On strong consistency of posterior distributions in finite mixture models with unknown number of components

Russell C. H. Cheng and Sujit K. Sahu
Faculty of Mathematical Studies
University of Southampton
Highfield, SO17 1BJ, UK.

August 22, 2002

SUMMARY

In analyzing finite mixture models with an unknown number of components, standard regularity conditions for consistency are not usually satisfied. Consistency is then a serious issue, as estimators can exhibit unusual behavior and actually be inconsistent. A number of methods have been suggested using the Bayesian approach, but even here consistency is a very desirable property if such methods are to have a sound theoretical foundation. In recent work Barron *et al.* (1999) give powerful general conditions under which the posterior distribution in Bayesian analysis is consistent. These conditions do not cover the case of mixture distributions. An ideal result would be for consistency to be implied in the mixture case when the components themselves individually satisfy these conditions. In this paper we give a general finite parametric mixture model formulation where this is the case.

Keywords: BAYESIAN INFERENCE, HELLINGER DISTANCE, KULLBACK-LEIBLER DISTANCE, ϵ-ENTROPY.

1 Introduction

Consider a family of probability density functions of the form

$$f^{(k)}(x) = \sum_{j=1}^{k} w_j f(x|\theta_j), \tag{1}$$

where the density f is given and the non-negative weights w_j sum to 1. The number of components, k, as well as the parameters (θ_j) and weights are all assumed unknown. We observe $X_1, \ldots, X_n$, a

random sample of size n from (1).

There has been much recent work on practical Bayesian analysis of finite parametric mixture models as given in (1). Normal mixtures have been particularly well studied. See for example Diebolt and Robert (1994), Mengersen and Robert (1996), Richardson and Green (1997), Roeder and Wasserman (1997) and Stephens (2000).

Theoretical studies from the classical viewpoint (see for example Redner, 1981; Hartigan,1985; Smith, 1989; Cheng and Traylor 1995; Feng and McCulloch 1996, Cheng and Liu, 2001) show that the problem of consistency is a difficult one. Despite the previously mentioned Bayesian work, consistency of the Bayesian approach has not really been adequately addressed, even though it is an important issue, given the known theoretical difficulties in the classical case. Although consistency is an asymptotic concept, it has implications even for finite samples. See for example Diaconis and Freedman (1986). They conclude that, although a Bayes rule will do well for most parameter values, oscillation at infinity will even show up with large, finite samples–in high-dimensional problems.

Roeder and Wasserman (1997) prove consistency for normal mixture models where k is allowed to grow at a certain rate with n, the number of observations. Walker and Hjort (2001) prove strong consistency of posterior distributions which are based on data dependent prior distributions and they illustrate the results with Dirichlet mixtures. Rates of convergence posterior distributions are discussed by several authors including Ghosal *et al.* (2000) and Shen and Wasserman (2001). Powerful general results on the consistency of the posterior distribution have been obtained by Barron *et al.* (1999) and Ghosal *et al.* (1999). These results do not cover finite mixture models but would be very suited to the mixture case were they to do so.

In this paper we give a mixture model formulation where the assumptions, as given by Barron *et al.*, automatically hold for the mixture model if the same assumptions are true for each component of the mixture model. Thus we can verify consistency of the posterior distribution for the mixture model simply by checking assumptions for the individual components.

In Section 2 we describe the base model and the assumptions. We formulate a mixture of base model in Section 2 and verify the assumptions for the mixture model. We verify the assumptions with an example of normal mixture model in Section 4.

2 Base Model and Assumptions

Suppose that the family $f(x|\theta)$, $\theta \in \Theta$ with prior $\pi(\theta)$, $\theta \in \Theta$ is given where θ is an m-dimensional parameter. We shall only consider continuous distributions, with $\pi(\theta)$ also continuous, and will assume that Θ is open. We call this the *base model.*

For any $\Theta_n \subset \Theta$ and $\delta > 0$ we shall write $\mathcal{H}(\Theta_n, \delta)$ as the δ-upper metric entropy of the set of

distributions $f(x|\theta), \theta \in \Theta_n$ as defined in Barron *et al.* (1999). They actually use the notation $\mathcal{H}(\mathcal{F}, \delta)$ where $\mathcal{F}$ is a set of distributions. We shall assume that the mapping $\theta \to f(x|\theta)$ is a bijection hence we can identify any member of the family uniquely by $f(x|\theta)$, so that our notation is justified. Let the following two assumptions of Barron *et al.* (1999) be true.

Assumption (A1). For every $\varepsilon > 0$, $\pi(N_\varepsilon) > 0$ where

$$N_\varepsilon = \{\theta : E_0 \log f(X|\theta) > E_0 \log f(X|\theta_0) - \varepsilon\},$$

with $E_0 \log f(X|\theta) = \int [\log f(x|\theta)] f(x|\theta_0) dx$. The set N_ε is called the ε Kullback-Leibler neighborhood of $f(x|\theta_0)$.

Assumption (A2). For every $\varepsilon > 0$, there exists a sequence $\{\Theta_n\}_{n=1}^{\infty}$ of subsets of Θ, and positive reals c, c_1, c_2, δ such that

$$c < \frac{1}{2}\varepsilon^2 - \varepsilon\sqrt{\delta},\ \delta < \varepsilon^2/4 \tag{2}$$

and such that

(i) $\pi(\Theta_n^c) \leq c_1 \exp(-nc_2)$ for all but finitely many n

(ii) $\mathcal{H}(\Theta_n, \delta) \leq nc$ for all but finitely many n.

Under these two conditions Barron *et al.* (1999) show that the posterior distribution is consistent in the strong sense of concentrating in Hellinger distance about the true model. This form of convergence is especially appropriate in the study of mixture models in that it eliminates giving excess posterior weight to incorrect 'spikey' distributions.

There is one difficulty that we wish to avoid. Assumption (A1) involves simultaneously both the prior and the form of the base model. This can make the assumption difficult to verify in particular instances. We shall replace the assumption by a more explicit version which handles the prior and base model separately, but which guarantees that Assumption (A1) is satisfied.

