The Annals of Statistics
1992, Vol. 20, No. 3, 1350–1360

CONSISTENT ESTIMATION OF A MIXING DISTRIBUTION

BY BRIAN G. LEROUX

University of Washington

A maximum-penalized-likelihood method is proposed for estimating a mixing distribution and it is shown that this method produces a consistent estimator, in the sense of weak convergence. In particular, a new proof of the consistency of maximum-likelihood estimators is given. The estimated number of components is shown to be at least as large as the true number, for large samples. Also, the large-sample limits of estimators which are constrained to have a fixed finite number of components are identified as distributions minimizing Kullback–Leibler divergence from the true mixing distribution. Estimation of a Poisson mixture distribution is illustrated using the distribution of traffic accidents presented by Simar.

1. Introduction. Given a family of densities $\{p(y, \theta): \theta \in \Theta\}$ with respect to a measure μ, a density of the form

$$p_F(y) = \int_\Theta p(y, \theta)\, dF(\theta) \tag{1}$$

is called a *mixture density* corresponding to the *mixing distribution* F. A *finite mixture density* is given by

$$p(y) = \sum_{j=1}^{m} \alpha_j p(y, \theta_j),$$

where $0 < \alpha_j \leq 1$ and $\sum_{j=1}^{m} \alpha_j = 1$. Mixture distributions have seen frequent application, especially mixtures of Poissons, binomials, normals and exponentials. The mixing distribution sometimes represents a physical reality, but otherwise, it can still provide an interpretive model for data. Parametric mixing distributions are often used to model distributional features such as overdispersion.

Much of the work on estimation of a mixing distribution has been concerned with maximum-likelihood procedures. For an observed random sample $y_1, \ldots, y_n$ from the density (1) the log-likelihood function is given by

$$l_n(F) = \sum_{i=1}^{n} \log p_F(y_i).$$

Early work on the maximum-likelihood estimator includes Kiefer and Wolfowitz (1956), Simar (1976), Laird (1978) and Lindsay (1983); the latter author proved that there is a maximum-likelihood estimate $\hat{F}$ with K or fewer components, where K is the number of distinct points in the sample. It is

Received December 1989; revised September 1991.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62F12.
Key words and phrases. Mixture distribution, maximum likelihood, maximum penalized likelihood, model selection.

easily seen that Lindsay's result holds if, for every i, the mapping $\theta \to p(y_i, \theta)$ is continuous and vanishes at ∞, that is, for every $\varepsilon > 0$ there is a compact $K_\varepsilon \subset \Theta$ for which $p_i(y_i, \theta) < \varepsilon$, $\theta \notin K_\varepsilon$; this includes mixtures of Poissons, binomials, exponentials and mean-parametrized normals. Lindsay's proof also establishes the existence of a constrained maximum-likelihood estimate $\hat{F}_m$ which maximizes the log-likelihood over all mixing distributions with m or fewer components.

Asymptotic results for maximum-likelihood estimators of the parameters defining a finite mixture are given by Sundberg (1974), Redner (1981), Redner and Walker (1984) and Hathaway (1985). To discuss the consistency of $\hat{F}$, the topology of weak convergence is imposed on the space of mixing distributions. Thus, $\hat{F}$ is consistent for the true mixing distribution F^* if, with probability 1, $\hat{F}$ converges weakly to F^* as $n \to \infty$ ($\hat{F} \to_w F^*$), that is, $\hat{F}(\theta) \to F^*(\theta)$ for all continuity points θ of F^*. Consistency in this sense implies, under mild conditions (see Lemma 2), that the estimated density $p_{\hat{F}}$ converges to the true density p_{F^*}. Consistency results for $\hat{F}$ are given by Kiefer and Wolfowitz (1956), Simar (1976), Jewell (1982), Heckman and Singer (1984) and Pfanzagl (1988).

On the subject of selecting the number of components, McLachlan and Basford (1988) discuss hypothesis-testing procedures and Henna (1985) gives a consistent estimator of the number of components.

In Section 2 maximum-penalized-likelihood methods are proposed for selecting the number of mixture components and an example of their use is given. Asymptotic results are presented in Section 3. These include the convergence of maximum-likelihood estimators constrained to have a fixed number of components, the large-sample behavior of the estimator of the number of components and the consistency, in the sense of weak convergence, of the mixing-distribution estimator obtained by the maximum-penalized-likelihood method.

2. Selecting the number of components. For those applications for which a finite mixing distribution is plausible, special consideration might be given to the problem of selecting the number of mixture components, m. Although $\hat{F}$ itself provides an estimate of m, it can include more components than are necessary for a good fit to the data. A procedure that penalizes overfitting might be preferable to maximum likelihood. On the other hand, if a continuous mixing distribution seems more physically meaningful, then an estimate of a mixing density function might be preferable.

