Bernoulli **12**(6), 2006, 1003–1017

Strong consistency of the maximum likelihood estimator for finite mixtures of location–scale distributions when the scale parameters are exponentially small

KENTARO TANAKA[1] and AKIMICHI TAKEMURA[2]

[1]*Department of Industrial Engineering and Management, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan. E-mail: tanaken@me.titech.ac.jp*
[2]*Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. E-mail: takemura@stat.t.u-tokyo.ac.jp*

In a finite mixture of location–scale distributions the maximum likelihood estimator does not exist because of the unboundedness of the likelihood function when the scale parameter of some mixture component approaches zero. In order to study the strong consistency of the maximum likelihood estimator, we consider the case where the scale parameters of the component distributions are restricted from below by c_n, where $\{c_n\}$ is a sequence of positive real numbers which tend to zero as the sample size n increases. We prove that under mild regularity conditions the maximum likelihood estimator is strongly consistent if the scale parameters are restricted from below by $c_n = \exp(-n^d)$, $0 < d < 1$.

Keywords: consistency; maximum likelihood estimator; mixture distribution

1. Introduction

In some finite mixture distributions the maximum likelihood estimator (MLE) does not exist. Let us consider the following example. Denote a normal mixture distribution with M components and parameter $\theta = (\alpha_1, \mu_1, \sigma_1^2, \ldots, \alpha_M, \mu_M, \sigma_M^2)$ by

$$f(x; \theta) = \sum_{m=1}^{M} \alpha_m \phi_m(x; \mu_m, \sigma_m^2),$$

where the α_m $(m = 1, \ldots, M)$ are non-negative real numbers that sum to one and the $\phi_m(x; \mu_m, \sigma_m^2)$ are normal densities. Let $x_1, \ldots, x_n$ denote a random sample of size $n \geqslant 2$ from the density $f(x; \theta_0)$. In view of the identifiability problem of mixture distributions discussed below, here θ_0 is a parameter value designating the true distribution. However, for

1350–7265

simplicity we henceforth just say that θ_0 is the true parameter. If we set $\mu_1 = x_1$, then the likelihood tends to infinity as $\sigma_1^2 \to 0$. Thus the MLE does not exist.

But when we restrict $\sigma_m \geqslant c$ $(m = 1, \ldots, M)$ by some positive real constant c, we can avoid the divergence of the likelihood. Furthermore, in this situation, it can be shown that the MLE is strongly consistent if the true parameter θ_0 is in the restricted parameter space.

On the other hand, the smaller σ_1^2 is, the smaller the contribution $\phi_1(x; \mu_1 = x_1, \sigma_1^2)$ makes to the likelihood at other observations $x_2, \ldots, x_n$. Therefore an interesting question here is whether we can decrease the bound $c = c_n$ to zero with the sample size n and yet guarantee the strong consistency of the MLE. If this is possible, a further question is how fast c_n can decrease to zero.

This question is similar to the (so far open) problem stated in Hathaway (1985), which treats mixtures of normal distributions with constraints imposed on the ratios of variances, while our restriction is imposed on the variances themselves. See also the discussion in Section 3.8.1 of McLachlan and Peel (2000).

In the above example, the normality of the component distributions is not essential and the same difficulty exists for finite mixtures of general location–scale distributions such as mixtures of uniform distributions. Furthermore, in this paper we allow each component to belong to different location–scale families. Let σ_m $(m = 1, \ldots, M)$ denote the scale parameters of the component distributions, and consider the restriction $\sigma_m \geqslant c_n$ $(m = 1, \ldots, M)$. Then a question of interest here is whether we can decrease the bound c_n to zero.

For the case of mixtures of uniform distributions, Tanaka and Takemura (2005) proved that the MLE is strongly consistent if $c_n = \exp(-n^d)$, $0 < d < 1$. Here d can be arbitrarily close to 1 but fixed. In this paper, we prove that the same result holds for general finite mixtures of location–scale distributions under very mild regularity conditions (Assumptions 1–4 below). As discussed in Section 5, the normal density satisfies the regularity conditions and our result implies that the MLE is strongly consistent for the finite normal mixture if $\sigma_m \geqslant c_n = \exp(-n^d)$, $0 < d < 1$, $m = 1, \ldots, M$.

Our framework is closely related to the sieve method (Grenander 1981), in which an objective function is maximized over a constrained subspace of the parameter space and then this subspace is expanded to the whole parameter space as the sample size increases. Some applications and consistency results for the method are given in Geman and Hwang (1982). An MLE based on a sieve is called a sieve MLE. The convergence rates of sieve MLEs for Gaussian mixture problems are studied in Genovese and Wasserman (2000) and Ghosal and van der Vaart (2001), and their ideas are very interesting. They obtain the convergence rates by bounding the Hellinger bracketing entropy of subsets of the function space and assume that the corresponding subsets of the parameter space are compact so that their bracketing entropy does not diverge. In the case of the sieve MLE, the approximating subspaces are usually taken to be compact, whereas we treat a sequence of non-compact subsets of the parameter space expanding to the whole parameter space as the sample size increases. Therefore results on sieve MLEs are not directly applicable in our framework.

The paper is organized as follows. In Section 2 we summarize some preliminaries. In Section 3 we state our main results (Theorems 1 and 2), giving their proofs in Section 4. We conclude in Section 5 with some discussions.

2. Preliminaries: strong consistency and identifiability of mixture distributions

A mixture of M densities with parameter $\theta = (\alpha_1, \mu_1, \sigma_1, \ldots, \alpha_M, \mu_M, \sigma_M)$ is defined by $f(x; \theta) \equiv \sum_{m=1}^{M} \alpha_m f_m(x; \mu_m, \sigma_m)$, where the α_m $(m = 1, \ldots, M)$, called the mixing weights, are non-negative real numbers that sum to one and the $f_m(x; \mu_m, \sigma_m)$, called the components of the mixture, are density functions. In this paper we consider the case where the components are location–scale densities with location parameter $\mu_m \in \mathbb{R}$ and scale parameter $\sigma_m > 0$, that is,

$$f_m(x; \mu_m, \sigma_m) = \frac{1}{\sigma_m} f_m\left(\frac{x - \mu_m}{\sigma_m}; 0, 1\right). \tag{1}$$