The assumption is similar to the set of classic assumptions given by Wald (1949). It requires a metric in the parameter space, the precise form of which is not especially important. To be specific, we shall use the metric Ξ where $\Xi(\theta, \phi) = \sup_i |\theta_i - \phi_i|$ where θ_i and ϕ_i are ith component of θ and ϕ respectively. This seems a natural metric in which to check the validity of assumptions. The metric generalizes in the obvious way to mixture models where the supremum is evaluated over all individual parameters that appear, including the weights w_i. We write $B(\theta, r) = \{\phi : \Xi(\phi, \theta) < r\}$. Note that in our formulation we only need properties of neighborhoods in subspaces where the number of components is known. There is thus no ambiguity when specifying such regions B. Further, let

$$\begin{aligned} f(x|\theta, r) &= \sup_{\phi \in B(\theta,r)} f(x|\phi), \quad f^*(x|\theta) = \max[f(x|\theta), 1], \quad f^*(x|\theta, r) = \max[f(x|\theta, r), 1], \\ g(x|\theta, r) &= \inf_{\phi \in B(\theta,r)} f(x|\phi), \quad g^*(x|\theta) = \min[f(x|\theta), 1], \quad g^*(x|\theta, r) = \min[g(x|\theta, r), 1]. \end{aligned}$$

Assumption (B1)

(i) For any $\theta_0, \theta_1 \in \Theta$. If $\theta_0 \neq \theta_1$, then $f(x|\theta_0) \neq f(x|\theta_1)$ for some x.

(ii) For any $\theta_0 \in \Theta$, $f(x|\theta)$ is continuous at θ_0 for almost all x.

(iii) For any $\theta_0, \theta_1 \in \Theta$, there exists $r > 0$ such that

$$E_0[\log f^*(X|\theta_1, r)] < \infty \text{ and } E_0[|\log g^*(X|\theta_1, r)|] < \infty. \tag{3}$$

(iv) $\pi(\theta_0)$ is continuous at θ_0, and $\pi(\theta_0) > 0$.

Wald (1949, Lemma 1) shows that under Assumptions (B1) (i) and (iii)

$$E_0[\log f(X|\theta)] < E_0[\log f(X|\theta_0). \tag{4}$$

We also have the following.

Lemma 1

Under Assumptions (B1) (ii)-(iii)

$$\lim_{\theta \to \theta_0} E_0[\log f(X|\theta)] = E_0[\log f(X|\theta_0)] < \infty \tag{5}$$

i.e. $E_0[\log f(X|\theta)]$ *is continuous at* θ_0.

Proof: From the definitions $|\log f(x|\theta)| = \log f^*(x|\theta) + |\log g^*(x|\theta)| \leq \log f^*(x|\theta, r) + |\log g^*(x|\theta, r)|$. We can therefore take $M(x|\theta, r) = [\log f^*(x|\theta, r) + |\log g^*(x|\theta, r)|]\ f(x|\theta_0)$ as a bounding function, using Assumption (B1) (iii) to select an appropriate r, to apply the dominated convergence theorem to give (5). □

Consider

$$g(\theta_0, r) = \inf_{\phi \in B(\theta, r)} E_0 \left[\log f(X|\phi)\right].$$

We then have the following.

Lemma 2

Under Assumptions (B1) (i)-(iii), given $\varepsilon > 0$, *there exists* $r > 0$ *such that*

$$g(\theta_0, r) > E_0[\log f(X|\theta_0)] - \varepsilon. \tag{6}$$

Proof: Under Assumptions (B1) (i), (iii) we have (4) so clearly $g(\theta_0, r) \leq E_0[\log f(X|\theta_0)]$. Suppose (6) is not true. Then there exists r_i with $r_i \to 0$, such that $g(\theta_0, r_i) < E_0[\log f(X|\theta_0)] - \varepsilon$ for all i. For each i we can therefore select $\theta_i \in B(\theta_0, r_i)$ for which $E_0[\log f(X|\theta_i)] < E_0[\log f(X|\theta_0)] - \varepsilon$. Clearly $\theta_i \to \theta_0$, and $\lim_i E_0[\log f(X|\theta_i]$, if it exists at all, is less than $E_0[\log f(X|\theta_0)] - \varepsilon$. This contradicts (5) which Lemma 1 shows will hold under Assumptions (B1)(ii),(iii). Thus (6) holds under Assumption (B1). □

Corollary 1

Assumption (A1) is true when Assumption (B1) is true.

Proof: Lemma 2 shows that under Assumption (B1)(i)(ii)(iii), given $\varepsilon > 0$, there exists r such that the models with $\theta \in B(\theta_0, r)$ belong to N_ε, the ε Kullback-Leibler neighborhood of θ_0. But by Assumption (B1) (iv), $\pi(\theta_0) > 0$ and $\pi(\theta)$ is continuous at θ_0. So $\pi(\theta) > p$, for some $p > 0$, for all $\theta \in B(\theta_0, \delta)$. Thus $\pi(B(\theta_0, \delta)) > 0$ for some $\delta > 0$ sufficiently small, and taking $\eta = \min(r, \delta) > 0$, we have $\pi(N_\varepsilon) \geq \pi(B(\theta_0, \eta)) > 0$. i.e. Assumption (A1) holds. □

3 Mixtures of the Base Model

Consider the following mixture model derived from the above base model:

$$f^{(k)}(x|w^k, \tilde{\theta}^k) = \sum_{j=1}^{k} w_j f(x|\theta_j^{(k)}) \tag{7}$$

where

$$w^k = (w_1, w_2, \ldots, w_k), \ \sum_{j=1}^{k} w_j = 1$$

and

$$\tilde{\theta}^k = (\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_k^{(k)}), \quad \theta_j^{(k)} = (\theta_{j1}^{(k)}, \ldots, \theta_{jm}^{(k)}) \in \Theta_j^{(k)}.$$

The density (7) is the previously defined mixture density (1) written in a fuller notation. We use the notation $\tilde{\theta}^k \in \Theta^k \equiv \Theta_i^{(k)} \times \Theta_i^{(k)} \times \cdots \times \Theta_k^{(k)}$, the Cartesian product of k copies of Θ, with $\Theta_j^{(k)}$ being the jth copy.

Non-uniqueness, with points in the parameter space *not* giving rise to a unique model can occur in two ways.

Firstly, it occurs because a direct switch of values of two of the mixture component vectors $\theta_i^{(k)}$ and $\theta_j^{(k)}$ leaves $f^{(k)}$ unchanged. This is simply stopped by imposing an ordering on the values of just *one* of the components of θ, the first say. Thus, in $\tilde{\theta}^k$, we require the (strict) ordering

$$\theta_{11}^{(k)} < \theta_{21}^{(k)} < \cdots < \theta_{k1}^{(k)}.$$

We define

$$S^k = \{\tilde{\theta}^k : \tilde{\theta}^k \in \Theta^k, \ \theta_{11}^{(k)} < \theta_{21}^{(k)} < \cdots < \theta_{k1}^{(k)}.$$

Non-uniqueness also occurs when some components of w^k are set equal to zero. This also has the additional effect of reducing the number of components in the model (7). To overcome this we define

$$W^k = \{w^k : \sum_{j=1}^{k} w_j = 1, \ w_j > 0 \text{ all } j\}$$

and the *model set* M^k as

$$M^k = \{f^{(k)}(x, w^k, \tilde{\theta}^k) : \ w^k \in W^k, \ \tilde{\theta}^k \in S^k\}.$$

This model set comprises all models $f^{(k)}$ with exactly k distinct components.

