
Inadequacy of interval estimates corresponding to variational Bayesian approximations

Bo Wang

Department of Statistics
University of Glasgow
Glasgow G12 8QQ
Scotland, U.K.

D. M. Titterington

Department of Statistics
University of Glasgow
Glasgow G12 8QQ
Scotland, U.K.

Abstract

In this paper we investigate the properties of the covariance matrices associated with variational Bayesian approximations, based on data from mixture models, and compare them with the true covariance matrices, corresponding to Fisher information matrices. It is shown that the covariance matrices from the variational Bayes approximations are normally ‘too small’ compared with those for the maximum likelihood estimator, so that resulting interval estimates for the parameters will be unrealistically narrow, especially if the components of the mixture model are not well separated.

1 INTRODUCTION

A standard paradigm for learning about the parameters of latent variable models from data is that of maximum likelihood. However, maximum likelihood is well known for its tendency to overfit the data. On the other hand, the Bayesian framework averages over all possible settings of the model parameters. As a result Bayesian inference does not suffer from overfitting, and, moreover, prior knowledge can be incorporated naturally. Unfortunately, for most models of interest involving missing data a full Bayesian analysis requires the computation of the posterior distribution for a collection of unknown quantities, including parameters and latent variables, which often leads to intractable calculations because complicated multiple integrations are involved. The use of Markov chain Monte Carlo methods for numerical integration helps to side-step this problem, but this is clearly quite expensive, in terms of time and storage. Moreover MCMC algorithms can still exhibit conceptual and technical difficulties, for example in the assessment of the convergence of the chain to its stationary distribution.

Recently, a deterministic approximate approach to the intractable Bayesian learning problem, the variational Bayesian approximation, has been introduced in the machine learning community, and is widely recognised to be effective and promising in a variety of models, such as hidden Markov models (MacKay (1997)), graphical models (Attias (1999, 2000)), mixture models (Humphreys and Titterington (2000); Penny and Roberts (2000)), mixtures of factor analysers (Ghahramani and Beal (2000)) and state space models (Ghahramani and Beal (2001); Beal (2003)). A general formulation of the variational approach is described in Jordan (2004). The variational Bayes approach facilitates analytical calculation of approximate posterior distributions over the hidden variables, parameters and structures. They are computed via an iterative algorithm that is closely related to the Expectation-Maximisation (EM) algorithm and so its convergence is guaranteed. Empirically, variational Bayesian approximations have often been shown to perform well in earlier contributions, but it has also been noticed that this approach may underestimate the spread of the posterior distributions for some particular examples (Humphreys and Titterington (2000); Consonni and Marin (2004)), so its validity has still to be assessed properly: exact theoretical analysis of the quality of the method needs to be studied.

Some initial investigations have been implemented by the authors in Wang and Titterington (2003) and Wang and Titterington (2004b). It was shown theoretically that the iterative algorithm for obtaining the variational Bayes approximation for the parameters of Gaussian mixture models converges locally to the maximum likelihood estimator at the rate of $O(1/n)$ in the large sample limit. Later in Wang and Titterington (2004a) we proved local convergence of variational approximation algorithms for more general models, namely exponential family models with missing values, and showed that the variational posterior distribution for the parameters is asymptotically normal with the same mean but a different covariance ma-

trix compared with those for the maximum likelihood estimator.

Since the maximum likelihood estimators and posterior distributions are also asymptotically normal (see for instance Walker (1969), Chen (1985) and Ghosal et al. (1995)), an interesting problem is how these two limiting normal distributions can be compared. From the early results on local convergence of variational approximations, one can note that they have the same mean (i.e. the true value). However, their covariance matrices do not appear to be equal. In the context of Gaussian mixture models, in this paper we study the covariance matrices associated with variational Bayesian approximations, which dictate the performance of variational Bayes approximations for interval estimates, and compare them with the true covariance matrices, as given in terms of Fisher information matrices. We show that the covariance matrices from the variational Bayes approximation are normally ‘too small’ compared with those for the MLE, so that resulting interval estimates for the parameters will be too narrow, especially if the components of the mixture model are not well separated. Some numerical examples illustrate the theoretical analysis.