In any case, a simple explanatory model for the data is provided by a small number of mixture components, and this could be useful in comparative studies involving estimates for two or more samples. In this regard, a finite mixture might uncover features that would remain hidden by fitting a parametric family of mixing distributions.

The elimination of unnecessary components might also lead to more precise estimates of the parameters in a finite mixture. Related to this possibility, Chen (1991) proves that the best possible rate of convergence of estimators for an overparametrized mixture (i.e., with more components than in the true

model) is $n^{-1/4}$, compared to the rate $n^{-1/2}$ which is achieved when the number of components is correctly specified.

The procedures for estimating the number of components are based on the general theory of model selection [see, e.g., Linhart and Zucchini (1986)]. Consider the sequence of nested models for the parameter F^* defined by the possible numbers of components. We propose the choice of $\hat{m}_n$ for m, where $\hat{m}_n$ maximizes

$$l_n(\hat{F}_m) - a_{mn} \tag{2}$$

over m; here a_{mn} is a penalty term, satisfying $a_{m+1,n} \geq a_{mn}$, which discourages the selection of a model with an excessive number of components and can depend on $y_1, \ldots, y_n$. The resulting mixing distribution estimator $\hat{F}_{\hat{m}_n}$ is called a *maximum-penalized-likelihood estimator*. We will consider, for example, the Akaike information criteria (AIC [Akaike (1973)]), given by $a_{mn} = \dim(\mathcal{F}_m)$, and the Bayesian information criterion (BIC [Schwarz (1978)]), given by $a_{mn} = (1/2)(\log n)\dim(\mathcal{F}_m)$, where $\dim(\mathcal{F}_m) = m - 1 + m \dim(\theta)$.

EXAMPLE. The following frequency distribution of the number of automobile accident claims in a single year for 9461 policies was given by Simar (1976):

Number of accident claims	0	1	2	3	4	5	6	7
Number of individuals	7840	1317	239	42	14	4	4	1

Simar fits a Poisson mixture to these data and claims the maximum-likelihood estimate is the distribution given by the following parameter values:

Probability	Rate
0.75997	0.08854
0.23617	0.58020
0.00370	3.17606
0.00016	3.66871

But the mean of this distribution, 0.21665, is not equal to the sample mean, 0.21435, and so this distribution cannot define a local maximum of the likelihood [see Lindsay (1981) or Titterington, Smith and Makov (1985), page 86].

We present below the constrained maximum-likelihood estimates for one, two and three components and the corresponding values of the log-likelihood (excluding the $y!$ term in the Poisson probability) and the penalized-likelihood criteria, AIC and BIC:

Number of components	Estimates		Log-likelihood	AIC	BIC
	Probability	Rate			
1	1	0.21435	−5151.38	−5152.38	−5155.96
2	0.93780	0.14694	−5008.56	−5011.56	−5022.29
	0.06220	1.23069			
3	0.41830	0.00000	−5001.30	−5006.30	−5024.19
	0.57302	0.33554			
	0.00868	2.54498			

The estimate with three components satisfies the directional-derivative inequality of Lindsay (1983), and so it must be the maximum-likelihood estimate (the reported estimate has directional derivative 0.04 at $\lambda = 0.1$, due to roundoff error).

BIC leads to a choice of two components, while three are indicated by AIC. The choice of the number of components might instead be based on a direct comparison of the fitted frequency distributions. The table below lists the fitted frequencies for the following estimates: (1) Poisson distribution, (2) constrained maximum-likelihood estimate with two components and (3) maximum-likelihood estimate:

Count	Observed frequency	(1)	(2)	(3)
0	7840	7635.6	7831.9	7840.0
1	1317	1636.7	1337.1	1317.0
2	239	175.4	212.9	239.1
3	42	12.5	57.5	42.1
4	14	0.7	16.6	13.3
5	4	0.0	4.0	5.9
6	4	0.0	0.8	2.4
7	1	0.0	0.1	0.9
8+	0	0.0	0.0	0.4

The fitted Poisson frequency distribution makes it clear that one component is certainly not sufficient. The two-component estimate provides a much improved fit, but might be judged inadequate because it underfits the number of individuals with six or more claims, if these large numbers of claims were considered especially important. On the other hand, the maximum-likelihood estimate might be overfitting.