Our basic assumption is that the number of components $k \leq K$, where K is known but the precise value of k is not known. The full model set under consideration is therefore the disjoint union

$$M = \bigcup_{k=1}^{K} M^k. \tag{8}$$

The k component parameter space is $\Omega^k = W^k \times S^k$ and the full parameter space is

$$\Omega = \bigcup_{k=1}^{K} \Omega^k$$

with the Ω^k disjoint.

Consider now the prior distribution. We shall assume a discrete prior

$$\rho_k > 0, \ k = 1, \ldots, K, \ \sum_{k=1}^{K} \rho_k = 1$$

for the number of components k.

For a given k, for simplicity, we assume the weight vector is uniformly distributed in W^k, and write this uniform distribution as

$$u(w^k), \ w^k \in W^k.$$

We have $\int_{W^k} u(w^k) dw^k = 1$ for each k. (Though we do not do so here, it will be clear that our results extend easily to the case where W^k follows the Dirichlet distribution.)

The prior $\pi(\theta)$ is assumed originally given. The natural prior on $\Theta^{(k)}$ is

$$\pi^{(k)}(\tilde{\theta}^k) = \prod_{j=1}^{k} \pi(\theta_j^{(k)}), \ \tilde{\theta}^k \in \Theta^{(k)}.$$

Under the restriction $\tilde{\theta}^k \in S^k$ this becomes

$$\begin{aligned} \pi^{(k)}(\tilde{\theta}^k | S^k) &= \left(\int_{S^k} \prod_{j=1}^{k} \pi(\theta_j^{(k)}) d\theta_j^{(k)} \right)^{-1} \prod_{j=1}^{k} \pi(\theta_j^{(k)}), \ \tilde{\theta}^k \in S^k \\ &= (k!) \prod_{j=1}^{k} \pi(\theta_j^{(k)}), \ \tilde{\theta}^k \in S^k \end{aligned}$$

by symmetry and the fact that $\pi(\theta)$ is a continuous density. We therefore take as our prior

$$\pi(w^k, \tilde{\theta}) = (k!) \, \rho_k \, u(w^k) \prod_{j=1}^{k} \pi(\theta_j^{(k)}), \ w^k \in W^k, \ \tilde{\theta} \in S^k, \ k = 1, 2, \ldots, K. \tag{9}$$

The following theorem is our main result and it follows from Theorems 2 and 3.

Theorem 1

Suppose that Assumptions (A1) and (A2) hold for each component density of the mixture density (7) and the associated prior distribution $\pi(\theta)$. Then the posterior distribution of the parameters in (7) when the prior is assumed to be (9) is strongly consistent.

3.1 Verifying Assumption (A1) for the Mixture Model

We shall show that Assumption (A1) induces a similar result for the family of mixture models M as defined in (8) when the prior is as given in (9).

Theorem 2

Suppose that Assumption (B1) holds for the base model and that, in the corresponding mixture model set, M given by (8), the true model has k components with true parameter values $(w_0^k, \tilde{\theta}_0^k)$. Then Assumption (B1) holds with $f(x|\theta)$ replaced by $f^{(k)}(x|w^k, \tilde{\theta}^k)$, θ_0 by $(w_0^k, \tilde{\theta}_0^k)$, Θ by Ω^k, and $\pi(\theta)$ by $\pi(w^k, \tilde{\theta}^k)$ as given in (9).

Proof: Assumptions (B1) (i),(ii) and (iv) for the mixture case follow directly from the definitions. The only part that needs particular consideration is (B1) (iii). For simplicity in notation we suppress the superscript in $\theta^{(k)}$.

We first need the following definitions:

$$f^{(k)}(x|w,\tilde{\theta},r) = \sup_{w' \in B(w,r),\ \tilde{\theta}' \in B(\tilde{\theta},r)} f^{(k)}(x|w',\tilde{\theta}'), \quad f^{(k)*}(x|w,\tilde{\theta},r) = \max[f^{(k)}(x|w,\tilde{\theta},r),\ 1],$$
$$g^{(k)}(x|w,\tilde{\theta},r) = \inf_{w' \in B(w,r),\ \tilde{\theta}' \in B(\tilde{\theta},r)} f^{(k)}(x|w',\tilde{\theta}'), \quad g^{(k)*}(x,w,\tilde{\theta},r) = \min[g^{(k)}(x|w,\tilde{\theta},r),\ 1].$$

Now we have,

$$\begin{aligned}
\max[f^{(k)}(x|w,\tilde{\theta},r),1] &= \max[\sup_{w' \in B(w,r),\ \tilde{\theta}' \in B(\tilde{\theta},r)} \sum_j w'_j f(x|\theta'_j), 1] \\
&\leq \max[\sup_{\tilde{\theta}' \in B(\tilde{\theta},r)} \sum_j f(x|\theta'_j), 1] \\
&\leq \max[\sum_j \sup_{\theta'_j \in B(\theta_j,r)} f(x|\theta'_j), 1] \\
&= \max[\sum_j f(x|\theta'_j, r), 1] \\
&\leq \max[k \max_j f(x|\theta_j, r), 1] \\
&\leq \max[k \max_j f^*(x|\theta_j, r), 1] \\
&= k \max_j f^*(x|\theta_j, r).
\end{aligned}$$

Hence

$$\begin{aligned}
\log[f^{(k)*}(x|w,\tilde{\theta},r)] &\leq \log k + \log[\max_j f^*(x|\theta_j, r)] \\
&\leq \log k + \sum_j \log[f^*(x|\theta_j, r)].
\end{aligned}$$

This last inequality holds as all the $f^*(x|\theta_j, r)$ are greater than unity so that the logs are all positive. Now take expectations, with respect to the true mixture model distribution $f^{(k)}(x|w_0^k, \tilde{\theta}_0^k)$, and denote this expectation by $E_{f_0^{(k)}}$.

