

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

Bo Wang and D. M. Titterington

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

Abstract

In this paper we propose a generalised iterative algorithm for calculating variational Bayesian estimates for a normal mixture model and we investigate its convergence properties. It is shown theoretically that the variational Bayes estimator converges locally to the maximum likelihood estimator at the rate of $O(1/n)$ in the large sample limit. We also demonstrate by numerical experiments that the generalised algorithm can be accelerated by suitable choice of step size.

Key words: Mixture model, Variational Bayes, Local convergence, Laplace's approximate

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. In addition, the actual 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

*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.

dependence, and one has to deal with issues such as convergence and storage of the MCMC realisations.

In the face of these difficulties, deterministic variational Bayes 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 Titterton (2000); Penny and Roberts (2000)), mixtures of factor analysers (Ghahramani and Beal (2000)) and state space models (Ghahramani and Beal (2001); Beal (2003)). Titterton (2004) gives a more extensive review and Jordan (2004) provides an overview of a general formulation of the approach in terms of convex analysis.

Let y denote observed data, x denote missing data, p and θ generically denote probability density and parameters, and $p(x, \theta|y)$ denote the posterior distribution of (x, θ) , given y . The variational Bayes approximation, $q(x, \theta|y)$, for $p(x, \theta|y)$, is defined as the minimiser of the Kullback-Leibler divergence between q and p

$$\int q(x, \theta|y) \log \frac{p(x, \theta|y)}{q(x, \theta|y)} dx d\theta, \quad (1)$$

with q restricted to have a special structure, usually corresponding to independence between θ and x . The minimisation of the Kullback-Leibler divergence (1) is equivalent to maximising the so-called negative free energy

$$\int q(x, \theta|y) \log \frac{p(x, \theta, y)}{q(x, \theta|y)} dx d\theta.$$

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 Bayes estimator: exact theoretical analysis of the quality of the method needs to be studied. Hall et al. (2002) initiated a discussion of these aspects and proved that, for certain Markov models, the parameter estimator obtained by maximising the lower bound function is asymptotically consistent provided the proportion of all values that are missing tends to zero. However, unfortunately this sufficient condition is not satisfied in the case of many problems, such as state space models and mixture models. In Wang and Titterton (2003a) we investigated the consistency properties of both so-called mean field and variational Bayes estimators in the context of linear state space models, in which the above sufficient condition obviously does not hold. 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. Later we studied the consistency property of variational Bayesian estimators for mixture models involving known component densities in Wang and Titterington (2003b). Since the likelihood equations have multiple solutions in general (see Duda and Hart (1973)), the variational Bayes algorithm may, like other optimum-seeking algorithms, converge to different limits if different starting values (or hyperparameters) are chosen. It was shown in Wang and Titterington (2003b) that, with probability 1 as the sample size grows large, the iterative algorithm for the variational Bayes approximation converges locally to the maximum likelihood estimator, in that very special context.

In this paper we investigate a generalised iterative algorithm for a more general mixture model. 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. Motivated by the earlier work on the EM algorithm, we propose a generalised iterative algorithm for calculating approximate Bayesian estimates, and prove theoretically that the variational Bayes 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, which is the main contribution of the paper. It is also demonstrated by numerical experiments that, for appropriate choice of step size, the generalised algorithm can be accelerated.

2 The mixture model and the variational approximation

We consider a model in which we have a mixture of m multivariate normal densities p_1, \dots, p_m with mean vectors μ_1, \dots, μ_m and precision (inverse covariance) matrices $\Gamma_1, \dots, \Gamma_m$, respectively. Thus the density of an observation is given by

$$p(y_i|\Theta) = \sum_{s=1}^m p_s(y_i|\Theta)p(s_i = s|\Theta), \quad (2)$$

where $y_i \in \mathbb{R}^d$ denotes the i th observed data vector, and s_i indicates the hidden component that generated it. The components are labelled by $s =$

$1, \dots, m$, and the component s has mixing coefficient $\pi_s = p(s_i = s|\Theta)$ 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 notational convenience, we define several vector spaces. For each s , $1 \leq s \leq m$, π_s , μ_s and Γ_s are elements of the vector spaces \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 spaces 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.

We use conjugate priors on the parameters Θ . 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 in the normal distribution.

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

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

In the variational Bayes approach, we use an approximate density $q(S, \Theta|Y)$ for $p(S, \Theta|Y)$, which factorises as

$$q(S, \Theta|Y) = q^{(S)}(S|Y)q^{(\Theta)}(\Theta|Y),$$

and such that the factors are chosen to maximise the *negative free energy*

$$\int \sum_{\{S\}} q(S, \Theta|Y) \log \frac{p(\Theta, S, Y)}{q(S, \Theta|Y)} d\Theta. \quad (3)$$

For the sake of simplification, we drop the dependence of $q(S, \Theta|Y)$, $q^{(S)}(S|Y)$ and $q^{(\Theta)}(\Theta|Y)$ on Y and write them as $q(S, \Theta)$, $q^{(S)}(S)$ and $q^{(\Theta)}(\Theta)$, respectively. As a result of the form of $p(\Theta, S, Y)$, it follows immediately that the optimal $q^{(S)}(S)$ and $q^{(\Theta)}(\Theta)$ must factorise as

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

As in Attias (1999, 2000), Humphreys and Titterton (2000) and Penny and Roberts (2000), the remaining details of the variational posteriors can be obtained by the following iterative procedure. In turn, we perform the following two stages.

(i) Optimise $q^{(\Theta)}(\Theta)$ for fixed $\{q_i^{(S)}(s_i), i = 1, \dots, n\}$. Since conjugate priors are used, these variational posteriors are functionally identical to the priors, with different hyperparameter values: the mixing coefficients $\boldsymbol{\pi}$ are jointly Dirichlet, with $q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi} : \lambda_1, \dots, \lambda_m)$; the precisions are independently Wishart, with $q(\Gamma_s) = \mathcal{W}(\Gamma_s : \nu_s, \Phi_s)$; and the means conditioned on the precisions are independently normal, with $q(\mu_s | \Gamma_s) = \mathcal{N}(\mu_s : \rho_s, \beta_s \Gamma_s)$. Here $\mathcal{D}(\boldsymbol{\pi} : \lambda_1, \dots, \lambda_m)$, $\mathcal{W}(\Gamma_s : \nu_s, \Phi_s)$ and $\mathcal{N}(\mu_s : \rho_s, \beta_s \Gamma_s)$ denote the relevant density functions. The hyperparameters are updated as follows:

$$\begin{aligned} \lambda_s &= \sum_{i=1}^n r_{is} + \lambda^0, & \rho_s &= \left(\sum_{i=1}^n r_{is} y_i + \beta^0 \rho^0 \right) / \left(\sum_{i=1}^n r_{is} + \beta_0 \right), & (4) \\ \beta_s &= \sum_{i=1}^n r_{is} + \beta^0, & \nu_s &= \sum_{i=1}^n r_{is} + \nu^0, \\ \Phi_s &= \sum_{i=1}^n r_{is} (y_i - \bar{\mu}_s)(y_i - \bar{\mu}_s)' \\ &+ \left[\left(\sum_{i=1}^n r_{is} \right) \beta^0 (\bar{\mu}_s - \rho^0)(\bar{\mu}_s - \rho^0)' \right] / \left(\sum_{i=1}^n r_{is} + \beta_0 \right) + \Phi^0, \end{aligned}$$

where

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

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

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

where

$$\begin{aligned} \tilde{\pi}_s &= \exp \left\{ \int q(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi} \right\}, \\ \tilde{\Gamma}_s &= \exp \left\{ \int q(\Gamma_s) \log |\Gamma_s| d\Gamma_s \right\}, \\ \bar{\Gamma}_s &= \nu_s \Phi_s^{-1}. \end{aligned}$$

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

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

$$r_{is} \propto \lambda^0 (\nu^0 \Phi^0)^{1/2} e^{-(y_i - \rho^0)' \nu^0 \Phi^0 (y_i - \rho^0) / 2 - d / (2\beta^0)}.$$

Remark 1. *Stage (i) of the procedure reveals the key fact that underlies the 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 is a complicated mixture of a large number of such conjugate distributions.*

3 Generalised iterative algorithm for calculating variational Bayesian estimates and its convergence

The Bayesian estimator $\tilde{\theta}$ of the parameter θ associated with a quadratic loss function is taken to be its posterior mean, i.e.

$$\tilde{\theta} = \int \theta p_{\text{pos}}(\theta) d\theta,$$

where p_{pos} is the posterior density of θ . Similarly we define the variational Bayesian estimator $\hat{\theta}$ as

$$\hat{\theta} = \int \theta q_{\text{pos}}(\theta) d\theta, \quad (5)$$

where q_{pos} is the variational posterior density of θ .

At the k th iteration of the iterative procedure (i) (ii), we define the corresponding approximations to the variational Bayesian estimates by

$$\pi_s^{(k)} = \frac{1}{n} \sum_{i=1}^n r_{is}^{(k-1)}, \quad (6a)$$

$$\mu_s^{(k)} = \left(\sum_{i=1}^n r_{is}^{(k-1)} y_i \right) / \left(\sum_{i=1}^n r_{is}^{(k-1)} \right), \quad (6b)$$

$$\Gamma_s^{(k)} = \left(\sum_{i=1}^n r_{is}^{(k-1)} \right) \left(\sum_{i=1}^n r_{is}^{(k-1)} (y_i - \mu_s^{(k-1)})(y_i - \mu_s^{(k-1)})' \right)^{-1}, \quad (6c)$$

where the notation for r now recognises the fact that the r -values change from iteration to iteration. Then the procedure given in the previous section suggests the following algorithm, which is clearly similar in character to the

EM algorithm itself: starting with some initial value $\Theta^{(0)}$, we define successive iterates inductively by

$$\pi_s^{(k+1)} = \frac{1}{n} \sum_{i=1}^n r_{is}^{(k)} \triangleq \Pi_s(\Theta^{(k)}), \quad (7a)$$

$$\mu_s^{(k+1)} = \left(\sum_{i=1}^n r_{is}^{(k)} y_i \right) / \left(\sum_{i=1}^n r_{is}^{(k)} \right) \triangleq M_s(\Theta^{(k)}), \quad (7b)$$

$$\Gamma_s^{(k+1)} = \left(\sum_{i=1}^n r_{is}^{(k)} \right) \left(\sum_{i=1}^n r_{is}^{(k)} (y_i - \mu_s^{(k)}) (y_i - \mu_s^{(k)})' \right)^{-1} \triangleq S_s(\Theta^{(k)}), \quad (7c)$$

where $r_{is}^{(k)}$ is updated as follows:

$$r_{is}^{(k)} = \gamma_{is}^{(k)} / \gamma_i^{(k)}, \quad \gamma_i^{(k)} = \sum_{s=1}^m \gamma_{is}^{(k)},$$

$$\gamma_{is}^{(k)} = \tilde{\pi}_s^{(k)} (\tilde{\Gamma}_s^{(k)})^{1/2} e^{-(y_i - \rho_s^{(k)})' \tilde{\Gamma}_s^{(k)} (y_i - \rho_s^{(k)}) / 2 - d / (2\beta_s^{(k)})},$$

in which

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

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

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

and

$$q^{(k)}(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi} : \lambda_1^{(k)}, \dots, \lambda_m^{(k)}),$$

$$q^{(k)}(\Gamma_s) = \mathcal{W}(\Gamma_s : \nu_s^{(k)}, \Phi_s^{(k)}),$$

$$q^{(k)}(\mu_s | \Gamma_s) = \mathcal{N}(\mu_s : \rho_s^{(k)}, \beta_s^{(k)} \Gamma_s),$$

and the hyperparameters in the variational posterior distributions are given by

$$\lambda_s^{(k)} = n\pi_s^{(k)} + \lambda^0, \quad \rho_s^{(k)} = (n\mu_s^{(k)}\pi_s^{(k)} + \beta^0 \rho^0) / (n\pi_s^{(k)} + \beta_0),$$

$$\beta_s^{(k)} = n\pi_s^{(k)} + \beta^0, \quad \nu_s^{(k)} = n\pi_s^{(k)} + \nu^0,$$

$$\Phi_s^{(k)} = n\pi_s^{(k)} (\Gamma_s^{(k)})^{-1} + n\pi_s^{(k)} \beta^0 (\mu_s^{(k)} - \rho^0) (\mu_s^{(k)} - \rho^0)' (n\pi_s^{(k)} + \beta_0)^{-1} + \Phi^0.$$

Remark 2. *In fact, the estimates given by (6) are not exactly the variational Bayesian estimates defined in (5). For example, from (4) at the k th iteration*

the variational Bayesian estimate of μ_s is

$$\hat{\mu}_s^{(k)} = \left(\sum_{i=1}^n r_{is}^{(k-1)} y_i + \beta^0 \rho^0 \right) / \left(\sum_{i=1}^n r_{is}^{(k-1)} + \beta^0 \right),$$

not (7b). For simplicity we neglect the constants given by the priors, because in this paper we investigate their properties in the context of large samples and the limiting behaviours of the two iterations are the same.

