begin{aligned}
& \langle B, \mathbb{E}(HR(B))\rangle \\
= \;& \mathbb{E}\big(\langle B, H\rangle R(B)\big) \\
= \;& \mathbb{E}\Big[\big(\langle \boldsymbol{u}, H_1\rangle + \langle \boldsymbol{v}, H_2\rangle + \langle \boldsymbol{W}, H_3\rangle\big)R(B)\Big] \\
= \;& \mathbb{E}\Big(\textstyle\sum_{s=1}^m \alpha_s^1 u_s + \sum_{s=1}^m (\alpha_s^2)' v_s + \sum_{s=1}^m \mathrm{tr}\{\alpha_s^3 W_s\}\Big)^2 \\
= \;& \mathbb{E}\Bigg(\textstyle\sum_{s=1}^m \pi_s^* \frac{\phi_s^*}{\phi^*}\Big[u_s \pi_s^{*-1} + (y-\mu_s^*)'\Gamma_s^* v_s \\
& + \mathrm{tr}\big\{\tfrac{1}{2}\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\}\Big]\Bigg)^2.
\end{aligned}$$

As a corollary of Schwarz's inequality it holds that, if $\eta_s \geq 0$ for $s = 1, \cdots, m$ and $\sum_{s=1}^m \eta_s = 1$, then $|\sum_{s=1}^m \xi_s\eta_s|^2 \leq \sum_{s=1}^m \xi_s^2\eta_s$ for all $\{\xi_s\}_{s=1,\cdots,m}$ (see Peters and Walker (1978)). Applying this result and noting that $\sum_{s=1}^m \pi_s^*\phi_s^*/\phi^* = 1$, we obtain

$$\begin{aligned}
& \langle B, \mathbb{E}(HR(B))\rangle \\
\leq \;& \mathbb{E}\Bigg(\textstyle\sum_{s=1}^m \pi_s^* \frac{\phi_s^*}{\phi^*}\Big[u_s \pi_s^{*-1} + (y-\mu_s^*)'\Gamma_s^* v_s \\
& + \mathrm{tr}\big\{\tfrac{1}{2}\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\}\Big]^2\Bigg) \\
= \;& \textstyle\sum_{s=1}^m \mathbb{E}\Bigg(\pi_s^* \frac{\phi_s^*}{\phi^*}\Big[u_s^2 \pi_s^{*-2} + [(y-\mu_s^*)'\Gamma_s^* v_s]^2 \\
& \quad + \big(\mathrm{tr}\big\{\tfrac{1}{2}\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\}\big)^2 + 2u_s\pi_s^{*-1}(y-\mu_s^*)'\Gamma_s^* v_s \\
& \quad + u_s\pi_s^{*-1}\mathrm{tr}\big\{\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\} \\
& \quad + (y-\mu_s^*)'\Gamma_s^* v_s \mathrm{tr}\big\{\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\}\Bigg) \\
= \;& \textstyle\sum_{s=1}^m \Bigg(u_s^2\pi_s^{*-1}\mathbb{E}(\frac{\phi_s^*}{\phi^*}) + v_s'\pi_s^*\Gamma_s^*\mathbb{E}[\frac{\phi_s^*}{\phi^*}(y-\mu_s^*)(y-\mu_s^*)']\Gamma_s^* v_s \\
& \quad + \mathbb{E}\Big[\pi_s^*\frac{\phi_s^*}{\phi^*}\big(\mathrm{tr}\big\{\tfrac{1}{2}\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\}\big)^2\Big] \\
& \quad + 2u_s\mathbb{E}[\tfrac{\phi_s^*}{\phi^*}(y-\mu_s^*)']\Gamma_s^* v_s \\
& \quad + u_s\mathrm{tr}\big\{\big[\Gamma_s^*\mathbb{E}[\tfrac{\phi_s^*}{\phi^*}(y-\mu_s^*)(y-\mu_s^*)']\Gamma_s^* - \mathbb{E}[\tfrac{\phi_s^*}{\phi^*}]\Gamma_s^*\big]W_s\big\} \\
& \quad + \mathrm{tr}\big\{\Gamma_s^*\mathbb{E}[\tfrac{\phi_s^*}{\phi^*}(y-\mu_s^*)(y-\mu_s^*)'(y-\mu_s^*)']\Gamma_s^* v_s\Gamma_s^* W_s\big\} \\
& \quad - \mathbb{E}[\tfrac{\phi_s^*}{\phi^*}(y-\mu_s^*)']\Gamma_s^* v_s \mathrm{tr}\big\{\Gamma_s^* W_s\big\}\Bigg) \\
= \;& \textstyle\sum_{s=1}^m \Big\{u_s^2\pi_s^{*-1} + v_s'\pi_s^*\Gamma_s^* v_s + \mathrm{tr}\big\{\tfrac{1}{2}\pi_s^* W_s\Gamma_s^{*2}W_s\big\}\Big\} \\
= \;& \langle B, \Psi^{-1}B\rangle,
\end{aligned}$$