It follows, from repeated application of (3) in evaluating components on the right-hand side, that

$$E_{f_0^{(k)}}\{\log[f^{(k)*}(X|w, \tilde{\theta}, r)]\} \leq \log k + \sum_j E_{f_0^{(k)}} \log[f^*(X|\theta_j, r)] < \infty \tag{10}$$

for some r sufficiently small

Consider now $g^{(k)}(x|w, \tilde{\theta}, r)$. We have

$$\begin{aligned} \min[g(x^{(k)}|w, \tilde{\theta}, r), 1] &= \min[\inf_{w' \in B(w,r),\ \tilde{\theta}' \in B(\tilde{\theta}, r)} \sum_i w'_j f(x|\theta'_j), 1] \\ &\geq \min[\inf_{\tilde{\theta}' \in B(\tilde{\theta}, r)} \sum_j w^* f(x|\theta'_j), 1] \end{aligned}$$

where $w^* = \min_j \inf_{w'_j \in B(w_j, r)} w'_j$. We now have

$$\begin{aligned} \min[\inf_{\tilde{\theta}' \in B(\tilde{\theta}, r)} \sum_j w^* f(x|\theta'_j), 1] &\geq \min[\sum_j w^* \inf_{\theta'_j \in B(\theta_j, r)} f(x|\theta'_j), 1] \\ &= \min[w^* \sum_j g(x|\theta_j, r), 1] \\ &\geq \min[kw^* \min_j g(x|\theta_j, r), 1] \\ &\geq \min[kw^* \min_j g^*(x|\theta_j, r), 1] \\ &= kw^* \min_j g^*(x|\theta_j, r) \end{aligned}$$

as $kw^* < 1$. Hence

$$\begin{aligned} \log[g^{(k)*}(x|w, \tilde{\theta}, r)] &\geq \log(kw^*) + \log[\min_j g^*(x|\theta_j, r)] \\ &\geq \log(kw^*) + \sum_j \log[g^*(x|\theta_j, r)]. \end{aligned}$$

This last inequality holds as all the $g^*(x|\theta_j, r)$ are less than unity so that the logs are all negative. In fact all the logs are negative so that

$$E_{f_0^{(k)}} |\log[g^{(k)*}(X|w, \theta, r)]| \leq |\log kw^*| + \sum E_{f_0^{(k)}} |\log[g^*(X|\theta_j, r)]| < \infty \tag{11}$$

for some r sufficiently small. The inequalities (10) and (11) together show that Assumption (B1)(iii), as given in the theorem, holds for the mixture model. □

3.2 Verifying Assumption (A2) for the Mixture Model

Lemmas 3 and 5 stated and proved below combine to give the following.

Theorem 3

Let $f(x|\theta)$, $\pi(\theta)$, $\theta \in \Theta$ and $\{\Theta_n\}_{n=1}^{\infty}$ be a base model, prior and a sequence of subsets satisfying Assumption (A2). Then Assumption (A2) is also satisfied if the base model is replaced by the model

set M given by (8), the prior $\pi(\theta)$ is replaced by $\pi(w, \tilde{\theta})$, $(w, \tilde{\theta}) \in \Omega$, as defined in (9), and the sequence $\{\Theta_n\}_{n=1}^{\infty}$ is replaced the sequence $\{\Omega_n\}$ as defined in (13).

Consider now Assumption (A2). When this holds, for all but finitely many n, there exists $\{\Theta_n\}_{n=1}^{\infty}$ and a corresponding upper δ-bracketing of Θ_n (for definition of which see Barron *et al.*, 1999) which we write as

$$f_{in}^{U}(x), \ i = 1, \ldots, I.$$

I will depend on n and δ, but we shall not need to make this explicit in the notation. We need to construct a corresponding set $\{\Omega_n\}_{n=1}^{\infty}$ and upper bracketing

$$g_{in}^{U}(x), \ i = 1, \ldots, M$$

for the mixture model.

We consider the construction of Ω_n first. Consider the Cartesian product

$$\Theta_{kn} = \Theta_n^{(1)} \times \Theta_n^{(2)} \times \cdots \times \Theta_n^{(k)} \tag{12}$$

where $\Theta_n^{(j)}$ $(= \Theta_n)$ is just the jth copy of Θ_n.

Let

$$\Omega_n = \bigcup_{k=1}^{K} \Omega_{kn} \tag{13}$$

where

$$\Omega_{kn} = W^k \times [\Theta_{kn} \cap S^k] \subseteq \Omega^k.$$

Lemma 3

Let $f(x|\theta)$ and $\pi(\theta)$, $\theta \in \Theta$ be a base model and prior satisfying Assumption (A2) (i). Then, for the model set M, defined in (8) with prior $\pi(w, \tilde{\theta})$, $(w, \tilde{\theta}) \in \Omega$, defined as in (9), the sequence $\{\Omega_n\}$ as defined in (13) satisfies Assumption (A2) (i), i.e. there exists c_1', c_2' such that

$$\pi(\Omega_n^c) \leq c_1' \exp(-nc_2') \text{ for all but finitely many } n.$$

Proof: The bulk of the argument focuses on a fixed k. Until we explicitly relax this, assume that k is given. To avoid an over elaborate notation the calculation of probabilities in what follows is conditional on a given k. We can recover the unconditional probabilities simply by multiplying by the appropriate ρ_k.

We wish to consider the complement of Ω_{kn}, where the total space is restricted to Ω^k and not Θ^k. However in some of the calculations we need to work with the complement treating Θ^k as being the total space. Where ambiguity may arise we shall write

$$\Omega_{kn}^c|\Omega^k \text{ and } \Omega_{kn}^c|\Theta^k$$

to indicate how the complement is being evaluated. This notation extends naturally to the evaluation of probabilities where $\pi(\Omega_{kn}^c|\Omega^k)$ has the usual interpretation of a conditional probability and can be evaluated as $\pi(\Omega_{kn}^c|\Omega^k) = \pi(\Omega_{kn}^c \cap \Omega^k|)/\pi(\Omega^k)$ using the product probability distribution $\pi(w^k, \tilde{\theta}^k) = u(w^k)\prod_{j=1}^k \pi(\theta_j)$ for calculations on the right hand side.