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 (7) can be rewritten as

$$\Theta^{(k+1)} = \begin{pmatrix} \Pi(\Theta^{(k)}) \\ M(\Theta^{(k)}) \\ S(\Theta^{(k)}) \end{pmatrix}. \quad (8)$$

More generally, we define the following generalised iterative algorithm for obtaining the variational Bayesian estimates. Given some initial value $\Theta^{(0)}$, successive iterates are defined recursively by

$$\Theta^{(k+1)} = (1 - \varepsilon)\Theta^{(k)} + \varepsilon \begin{pmatrix} \Pi(\Theta^{(k)}) \\ M(\Theta^{(k)}) \\ S(\Theta^{(k)}) \end{pmatrix} \triangleq \Phi_n^\varepsilon(\Theta^{(k)}), \quad (9)$$

for $k = 1, 2, \dots$ and some $\varepsilon > 0$. Obviously, when $\varepsilon = 1$ algorithm (9) becomes (8).

Suppose that the true value of the parameter Θ is Θ^* . Then in the Appendix we establish the following theorem.

Theorem 1. *With probability 1 as n approaches infinity, the iterative procedure (9) converges locally to the true value Θ^* whenever $0 < \varepsilon < 2$, that is, the iterative procedure (9) converges to the true value Θ^* whenever $0 < \varepsilon < 2$ and the starting values are sufficiently near to Θ^* .*

Intuitively, as the sample size gets large, the joint posterior distribution $p(S, \Theta|Y)$ of S and Θ becomes less dependent on the prior distribution of Θ , and thus the factorised distribution $q^{(S)}(S|Y)q^{(\Theta)}(\Theta|Y)$ provides a more

accurate approximation to the true posterior $p(S, \Theta|Y)$ when the sample size becomes larger. Theorem 1 verifies that in mixture models the use of a factorised form $q^{(S)}(S|Y)q^{(\Theta)}(\Theta|Y)$ for the posterior distributions of Θ and S does not lead to bias for large samples, because the correlation comes from the prior distribution of the parameters, but not the model itself, as would be the case for hidden Markov chain models, for instance, for which we proved in Wang and Titterton (2003a) that the variational Bayes estimators for linear state space models are not always asymptotically consistent as the sample size becomes large, because they destroy the intrinsic correlations between the states of the models.

Unfortunately, for mixture models with unknown parameters in the components, because the negative free energy (3) may be multi-modal, the variational Bayes algorithm may converge to different limits if different starting values (or hyperparameters) are chosen. Therefore only the local convergence property is proved here.

4 The convergence rate of the variational Bayes estimator

It is known that in general the (non-variational) Bayes estimator and the MLE get closer to each other at rate $O(1/n)$. In this section we estimate the rate at which the variational Bayes estimator converges to the maximum likelihood estimator (MLE). Suppose 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 $s = 1, \dots, m$,

$$\begin{aligned} L_s^n(\Theta) &\triangleq \pi_s - \frac{1}{n} \sum_{i=1}^n \frac{\psi_{is} \pi_s}{\psi_i} = 0, \\ \bar{L}_s^n(\Theta) &\triangleq \mu_s - \left\{ \frac{1}{n} \sum_{i=1}^n y_i \frac{\psi_{is}}{\psi_i} \right\} / \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\psi_{is}}{\psi_i} \right\} = 0, \\ \tilde{L}_s^n(\Theta) &\triangleq \Gamma_s - \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\psi_{is}}{\psi_i} \right\} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\psi_{is}}{\psi_i} (y_i - \mu_s)(y_i - \mu_s)' \right\}^{-1} = 0, \end{aligned}$$

where

$$\psi_{is} = |\Gamma_s|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y_i - \mu)' \Gamma_s (y_i - \mu_s) \right\}, \quad \text{and} \quad \psi_i = \sum_{s=1}^m \pi_s \psi_{is}.$$

Denote by $\hat{\Theta}^n$ the variational Bayes estimator, which is the equilibrium point of the iteration (9) in the neighbourhood of the true value; that is, from (9), at the equilibrium point, $\hat{\Theta}^n$ yields

$$\hat{\Theta}^n - \begin{pmatrix} \Pi(\hat{\Theta}^n) \\ M(\hat{\Theta}^n) \\ S(\hat{\Theta}^n) \end{pmatrix} = 0, \quad (10)$$

and the hyperparameters in the variational posterior distributions are, correspondingly, given by

$$\begin{aligned} \hat{\lambda}_s^n &= n\hat{\pi}_s^n + \lambda^0, & \hat{\rho}_s^n &= (n\hat{\mu}_s^n\hat{\pi}_s^n + \beta^0\rho^0)/(n\hat{\pi}_s^n + \beta_0), \\ \hat{\beta}_s^n &= n\hat{\pi}_s^n + \beta^0, & \hat{\nu}_s^n &= n\hat{\pi}_s^n + \nu^0, \\ \hat{\Phi}_s^n &= n\hat{\pi}_s^n(\hat{\Gamma}_s^n)^{-1} + n\hat{\pi}_s^n\beta^0(\hat{\mu}_s^n - \rho^0)(\hat{\mu}_s^n - \rho^0)'(n\hat{\pi}_s^n + \beta_0)^{-1} + \Phi^0. \end{aligned}$$

Hence it follows from (10) that

$$\begin{aligned} 0 &= \hat{\pi}_s^n - \frac{1}{n} \sum_{i=1}^n \frac{\hat{\gamma}_{is}}{\hat{\gamma}_s} \\ &= L_s^n(\hat{\Theta}^n) + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\hat{\pi}_s \hat{\psi}_{is}}{\hat{\psi}_s} - \frac{\hat{\gamma}_{is}}{\hat{\gamma}_s} \right\} \\ &= L_s^n(\hat{\Theta}^n) + \frac{1}{n} \sum_{i=1}^n \frac{\hat{\pi}_s \hat{\psi}_{is} \hat{\gamma}_s - \hat{\psi}_s \hat{\gamma}_{is}}{\hat{\psi}_s \hat{\gamma}_s}. \end{aligned} \quad (11)$$

According to Appendix A we have that

$$\hat{\gamma}_{is} = \hat{\pi}_s \hat{\psi}_{is}^n + O(1/n), \quad \hat{\gamma}_s = \hat{\psi}_s^n + O(1/n),$$

so the second term of (11) is of order $O(1/n)$. From Taylor's expansion the first term can be rewritten as

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

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

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

Similarly, we have

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

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

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

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 $B \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$, $\nabla \mathcal{L}^n(\tilde{\Theta}^n + \lambda(\hat{\Theta}^n - \tilde{\Theta}^n))(B)$ converges to $\Psi \mathbb{E}(HR(B))$, which is positive definite, where Ψ , H and $R(\cdot)$ are as defined in Appendix. Therefore, we obtain that $\hat{\Theta}^n = \tilde{\Theta}^n + O(1/n)$.