2 THE MIXTURE MODEL AND THE VARIATIONAL APPROXIMATION

We consider a model in which we have a mixture of m multivariate Gaussian densities p_1, p_2, \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

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

where $y_i \in \mathbb{R}^d$ denotes the i th observed data vector, and s_i indicates the hidden component that generated it. The components are labelled by $s = 1, 2, \dots, m$, and component s has mixing coefficient $\pi_s = p(s_i = s|\Theta)$, for any i and $s = 1, 2, \dots, m-1$. Consequently $\pi_m \triangleq p(s_i = m|\Theta) = 1 - \sum_{s=1}^{m-1} \pi_s$. We write the parameters collectively as

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_{m-1} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \text{vec}(\Gamma_1) \\ \vdots \\ \text{vec}(\Gamma_m) \end{pmatrix},$$

and $\Theta = (\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\Gamma}')$. Here $\text{vec}(A)$ is defined as the stacked columns of A .

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 Gaussian, 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 Gaussian distribution.

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 approximating density $q(S, \Theta)$ for $p(S, \Theta|Y)$, which factorises as

$$q(S, \Theta) = q^{(S)}(S)q^{(\Theta)}(\Theta), \quad (2)$$

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

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

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

$$q^{(S)}(S) = \prod_{i=1}^n q_i^{(S)}(s_i),$$

$$q^{(\Theta)}(\Theta) = q(\boldsymbol{\pi}) \prod_{s=1}^m q(\mu_s|\Gamma_s)q(\Gamma_s).$$

As in Attias (1999, 2000), Ghahramani and Beal (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 Gaussian, 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:

$$\lambda_s = \sum_{i=1}^n r_{is} + \lambda^0, \quad (4)$$

$$\rho_s = \left(\sum_{i=1}^n r_{is} y_i + \beta^0 \rho^0 \right) / \left(\sum_{i=1}^n r_{is} + \beta_0 \right), \quad (5)$$

$$\beta_s = \sum_{i=1}^n r_{is} + \beta^0, \quad \nu_s = \sum_{i=1}^n r_{is} + \nu^0, \quad (6)$$

and

$$\begin{aligned} \Phi_s &= \Phi^0 + \sum_{i=1}^n r_{is}(y_i - \bar{\mu}_s)(y_i - \bar{\mu}_s)' \\ &+ \left[\left(\sum_{i=1}^n r_{is} \right) \beta^0 (\bar{\mu}_s - \rho^0) (\bar{\mu}_s - \rho^0)' \right] / \left(\sum_{i=1}^n r_{is} + \beta_0 \right), \end{aligned} \quad (7)$$

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

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

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 ,

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

3 THE CONVERGENCE OF VARIATIONAL BAYES APPROXIMATIONS AND ASSOCIATED COVARIANCE MATRICES

Suppose that the true value of the parameter Θ is Θ^* . At the k th iteration of the iterative procedure (i) (ii), we define the variational Bayesian estimates $\boldsymbol{\pi}^{(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Gamma}^{(k)}$ of the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ as their variational posterior means corresponding to the distributions $q(\boldsymbol{\pi})$, $q(\boldsymbol{\mu}_s | \Gamma_s)$ and $q(\Gamma_s)$ at the current iteration, thus the iterative procedure (i) (ii) suggests the following algorithm: starting with some initial values $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\Gamma}^{(0)}$, the variational Bayesian estimates are computed recursively by

$$\boldsymbol{\pi}_s^{(k+1)} = M_1(\boldsymbol{\pi}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Gamma}^{(k)}), \quad (8a)$$

$$\boldsymbol{\mu}_s^{(k+1)} = M_2(\boldsymbol{\pi}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Gamma}^{(k)}), \quad (8b)$$

$$\boldsymbol{\Gamma}_s^{(k+1)} = M_3(\boldsymbol{\pi}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Gamma}^{(k)}), \quad (8c)$$

where the maps M_1 , M_2 and M_3 represent the iterative procedure in (i) (ii).

In Wang and Titterton (2004b), the following convergence property of the variational Bayes estimates defined by (8) has been proved.

Lemma 1. *With probability 1 as n approaches infinity, $\boldsymbol{\pi}^{(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Gamma}^{(k)}$ converge locally to the true values $\boldsymbol{\pi}^*$, $\boldsymbol{\mu}^*$ and $\boldsymbol{\Gamma}^*$; that is, they converge to the true values whenever the starting values are sufficiently near to $\boldsymbol{\pi}^*$, $\boldsymbol{\mu}^*$ and $\boldsymbol{\Gamma}^*$.*

Remark 1. *For general mixture models, 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 local convergence was proved.*