As mentioned above, we allow $f_m(x; \mu_m, \sigma_m)$ to belong to different families. For example, $f_1(x; \mu_1, \sigma_1)$ may be a normal density, $f_2(x; \mu_2, \sigma_2)$ may be a uniform density, and so on. Let $\Omega_m = \mathbb{R} \times (0, \infty)$ denote the parameter space of the mth component (μ_m, σ_m) and let Θ denote the entire parameter space:

$$\Theta \equiv \left\{(\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M \,|\, \sum_{m=1}^{M} \alpha_m = 1, \alpha_m \geqslant 0\right\} \times \prod_{m=1}^{M} \Omega_m.$$

Let $\mathcal{K}$ be a subset of $\{1, 2, \ldots, M\}$ and let $|\mathcal{K}|$ denote the number of elements in $\mathcal{K}$. Denote by $\theta_{\mathcal{K}}$ a subvector of $\theta \in \Theta$ consisting of the components in $\mathcal{K}$. Then the parameter space of subprobability measures consisting of the components in $\mathcal{K}$ is

$$\overline{\Theta}_{\mathcal{K}} \equiv \left\{\theta_{\mathcal{K}} \,|\, \theta \in \Theta, \sum_{m \in \mathcal{K}} \alpha_m \leqslant 1\right\}.$$

The corresponding density and the set of subprobability densities are denoted by

$$f_{\mathcal{K}}(x; \theta_{\mathcal{K}}) \equiv \sum_{k \in \mathcal{K}} \alpha_k f_k(x; \mu_k, \sigma_k), \qquad \mathcal{G}_{\mathcal{K}} \equiv \{f_{\mathcal{K}}(x; \theta_{\mathcal{K}}) \,|\, \theta_{\mathcal{K}} \in \overline{\Theta}_{\mathcal{K}}\}.$$

Furthermore, denote the set of subprobability densities with no more than K components by

$$\mathcal{G}_K \equiv \bigcup_{|\mathcal{K}| \leqslant K} \mathcal{G}_{\mathcal{K}} \qquad (1 \leqslant K \leqslant M). \tag{2}$$

We now briefly discuss identifiability of parameters. In mixture models, different parameters may designate the same distribution. When the component densities belong to a common location–scale family, we can permute the labels of the components and the distribution remains the same. A mixture model of $K - 1$ components can be obtained by setting one weight $\alpha_m = 0$ (with arbitrary μ_m and σ_m) in a model with K components. These are trivial cases of unidentifiability of parameters. However, there are more complicated cases. Let $U(x; a, b)$ denote the uniform density on the interval $[a, b]$. Then, for example, $\frac{1}{3}U(x; -1, 1) + \frac{2}{3}U(x; -2, 2)$ and $\frac{1}{2}U(x; -2, 1) + \frac{1}{2}U(x; -1, 2)$ represent the same distribution (Everitt and Hand 1981). In this case the limiting behaviour of the MLE is not obvious, although the estimated density should be consistent. Therefore, we first give a definition of consistency in terms of the estimated density.

Let $f_0(x) = f(x;\, \theta_0)$ denote the true density and let $\hat{f}_n(x) = f(x;\, \hat{\theta}_n)$ denote the estimated density.

Definition 1. *An estimator $\hat{f}_n$ is strongly consistent if*

$$\mathrm{Prob}\Big(\lim_{n\to\infty} ||\hat{f}_n - f_0|| = 0\Big) = 1,$$

where $\|\cdot\|$ is the L_1-norm.

Although Definition 1 is conceptually simple, in order to prove the strong consistency of the MLE we work with the location and scale parameters in (1) and the mixing weights. In order to deal with the identifiability problem let us introduce a distance between two sets of parameters. Let $\mathrm{dist}(\theta,\, \theta')$ denote the ordinary Euclidean distance (or any other equivalent distance) between two parameter vectors $\theta, \theta' \in \Theta$. For $U,\ V \subset \Theta$ define $\mathrm{dist}(U,\, V) \equiv \inf_{\theta\in U} \inf_{\theta'\in V} \mathrm{dist}(\theta,\, \theta')$. For a parameter θ, let

$$\Theta(\theta) \equiv \{\theta' \in \Theta \,|\, f(x;\, \theta') = f(x;\, \theta) \quad \forall x\}.$$

Then $\Theta_0 = \Theta(\theta_0)$ denotes the set of true parameters. Since our densities are continuous with respect to θ, by Scheffé's theorem (Theorem 16.12 of Billingsley 1995) $\mathrm{dist}(\Theta(\hat{\theta}_n),\, \Theta_0) \to 0$ implies $\|\hat{f}_n - f_0\| \to 0$.

3. Main results

We assume the following regularity conditions for strong consistency of the MLE.

Assumption 1. *There exist real constants $v_0, v_1 > 0$ and $\beta > 1$ such that*

$$f_m(x;\, \mu_m = 0,\, \sigma_m = 1) \leqslant \min\{v_0,\, v_1 \cdot |x|^{-\beta}\}$$

for all m.

This assumption means that the f_m $(m = 1, \ldots, M)$ are bounded and that their tails decrease to zero as fast as or faster than $|x|^{-\beta}$, which is a very mild condition.

The following three regularity conditions are standard conditions assumed in discussing strong consistency of the MLE. Let Γ denote any compact subset of Θ.

Assumption 2. *For $\theta \in \Theta$ and any positive real number ρ, let*

$$f(x;\, \theta,\, \rho) \equiv \sup_{\mathrm{dist}(\theta',\theta)\leqslant\rho} f(x;\, \theta').$$

For each $\theta \in \Gamma$ and sufficiently small ρ, $f(x;\, \theta,\, \rho)$ is measurable.

Assumption 3. *For each $\theta \in \Gamma$, if $\lim_{j\to\infty} \theta^{(j)} = \theta$ ($\theta^{(j)} \in \Gamma$) then $\lim_{j\to\infty} f(x;\,\theta^{(j)}) = f(x;\,\theta)$ except on a set which is a null set and does not depend on the sequence $\{\theta^{(j)}\}_{j=1}^{\infty}$.*

Assumption 4.

$$\int |\log f(x;\,\theta_0)| f(x;\,\theta_0)\mathrm{d}x < \infty.$$

Let $\mathrm{E}_0[\cdot]$ denote the expectation under the true parameter θ_0. The following theorem is essential to our argument as well as being of some independent interest.

Theorem 1. *Suppose that Assumptions 1–4 are satisfied and $f_0 \in \mathcal{G}_M \backslash \mathcal{G}_{M-1}$, where $\mathcal{G}_M$ and $\mathcal{G}_{M-1}$ are defined in* (2). *Then there exist real constants $\kappa, \lambda > 0$ such that*

$$\mathrm{E}_0[\log\{g(x) + \kappa\}] + \lambda < \mathrm{E}_0[\log f(x;\,\theta_0)] \tag{3}$$

for all $g \in \mathcal{G}_{M-1}$.