5 Experiments

As an example to demonstrate the convergence properties of the algorithm and the effect of the step size we consider the simple mixture of two known univariate normal densities $p_1(\cdot)$ and $p_2(\cdot)$ with means of μ_1 and μ_2 ; both have unit variance. The mixing coefficients are π and $1 - \pi$, respectively. The parameter π has a Beta prior distribution $\text{Beta}(a_0, b_0)$. We let f_{1i} and f_{2i} denote the known conditional densities evaluated at the observation y_i .

At the $(k + 1)$ st iteration, the variational posteriors of the hidden states and the parameter are

$$\begin{aligned} q_i(s_i) &= r_i^{s_i}(1 - r_i)^{1-s_i}, \quad \text{for } s_i \in \{0, 1\}, \quad i = 1, \dots, n, \\ q(\pi) &= \text{Beta}(\pi; a, b) = \pi^{a-1}(1 - \pi)^{b-1}/B(a, b), \quad 0 < \pi < 1, \end{aligned}$$

where

$$\begin{aligned} r_i &= p_{1i}\tilde{\pi}_1/(p_{1i}\tilde{\pi}_1 + p_{2i}\tilde{\pi}_2), \quad a = n\pi^{(k)} + a^0, \quad b = n - n\pi^{(k)} + b^0, \\ \tilde{\pi}_1 &= \exp\left\{\int q(\pi) \log \pi d\pi\right\}, \quad \tilde{\pi}_2 = \exp\left\{\int q(\pi) \log(1 - \pi) d\pi\right\}, \end{aligned}$$

and $\text{Beta}(\pi; a, b)$ denotes the $\text{Beta}(a, b)$ density function.

Therefore, the negative free energy (3) is

$$\begin{aligned} & \int \sum_{\{S\}} q(S, \pi) \log \frac{p(\pi, S, Y)}{q(S, \pi)} d\pi \\ &= \int \sum_{\{S\}} \prod_{i=1}^n q_i(s_i) q(\pi) \log p(\pi, S, Y) d\pi - \sum_{i=1}^n \sum_{\{s_i\}} q_i(s_i) \log q_i(s_i) \\ & \quad - \int q(\pi) \log q(\pi) d\pi. \end{aligned}$$

It is easily derived that

$$\begin{aligned} & \sum_{\{s_i\}} q_i(s_i) \log q_i(s_i) = r_i \log r_i + (1 - r_i) \log(1 - r_i), \\ & \int q(\pi) \log q(\pi) d\pi = (a - 1) \log \tilde{\pi}_1 + (b - 1) \log \tilde{\pi}_2 - \log B(a, b), \\ & \int \sum_{\{S\}} \prod_{i=1}^n q_i(s_i) q(\pi) \log p(\pi, S, Y) d\pi = \sum_{i=1}^n r_i \log p_{1i} + \sum_{i=1}^n (1 - r_i) \log p_{2i} \\ & \quad + \sum_{i=1}^n [(a_0 + r_i - 1) \log \tilde{\pi}_1 + (b_0 - r_i) \log \tilde{\pi}_2] - \log B(a_0, b_0). \end{aligned}$$

Hence, we obtain

$$\begin{aligned} & \int \sum_{\{S\}} q(S, \pi) \log \frac{p(\pi, S, Y)}{q(S, \pi)} d\pi \\ &= \sum_{i=1}^n \left[r_i \log p_{1i} + (1 - r_i) \log p_{2i} - r_i \log r_i - (1 - r_i) \log(1 - r_i) \right] \\ & \quad + (na_0 - n + \sum_{i=1}^n r_i - a + 1) \log \tilde{\pi}_1 + (nb_0 - \sum_{i=1}^n r_i - b + 1) \log \tilde{\pi}_2 \\ & \quad + \log B(a, b) - \log B(a_0, b_0). \end{aligned}$$

We generate a sample of total size 200 observations using $\pi = 0.65$, $\mu_1 = 2.5$ and $\mu_2 = 1.5$. For different sample sizes made up of the first n of the generated values the MLE and the variational Bayes estimate based on a Beta prior distribution for π with $a_0 = b_0 = 1$ are computed using the first n observations, and are plotted in Figure 1. It turns out that, when the sample size is small, there is a gap between the two estimates, but they come together very quickly as the sample size grows.