Denote by \otimes the Kronecker product. By (4)-(7) and the convergence property given by Lemma 1, one can easily obtain that, as n tends to infinity, $n\text{Cov}(\boldsymbol{\pi}) \rightarrow$

$$\begin{pmatrix} \pi_1^*(1 - \pi_1^*) & & & -\pi_s^* \pi_k^* \\ & \ddots & & \\ -\pi_s^* \pi_k^* & & & \pi_{m-1}^*(1 - \pi_{m-1}^*) \end{pmatrix} \triangleq \Lambda,$$

$$\begin{aligned} n\text{Cov}(\Gamma_s) &= 2n\nu_s^{(k)} (\Phi_s^{(k)})^{-1} \otimes (\Phi_s^{(k)})^{-1} \\ &= 2\nu_s^{(k)} (\Phi_s^{(k)} \otimes \Phi_s^{(k)})^{-1} \rightarrow 2\pi_s^{*-1} (\Gamma_s^* \otimes \Gamma_s^*), \end{aligned}$$

$$\begin{aligned} \mathbb{E}(\boldsymbol{\mu}_s) &= \int \boldsymbol{\mu}_s q^{(k)}(\boldsymbol{\mu}_s) d\boldsymbol{\mu}_s \\ &= \int \boldsymbol{\mu}_s q^{(k)}(\boldsymbol{\mu}_s | \Gamma_s) q^{(k)}(\Gamma_s) d\Gamma_s d\boldsymbol{\mu}_s = \boldsymbol{\rho}_s^{(k)}, \end{aligned}$$

and it follows that

$$\begin{aligned} n\text{Cov}(\boldsymbol{\mu}_s) &= n \int (\boldsymbol{\mu}_s - \boldsymbol{\rho}_s^{(k)}) (\boldsymbol{\mu}_s - \boldsymbol{\rho}_s^{(k)})' q^{(k)}(\boldsymbol{\mu}_s) d\boldsymbol{\mu}_s \\ &= n \int (\boldsymbol{\mu}_s - \boldsymbol{\rho}_s^{(k)}) (\boldsymbol{\mu}_s - \boldsymbol{\rho}_s^{(k)})' q^{(k)}(\boldsymbol{\mu}_s | \Gamma_s) q^{(k)}(\Gamma_s) d\Gamma_s d\boldsymbol{\mu}_s \\ &= n \int (\beta_s^{(k)} \Gamma_s^{(k)})^{-1} q^{(k)}(\Gamma_s) d\Gamma_s \\ &= n(\beta_s^{(k)})^{-1} (\nu_s^{(k)} - m - 1)^{-1} \Phi_s^{(k)} \rightarrow \pi_s^{*-1} \Gamma_s^{*-1}. \end{aligned}$$

Moreover, letting $\boldsymbol{\mu}_s^j$ and $\Gamma_s^{t\tau}$ denote any elements of $\boldsymbol{\mu}_s$ and Γ_s , respectively, we have

$$\begin{aligned} n\text{Cov}(\boldsymbol{\mu}_s^j, \Gamma_s^{t\tau}) &= n \int [\boldsymbol{\mu}_s^j - \mathbb{E}(\boldsymbol{\mu}_s^j)] [\Gamma_s^{t\tau} - \mathbb{E}(\Gamma_s^{t\tau})] \\ &\quad \cdot q^{(k)}(\boldsymbol{\mu}_s) q^{(k)}(\Gamma_s) d\boldsymbol{\mu}_s^j d\Gamma_s^{t\tau} \\ &= n \int [\boldsymbol{\mu}_s^j - \boldsymbol{\rho}_s^{(k);j}] [\Gamma_s^{t\tau} - \mathbb{E}(\Gamma_s^{t\tau})] \\ &\quad \cdot q^{(k)}(\boldsymbol{\mu}_s | \Gamma_s) (q^{(k)}(\Gamma_s))^2 d\boldsymbol{\mu}_s^j d\Gamma_s^{t\tau} = 0, \end{aligned}$$

and the other covariances between $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ are zero, by assumption (2).

Define

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

Then the covariance matrix of Θ associated with the variational posterior distributions is such that

$$n\text{Cov}(\Theta) \rightarrow \begin{pmatrix} \Lambda & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Omega & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma \end{pmatrix} \triangleq \Psi. \quad (9)$$

4 COMPARISON OF VARIATIONAL COVARIANCE MATRICES WITH FISHER INFORMATION MATRICES

In this section we first give an explicit expression for the Fisher information matrix associated with our mixture model, and then compare it with the covariance matrix associated with variational Bayes approximations, which is crucial for the performance of interval estimates based on variational Bayes approximations.

In the sequel, we denote by y any random vector distributed according to the probability density of the form (1). Thus the Fisher information matrix per observation is given by

$$I(\Theta) = \int_{\mathbb{R}^d} [\nabla \log p(y|\Theta)] [\nabla \log p(y|\Theta)]' p(y|\Theta) dy. \quad (10)$$

The Fisher information matrix plays an important role in determining the asymptotic distribution of maximum likelihood estimators. Under quite mild conditions, Redner and Walker (1984) stated the following property of asymptotic normality for the maximum likelihood estimator for mixture models.