We shall make repeated use of the elementary identity

$$(A_1 \times A_2)^c = (A_1^c \times \Omega_2) \cup (\Omega_1 \times A_2^c) \tag{14}$$

where $A_1 \subseteq \Omega_1$ and $A_2 \subseteq \Omega_2$, and Ω_1 and Ω_2 are respectively the total spaces used in evaluating the complements A_1^c and A_2^c. Moreover if π_1 and π_2 are probability distributions defined on Ω_1 and Ω_2 and $\pi(a_1 \times a_2) = \pi_1(a_1)\pi_2(a_2)$ for $a_1 \in A_1$, $a_2 \in A_2$ then

$$\begin{aligned} \pi((A_1 \times A_2)^c \cap B) &\leq \pi((A_1 \times A_2)^c) \qquad (15)\\ &= \pi((A_1^c \times \Omega_2) \cup (\Omega_1 \times A_2^c)) \\ &\leq \pi(A_1^c \times \Omega_2) + \pi(\Omega_1 \times A_2^c) \\ &= \pi_1(A_1^c) + \pi_2(A_2^c), \qquad (16) \end{aligned}$$

and

$$\begin{aligned} \Omega_{kn}^c|\Omega^k &= (W^k \times [\Theta_{kn} \cap S^k])^c|\Omega^k \\ &= W^k \times \{[\Theta_{kn} \cap S^k]^c|S^k\} \text{ by (14).} \end{aligned}$$

As $\pi(W^k) = 1$, we have therefore

$$\begin{aligned} \pi(\Omega_{kn}^c|\Omega^k) &= \pi([\Theta_{kn} \cap S^k]^c|S^k) \\ &= \pi(\Theta_{kn}^c \cap S^k|\Theta^k)/\pi(S^k|\Theta^k) \\ &\leq \pi(\Theta_{kn}^c|\Theta^k)/\pi(S^k|\Theta^k). \end{aligned}$$

Repeated application of (16) to the numerator on the right hand side then yields

$$\pi(\Omega_{kn}^c|\Omega^k) \leq k\pi(\Theta_n^c)/\pi(S^k|\Theta^k) = k(k!)\pi(\Theta_n^c|\Theta).$$

By Assumption (B2)(i) we therefore have

$$\pi(\Omega_{kn}^c|\Omega^k) \leq k(k!)c_1 \exp(-nc_2) \tag{17}$$

Now the Ω_{kn} are disjoint, so

$$\Omega_n^c = \bigcup_{k=1}^K (\Omega_{kn}^c|\Omega^k)$$

where the members of the union are disjoint. As stated initially, our argument throughout, has been conditional on given k. In particular the probability in (17) is conditional on given k. From the law of

total probability we have therefore

$$\begin{aligned}\pi(\Omega_n^c|\Omega) &= \sum_{k=1}^{K} \rho_k \pi(\Omega_{kn}^c|\Omega^k) \\ &\leq \sum_{k=1}^{K} \rho_k k(k!) c_1 \exp(-nc_2) \\ &= c_1' \exp(-nc_2),\end{aligned}$$

where, as required, c_1' depends on K but not on n. This completes the proof of Lemma 3. □

We now construct a bracketing set for the mixture model.

Let

$$0 < w_{1n} < w_{2n} < \cdots < w_{J-1,n} < w_{Jn} = 1$$

be a division of $(0, 1]$ into J subintervals $\omega_j = (w_{j-1,n}, w_{jn}]$, $j = 1, \ldots, J$, (with $w_{0n} = 0$) each of width $\eta > 0$. i.e.

$$(w_{jn} - w_{j-1,n}) = \eta \text{ for } j = 1, \ldots, J.$$

We can make η as small as we wish by choosing J sufficiently large.

For any $f^{(k)}(x|w^k, \tilde{\theta}^k) \in \Omega_{kn}$, consider the weight vector w^k. Each component w_i of w^k will fall into one of the J subintervals. Suppose $w_j \in \omega_{i_j} = (w_{i_j-1,n}, w_{i_j,n}]$. We focus on the upper limit of this interval, which, to avoid over complex notation, we write as $w_{jn}^U = w_{i_j,n}$.

Clearly we have

$$0 \leq w_{jn}^U - w_j \leq \eta, \quad i = 1, \ldots, k. \tag{18}$$

Next consider the upper bracketing set for the base model. By assumption, given $\delta > 0$, there exists $f_{in}^U(x)$, $i = 1, \ldots, I(n, \delta)$, such that for any $\theta \in \Theta$, $f(x|\theta) \leq f_{in}^U(x)$, all x, some i. We define

$$T_{in} = \{\theta : \theta \in \Theta,\ f(x|\theta) \leq f_{in}^U(x),\ \text{all } x\}\ \ i = 1, \ldots, I(n, \delta).$$

Thus for any given $\tilde{\theta}^k = (\theta_1, \theta_2, \ldots, , \theta_k)$, each component vector $\theta_j \in T_{l_j}$ for some l_j so that

$$f(x|\theta_j) \leq f_{l_j,n}^U(x) \quad \text{all } x. \tag{19}$$

Now define

$$f_{w^k,\tilde{\theta}^k}^U(x) = \sum_{j=1}^{k} w_{jn}^U f_{l_j,n}^U(x). \tag{20}$$

We have

$$\begin{aligned}f_{w^k,\tilde{\theta}^k}^U(x) - f^{(k)}(x|w^k, \tilde{\theta}^k) &= \sum_{j=1}^{k} [w_{jn}^U f_{l_j,n}^U(x) - w_j f(x|\theta_j)] \\ &= \sum_{j=1}^{k} [(w_{jn}^U - w_j) f_{l_j,n}^U(x) + w_j (f_{l_j,n}^U(x) - f(x|\theta_j))] \\ &\geq 0 \text{ using (18) and (19).}\end{aligned}$$

Moreover using (18) we also have

$$\begin{aligned}\int f^U_{w^k,\tilde{\theta}^k}(x)dx &\leq \int \sum_{j=1}^{k}(w_j+\eta)f^U_{l_j,n}(x)dx \\ &\leq 1+\delta+k(1+\delta)\eta \\ &\leq 1+\delta+K(1+\delta)\eta \\ &= 1+\delta', \text{ say.}\end{aligned}$$

Thus $\{f^U_{w^k,\tilde{\theta}^k}(x),\ w^k \in W^k,\ \tilde{\theta}^k \in S^k,\ k=1,\ldots,K\}$ forms an $\delta'-$upper bracketing of Ω_n where

$$\delta' = \delta + K(1+\delta)\eta.$$

Moreover the number of distinct $f^U_{w^k,\tilde{\theta}^k}(x)$ is finite as each function depends on $w^k, \tilde{\theta}^k$ only through the combination of subintervals ω_i that each of the components w_j fall in, and through the combination of T_l to which each component vector θ_j of $\tilde{\theta}^k$ belongs. For given k, the number of distinct combinations of ω and T subintervals is $(IJ)^k$. The total number of brackets is therefore $\sum_{k=1}^{K}(IJ)^k$. Here I depends on n and δ, and $J = O(\eta^{-1})$ does not depend on n or δ.