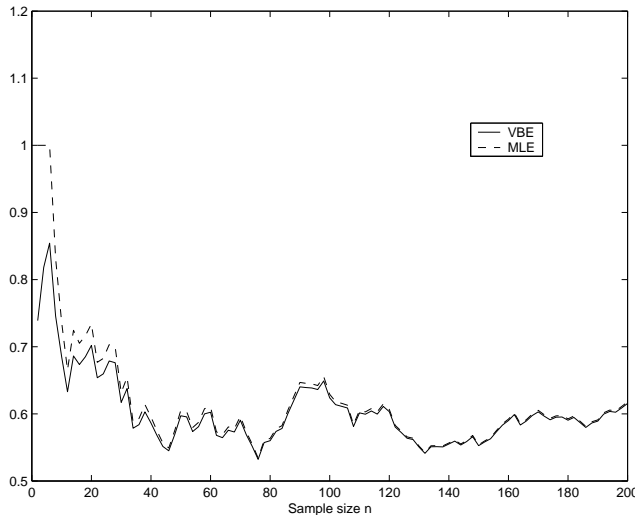


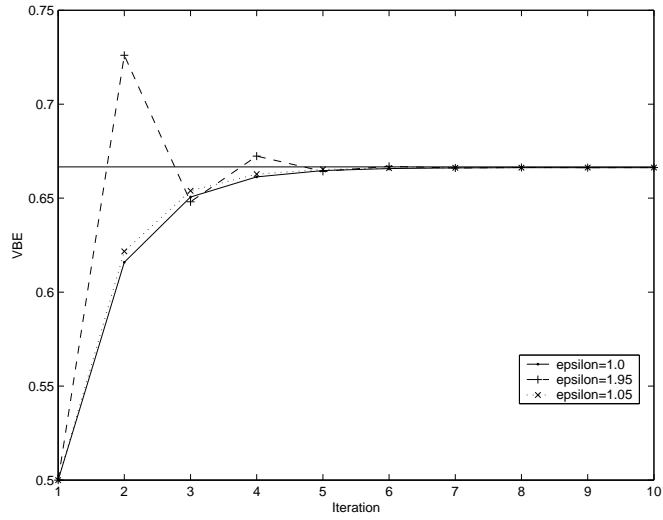
Figure 1: Variational Bayesian estimate of a mixing weight and MLE plotted against the sample size

Fixing the sample size at $n = 200$ and $\pi = 0.65$, we generate a sample based on different choices for the means of the normal distributions and compute the variational Bayesian estimate by the iterative algorithm for different values of ε . In Figure 2 we use $\mu_1 = 4.0$ and $\mu_2 = 1.5$, so that the two mixture components are widely separated. In this case changing ε from 1 does not accelerate the algorithm significantly; the slightly better ε is near 1 while ε near 2 hinders the convergence of the algorithm, a little.

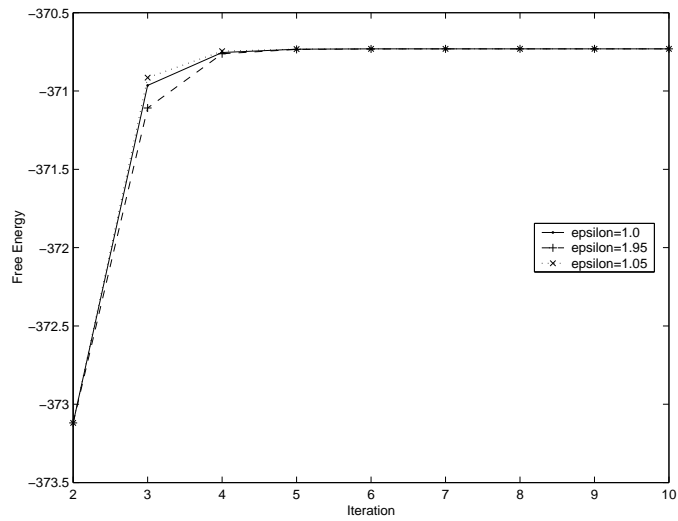
However, if the two components in the mixture are very similar, a choice for ε near 2 can improve the convergence rate considerably while the algorithm is accelerated only slightly for ε near 1. This is illustrated in Figure 3 for the case of $\mu_1 = 2.3$ and $\mu_2 = 1.5$.

6 Conclusion

Exact theoretical analysis of the quality of variational Bayes approximations is an important issue. In this paper we have investigated iterative algorithms for estimating parameters in normal mixture models. The results of this paper are twofold. First we proposed a generalised algorithm for obtaining the variational Bayesian estimates, and demonstrated by numerical experiments that for appropriate step size the generalised algorithm can be accelerated; if

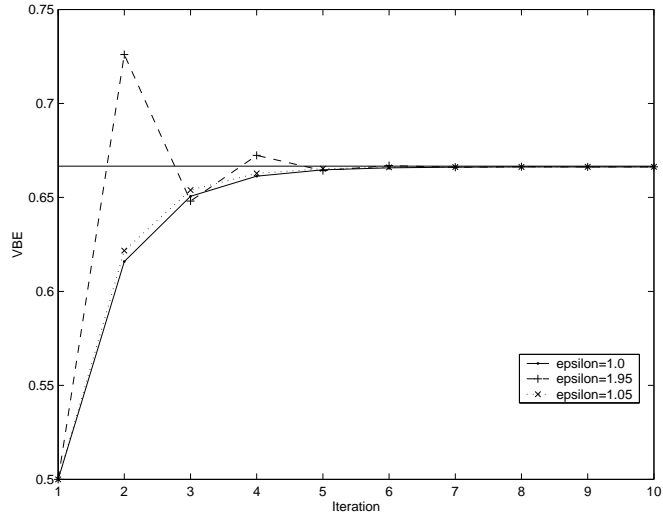


(a)

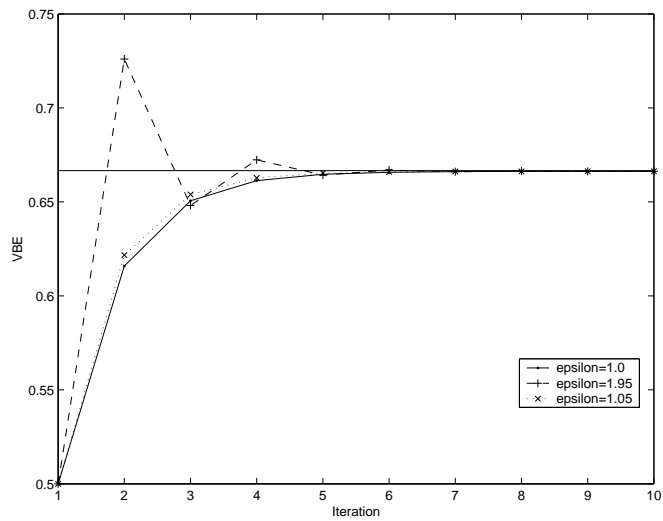


(b)

Figure 2: Variational Bayes estimates (a) and the corresponding negative free energies (b) for different step sizes ε for the example with widely separated components. The solid horizontal line in 2(a) indicates the MLE. The number of iterations to convergence of the algorithm for $\varepsilon = 1.0, 1.95, 1.05$ are 10, 9, 9 to achieve the same accuracy.



(a)



(b)

Figure 3: Variational Bayes estimates (a) and the corresponding negative free energies (b) for different step sizes ε for nearly identical components. The solid horizontal line in 3(a) indicates the MLE. The iteration times to convergence for $\varepsilon = 1.0, 1.95, 1.05$ are 44, 26, 43 to achieve the same accuracy.

the components in the mixture model are widely separated, the optimal ε is 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 estimate.