Theorem 1. *Let $\tilde{\Theta}^n$ be the strongly consistent MLE of the parameter Θ . Then $\sqrt{n}(\tilde{\Theta}^n - \Theta^*)$ is asymptotically normally distributed with mean zero and covariance matrix $I(\Theta^*)^{-1}$.*

Let $L(\Theta) = \log p(y|\Theta)$ and, for $s = 1, \dots, m$, let

$$\alpha_s = \frac{p_s(y|\Theta)}{p(y|\Theta)}, \quad \delta_s = y - \mu_s, \\ \sigma_s = \text{vec}[\Gamma_s^{-1} - (y - \mu_s)(y - \mu_s)'].$$

One should bear in mind the dependencies of α_s , δ_s and σ_s on Θ or its components, which are omitted for the sake of clarity.

After a straightforward calculation we obtain

$$\frac{\partial L}{\partial \pi_s} = \alpha_s - \alpha_m, \quad s = 1, \dots, m-1, \\ \frac{\partial L}{\partial \mu_s} = \pi_s \alpha_s \Gamma_s \delta_s, \quad s = 1, \dots, m, \\ \frac{\partial L}{\partial \text{vec}(\Gamma_s)} = \frac{1}{2} \pi_s \alpha_s \sigma_s, \quad s = 1, \dots, m.$$

Let

$$Q = \begin{pmatrix} \alpha_1 - \alpha_m \\ \vdots \\ \alpha_{m-1} - \alpha_m \\ \pi_1 \alpha_1 \Gamma_1 \delta_1 \\ \vdots \\ \pi_m \alpha_m \Gamma_m \delta_m \\ \frac{1}{2} \pi_1 \alpha_1 \sigma_1 \\ \vdots \\ \frac{1}{2} \pi_m \alpha_m \sigma_m \end{pmatrix}.$$

Then the Fisher information matrix (10) can be rewritten as

$$I(\Theta) = \int_{\mathbb{R}^d} Q Q' p(y|\Theta) dy = \mathbb{E}[Q Q']. \quad (11)$$

The following lemma is a corollary of Schwarz's inequality, which has been used in Peters and Walker (1978).

Lemma 2. *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 any $\{\xi_s\}_{s=1, \dots, m}$.

Moreover, after a tedious calculation the following equalities can be verified.

Lemma 3. *At the true value Θ^* , we have that, for $s = 1, \dots, m$,*

$$\mathbb{E}(\alpha_s) = 1, \quad \mathbb{E}(\alpha_s \delta_s) = 0, \quad (12a)$$

$$\mathbb{E}(\alpha_s \sigma_s) = 0, \quad \mathbb{E}(\alpha_s \delta_s \sigma_s') = 0, \quad (12b)$$

$$\mathbb{E}(\alpha_s \sigma_s \sigma_s') = 2(\Gamma_s^* \otimes \Gamma_s^*)^{-1}. \quad (12c)$$

Now we state the main result of this section.

Theorem 2. *If Ψ is defined as in (9), then the Fisher information matrix satisfies*

$$I(\Theta^*)^{-1} \geq \Psi, \quad (13)$$

by which it is meant that $I(\Theta^*)^{-1} - \Psi$ is nonnegative definite.

Proof. Obviously, Ψ is positive definite, and thus it is sufficient to show that

$$\Theta' I(\Theta^*) \Theta \leq \Theta' \Psi^{-1} \Theta = \mathbf{u}' \Lambda^{-1} \mathbf{u} + \mathbf{v}' \Omega^{-1} \mathbf{v} + \mathbf{W}' \Sigma^{-1} \mathbf{W}$$

for any

$$\Theta = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{W} \end{pmatrix} = \begin{pmatrix} u_1 \\ \vdots \\ u_{m-1} \\ v_1 \\ \vdots \\ v_m \\ \text{vec}(W_1) \\ \vdots \\ \text{vec}(W_m) \end{pmatrix},$$

where u_s , v_s and W_s are elements of the vector spaces \mathbb{R} , \mathbb{R}^d and the set of all real, symmetric $d \times d$ matrices, respectively, for each s .