We now state the main theorem of this paper.

Theorem 2. *Suppose that Assumptions 1–4 are satisfied and $f_0 \in \mathcal{G}_M \backslash \mathcal{G}_{M-1}$, where $\mathcal{G}_M$ and $\mathcal{G}_{M-1}$ are defined in* (2). *Let $c_0 > 0$ and $0 < d < 1$. If $c_n = c_0 \cdot \exp(-n^d)$ and*

$$\Theta_n \equiv \{\theta \in \Theta \,|\, \sigma_m \geqslant c_n,\, (m = 1, \ldots, M)\},$$

then

$$\mathrm{Prob}\Big(\lim_{n\to\infty} \mathrm{dist}(\Theta(\hat{\theta}_n),\, \Theta_0) = 0\Big) = 1,$$

where $\hat{\theta}_n$ is an MLE restricted to Θ_n.

As remarked at the end of the previous section, Theorem 2 implies the following corollary.

Corollary 1. *Under the assumptions of Theorem 2, $\hat{f}_n$ is strongly consistent in the sense of Definition 1.*

4. Proofs

In this section, we prove the theorems stated in Section 3. In Section 4.1 we state some lemmas required for Theorems 1 and 2. Then, in Section 4.2 we prove Theorem 1, which is also essential for Theorem 2. Finally, in Section 4.3 we prove Theorem 2. Our proofs are basically along the same lines as in the case of finite mixtures of uniform distributions in Tanaka and Takemura (2005). Therefore we omit the proofs of the lemmas. Full details of our proofs can be found in the preprint version of this paper (Tanaka and Takemura 2006).

4.1. Notation and some lemmas

Fix an arbitrary $\kappa_0 > 0$, which corresponds to κ in Theorem 1. Define $\tilde{\beta}$ and $\nu(y)$, $y > 0$, as

$$\tilde{\beta} \equiv \frac{\beta - 1}{\beta}, \qquad \nu(y) \equiv \left(\frac{\upsilon_1}{\kappa_0}\right)^{1/\beta} y^{\tilde{\beta}}, \tag{4}$$

where υ_1 and β are given in Assumption 1. Noting that $\upsilon_1 \cdot (\nu(y))^{-\beta} = \kappa_0/y$, the following lemma is easily proved.

Lemma 1. *Under Assumption 1, for arbitrary $\kappa_0 > 0$ each component density $f_m(x; \mu, \sigma)$ is bounded by a step function*

$$f_m(x; \mu, \sigma) \leqslant \max\left\{1_{[\mu-\nu(\sigma),\mu+\nu(\sigma))}(x) \cdot \frac{\upsilon_0}{\sigma}, \kappa_0\right\} \leqslant 1_{[\mu-\nu(\sigma),\mu+\nu(\sigma))}(x) \cdot \frac{\upsilon_0}{\sigma} + \kappa_0,$$

where $1_U(x)$ denotes the indicator function of $U \subset \mathbb{R}$.

From Lemma 1,

$$\sum_{m=1}^{M} \alpha_m f_m(x; \mu_m, \sigma_m) \leqslant \sum_{m=1}^{M} 1_{[\mu_m-\nu(\sigma_m),\mu_m+\nu(\sigma_m))}(x) \cdot \frac{\upsilon_0}{\sigma_m} + \kappa_0. \tag{5}$$

The right-hand side of (5) is a step function. We look at this step function where the density $f(x; \theta)$ is high, that is, the scale parameter of some component is small.

For a given choice of $\kappa_0 > 0$, choose $c_0 > 0$ such that

$$c_0 < \frac{\upsilon_0}{\kappa_0(M+1)}.$$

Below we will impose additional constraints on κ_0 and c_0 to make them sufficiently small to satisfy other conditions. For each θ, let

$$\mathcal{K}_{\sigma \leqslant c_0} = \mathcal{K}_{\sigma \leqslant c_0}(\theta) \equiv \{m \mid 1 \leqslant m \leqslant M, \sigma_m \leqslant c_0\}$$

denote the set of components with $\sigma_m \leqslant c_0$ and define

$$J(\theta) \equiv \bigcup_{m \in \mathcal{K}_{\sigma \leqslant c_0}} [\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m)).$$

On $J(\theta)$ the density $f(x; \theta)$ is high. Now dividing $J(\theta)$ according to the height of the step function on the right-hand side of (5), for $x \in J(\theta)$ we can write the right-hand side of (5) as

$$1_{J(\theta)}(x) \cdot \left\{\sum_{m=1}^{M} 1_{[\mu_m-\nu(\sigma_m),\mu_m+\nu(\sigma_m))}(x) \cdot \frac{\upsilon_0}{\sigma_m} + \kappa_0\right\} = \sum_{t=1}^{T(\theta)} H(J_t(\theta)) \cdot 1_{J_t(\theta)}(x),$$

where $J_t(\theta)$ $(t = 1, \ldots, T(\theta))$ are disjoint intervals, $[\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m))$ $(m \in \mathcal{K}_{\sigma \leqslant c_0})$ are unions of some of the $J_t(\theta)$s and the height $H(J_t(\theta))$ for each t is defined by any $x \in J_t(\theta)$ as

$$H(J_t(\theta)) \equiv \sum_{m=1}^{M} 1_{[\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m))}(x) \cdot \frac{v_0}{\sigma_m} + \kappa_0.$$

For $x \in J_t(\theta)$, there is at least one $m = m_t$ such that $x \in [\mu_m - \nu(\sigma_m), \mu_m + \nu(\sigma_m))$ and $H(J_t(\theta)) \geqslant v_0/c_0 + \kappa_0$. Also note that the total number $T(\theta)$ of $J_t(\theta)$s satisfies $T(\theta) \leqslant 2M$, because the change in the height can only occur at $\mu_m - \nu(\sigma_m)$ or $\mu_m + \nu(\sigma_m)$.

By (5) we have the following lemma for $x \in J(\theta)$.