We have therefore shown the following.

Lemma 4

Let $f(x|\theta)$, $\theta \in \Theta$ be a base model and a sequence $\{\Theta_n\}_{n=1}^{\infty}$ with upper δ bracketing $f^U_{in}(x)$ $i = 1,\ldots,I(n,\delta)$. For the mixture model set M in (8) define the sequence $\{\Omega_n\}$ as in (13). Then there is exists a δ'-upper bracketing of $\{\Omega_n\}$ with $\delta' = \delta + K(1+\delta)\eta$. The number of brackets is $L = \sum_{k=1}^{K}(IJ)^k$ where $\log I \leq nc$, with c satisfying (2), and $J = \eta^{-1}$.

This lemma enables us to extend Assumption (A2) (ii) to the mixture model.

Lemma 5

Let $f(x|\theta)$, $\theta \in \Theta$ be a base model. Suppose that, given $\varepsilon > 0$, there exists $\{\Theta_n\}$ satisfying Assumption 2(ii). For the mixture model set M (8) define the sequence $\{\Omega_n\}$ as in (13). Then, given $\varepsilon > 0$, there exists c and δ satisfying (2) such that

$$\mathcal{H}(\Omega_n, \delta) \leq nc \text{ for all } n > N, \text{ some finite } N.$$

Proof: By Assumption (B2) (ii), applied to the base model, for any $\varepsilon' > 0$ we can find $\{\Theta_n\}_{n=1}^{\infty}$, c', δ' such that

$$c' < \frac{1}{2}\varepsilon'^2 - \varepsilon'\sqrt{\delta'},\ \delta' < \varepsilon'^2/4 \tag{21}$$

and such that

$$\mathcal{H}(\Theta_n, \delta') \leq nc' \text{ for all but finitely many } n \tag{22}$$

Now by Lemma 4, given $\delta' > 0$ we can find a $\delta-$upper bracketing of the mixture model set M with

$$\delta = \delta' + K(1+\delta')\eta \tag{23}$$

and which contains no more than $L = \sum_{k=1}^{K}(IJ)^k$ functions. Moreover, the infimum $\mathcal{H}$ being discrete is achieved by appropriate choice of bracketing thus $\mathcal{H}(\Theta_n, \delta') = \log I$. As $\sum_{k=1}^{K}(IJ)^k < K(IJ)^K$, we have therefore

$$\mathcal{H}(\Omega_n, \delta) \leq \log L < K\mathcal{H}(\Theta_n, \delta') + K\log\eta^{-1} + \log K.$$

With (22) this gives

$$\begin{aligned} \mathcal{H}(\Omega_n, \delta) &< nKc' + K\log\eta^{-1} + \log K \\ &\leq n(Kc' + a) \text{ for all } n > N, \text{ some } N, \end{aligned}$$

where a can be made as small as we like irrespective of the choice of η by choosing $N(a)$ sufficiently large. Let

$$c = Kc' + a. \tag{24}$$

Given $\varepsilon > 0$, we wish to select ε', c', δ' satisfying (21) for which c and δ, as defined in (24) and (23), will satisfy (2). Now if (21) is satisfied, we have

$$\begin{aligned} c &= Kc' + a \\ &< K(\frac{1}{2}\varepsilon'^2 - \varepsilon'\sqrt{\delta'}) + a. \end{aligned}$$

The condition(2) which can be written as $\varepsilon/2 - \varepsilon\sqrt{\delta} - c > 0$ is then satisfied if

$$\frac{\varepsilon^2}{2} - \varepsilon\sqrt{\delta' + K\eta(1+\delta')} - [K(\frac{1}{2}\varepsilon'^2 - \varepsilon'\sqrt{\delta'}) + a] > 0.$$

This holds a fortiori if

$$\frac{\varepsilon^2}{2} - \varepsilon\sqrt{\delta' + K\eta(1+\delta')} - [\frac{K}{2}\varepsilon'^2 + a] > 0. \tag{25}$$

We are free first to choose ε'. Let us take $\varepsilon' = \varepsilon/(2\sqrt{K})$. Then the left hand side of (25) becomes

$$\frac{3\varepsilon^2}{8} - \varepsilon\sqrt{\delta' + K\eta(1+\delta')} - a.$$

But given ε' we know there exists $\delta' < \varepsilon'^2/4$. With this δ' we can then choose η as small as we like, so that $\sqrt{\delta' + K\eta(1+\delta')}$ is arbitrarily close to $\sqrt{\delta'} = \varepsilon/(4\sqrt{K}) < \varepsilon/4$; i.e. we can select η sufficiently small so that $\sqrt{\delta' + K\eta(1+\delta')} < \varepsilon/4$. Thus

$$\frac{3\varepsilon^2}{8} - \varepsilon\sqrt{\delta' + K\eta(1+\delta')} - a > \frac{\varepsilon^2}{8} - a.$$

Taking $a < \varepsilon^2/8$ therefore ensures (25) is satisfied. This completes the proof of Lemma 5. □

4 Example: Normal Mixture Model

As an application of the above results we consider mixture models with the normal component

$$f_N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}), \tag{26}$$

and $\Theta = \{(\mu,\sigma) : -\infty < \mu < \infty,\ 0 < \sigma < \infty\}$. The main condition that we need to verify is Assumption A2(ii) and this is what we now consider. Wong and Shen (1995) give a related result, but only for a very special two component case. Their result appears to depend on Equation (43) given in Kolmogorov and Tihomirov (1959), but it is not clear how this equation can be applied without significant manipulation, and details are not given. We therefore derive our result from first principles making use of the much more explicit Equation (11) in Kolmogorov and Tihomirov (1959) which shows that $\mathcal{H}(A,\varepsilon)$, the *minimal ε - entropy* of a set of functions A defined on $\Delta = [a,b]$, satisfies

$$\mathcal{H}(A,\varepsilon) = \frac{|\Delta|L}{\varepsilon} + \log_2 \frac{C}{\varepsilon} + O(1),\ \text{as } \varepsilon \to 0, \tag{27}$$

under the metric

$$\rho(f(u),\ g(u)) = \sup_{u\in\Delta} |f(u) - g(u)|,$$

provided each function $f \in A$ satisfies the Lipshitz condition

$$|f(u) - f(u')| \le L\,|u - u'|,\ \text{for } u,\ u' \in \Delta$$

and the bound

$$|f(u)| \le C,\ \text{for } u \in \Delta.$$

To apply the result in the normal density case which has unbounded support, we first apply the monotone transformation

$$x(u) = \begin{cases} \log(1+u) & \text{for } -1 \le u \le 0 \\ -\log(1-u) & \text{for } 0 \le u \le 1 \end{cases}$$

to the normal density (26). Under this transform the density becomes

$$h(u \mid \mu,\sigma) = \begin{cases} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\frac{1}{1+u}\exp(-\frac{1}{2\sigma^2}(\log(1+u)-\mu)^2), & \text{for } -1 \le u \le 0 \\ \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\frac{1}{1-u}\exp(-\frac{1}{2\sigma^2}(-\log(1-u)-\mu)^2), & \text{for } 0 \le u \le 1. \end{cases}$$