In fact, by (11) one has that

$$\begin{aligned} \Theta' I(\Theta^*) \Theta &= \Theta' \mathbb{E}(QQ') \Theta \\ &= \mathbb{E} \left\{ \sum_{s=1}^{m-1} (\alpha_s - \alpha_m) u_s + \sum_{s=1}^m \pi_s^* \alpha_s \delta'_s \Gamma_s^* v_s \right. \\ &\quad \left. + \sum_{s=1}^m \frac{1}{2} \pi_s^* \alpha_s \sigma'_s \text{vec}(W_s) \right\}^2 \\ &= \mathbb{E} \left\{ \sum_{s=1}^m \pi_s^* \alpha_s \left[u_s \pi_s^{*-1} + \delta'_s \Gamma_s^* v_s + \frac{1}{2} \sigma'_s \text{vec}(W_s) \right] \right\}^2, \end{aligned}$$

where we have defined $u_m = -\sum_{s=1}^{m-1} u_s$.

Noting that $\sum_{s=1}^m \pi_s^* \alpha_s = 1$ and applying Lemma 2, we have

$$\begin{aligned} &\Theta' \mathbb{E}(QQ') \Theta \\ &\leq \mathbb{E} \left\{ \sum_{s=1}^m \pi_s^* \alpha_s \left[u_s \pi_s^{*-1} + \delta'_s \Gamma_s^* v_s + \frac{1}{2} \sigma'_s \text{vec}(W_s) \right] \right\}^2 \\ &= \sum_{s=1}^m \mathbb{E} \left\{ u_s^2 \pi_s^{*-1} \alpha_s + \pi_s^* \alpha_s [\delta'_s \Gamma_s^* v_s]^2 \right. \\ &\quad \left. + \frac{1}{4} \pi_s^* \alpha_s [\sigma'_s \text{vec}(W_s)]^2 + 2u_s \alpha_s \delta'_s \Gamma_s^* v_s \right. \\ &\quad \left. + u_s \alpha_s \sigma'_s \text{vec}(W_s) + \pi_s^* \alpha_s \delta'_s \Gamma_s^* v_s \sigma'_s \text{vec}(W_s) \right\} \\ &= \sum_{s=1}^m \mathbb{E} \{ u_s^2 \pi_s^{*-1} \alpha_s \} + \sum_{s=1}^m \mathbb{E} \left\{ \pi_s^* \alpha_s [\delta'_s \Gamma_s^* v_s]^2 \right\} \\ &\quad + \sum_{s=1}^m \mathbb{E} \left\{ \frac{1}{4} \pi_s^* \alpha_s [\sigma'_s \text{vec}(W_s)]^2 \right\} \\ &\triangleq I_1 + I_2 + I_3, \end{aligned}$$

where the last equality holds since the cross terms average to zero, by (12).

Clearly, one has

$$I_1 = \sum_{s=1}^m \mathbb{E} \{ \alpha_s u_s^2 \pi_s^{*-1} \} = \sum_{s=1}^m u_s^2 \pi_s^{*-1}.$$

Note that $u_m = -\sum_{s=1}^{m-1} u_s$ and

$$\Lambda^{-1} = \begin{pmatrix} \pi_1^{*-1} + \pi_m^{*-1} & & & \\ & \ddots & & \\ & & \pi_m^{*-1} & \\ \pi_m^{*-1} & & & \pi_{m-1}^{*-1} + \pi_m^{*-1} \end{pmatrix},$$

from which it can be easily checked that $\mathbf{u}' \Lambda^{-1} \mathbf{u} = I_1$.

By (12),

$$I_2 = \sum_{s=1}^m \mathbb{E} \left\{ \pi_s^* \alpha_s [\delta'_s \Gamma_s^* v_s]^2 \right\} = \sum_{s=1}^m v'_s \pi_s^* \Gamma_s^* v_s = \mathbf{v}' \Omega^{-1} \mathbf{v}.$$

Finally,

$$\begin{aligned} I_3 &= \sum_{s=1}^m \mathbb{E} \left\{ \frac{1}{4} \pi_s^* \alpha_s [\sigma'_s \text{vec}(W_s)]^2 \right\} \\ &= \sum_{s=1}^m \mathbb{E} \left\{ \frac{1}{4} \pi_s^* \alpha_s \left[\text{tr} \{ [\Gamma_s^{*-1} - (y - \mu_s^*)(y - \mu_s^*)'] W_s \} \right]^2 \right\} \\ &= \sum_{s=1}^m \frac{1}{4} \pi_s^* \alpha_s \left\{ (\text{tr} \{ \Gamma_s^{*-1} W_s \})^2 + (\text{tr} \{ (y - \mu_s^*)(y - \mu_s^*)' W_s \})^2 \right. \\ &\quad \left. - 2 \text{tr} \{ \Gamma_s^{*-1} W_s \} \text{tr} \{ (y - \mu_s^*)(y - \mu_s^*)' W_s \} \right\} \\ &= \sum_{s=1}^m \frac{1}{4} \pi_s^* \alpha_s \left\{ \mathbb{E} \left[\alpha_s (\text{tr} \{ (y - \mu_s^*)(y - \mu_s^*)' W_s \})^2 \right] \right. \\ &\quad \left. - (\text{tr} \{ \Gamma_s^{*-1} W_s \})^2 \right\}. \end{aligned}$$