Lemma 2. *Under Assumption 1, for each* $x \in J(\theta)$,

$$\sum_{m=1}^{M} \alpha_m f_m(x; \mu_m, \sigma_m) \leqslant \sum_{t=1}^{T(\theta)} H(J_t(\theta)) \cdot 1_{J_t(\theta)}(x).$$

A density can be high only in a small region, and we want to have some explicit bound on the length $W(J_t(\theta))$ of $J_t(\theta)$ in terms of its height $H(J_t(\theta))$. Let

$$v_2 \equiv 2\left(\frac{v_1}{\kappa_0}\right)^{1/\beta} (v_0 \cdot (M+1))^{\tilde{\beta}}, \qquad \xi(y) \equiv v_2 \cdot \left(\frac{1}{y}\right)^{\tilde{\beta}}, \quad y > 0, \tag{6}$$

where v_0, v_1 and β are given in Assumption 1 and $\tilde{\beta}$ is defined in (4).

Lemma 3. *Under Assumption 1, the length* $W(J_t(\theta))$ *of* $J_t(\theta)$ *for each* t *is bounded by*

$$W(J_t(\theta)) \leqslant v_2 \cdot \left(\frac{1}{H(J_t(\theta))}\right)^{\tilde{\beta}} = \xi(H(J_t(\theta))).$$

So far we have been concerned with bounding the density at its peaks. Now we consider bounding the tail of the true density $f(x; \theta_0)$. Write $\bar{\mu}_0 \equiv \max(|\mu_{01}|, \ldots, |\mu_{0M}|)$ and $\theta_0 = (\alpha_{01}, \mu_{01}, \sigma_{01}, \ldots, \alpha_{0M}, \mu_{0M}, \sigma_{0M})$. Let

$$u_0 \equiv \sup_x f(x; \theta_0), \qquad u_1 \equiv \max\left(u_0 \cdot (2\bar{\mu}_0)^{\beta}, 2^{\beta} v_1 \sum_{m=1}^{M} \alpha_{0m}\sigma_{0m}^{\beta-1}\right). \tag{7}$$

Lemma 4. *Under Assumption 1,*

$$f(x; \theta_0) \leqslant \min\{u_0, u_1 \cdot |x|^{-\beta}\}, \qquad \forall x \in \mathbb{R}.$$

Based on Lemma 4, we can bound the behaviour of the minimum and the maximum of the sample. Let $x_1, \ldots, x_n$ denote a random sample of size n from $f(x; \theta_0)$ and let $x_{n,1} \equiv \min\{x_1, \ldots, x_n\}$, $x_{n,n} \equiv \max\{x_1, \ldots, x_n\}$. The following lemma follows from the Borel–Cantelli lemma.

Lemma 5. *For any real constant* $A_0 > 0$ *and* $\zeta > 0$, *define* $A_n \equiv A_0 \cdot n^{(2+\zeta)/(\beta-1)}$. *Then*

$$\mathrm{Prob}(x_{n,1} < -A_n \text{ or } x_{n,n} > A_n \textit{ infinitely often}) = 0.$$

Finally, we consider subprobability densities in $\mathcal{G}_{\mathcal{K}}$. For any positive real number ρ, let

$$f_{\mathcal{K}}(x;\, \theta_{\mathcal{K}},\, \rho) \equiv \sup_{\mathrm{dist}(\theta'_{\mathcal{K}}, \theta_{\mathcal{K}}) \leqslant \rho} f_{\mathcal{K}}(x;\, \theta'_{\mathcal{K}}) \qquad (\theta'_{\mathcal{K}} \in \overline{\Theta}_{\mathcal{K}}).$$

The following lemma follows from the bounded convergence theorem.

Lemma 6. *Let $\Gamma_{\mathcal{K}}$ denote any compact subset of $\overline{\Theta}_{\mathcal{K}}$. For any real constant $\kappa \geqslant 0$ and any point $\theta_{\mathcal{K}} \in \Gamma_{\mathcal{K}}$, the following equality holds under Assumptions 1 and 3:*

$$\lim_{\rho \to 0} \mathrm{E}_0[\log\{f_{\mathcal{K}}(x;\, \theta_{\mathcal{K}},\, \rho) + \kappa\}] = \mathrm{E}_0[\log\{f_{\mathcal{K}}(x;\, \theta_{\mathcal{K}}) + \kappa\}].$$

4.2. Proof of Theorem 1

In this section we prove Theorem 1 by contradiction. Fix an arbitrary proper subset $\mathcal{L}$ of $\{1, \ldots, M\}$. It suffices to prove that (3) holds for all $g \in \mathcal{G}_{\mathcal{L}}$. Suppose that (3) does not hold for some $\mathcal{G}_{\mathcal{L}}$. Then for any $\lambda,\, \kappa > 0$, there exists $g \in \mathcal{G}_{\mathcal{L}}$ such that

$$\mathrm{E}_0[\log\{g(x) + \kappa\}] + \lambda \geqslant \mathrm{E}_0[\log f(x;\, \theta_0)].$$

Here, let $\{\lambda_j\}$, $\{\kappa_j\}$ be positive sequences which decrease to zero. Then for each $\lambda_j,\, \kappa_j > 0$, there exists $g_j \in \mathcal{G}_{\mathcal{L}}$ such that $\mathrm{E}_0[\log\{g_j(x) + \kappa_j\}] + \lambda_j \geqslant \mathrm{E}_0[\log f(x;\, \theta_0)]$. It follows that

$$\liminf_{j \to \infty} \mathrm{E}_0[\log\{g_j(x) + \kappa_j\}] + \lambda_j \geqslant \mathrm{E}_0[\log f(x;\, \theta_0)]. \tag{8}$$

Now g_j can be written as $g_j(x) = f_{\mathcal{L}}(x;\, \theta_{\mathcal{L}}^{(j)})$. Then the following lemma holds by compactification argument.

Lemma 7. *There exist a subsequence of $\{\theta_{\mathcal{L}}^{(j)}\}_{j=1}^{\infty} \equiv \{\{\alpha_m^{(j)},\, \mu_m^{(j)},\, \sigma_m^{(j)} \mid m \in \mathcal{L}\}\}_{j=1}^{\infty}$ and disjoint subsets $\mathcal{K}_{\sigma\downarrow 0}$, $\mathcal{K}_{\sigma\uparrow\infty}$, $\mathcal{K}_{|\mu|\uparrow\infty} \subset \mathcal{L}$ such that along the subsequence*

$$\sigma_m^{(j)} \to 0 \quad \textit{for } m \in \mathcal{K}_{\sigma\downarrow 0}, \qquad \sigma_m^{(j)} \to \infty \quad \textit{for } m \in \mathcal{K}_{\sigma\uparrow\infty},$$

$$\sigma_m^{(j)} \textit{ converges to a finite value and } |\mu_m^{(j)}| \to \infty \quad \textit{for } m \in \mathcal{K}_{|\mu|\uparrow\infty},$$

$$(\alpha_m^{(j)},\, \mu_m^{(j)},\, \sigma_m^{(j)}) \textit{ converges to a finite point } (\alpha_m^{(\infty)},\, \mu_m^{(\infty)},\, \sigma_m^{(\infty)}) \quad \textit{for } m \in \mathcal{K}_R,$$

where $\mathcal{K}_R \equiv \mathcal{L} \backslash \{\mathcal{K}_{\sigma\downarrow 0} \cup \mathcal{K}_{\sigma\uparrow\infty} \cup \mathcal{K}_{|\mu|\uparrow\infty}\}$.