Secondly, we proved theoretically that the variational Bayes 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 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 be hold for other models. For example, we proved in Wang and Titterington (2003a) that the variational Bayes estimators for linear state space models are not always asymptotically consistent as the ‘sample size’ becomes large, because the factorised form destroys the intrinsic correlations between the hidden states in the models.

Since we have proved that the means of the variational posterior distribution converge to the maximum likelihood estimators, an interesting question arises naturally: do the variances (covariances) of the variational posteriors have the same properties, 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 (2004) we examine this problem and investigate the resulting performance of variational Bayes approximations in this context for interval estimation. It turns out that the covariance matrices corresponding to the variational Bayes approximation are normally ‘too small’ compared with those for the MLE, and therefore the variational Bayes approximations are unrealistically narrow.

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

Appendix: Proof of Theorem 1

In order that the derivative in the vector space $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$ makes sense, we endow them 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

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

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

$$\begin{aligned} \|\boldsymbol{\pi}\| &= \sum_{s=1}^m |\pi_s|, \text{ for } \boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_m \end{pmatrix} \in \mathcal{A}, \\ \|\boldsymbol{\mu}\| &= \sum_{s=1}^m \|\mu_s\|, \text{ for } \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} \in \mathcal{M}, \\ \|\boldsymbol{\Gamma}\| &= \sum_{s=1}^m \|\Gamma_s\|, \text{ for } \boldsymbol{\Gamma} = \begin{pmatrix} \Gamma_1 \\ \vdots \\ \Gamma_m \end{pmatrix} \in \mathcal{T}, \\ \|\Theta\| &= \|\boldsymbol{\pi}\| + \|\boldsymbol{\mu}\| + \|\boldsymbol{\Gamma}\|, \text{ for } \Theta = \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\mu} \\ \boldsymbol{\Gamma} \end{pmatrix} \in \mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}. \end{aligned}$$

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

$$\|\Phi\| = \sup_{\|B\|=1} \|\Phi(B)\|. \quad (12)$$

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 , which is a linear operator on $\mathcal{A} \oplus \mathcal{M} \oplus \mathcal{T}$; see Chapter X of Bhatia (1997).

Appendix A. First of all we prove the following lemma, which is a variant of the Laplace's approximation; see Chapter 4 of Evans and Swartz (2000).

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

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

Proof. From Taylor expansion we have

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

and thus

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

□

Appendix B. We now show that, if $\{X_n\}$ is a sequence of independent and identically distributed random variables and $F_n(\cdot) \rightarrow F_0(\cdot)$ uniformly, then, with probability 1,

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

In fact, we have that

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

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

Appendix C. Proof of the theorem: 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

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

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

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

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.

To obtain the limits of these derivatives as n tends to infinity, from (7) 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 $\boldsymbol{\Gamma}^*$.

Since, at Θ^* ,

$$\begin{aligned} q^*(\boldsymbol{\pi}) &= \mathcal{D}(\boldsymbol{\pi} : \lambda_1^*, \dots, \lambda_m^*), \\ q^*(\Gamma_s) &= \mathcal{W}(\Gamma_s : \nu_s^*, \Phi_s^*), \\ q^*(\mu_s | \Gamma_s) &= \mathcal{N}(\mu_s : \rho_s^*, \beta_s^* \Gamma_s), \end{aligned}$$

and

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

it is obvious that, as n tends to infinity, the mean of π_s corresponding to the density $q^*(\boldsymbol{\pi})$ is

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

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

$$\begin{aligned}& -\lambda_s^*\lambda_t^* / \left[\left(\sum_{s=1}^m \lambda_s^* \right)^2 \left(\sum_{s=1}^m \lambda_s^* + 1 \right) \right] \\ &= - (n\pi_s^* + \lambda^0)(n\pi_t^* + \lambda^0) / [(n + m\lambda^0)^2(n + m\lambda^0 + 1)] \\ &= O\left(\frac{1}{n}\right) \rightarrow 0,\end{aligned}$$

and the variance of π_s is

$$\lambda_s^* \left(\sum_{s=1}^m \lambda_s^* - \lambda_s^* \right) / \left[\left(\sum_{s=1}^m \lambda_s^* \right)^2 \left(\sum_{s=1}^m \lambda_s^* + 1 \right) \right] = O\left(\frac{1}{n}\right) \rightarrow 0;$$

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

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

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 Kronecker product.

After a straightforward calculation we obtain

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

$$\begin{aligned}
\nabla_{\pi_j} M_s(\Theta^*) &= \left[\left(\sum_{i=1}^n r_{is} \right) \left(\sum_{i=1}^n y_i \nabla_{\pi_j} r_{is} \right) - \left(\sum_{i=1}^n r_{is} y_i \right) \left(\sum_{i=1}^n \nabla_{\pi_j} r_{is} \right) \right] / \left(\sum_{i=1}^n r_{is} \right)^2, \\
\nabla_{\mu_j} M_s(\Theta^*) &= \left[\left(\sum_{i=1}^n r_{is} \right) \left(\sum_{i=1}^n y_i \nabla_{\mu_j} r_{is} \right) - \left(\sum_{i=1}^n r_{is} y_i \right) \left(\sum_{i=1}^n \nabla_{\mu_j} r_{is} \right) \right] / \left(\sum_{i=1}^n r_{is} \right)^2, \\
\nabla_{\Gamma_j} M_s(\Theta^*) &= \left[\left(\sum_{i=1}^n r_{is} \right) \left(\sum_{i=1}^n y_i \nabla_{\Gamma_j} r_{is} \right) - \left(\sum_{i=1}^n r_{is} y_i \right) \left(\sum_{i=1}^n \nabla_{\Gamma_j} r_{is} \right) \right] / \left(\sum_{i=1}^n r_{is} \right)^2, \\
\nabla_{\pi_j} S_s(\Theta^*) &= \left(\sum_{i=1}^n \nabla_{\pi_j} r_{is} \right) \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1} \\
&\quad - \left(\sum_{i=1}^n r_{is} \right) \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1} \left(\sum_{i=1}^n (y_i - \mu_s^*) (y_i - \mu_s^*)' \nabla_{\pi_j} r_{is} \right) \\
&\quad \times \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1}, \\
\nabla_{\mu_j} S_s(\Theta^*) v_j &= \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1} \left(\sum_{i=1}^n \nabla_{\mu_j} r_{is} v_j \right) \\
&\quad - \left(\sum_{i=1}^n r_{is} \right) \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1} \left(\sum_{i=1}^n [(y_i - \mu_s^*) (y_i - \mu_s^*)' \nabla_{\mu_j} r_{is} v_j \right. \\
&\quad \left. - r_{is} v_j (y_i - \mu_s^*)' \delta_{sj} - r_{is} (y_i - \mu_s^*) v_j' \delta_{sj}] \right) \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1}, \\
\nabla_{\Gamma_j} S_s(\Theta^*) W_j &= \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1} \left(\sum_{i=1}^n \nabla_{\Gamma_j} r_{is} W_j \right) \\
&\quad - \left(\sum_{i=1}^n r_{is} \right) \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1} \left(\sum_{i=1}^n (y_i - \mu_s^*) (y_i - \mu_s^*)' \nabla_{\Gamma_j} r_{is} W_j \right) \\
&\quad \times \left(\sum_{i=1}^n r_{is} (y_i - \mu_s^*) (y_i - \mu_s^*)' \right)^{-1},
\end{aligned}$$