By expanding the matrices into expressions involving their components and noting (12c), one can check that

$$\begin{aligned} &\mathbb{E} \left[\alpha_s (\text{tr} \{ (y - \mu_s^*)(y - \mu_s^*)' W_s \})^2 \right] \\ &= 2 \text{tr} \{ (W_s \Gamma_s^{*-1})^2 \} + (\text{tr} \{ \Gamma_s^{*-1} W_s \})^2. \end{aligned}$$

Therefore,

$$\begin{aligned} I_3 &= \sum_{s=1}^m \frac{1}{2} \pi_s^* \text{tr} \{ W_s \Gamma_s^{*-2} W_s \} \\ &= \sum_{s=1}^m \frac{1}{2} \pi_s^* [\text{vec}(W_s)]' (\Gamma_s^* \otimes \Gamma_s^*)^{-1} \text{vec}(W_s) \\ &= \mathbf{W}' \Sigma^{-1} \mathbf{W}. \end{aligned}$$

The proof is complete. \square

Table 1: The Fisher information (FI) matrices and the inverse of the variational covariance (IVC) matrices corresponding to Figure 1. Each cell contains a 2×2 matrix.

	(1)	(2)	(3)	(4)
FI	5.83 2.50	2.47 1.29	6.08 3.77	0.00 0.00
	2.50 5.83	1.29 0.91	3.77 5.50	0.00 11.11
IVC	5.83 2.50	5.83 3.33	6.67 3.33	4.50 2.50
	2.50 5.83	3.33 6.67	3.33 5.83	2.50 12.50

By Lemma 2 the equality in (13) holds if and only if the mixture model (1) has only one component. In other cases, there must exist overlapping. If the components are well separated or have smaller overlaps, the mixture distribution can be regarded approximately as multinomial. In this case, for a given observation y_i , there exists one $p_s(y_i)$ which is far larger than the others, and therefore the inverse of Fisher information matrix is close to the covariance matrix of the variational posterior distribution. Theorem 2 shows that if overlapping exists between the components of a mixture model then the covariance matrix from the variational Bayes approximation is ‘too small’ compared with that for the MLE, so that resulting interval estimates for the parameters will be too narrow.

5 NUMERICAL EXPERIMENTS

In this section we demonstrate our results with some simple mixtures of normal densities.

First we consider mixtures of three known univariate normal densities $p_1(\cdot)$, $p_2(\cdot)$ and $p_3(\cdot)$ with means μ_1 , μ_2 and μ_3 ; all have unit variance. The mixing coefficients are π_1 , π_2 and $1 - \pi_1 - \pi_2$, respectively. For different values of these parameters, we compute the corresponding Fisher information matrices and the covariance matrices of the variational posteriors. The mixture densities of some typical cases are plotted in Figure 1, and the corresponding Fisher information matrices and the inverses of variational covariance matrices are described in Table 1. Obviously, if the components in the mixture models are widely separated, these two matrices are very similar, whereas, if the components are nearly identical, they are very different. The latter behaviour is reflected particularly by the case (4) in Figure 1, where $p_1(\cdot)$ and $p_2(\cdot)$ are completely identical.

Next we consider a more general mixture model of two unknown normal densities. Their means, precisions and mixing coefficients are μ_1 , Γ_1 , π and μ_2 , Γ_2 , $1 - \pi$,

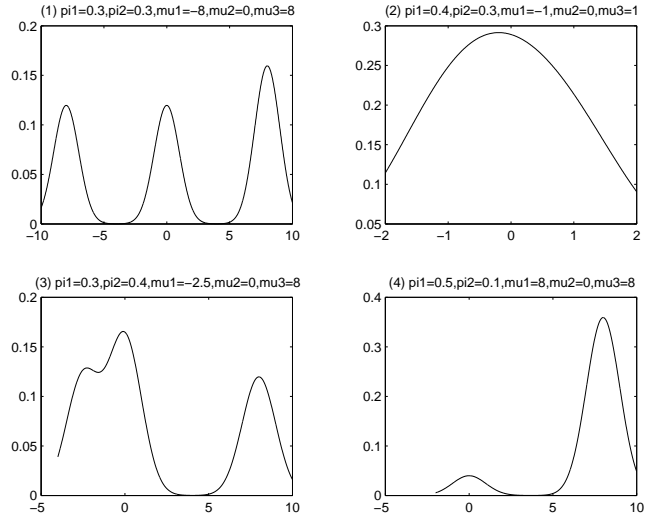


Figure 1: Some typical mixture densities based on different values of the parameters.