From Lemma 7, we define g_∞ as $g_\infty(x) \equiv \sum_{m \in \mathcal{K}_R} \alpha_m^{(\infty)} f_m(x;\, \mu_m^{(\infty)},\, \sigma_m^{(\infty)}) \in \mathcal{G}_{\mathcal{K}_R}$. For notational simplicity and without loss of generality, we replace the original sequence with this subsequence, because (8) holds for this subsequence as well. Furthermore, by considering the sequence $\{\theta_{\mathcal{L}}^{(j)}\}_{j=j_0}^{\infty}$ where j_0 is sufficiently large and replacing j by $j - j_0$ if necessary, we can assume without loss of generality that there exist sufficiently small real constants $\kappa_0 > 0$ and $c_0 > 0$ such that

$$\mathrm{E}_0[\log f(x;\,\theta_0)] - \mathrm{E}_0[\log\{g_\infty(x) + 3\kappa_0\}] > 0, \qquad \kappa_0 < \frac{v_0}{c_0(M+1)},$$

$$\sigma_m^{(j)} < c_0 \quad (m \in \mathcal{K}_{\sigma\downarrow 0}), \qquad \sigma_m^{(j)} > \frac{v_0}{\kappa_0} \quad (m \in \mathcal{K}_{\sigma\uparrow\infty}),$$

$$c_0 \leqslant \sigma_m^{(j)} \leqslant \frac{v_0}{\kappa_0} \qquad (m \in \mathcal{K}_{|\mu|\uparrow\infty}), \text{ for all } j. \tag{9}$$

From Lemmas 1 and 2, we have

$$\begin{aligned}
&\mathrm{E}_0[\log\{f_{\mathcal{L}}(x;\,\theta_{\mathcal{L}}^{(j)}) + \kappa_j\} + \lambda_j] \\
&\quad \leqslant \int 1_{J(\theta_{\mathcal{L}}^{(j)})}(x)\cdot\log\left\{\sum_{t=1}^{T(\theta_{\mathcal{L}}^{(j)})} H(J_t(\theta_{\mathcal{L}}^{(j)}))\cdot 1_{J_t(\theta_{\mathcal{L}}^{(j)})}(x) + \kappa_j\right\} f(x;\,\theta_0)\mathrm{d}x \\
&\qquad + \int 1_{\mathbb{R}\setminus J(\theta_{\mathcal{L}}^{(j)})}(x)\cdot\log\left\{f_{\mathcal{K}_\sigma>0}(x;\,\theta_{\mathcal{K}_\sigma>0}^{(j)}) + \kappa_0 + \kappa_j\right\} f(x;\,\theta_0)\mathrm{d}x + \lambda_j,
\end{aligned} \tag{10}$$

where $\mathcal{K}_{\sigma>0} \equiv \mathcal{L}\setminus\mathcal{K}_{\sigma\downarrow 0}$.

Now we evaluate the first term on the right-hand side of (10). From Lemma 3,

$$\begin{aligned}
&\int 1_{J(\theta_{\mathcal{L}}^{(j)})}(x)\cdot\log\left\{\sum_{t=1}^{T(\theta_{\mathcal{L}}^{(j)})} H(J_t(\theta_{\mathcal{L}}^{(j)}))\cdot 1_{J_t(\theta_{\mathcal{L}}^{(j)})}(x) + \kappa_j\right\} f(x;\,\theta_0)\mathrm{d}x \\
&\quad \leqslant \sum_{t=1}^{T(\theta_{\mathcal{L}}^{(j)})} W(J_t(\theta_{\mathcal{L}}^{(j)}))\cdot\log\left\{H(J_t(\theta_{\mathcal{L}}^{(j)})) + \kappa_j\right\}\cdot u_0 \to 0 \qquad (n\to\infty),
\end{aligned} \tag{11}$$

where $u_0 = \sup_x f(x;\,\theta_0)$ defined in (7). Next we evaluate the second term on the right-hand side of (10). Let $A^{(j)} \equiv \min_{m\in\mathcal{K}_{|\mu|\uparrow\infty}}\{\min\{|\mu_m^{(j)} + \nu(\sigma_m^{(j)})|,\, |\mu_m^{(j)} - \nu(\sigma_m^{(j)})|\}\}$. Then $f_{\mathcal{K}_{\sigma\uparrow\infty}}(x;\,\theta_{\mathcal{K}_{\sigma\uparrow\infty}}^{(j)}) + f_{\mathcal{K}_{|\mu|\uparrow\infty}}(x;\,\theta_{\mathcal{K}_{|\mu|\uparrow\infty}}^{(j)}) \leqslant \kappa_0$ for $x \in [-A^{(j)},\, A^{(j)}]\setminus J(\theta^{(j)})$. Therefore, the following inequality holds:

$$\begin{aligned}
&\int 1_{\mathbb{R}\setminus J(\theta_{\mathcal{L}}^{(j)})}(x)\cdot\log\{f_{\mathcal{K}_{\sigma>0}}(x;\,\theta_{\mathcal{K}_{\sigma>0}}^{(j)}) + \kappa_0 + \kappa_j\} f(x;\,\theta_0)\mathrm{d}x \\
&\quad \leqslant \int 1_{[-A^{(j)},A^{(j)}]\setminus J(\theta_{\mathcal{L}}^{(j)})}(x)\cdot\log\left\{f_{\mathcal{K}_R}(x;\,\theta_{\mathcal{K}_R}^{(j)}) + 2\kappa_0 + \kappa_j\right\} f(x;\,\theta_0)\mathrm{d}x \\
&\qquad + \int 1_{\mathbb{R}\setminus\{[-A^{(j)},A^{(j)}]\cup J(\theta_{\mathcal{L}}^{(j)})\}}(x)\cdot\log\{f_{\mathcal{K}_{\sigma>0}}(x;\,\theta_{\mathcal{K}_{\sigma>0}}^{(j)}) + \kappa_0 + \kappa_j\} f(x;\,\theta_0)\mathrm{d}x \\
&\quad \equiv I_1^{(j)} + I_2^{(j)}
\end{aligned} \tag{12}$$

(say). By the bounded convergence theorem, we obtain

$$I_1^{(j)} \to \int \log\{g_\infty(x) + 2\kappa_0\} f(x;\, \theta_0)\mathrm{d}x, \qquad I_2^{(j)} \to 0. \tag{13}$$

From (10)–(13), we have

$$\mathrm{E}_0[\log f(x;\, \theta_0)] \leqslant \limsup_{j\to\infty} \mathrm{E}_0[\log\{g_j(x) + \kappa_j\}] + \lambda_j \leqslant \mathrm{E}_0[\log\{g_\infty(x) + 2\kappa_0\}].$$

This is a contradiction to (9). This completes the proof of Theorem 1.