The support $\Delta = [-1,1]$ is thus finite. We consider the set A_n defined by

$$A_n = \{h(\cdot,\mu,\sigma) : (\mu,\sigma) \in \Theta_n\}$$

where

$$\Theta_n = \{(\mu,\sigma) : -c_n < \mu < c_n,\ 0 < a_n < \sigma < b_n\}.$$

Let Γ_n be an *$\varepsilon-$covering of A_n*, as defined by Kolmogorov and Tihomirov (1959). Then, by definition, for any $h \in A_n$, there is an element $\phi \in \Gamma_n$ such that

$$\phi(u) - 2\varepsilon \le h(u) \le \phi(u),\ u \in \Delta.$$

We therefore have

$$\int_{-1}^{1}(\phi(u)-2\varepsilon)du \leq \int_{-1}^{1} h(u)du = 1.$$

Thus

$$\int_{-1}^{1}\phi(u)du \leq 1+4\varepsilon,$$

so that, with $\delta = 4\varepsilon$, the set Γ_n forms a $\delta-$*upper bracketing of* A_n and $(\log 2)\mathcal{H}(A_n,\, \varepsilon = \delta/4)$ is the $\delta-$*upper metric entropy of* A_n, as defined by Barron *et al.* (1999). [The factor $\log 2$ is needed as Kolmogorov and Timohirov define entropy using logarithms to base 2.] To apply (27) to assess the magnitude of $\mathcal{H}(A_n,\, \varepsilon)$ we need to determine a bound C and a Lipshitz parameter L applicable to all $h \in A_n$.

We need only consider $h(u \mid \mu, \sigma)$, for $0 \leq u \leq 1$. By symmetry, the properties of $h(u \mid \mu, \sigma)$ for $-1 \leq u \leq 0$ are essentially identical and so do not have to be separately considered. From now on we restrict attention to $0 \leq u \leq 1$. We also, for simplicity, write $h(u)$ for $h(u \mid \mu, \sigma)$ when there is no ambiguity.

The first and second derivatives of h are:

$$\frac{dh(u)}{du} = h_1(u) = \frac{\sigma^2+\mu+\log(1-u)}{\sqrt{2\pi}\sigma^3(1-u)^2}\exp\left(-\frac{1}{2}\frac{(\log(1-u)+\mu)^2}{\sigma^2}\right)$$

and

$$\frac{d^2h(u)}{du^2} = h_2(u) = \frac{q(u \mid \mu, \sigma)}{\sqrt{2\pi}\sigma^5(1-u)^3}\exp\left(-\frac{1}{2}\frac{(\log(1-u)+\mu)^2}{\sigma^2}\right)$$

where

$$q(u \mid \mu, \sigma) = \log^2(1-u) + (2\mu+3\sigma^2)\log(1-u) + 2\sigma^4 + 3\sigma^2\mu - \sigma^2 + \mu^2.$$

Elementary use of these formulas shows that the maximum of $h(u \mid \mu, \sigma)$ is at $u_{\max} = 1 - e^{-\sigma^2-\mu}$ with maximum value

$$h(u_{\max} \mid \mu, \sigma) = \frac{1}{2}\frac{\sqrt{2}}{\sqrt{\pi}\sigma}e^{\frac{1}{2}\sigma^2+\mu}. \tag{28}$$

This is a convex function of σ so, for $(\mu, \sigma) \in \Theta_n$, $h(u_{\max} \mid \mu, \sigma)$ is maximized at either $\sigma = a_n$ or at $\sigma = b_n$, with $\mu = c_n$.

Consider now the Lipshitz property. The maximum magnitude of the slope $h_1(u)$ is either at a stationary point satisfying $h_2(u) = 0$ or else is at one of the limits $u = 0$ or $u = 1$. From the expression for $h_2(u)$, this is zero only when $q(u \mid \mu,\, \sigma) = 0$. This has solutions

$$u_1 = 1 - \exp(-\frac{3}{2}\sigma^2 - \mu + \frac{1}{2}\sqrt{(\sigma^4+4\sigma^2)}) \text{ and } u_2 = 1 - \exp(-\frac{3}{2}\sigma^2 - \mu - \frac{1}{2}\sqrt{(\sigma^4+4\sigma^2)})$$

with corresponding values of the slope being

$$h_1(u_1) = \frac{\left(\sqrt{(\sigma^2+4)}-\sigma\right)}{2\sqrt{2\pi}\sigma^2}\exp\left(\frac{7}{4}\sigma^2 - \frac{1}{4}\sigma\sqrt{(\sigma^2+4)} - \frac{1}{2} + 2\mu\right)$$

and

$$h_1(u_2) = -\frac{\left(\sqrt{(\sigma^2+4)}+\sigma\right)}{2\sqrt{2\pi}\sigma^2}\exp\left(\frac{7}{4}\sigma^2+\frac{1}{4}\sigma\sqrt{(\sigma^2+4)}-\frac{1}{2}+2\mu\right).$$

Further elementary consideration of the terms involving σ in these expressions shows that the maximum of these expressions, subject to $a_n \le \sigma \le b_n$ is either at $\sigma = a_n$ or at $\sigma = b_n$. In either case the maximum with respect to μ, subject to $-c_n \le \mu \le c_n$, occurs at $\mu = c_n$.