respectively. We compute the Fisher information matrices and the covariance matrices of the variational posteriors by using two sets of values of the parameters, namely, $(\pi, \mu_1, \Gamma_1, \mu_2, \Gamma_2) = (0.1, 1, 1, 0, 1)$ and $(0.5, 6, 1, 0, 1)$. There is large overlap between the two components for the first set of the parameters while they are well separated for the second. For the first set, the Fisher information matrix, denoted by I^* , is

$$\begin{pmatrix} 1.1542 & 0.1505 & 0.7456 & -0.0363 & -0.2612 \\ 0.1505 & 0.0259 & 0.0606 & -0.0134 & -0.0539 \\ 0.7456 & 0.0606 & 0.7723 & 0.0167 & 0.0774 \\ -0.0363 & -0.0134 & 0.0167 & 0.0152 & 0.0198 \\ -0.2612 & -0.0539 & 0.0774 & 0.0198 & 0.3646 \end{pmatrix}$$

and the inverse of the variational covariance matrix is

$$\Psi^{-1} = \text{diag}(11.1111, 0.1000, 0.9000, 0.0500, 0.4500).$$

Evaluated at a couple of arbitrary vectors $\Theta = (0.8, 4, 3, 2, 1)'$ and $(1, 1, 1, 1, 1)'$, for illustrative purposes, $\Theta' I^* \Theta$ ($\Theta' \Psi^{-1} \Theta$) are equal to 14.0885 and 3.7435 (17.4611 and 12.6111), respectively. For the second set, the Fisher information matrix I^* is

$$\begin{pmatrix} 3.9834 & 0.0125 & 0.0125 & 0.0170 & -0.0170 \\ 0.0125 & 0.4905 & -0.0091 & -0.0133 & 0.0122 \\ 0.0125 & -0.0091 & 0.4905 & -0.0122 & 0.0133 \\ 0.0170 & -0.0133 & -0.0122 & 0.2308 & 0.0157 \\ -0.0170 & 0.0122 & 0.0133 & 0.0157 & 0.2308 \end{pmatrix}$$

and the inverse of the variational covariance matrix is

$$\Psi^{-1} = \text{diag}(4.0000, 0.5000, 0.5000, 0.2500, 0.2500).$$

Evaluated at the same vectors Θ , $\Theta' I^* \Theta$ ($\Theta' \Psi^{-1} \Theta$) are equal to 15.7931 and 5.4889 (16.3100 and 5.5000), respectively.

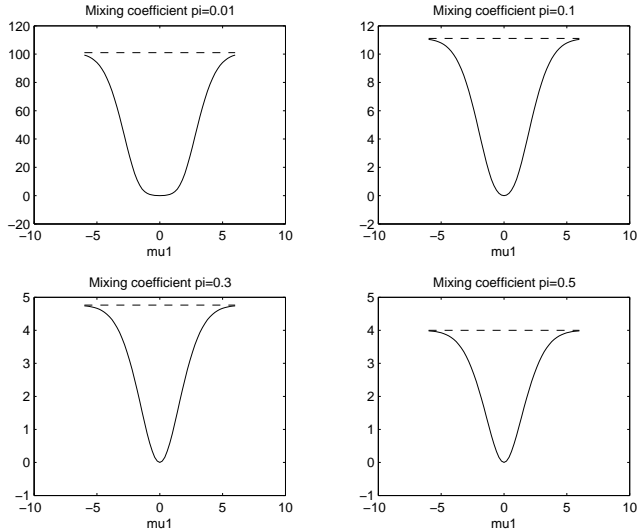


Figure 2: The inverses of the variances associated with the variational Bayes approximation and Fisher information for different mixing coefficients. The solid lines denote Fisher information and the dashed horizontal lines indicate the inverses of the variances for the variational Bayes approximation.