where the second-last equality comes from (22)-(26) and the fact that

$$\mathbb{E}\Big[\pi_s^* \frac{\phi_s^*}{\phi^*}\big(\mathrm{tr}\big\{\frac{1}{2}\big[\Gamma_s^*(y-\mu_s^*)(y-\mu_s^*)'\Gamma_s^* - \Gamma_s^*\big]W_s\big\}\big)^2\Big] = \mathrm{tr}\big\{\frac{1}{2}\pi_s^* W_s \Gamma_s^{*2} W_s\big\},$$

which can be verified by expressing the matrices in terms of their components and carrying out a straightforward simplification.

References

Attias, H. (1999). "Inferring parameters and structure of latent variable models by variational Bayes." In Prade, H. and Laskey, K. (eds.), *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, 21–30. Stockholm, Sweden: Morgan Kaufmann Publishers. 625, 629

— (2000). "A variational Bayesian framework for graphical models." In Solla, S., Leen, T., and Muller, K.-R. (eds.), *Advances in Neural Information Processing Systems 12*, 209–215. Cambridge, MA: MIT Press. 625, 629

Beal, M. J. (2003). "Variational Algorithms for Approximate Bayesian Inference." Ph.D. thesis, University College London. 625

Bhatia, R. (1997). *Matrix Analysis*. New York: Springer-Verlag. 636, 637

Celeux, G., Hurn, M., and Robert, C. P. (2000). "Computational and inferential difficulties with mixture posterior distributions." *Journal of the American Statistical Association*, 95(957-979). 625

Corduneanu, A. and Bishop, C. M. (2001). "Variational Bayesian model selection for mixture distributions." In Richardson, T. and Jaakkola, T. (eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, 27–34. Morgan Kaufmann. 625, 629

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley. 632

Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. New York: Oxford University Press. 641, 644

Ghahramani, Z. and Beal, M. J. (2000). "Variational inference for Bayesian mixtures of factor analysers." In Solla, S., Leen, T., and Muller, K.-R. (eds.), *Advances in Neural Information Processing Systems 12*, 449–455. Cambridge, MA: MIT Press. 625

— (2001). "Propagation algorithms for variational Bayesian learning." In Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, 507–513. Cambridge, MA: MIT Press. 625

Hall, P., Humphreys, K., and Titterington, D. M. (2002). "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. 626

Humphreys, K. and Titterington, D. M. (2000). "Approximate Bayesian inference for simple mixtures." In Bethlehem, J. G. and van der Heijden, P. G. M. (eds.), *COMPSTAT2000*, 331–336. Heidelberg: Physica-Verlag. 625, 629

Jordan, M. I. (2004). "Graphical Models." *Statistical Science*, 19(1): 140–155. 625

MacKay, D. J. C. (1997). "Ensemble learning for hidden Markov models." Technical report, Cavendish Laboratory, University of Cambridge. 625

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. 627

Penny, W. D. and Roberts, S. J. (2000). "Variational Bayes for 1-dimensional mixture models." Technical Report PARG-2000-01, Oxford University. 625, 629

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 J. Appl. Math.*, 35: 362–378. 627, 632, 635, 641, 647

Redner, R. A. and Walker, H. F. (1984). "Mixture densities, Maximum likelihood and the EM Algorithm." *SIAM Review*, 26: 195–239. 633

Salakhutdinov, R. and Roweis, S. (2003). "Adaptive overrelaxed bound optimization methods." In Fawcett, T. and Mishra, N. (eds.), *Proceedings of the Twentieth International Conference on Machine Learning*, 664–671. AAAI Press. 627

Titterington, D. M. (2004). "Bayesian methods for neural networks and related models." *Statistical Science*, 19(1): 128–139. 625

Ueda, N. and Ghahramani, Z. (2003). "Bayesian model search for mixture models based on optimizing variational bounds." *Neural Networks*, 15: 1223–1241. 625, 629

Wang, B. and Titterington, D. M. (2003). "Local convergence of variational Bayes estimators for mixing coefficients." Technical Report 03-4, University of Glasgow. Http://www.stats.gla.ac.uk/Research/TechRep2003/03-4.pdf. 627

— (2004). "Lack of consistency of mean field and variational Bayes approximations for state space models." *Neural Processing Letters*, 20: 151–170. 626, 635

— (2005). "Inadequacy of interval estimates corresponding to variational Bayesian approximations." In Cowell, R. G. and Ghahramani, Z. (eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6-8, 2005, Savannah Hotel, Barbados*, 373–380. Society for Artificial Intelligence and Statistics. (Available electronically at http://www.gatsby.ucl.ac.uk/aistats/). 635