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

Now we study their limits with the help of Appendix B. All of the corresponding functions $F_n(X_i)$, such as r_{is} and their derivatives, are of the following form

$$a_n (y_i - b_n)^k e^{-(y_i - b_n)' C_n (y_i - b_n) / 2}, \quad (13)$$

where $k \in \{0, 1, 2, 3, 4\}$ and, as n tends to infinity, it follows from Appendix D that

$$a_n \rightarrow a_0, \quad b_n \rightarrow b_0, \quad C_n \rightarrow C_0.$$

Since the first-order derivatives of the function (13) with respect to a_n , b_n and c_n are bounded in y_i , (13) converges uniformly in y_i to

$$a_0(y_i - b_0)^k e^{-(y_i - b_0)'C_0(y_i - b_0)/2}.$$

Moreover, if we denote by y any random vector distributed according to the probability density of the form (2) and let

$$\phi_s = |\Gamma_s^*|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2}(y - \mu_s^*)' \Gamma_s^* (y - \mu_s^*) \right\}, \quad \text{and} \quad \phi = \sum_{s=1}^m \pi_s^* \phi_s,$$

we have that

$$\mathbb{E} \left(\frac{\phi_s}{\phi} \right) = \int \frac{\phi_s}{\phi} p(y|\Theta^*) dy = 1, \quad (14a)$$

$$\mathbb{E} \left(\frac{\phi_s}{\phi} (y - \mu_s^*) \right) = \int \frac{\phi_s (y - \mu_s^*)}{\phi} p(y|\Theta^*) dy = 0, \quad (14b)$$

$$\mathbb{E} \left(\frac{\phi_s}{\phi} (y - \mu_s^*) (y - \mu_s^*)' \right) = \int \phi_s (y - \mu_s^*) (y - \mu_s^*)' dy = \Gamma_s^{*-1}, \quad (14c)$$

$$\mathbb{E} \left(\frac{\phi_s}{\phi} (y - \mu_s^*) (y - \mu_s^*)' (y - \mu_s^*) \right) = 0, \quad (14d)$$

$$\mathbb{E} \left(\frac{\phi_s}{\phi} [(y - \mu_s^*) (y - \mu_s^*)']^2 \right) = 3\Gamma_s^{*-2}. \quad (14e)$$

For $s = 1, \dots, m$, we introduce the notation

$$\alpha_s^1 = \frac{\phi_s}{\phi}, \quad \alpha_s^2 = \pi_s^* \frac{\phi_s}{\phi} \Gamma_s^* (y - \mu_s^*),$$

$$\alpha_s^3 = \frac{1}{2} \pi_s^* \frac{\phi_s}{\phi} [\Gamma_s^* (y - \mu_s^*) (y - \mu_s^*)' \Gamma_s^* - \Gamma_s^*],$$

$$\Lambda = \text{diag}(\pi_s^*), \quad \Omega = \text{diag}(\pi_s^{*-1} \Gamma_s^{*-1}), \quad \Sigma = \text{diag}(2\pi_s^{*-1} \Gamma_s^{*-2}).$$

Then, applying Appendices A, B and D and after a very careful calculation, we obtain

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

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

$$\nabla_{\Gamma}\Pi(\Theta^*)\mathbf{W} \rightarrow -\Lambda\mathbb{E} \begin{pmatrix} \alpha_1^1 \\ \vdots \\ \alpha_m^1 \end{pmatrix} \left\{ \sum_{s=1}^m \text{tr}\{\alpha_s^3 W_s\} \right\},$$

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

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

$$\nabla_{\Gamma}M(\Theta^*)\mathbf{W} \rightarrow -\Omega\mathbb{E} \begin{pmatrix} \alpha_1^2 \\ \vdots \\ \alpha_m^2 \end{pmatrix} \left\{ \sum_{s=1}^m \text{tr}\{\alpha_s^3 W_s\} \right\},$$

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

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

$$\nabla_{\Gamma}S(\Theta^*)\mathbf{W} \rightarrow I_{md}\mathbf{W} - \Sigma\mathbb{E} \begin{pmatrix} \alpha_1^3 \\ \vdots \\ \alpha_m^3 \end{pmatrix} \left\{ \sum_{s=1}^m \text{tr}\{\alpha_s^3 W_s\} \right\}.$$

Set

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

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

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 of \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 $\langle A, B\rangle = \text{tr}\{AB\}$. 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 (12) is actually 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 E.

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

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

the iterative procedure (9) converges locally to the true value Θ^* as n approaches infinity .