To clarify the dependence of the differences between the inverses of the variational covariance and Fisher information matrices on the overlaps between the components, now we use a mixture model of two known normal densities with means μ_1 and μ_2 with unit variance. The mixing coefficients are π and $1 - \pi$, respectively. The parameter π is given a Beta prior distribution $\text{Beta}(1, 1)$; i.e. $\pi \sim \text{Un}(0, 1)$. To compare the variance associated with the variational Bayes approximation with Fisher information, we fix one component to have mean zero and compute the inverse of the variance and Fisher information with the other component having varying mean μ_1 . The results are plotted in Figure 2 for different mixing coefficients π . The inverses of the variances associated with the variational Bayes approximation do not vary with the changes of μ_1 , whereas the Fisher informations do. And the differences between them become larger as μ_1 is closer to zero, the mean of the first component.

We investigate the performance of interval estimates based on the variational approximation using two empirical experiments. We fix the mixing coefficient at $\pi^* = 0.65$ and one component to have mean zero and unit variance within a mixture model of two normal densities. Independent random samples, each of size $n = 50$, are selected from the mixture model with the mean of the other component equal to $\mu_2 = 3.0$ and 1.0, and with unit variance. For each sample we calculate the variational Bayesian estimate $\hat{\pi}$ as given by

(8), the variational variance Λ , the maximum likelihood estimate $\tilde{\pi}$ and the Fisher information $I(\pi^*)$, and these are used to form approximate 95% confidence intervals given by $\hat{\pi} \pm 1.96\sqrt{\Lambda/n}$ and $\tilde{\pi} \pm 1.96/\sqrt{nI(\pi^*)}$. For $\mu_2 = 3.0$, a total of 100 samples are generated and the resulting 100 confidence intervals are computed. It turns out that 91 out of these 100 intervals do include the true value if the variational approximation is used, and 92 by the MLE method. Both proportions are close to the nominal confidence coefficient of 0.95. For $\mu_2 = 1.0$, the same number of confidence intervals are generated. Among these 100 intervals, only 68 of those based on the variational approximation include the true value, while this number is 92 from the MLE. In this case the resulting interval estimates are obviously too narrow.

Since the variational approaches provide good approximations for point estimates but poor approximations for interval estimates, a question of interest is whether or not the performance can be improved if we substitute the variational covariance matrix by the inverse of Fisher information for interval estimates. Theoretically, by this approach the resulting intervals would be very close to those obtained by MLE when the sample size is large, since the variational Bayes estimator converges to the maximum likelihood estimator. In order to verify this point, we use the same independent random samples as in the previous numerical examples to generate approximate 95% confidence interval given by $\hat{\pi} \pm 1.96/\sqrt{nI(\pi^*)}$. For $\mu_2 = 3.0$, 93 out of 100 intervals include the true value, whereas 96 intervals contain the true value if μ_2 is 1.0. It turns out that the approach of substituting the variational covariance matrix in the inverse of Fisher information does refine the interval estimates.

6 CONCLUSION

Exact theoretical analysis of the quality of variational Bayes approximations is an important issue. Having proved the properties of local convergence and asymptotic normality in Wang and Titterton (2003, 2004b,a), in this paper we examined the covariance matrices associated with variational Bayesian approximations and the resulting performance of variational Bayes approximations for interval estimates, by comparing them with the true covariances, given in terms of Fisher information matrices. It has been shown that the covariance matrices corresponding to the variational Bayes approximation are normally ‘too small’ compared with those for the MLE, so that resulting interval estimates for the parameters will be too narrow if the components of the mixture model are not well separated. Finally the theoretical analysis was reinforced by some numerical examples, which also sug-

gested an idea leading to the refinement of variational Bayes approximations for interval estimates by substituting the variational covariance by the ‘usual’ true covariance - the inverse of Fisher information. The arguments in the paper can be extended to non-Gaussian mixture models, such as mixtures of exponential family distributions, without any technical difficulty.

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.

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.
- Chen, C.-F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *J. R. Statist. Soc. B*, 47:540–546.
- Consonni, G. and Marin, J.-M. (2004). A note on variational approximate bayesian inference for latent variable models. Preprint.
- 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.
- Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *The Annals of Statistics*, 23(6):2145–2152.
- Humphreys, K. and Titterton, 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.
- 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.
- Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *J. R. Statist. Soc. B*, 31:80–88.
- Wang, B. and Titterton, 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>.
- Wang, B. and Titterton, D. M. (2004a). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In Chickering, M. and Halpern, J., editors, *Proceedings of the twentieth conference on Uncertainty in Artificial Intelligence*, pages 577–584, Banff, Canada. AUAI Press.
- Wang, B. and Titterton, D. M. (2004b). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Technical Report 04-3, University of Glasgow. <http://www.stats.gla.ac.uk/Research/TechRep2004/04-3.pdf>.