4.3. Proof of Theorem 2

We choose real constants κ and λ to satisfy (3) by using Theorem 1. Having chosen these constants, from now on we proceed along the lines of the proof in Tanaka and Takemura (2005), although the details of the proof here are much more complicated. For the sake of readability, we divide our proof into further sections.

4.3.1. Setting up constants

For κ, λ satisfying (3), let κ_0, λ_0 be real constants such that $0 < 4\kappa_0 \leqslant \kappa$, $0 < 4\lambda_0 \leqslant \lambda$. Note that $4\kappa_0$, $4\lambda_0$ also satisfy (3). Define

$$B \equiv \frac{v_0}{\kappa_0} > \max\{\sigma_{01}, \ldots, \sigma_{0M}\}.$$

If $\sigma_m \geqslant B$, then the density of the mth component is almost flat and makes little contribution to the likelihood.

Because $\{c_n\}$ is decreasing to zero, by replacing c_0 by some c_n if necessary, we can assume without loss of generality that c_0 is sufficiently small to satisfy the following conditions,

$$\begin{aligned}
&\left(\frac{v_0}{c_0}\right)^{\tilde{\beta}} > \mathrm{e}, \qquad c_0 < \min\{\sigma_{01}, \ldots, \sigma_{0M}\}, \qquad \kappa_0 < \frac{v_0}{c_0(M+1)},\\
&3M \cdot u_0 \cdot 2\nu(c_0) \cdot |\log \kappa_0| < \lambda_0, \\
&3 \cdot 2M \cdot u_0 \cdot \xi(v_0/c_0) \cdot \log(v_0/c_0) < \lambda_0,
\end{aligned} \tag{14}$$

where $\tilde{\beta}$, $\nu(\cdot)$ and $\xi(\cdot)$ are defined in (4) and (6).

For any subset $V \subset \mathbb{R}$, let $P_0(V)$ denote the probability of V under the true density

$$P_0(V) \equiv \int_V f(x;\, \theta_0)\mathrm{d}x.$$

Let $A_0 > 0$ be a positive constant which satisfies

$$P_0(\mathcal{A}_0) \cdot \log\left(\frac{v_0/c_0 + 2\kappa_0}{3\kappa_0}\right) < \lambda_0, \tag{15}$$

where $\mathcal{A}_0 \equiv (-\infty, -A_0] \cup [A_0, \infty)$. Let $A_n \equiv A_0 \cdot n^{(2+\xi)/(\beta-1)}$ as in Lemma 5. Define a subset Θ'_n of Θ_n in Theorem 2 by

$$\Theta'_n \equiv \{\theta \in \Theta_n \mid \exists m \text{ s.t. } c_n \leqslant \sigma_m \leqslant c_0 \text{ or } |\mu_m| > A_0\} \subset \Theta_n.$$

4.3.2. Partitioning the parameter space

In view of theorems in Wald (1949) and Redner (1981), for the strong consistency of the MLE on Θ_n under Assumption 1–4, it suffices to prove that

$$\lim_{n\to\infty} \frac{\sup_{\theta\in\Theta'_n} \prod_{i=1}^{n} f(x_i;\, \theta)}{\prod_{i=1}^{n} f(x_i;\, \theta_0)} = 0, \qquad \text{almost everywhere (a.e.).}$$

In our proof, we consider finer and finer finite coverings of Θ'_n.

Let $\theta \in \Theta'_n$. Let $\mathcal{K}_{\sigma\leqslant c_0}, \mathcal{K}_{\sigma\geqslant B}, \mathcal{K}_{|\mu|\geqslant A_0}$ represent disjoint subsets of $\{1, \ldots, M\}$ and define

$$\mathcal{K}_R \equiv \{1, \ldots, M\}\backslash\{\mathcal{K}_{\sigma\leqslant c_0} \cup \mathcal{K}_{\sigma\geqslant B} \cup \mathcal{K}_{|\mu|\geqslant A_0}\}.$$

For any given $\mathcal{K}_{\sigma\leqslant c_0}, \mathcal{K}_{\sigma\geqslant B}, \mathcal{K}_{|\mu|\geqslant A_0}$, we define a subset of Θ'_n by

$$\Theta'_{n,\mathcal{K}} \equiv \{\theta \in \Theta'_n \mid \sigma_m \leqslant c_0 \ (m \in \mathcal{K}_{\sigma\leqslant c_0});\ \sigma_m \geqslant B \ (m \in \mathcal{K}_{\sigma\geqslant B});$$
$$c_0 < \sigma_m < B,\ |\mu_m| \geqslant A_0 \ (m \in \mathcal{K}_{|\mu|\geqslant A_0});\ c_0 < \sigma_m < B,\ |\mu_m| < A_0 \ (m \in \mathcal{K}_R)\}.$$

As above, it suffices to prove that for each choice of disjoint subsets $\mathcal{K}_{\sigma\leqslant c_0}, \mathcal{K}_{\sigma\geqslant B}, \mathcal{K}_{|\mu|\geqslant A_0}$, the ratio of the supremum of the likelihood over $\Theta'_{n,\mathcal{K}}$ to the likelihood at θ_0 converges to zero almost everywhere. We fix $\mathcal{K}_{\sigma\leqslant c_0}, \mathcal{K}_{\sigma\geqslant B}, \mathcal{K}_{|\mu|\geqslant A_0}$, from now on.

Next we consider coverings of $\overline{\Theta}_{\mathcal{K}_R}$. The following lemma follows from Lemma 6 and compactness of $\overline{\Theta}_{\mathcal{K}_R}$.