Appendix D. 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

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

and

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

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

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

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

$$\begin{aligned}\nabla_{\pi_s} \int q(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi} &= \nabla_{\pi_s} \left[\Gamma^{-1}(\lambda_s) \int_0^\infty z^{\lambda_s-1} e^{-z} \log z dz \right] \\ &= n\Gamma^{-2}(\lambda_s) \left[\int_0^\infty z^{\lambda_s-1} e^{-z} \log^2 z dz \int_0^\infty z^{\lambda_s-1} e^{-z} dz \right. \\ &\quad \left. - \left(\int_0^\infty z^{\lambda_s-1} e^{-z} \log z dz \right)^2 \right].\end{aligned}\tag{15}$$

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 Laplace's approximation yields, see for example Chapter 4 of Evans and Swartz (2000),

$$\begin{aligned} \int_0^{\infty} e^{-(\lambda_s-1)(u-\log u)} h(u) du &= (2\pi)^{1/2} e^{-(\lambda_s-1)} \left\{ h(\hat{u}) (\lambda_s - 1)^{-1/2} \right. \\ &\quad \left. + (\lambda_s - 1)^{-3/2} [a_1 h(\hat{u}) - a_2 h'(\hat{u}) + a_3 h''(\hat{u})] + o((\lambda_s - 1)^{-3/2}) \right\}, \end{aligned}$$

where

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

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

$$\begin{aligned} \int_0^{\infty} z^{\lambda_s-1} e^{-z} dz &= (2\pi)^{1/2} e^{-(\lambda_s-1)} (\lambda_s - 1)^{\lambda_s} \\ &\quad \left\{ (\lambda_s - 1)^{-1/2} + a_1 (\lambda_s - 1)^{-3/2} + o((\lambda_s - 1)^{-3/2}) \right\}, \end{aligned}$$

$$\begin{aligned} \int_0^{\infty} z^{\lambda_s-1} e^{-z} \log z dz &= (2\pi)^{1/2} e^{-(\lambda_s-1)} (\lambda_s - 1)^{\lambda_s} \\ &\quad \left\{ (\lambda_s - 1)^{-1/2} \log(\lambda_s - 1) + (\lambda_s - 1)^{-3/2} [a_1 \log(\lambda_s - 1) - a_2 - a_3] \right. \\ &\quad \left. + o((\lambda_s - 1)^{-3/2}) \right\}, \end{aligned}$$

$$\begin{aligned} \int_0^{\infty} z^{\lambda_s-1} e^{-z} \log^2 z dz &= (2\pi)^{1/2} e^{-(\lambda_s-1)} (\lambda_s - 1)^{\lambda_s} \\ &\quad \left\{ (\lambda_s - 1)^{-1/2} \log^2(\lambda_s - 1) + (\lambda_s - 1)^{-3/2} [a_1 \log^2(\lambda_s - 1) - 2a_2 \log(\lambda_s - 1) \right. \\ &\quad \left. + 2a_3(1 - \log(\lambda_s - 1))] + o((\lambda_s - 1)^{-3/2}) \right\}. \end{aligned}$$

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

$$\begin{aligned} \nabla_{\pi_s} \int q(\boldsymbol{\pi}) \log \pi_s d\boldsymbol{\pi} &\sim \frac{2a_3 n (\lambda_s - 1)^{-2} + o((\lambda_s - 1)^{-1})}{(\lambda_s - 1)^{-1} + 2a_1 (\lambda_s - 1)^{-2} + o((\lambda_s - 1)^{-2})} \\ &\rightarrow \frac{1}{\pi_s}. \end{aligned}$$

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

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

Along the same lines as (15), we obtain

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

and it is easy to show that

$$\nabla_{\pi_s} \log |\Phi_s| = \text{tr} \left(\Phi_s^{-1} \nabla_{\pi_s} \Phi_s \right) \rightarrow \text{tr} \left([\pi_s \Gamma_s^{-1}]^{-1} \Gamma_s^{-1} \right) = \frac{d}{\pi_s}.$$

Therefore,

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

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

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

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

$$\begin{aligned} & \langle B, \mathbb{E}(HR(B)) \rangle \\ &= \mathbb{E} \left(\sum_{s=1}^m \alpha_s^1 u_s + \sum_{s=1}^m (\alpha_s^2)' v_s + \sum_{s=1}^m \text{tr} \{ \alpha_s^3 W_s \} \right)^2 \\ &= \mathbb{E} \left(\sum_{s=1}^m \pi_s^* \frac{\phi_{is}}{\phi_i} \left[u_s \pi_s^{*-1} + (y_i - \mu_s^*)' \Gamma_s^* v_s \right. \right. \\ & \quad \left. \left. + \text{tr} \left\{ \frac{1}{2} [\Gamma_s^* (y_i - \mu_s^*) (y_i - \mu_s^*)' \Gamma_s^* - \Gamma_s^*] W_s \right\} \right] \right)^2. \end{aligned}$$

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

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

where the second-last equality comes from (14) and the fact

$$\mathbb{E} \left[\pi_s^* \frac{\phi_s}{\phi} \left(\text{tr} \left\{ \frac{1}{2} [\Gamma_s^* (y - \mu_s^*) (y - \mu_s^*)' \Gamma_s^* - \Gamma_s^*] W_s \right\} \right)^2 \right] = \text{tr} \left\{ \frac{1}{2} \pi_s^* W_s \Gamma_s^{*2} W_s \right\}$$

which can be verified by expanding the matrices into the expression of their components and after a straightforward reduction.

References

- Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In Prade, H. and Laskey, K., editors, *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, pages 21–30, Stockholm, Sweden. Morgan Kaufmann Publishers.
- Attias, H. (2000). A variational Bayesian framework for graphical models. In Solla, S., Leen, T., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, Cambridge, MA.
- Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London.
- Bhatia, R. (1997). *Matrix Analysis*. Springer-Verlag, New York.
- Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley, New York.
- Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, New York.
- Ghahramani, Z. and Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In Solla, S., Leen, T., and Muller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, Cambridge, MA.
- Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 507–513. MIT Press, Cambridge, MA.
- Hall, P., Humphreys, K., and Titterton, 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.

- Humphreys, K. and Titterington, D. M. (2000). Approximate Bayesian inference for simple mixtures. In Bethlehem, J. G. and van der Heijden, P. G. M., editors, *COMPSTAT2000*, pages 331–336. Physica-Verlag, Heidelberg.
- Jordan, M. I. (2004). Graphical models. *Statistical Science*, 19(1):140–155.
- MacKay, D. J. C. (1997). Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.
- Penny, W. D. and Roberts, S. J. (2000). Variational Bayes for 1-dimensional mixture models. Technical Report PARG-2000-01, Oxford University.
- Peters, B. C. and Walker, H. F. (1978). An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.*, 35:362–378.
- Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–239.
- Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, (1):128–139.
- Wang, B. and Titterington, D. M. (2003a). Lack of consistency of mean field and variational Bayes approximations for state space models. Technical Report 03-5, University of Glasgow. <http://www.stats.gla.ac.uk/Research/TechRep2003/03-5.pdf>.
- Wang, B. and Titterington, D. M. (2003b). 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>.
- Wang, B. and Titterington, D. M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. Technical Report 04-19, University of Glasgow. http://www.stats.gla.ac.uk/Research/TechRep2004/04_19.pdf.