Consider now $h_1(u)$ at its two limits $u = 0$ and $u = 1$. We have immediately

$$\lim_{u\uparrow 1} h_1(u) = 0$$

and

$$h_1(0) = \lim_{u\downarrow 0} h_1(u) = \left(\frac{1}{2}\sigma^2+\frac{1}{2}\mu\right)\frac{\sqrt{2}}{\sqrt{\pi}\sigma^3}e^{-\frac{1}{2}\frac{\mu^2}{\sigma^2}}.$$

Elementary considerations show that, for fixed σ, this latter limit is largest when

$$\mu = \mu_1(\sigma) = \left(-\frac{1}{2}\sigma+\frac{1}{2}\sqrt{(\sigma^2+4)}\right)\sigma$$

with value

$$h_1(0 \mid \mu_1(\sigma),\sigma) = \frac{\left(\sigma+\sqrt{(\sigma^2+4)}\right)}{2\sqrt{2\pi}\sigma^2}\exp\left(\frac{1}{4}\sigma-\frac{1}{4}\sqrt{(\sigma^2+4)}\right).$$

This expression decreases as σ decreases, and so is maximized when $\sigma = a_n$. Thus, subject to $a_n \le \sigma \le b_n$,

$$h_1(0 \mid \mu,\sigma) < \frac{\left(a_n+\sqrt{(a_n^2+4)}\right)}{2\sqrt{2\pi}a_n^2}\exp\left(\frac{1}{4}a_n-\frac{1}{4}\sqrt{(a_n^2+4)}\right). \tag{29}$$

Thus, if we restrict μ,σ to lie in Θ_n, the maximum magnitude of h_1 cannot exceed the largest of the five values:

$$\begin{aligned}
&h_1(0 \mid \mu_1(a_n), a_n),\\
&h_1(u_1(c_n,a_n) \mid c_n,a_n),\ h_1(u_2(c_n,a_n) \mid c_n,a_n),\\
&h_1(u_1(c_n,b_n) \mid c_n,b_n),\ h_1(u_2(c_n,b_n) \mid c_n,b_n).
\end{aligned} \tag{30}$$

Note that, depending on the values of a_n, b_n and c_n, not all the u_1 or u_2 values may lie in the valid range $0 \le u \le 1$, so that the largest of these values is not necessarily attainable in Θ_n. Irrespective of this, the largest of the five values furnishes a bound which, is definitely satisfied by the Lipshitz parameter in Θ_n and this is all that we require.

Barron *et al.* (1999) give a strengthened, but less complicated form of Assumption A2(ii) which takes $\delta = \eta^2/16$ $(= 4\varepsilon)$ when A2(ii) reduces to

$$(\log 2)\mathcal{H}(A_n,\eta^2/16) \le n\eta^2/5.$$

Combining with the Kolmogorov and Tihomirov result (27) shows that we need to select a_n, b_n, c_n so that

$$\mathcal{H}(A_n, \frac{\eta^2}{16}) \leq \frac{64|\Delta|L}{\eta^2} + \log_2 C - 2\log_2 \eta + 6 \leq \frac{n\eta^2}{5\log 2}. \tag{31}$$

Inspection of the values in (30) shows that, for a_n small, and b_n and c_n both large, the largest value corresponds to either

$$h_1(u_2(c_n, a_n)|c_n, a_n) = O(\frac{e^{2c_n}}{a_n^2})$$

or

$$h_1(u_2(c_n, b_n)|c_n, b_n) = O(\frac{e^{2c_n + 2b_n^2}}{b_n}).$$

Similarly from (28) the largest value corresponds to either

$$h(u_{\max} \mid c_n, a_n) = O(\frac{e^{c_n}}{a_n})$$

or to

$$h(u_{\max} \mid c_n, b_n) = O(\frac{e^{c_n + b_n^2/2}}{b_n}).$$

If we set $a_n = n^{-1/4}$, $b_n = (\frac{1}{4}\log n)^{\frac{1}{2}}$, $c_n = \frac{1}{4}\log n$, then

$$L = O[(\frac{1}{4}\log n)^{-1} n] \text{ and } C = O[(\frac{1}{4}\log n)^{-1} n^{3/8}]$$

and (31) is then satisfied for all n sufficiently large.

REFERENCES

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of Posterior Distributions in non-parameteric problems, *Annals of Statistics*, **27**, 536–561.

Cheng, R.C.H. and Traylor, L. (1995). Non-regular maximum likelihood problems, *Journal of the Royal Statistical Society*, B, **57**, 3–44.

Cheng, R.C.H. and Liu, W.B. (2001). The Consistency of Estimators in Finite Mixture Models, *Scandinavian Journal of Statistics,* **28**, 603–616.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates: special invited paper, with discussion, *Annals of Statistics*, **14**, 1–26.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society*, B, **56**, 363–375.

Feng, Z.D. and McCulloch, C.E. (1996). Using bootstrap likelihood ratios in finite mixture models, *Journal of the Royal Statistical Society*, B, **58**, 593–608.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior Consistency of Dirichlet Mixtures in Density Estimation, *Annals of Statistics*, **27**, 143–158.

Ghosal, S., Ghosh, J. K. and Van der Vaart (2000). Convergence rates of posterior distributions, *Annals of Statistics*, **28**, 500–531.

Hartigan, J.A. (1985). A failure of likelihood asymptotics for the mixture model, in *Proc. Berkeley Symp. in Honor of J. Neyman and J. Kiefer* (eds L. LeCam and R.A. Olshen), Vol. II, pp. 807–810. New York: Wadsworth.

Kolmogorov, A.N. and Tihomirov, V.M. (1959). $\varepsilon-$entropy and $\varepsilon-$capacity in function spaces, *Uspekhi Mat. Nauk* **14** 3-86 [in Russian; English transl. *Amer. Math. Soc. Transl. Ser. 2* **17**, 277–364 (1961)].

Mengersen, K. and Robert, C. P. (1996). Testing for mixtures: A Bayesian Entropic Approach (with discussion), In *Bayesian Statistics 5*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford: Oxford University Press, pp. 255–276.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions, in *Annals of Statistics*, Vol. 9, pp. 225–228.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian Density Estimation Using Mixture of Normals, *Journal of the American Statistical Association*, **92**, 894–902.

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion), *Journal of the Royal Statistical Society*, B, **59**, 473–484.

Shen, X.T. and Wasserman, L. (2001). Rates of convergence of posterior distributions, *Annals of Statistics*, **29**, 687–714.

Smith, R.L. (1989). A survey of nonregular problems, in *Proc. Int. Statist. Inst. Conf. 47th Session, Paris,* pp.353-372.

Stephens, M. (2000). Bayesian Analysis of Mixture Models with an Unknown Number of Components– an alternative to reversible jump methods, *Annals of Statistics*, **28**, 40–74.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate, *Annals of Mathematical Statistics*, **20**, 595–601.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency, *Journal of the Royal Statistical Society*, B, **63**, 811–821.

Wong, W.H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs, *Annals of Statistics*, **23**, 339–362.