Lemma 8. *Let $\mathcal{B}(\theta, \rho(\theta))$ denote the open ball with centre θ and radius $\rho(\theta)$. Then $\overline{\Theta}_{\mathcal{K}_R}$ can be covered by a finite number of balls $\mathcal{B}(\theta^{(1)}_{\mathcal{K}_R}, \rho(\theta^{(1)}_{\mathcal{K}_R})), \ldots, \mathcal{B}(\theta^{(S)}_{\mathcal{K}_R}, \rho(\theta^{(S)}_{\mathcal{K}_R}))$ such that*

$$\mathrm{E}_0[\log\{f_{\mathcal{K}_R}(x;\, \theta^{(s)}_{\mathcal{K}_R}, \rho(\theta^{(s)}_{\mathcal{K}_R})) + \kappa_0\}] + \lambda_0 < \mathrm{E}_0[\log f(x;\, \theta_0)] \qquad (s = 1, \ldots, S).$$

Based on Lemma 8 we partition $\Theta'_{n,\mathcal{K}}$. Define a subset of $\Theta'_{n,\mathcal{K}}$ by

$$\Theta'_{n,\mathcal{K},s} \equiv \{\theta \in \Theta'_{n,\mathcal{K}} \mid \theta_{\mathcal{K}_R} \in \mathcal{B}(\theta^{(s)}_{\mathcal{K}_R}, \rho(\theta^{(s)}_{\mathcal{K}_R}))\}.$$

Then $\Theta'_{n,\mathcal{K}}$ is covered by $\Theta'_{n,\mathcal{K},1}, \ldots, \Theta'_{n,\mathcal{K},S}$. Again it suffices to prove that for each choice of $\mathcal{K}_{\sigma\leqslant c_0}, \mathcal{K}_{\sigma\geqslant B}, \mathcal{K}_{|\mu|\geqslant A_0}$, s the ratio of the supremum of the likelihood over $\Theta'_{n,\mathcal{K},s}$ to the likelihood at θ_0 converges to zero almost everywhere. We fix $\mathcal{K}_{\sigma\leqslant c_0}, \mathcal{K}_{\sigma\geqslant B}, \mathcal{K}_{|\mu|\geqslant A_0}$ and s from now on. Then it suffices to prove the following inequality, which is a new intermediate goal of our proof:

$$\operatorname{im}\sup_{n\to\infty} \frac{1}{n} \sup_{\theta\in\Theta'_{n,\mathcal{K},s}} \sum_{i=1}^{n} \log f(x_i;\, \theta)\} < \mathrm{E}_0[\log f(x;\, \theta_0)], \qquad \text{a.e.} \tag{16}$$

4.3.3. Bounding the likelihood by four terms

In this section we bound the log-likelihood function of the left-hand side of (16) by four terms depending on the positions of the observations $x_1, \ldots, x_n$. Let $R_n(V)$ denote the number of observations which belong to a set $V \subset \mathbb{R}$.

Lemma 9. *For $\theta \in \Theta'_{n,\mathcal{K},s}$,*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\, \theta)\} \leqslant{} & \frac{1}{n}\sum_{i=1}^{n} \log\{f_{\mathcal{K}_R}(x_i;\, \theta_{\mathcal{K}_R},\, \rho(\theta_{\mathcal{K}_R})) + 3\kappa_0\} \\ & + \frac{1}{n} R_n(\mathcal{A}_0)\cdot \log\left(\frac{Mv_0/c_0 + 2\kappa_0}{3\kappa_0}\right) \\ & + \frac{1}{n} R_n(J(\theta))\cdot(-\log\kappa_0) + \frac{1}{n}\sum_{x_i\in J(\theta)} \log f(x_i;\, \theta).\end{aligned} \tag{17}$$

From Lemma 8 and the strong law of large numbers, the first term on the right-hand side of (17) converges to the expectation of a density which has fewer than M components and the expectation is less than that of the true density by Theorem 1. The second term converges to a small value because the relative frequency on $\mathcal{A}_0$ is very small. The third term also converges to a small value because the relative frequency on $J(\theta)$ is very small. The fourth term is somewhat complicated. The component in $\mathcal{K}_{\sigma\leqslant c_0}$ may have high peaks. However, the peaks are very narrow and the relative frequency on the interval is very small. Hence, the fourth term makes little contribution to the likelihood. Therefore, the mean log-likelihood (the left-hand side of (17)) converges to a value which is less than that of the true density. In the following we consider the details.

The first and second terms are easy. For the first, by Lemma 8 and the strong law of large numbers we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log\{f_{\mathcal{K}_R}(x_i;\, \theta_{\mathcal{K}_R},\, \rho(\theta_{\mathcal{K}_R})) + 4\kappa_0\} < \mathrm{E}_0[\log f(x;\, \theta_0)] - 4\lambda_0, \qquad \text{a.e.} \tag{18}$$

For the second, by (15) and the strong law of large numbers we have

$$\lim_{n\to\infty} \frac{1}{n} R_n(\mathcal{A}_0)\cdot\log\left(\frac{v_0/c_0 + 2\kappa_0}{3\kappa_0}\right) < \lambda_0, \qquad \text{a.e.} \tag{19}$$

Note that we have $-4\lambda_0$ from the first term and λ_0 from the second term. In the rest of our proof we show that both the third and fourth terms can be bounded by λ_0.

4.3.4. Bounding the third term

The third term can be bounded by dividing the interval $[-A_n, A_n]$ into short intervals of length $2\nu(c_0)$.

Lemma 10.

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta'_{n,\mathcal{K},s}} \frac{1}{n} R_n(J(\theta)) \leqslant 3M \cdot u_0 \cdot 2\nu(c_0), \qquad \text{a.e.}$$

By this lemma and (14) we have

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta'_{n,\mathcal{K},s}} \frac{1}{n} R_n(J(\theta)) \cdot (-\log\kappa_0) \leqslant 3M \cdot u_0 \cdot 2\nu(c_0) \cdot |\log\kappa_0| < \lambda_0, \qquad \text{a.e.} \tag{20}$$

This bounds the third term on the right-hand side of (17) from above.

4.3.5. Bonding the fourth term

Finally, we bound the fourth term on the right-hand side of (17) from above. From Lemma 2 we have

$$1_{J(\theta)}(x) \cdot \sum_{m=1}^{M} \alpha_m f_m(x; \mu_m, \sigma_m) \leqslant \sum_{t=1}^{T(\theta)} H(J_t(\theta)) \cdot 1_{J_t(\theta)}(x), \qquad x \in J(\theta). \tag{21}$$

We now classify the intervals $J_t(\theta)$, $t = 1, \ldots, T(\theta)$, by the height $H(J_t(\theta))$. Let $c'_n \equiv c_0 \cdot \exp(-n^{1/4})$ and define $\tau_n(\theta)$ and $\tau'_n(\theta)$ by

$$\tau_n(\theta) \equiv \left\{ t \in \{1, \ldots, T(\theta)\} \mid H(J_t(\theta)) \leqslant M \frac{v_0}{c'_n} \right\}, \qquad \tau'_n(\theta) \equiv \{1, \ldots, T(\theta)\} \backslash \tau_n(\theta).$$

Now suppose that

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta'_{n,\mathcal{K},s}} \left[\sum_{t=1}^{T(\theta)} \frac{1}{n} R_n(J_t(\theta)) \log H(J_t(\theta)) \right.$$
$$\left. -3 \left\{ \sum_{t\in\tau_n(\theta)} u_0 \cdot \xi(H(J_t(\theta))) \cdot \log H(J_t(\theta)) + \sum_{t\in\tau'_n(\theta)} \frac{2}{n} \log H(J_t(\theta)) \right\} \right] \leqslant 0, \qquad \text{a.e.} \tag{22}$$

Then, from (14), (21) and (22), the fourth term on the right-hand side of (17) is bounded from above by

$$\limsup_{n\to\infty} \frac{1}{n} \sup_{\theta\in\Theta'_{n,\mathcal{K},s}} \sum_{x_i\in J(\theta)} \log f(x_i; \theta) \leqslant \lambda_0, \qquad \text{a.e.} \tag{23}$$

Combining (17)–(20) and (23), we obtain (16). Therefore it suffices to prove (22), which is a new goal of our proof.

We now consider a further finite covering of $\Theta'_{n,\mathcal{K},s}$. For any T $(1 \leqslant T \leqslant 2M)$ and $\tau \subset \{1, \ldots, T(\theta)\}$, define a subset of $\Theta_{n,\mathcal{K},s}$ by

$$\Theta'_{n,\mathcal{K},s,T,\tau} \equiv \{\theta \in \Theta'_{n,\mathcal{K},s} \mid T(\theta) = T,\ \tau_n(\theta) = \tau\}.$$

Then (22) is derived from the following two lemmas.

Lemma 11. *For* $\tau' = \{1, \ldots, T\}\backslash\tau$,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta'_{n,\mathcal{K},s,T,\tau}} \left[\sum_{t\in\tau'} \frac{1}{n} R_n(J_t(\theta)) \cdot \log H(J_t(\theta)) - 3\sum_{t\in\tau'} \frac{2}{n} \log H(J_t(\theta))\right] \leqslant 0, \qquad \text{a.e.}$$

Lemma 12.

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta'_{n,\mathcal{K},s,T,\tau}} \left[\operatorname*{sum}_{t\in\tau} \frac{1}{n} R_n(J_t(\theta)) \cdot \log H(J_t(\theta)) \right.$$
$$\left. -\ 3\sum_{t\in\tau} u_0 \cdot \xi(H(J_t(\theta))) \cdot \log H(J_t(\theta))\right] \leqslant 0, \qquad \text{a.e.}$$

This completes the proof of Theorem 2.

5. Discussions

In this paper we consider the strong consistency of the MLE for mixtures of location–scale distributions. We treat the case where the scale parameters of the component distributions are restricted from below by $c_n = \exp(-n^d)$, $0 < d < 1$, and give the regularity conditions for the strong consistency of MLE.

As in the case of the uniform mixture in Tanaka and Takemura (2005), it is readily verified that if c_n decreases to zero faster than $\exp(-n)$, then the consistency of the MLE fails. Therefore the rate of $c_n = \exp(-n^d)$, $0 < d < 1$, obtained in this paper is almost the lower bound of the order of c_n which maintains strong consistency.

Although we treat the univariate case in this paper, it is clear that the result obtained here can be extended to the multivariate case under the condition that components are bounded and their tails decrease to zero fast enough if the minimum singular values of the scale matrices of the components are restricted from below by c_n.

Finally, let us consider some sufficient conditions for the regularity conditions. For $\theta_m \in \Omega_m$ and any positive real number ρ, let $f_m(x;\theta_m,\rho) \equiv \sup_{\mathrm{dist}(\theta'_m,\theta_m)\leqslant\rho} f_m(x;\theta'_m)$. Let Γ be any compact subset of Ω_m. Consider the following two conditions.

Assumption 5. *For each* $\theta_m \in \Gamma$ *and sufficiently small* ρ, $f_m(x;\theta_m,\rho)$ *is measurable.*

Assumption 6. *For each* $\theta_m \in \Gamma$, *if* $\lim_{j\to\infty}\theta_m^{(j)} = \theta_m$, *then* $\lim_{j\to\infty} f_m(x;\theta_m^{(j)}) = f_m(x;\theta_m)$ *for all* x.

If Assumptions 5 and 6 hold, then it is easily verified that Assumptions 2 and 3 hold. Thus Assumptions 1 and 4–6 are sufficient conditions for regularity conditions and Assumptions 5 and 6 are checked more easily. For example, a finite mixture density which consists of a normal, t and uniform on an open interval satisfies Assumptions 1 and 4–6.

References

Billingsley, P. (1995) *Probability and Measure*, 3rd edn. New York: Wiley.

Everitt, B.S. and Hand, D.J. (1981) *Finite Mixture Distributions*. London: Chapman & Hall.

Geman, S. and Hwang, C.-R. (1982) Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, **10**, 401–414.

Genovese, C.R. and Wasserman, L. (2000) Rates of convergence for the Gaussian mixture sieve. *Ann. Statist*., **28**, 1105–1127.

Ghosal. S. and van der Vaart, A.W. (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, **29**, 1233–1263.

Grenander, U. (1981) *Abstract Inference*. New York: Wiley.

Hathaway, R.J. (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, **13**, 795–800.

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

Redner, R. (1981) Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.*, **9**, 225–228.

Tanaka, K. and Takemura, A. (2005) Strong consistency of MLE for finite uniform mixtures when the scale parameters are exponentially small. *Ann. Inst. Statist. Math.*, **57**, 1–19.

Tanaka, K. and Takemura, A. (2006) Strong consistency of MLE for finite mixtures of location–scale distributions when the scale parameters are exponentially small. http://arXiv.org/abs/math.ST/0605148.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist*., **20**, 595–601.

Received September 2004 and revised May 2006