3. Consistency results. The following conditions will be referred to in the sequel:

1. $p(y, \theta)$ is continuous on $E \times \Theta$, where E and Θ are Borel subsets of Euclidean spaces.
2. For any compact $C \subset E$ and $\varepsilon > 0$, there exist $a, b \in \Theta$ such that $p(y, \theta) < \varepsilon$, $\theta \in \Theta \setminus [a, b]$, $y \in C$ ($A \setminus B$ denotes the set of points in A and not in B).
3. There are Borel sets $Z \subset E$ and $\Omega \subset \Theta$ such that $\mu(Z) > 0$, $\int_\Omega dF^* > 0$, $p(y, \theta) = 0$ on $Z \times (\Theta \setminus \Omega)$, and $p(y, \theta) > 0$ on $E \times \Omega$.
4. $p(y, \theta) \leq h(y)$, $\theta \in \Theta$, $y \in E$, where h is continuous on E and $\int p_{F^*} |\log h| \, d\mu < \infty$.
5. $\int p_{F^*}(y)[\log p(y, \theta)]^- \, d\mu(y) < \infty$, $\theta \in \Omega$ ($x^- = \max\{-x, 0\}$).

Conditions 1 and 2 together are slightly stronger than the sufficient condition for the existence and finite characterization of $\hat{F}$ mentioned in Section 1. Condition 3 is satisfied for densities which are strictly positive over their entire range. The integrability conditions, 4 and 5, are slightly stronger than the requirement of finite entropy, that is, $\int p_{F^*} |\log p_{F^*}| \, d\mu < \infty$, which is frequently imposed in large-sample studies of maximum-likelihood estimators.

EXAMPLE 1 (Poisson). $p(y,\theta) = \theta^y e^{-\theta}$, $\theta \geq 0$, $y = 0, 1, \ldots$, and $d\mu(y) = 1/y!$. Condition 3 is satisfied with $Z = \{1, 2, \ldots\}$, $\Omega = (0, \infty)$, provided $dF^*(0) < 1$. Condition 4 is satisfied with $h(y) = y^y e^{-y}$, if $\sum p_{F^*}(y) y \log y < \infty$; a sufficient condition for this is $\int \theta^2 \, dF^*(\theta) < \infty$, and condition 5 is then also satisfied.

EXAMPLE 2 (Exponential). $p(y,\theta) = \theta e^{-y\theta}$, $y > 0$, $\theta > 0$, and μ is Lebesgue measure. Condition 4 is satisfied with $h(y) = e^{-1} \max\{y^{-1}, 1\}$, if $\int_0^1 p_{F^*}(y) \log(1/y) \, dy < \infty$; a sufficient condition for this is $\int \theta \, dF^*(\theta) < \infty$. Condition 5 is satisfied if $\int p_{F^*}(y) y \, dy < \infty$, or equivalently, $\int (1/\theta) \, dF^*(\theta) < \infty$.

EXAMPLE 3 (Normal mean). $p(y,\theta) = e^{-(y-\theta)^2/2\sigma^2}$, $y, \theta \in (-\infty, \infty)$, and μ is $(2\pi\sigma^2)^{-1/2}$ times Lebesgue measure. No restrictions on F^* are necessary in this case.

The family of mixture densities given by (1) is *identifiable* if

$$\int p(y,\theta) \, dF_1(\theta) = \int p(y,\theta) \, dF_2(\theta) \quad \text{a.e. } d\mu(y) \Rightarrow F_1 = F_2. \tag{3}$$

This holds in many cases, including the above three examples. Identifiability implies that the Kullback–Leibler divergence from p_{F_1} to p_{F_2}, that is, $K(F_1, F_2) = \int p_{F_1} \log(p_{F_1}/p_{F_2}) \, d\mu$, is positive if F_1 and F_2 are different; this property is stated below for future reference.

LEMMA 1. *Let F_1 be a distribution function and F_2 be a subdistribution function (which corresponds to a measure with total mass of 1 or less) on Θ. Then $K(F_1, F_2) \geq 0$ and, if the identifiability property (3) holds, then $K(F_1, F_2) > 0$ if $F_2 \neq F_1$.*

3.1. *Consistency of the maximum-likelihood estimator.* A new proof of the consistency of $\hat{F}$ will be given based on the following technical result.

LEMMA 2. *If conditions 1 and 2 hold, then the following are true:*

(i) *For any subdistribution function F on Θ, p_F is continuous on E.*

(ii) *If F_k and F are subdistribution functions on Θ and $F_k(\theta) \to F(\theta)$ for all continuity points θ of F, then $p_{F_k} \to p_F$ uniformly on compact subsets of E.*

PROOF. (i) is straightforward. The proof of (ii) involves establishing that the sequence $\{p_{F_k}\}$ is equicontinuous on a compact subset of E and using the Ascoli–Arzela theorem. □

THEOREM 1. *If conditions 1–5 and the identifiability property (3) hold, then $\hat{F} \to_w F^*$ as $n \to \infty$, with probability 1.*

PROOF. It is well known that the empirical distribution based on a random sample from a distribution function on a Euclidean space converges weakly to that distribution function, with probability 1. Therefore, with probability 1, the empirical distribution H_n based on $Y_1, \ldots, Y_n$ converges weakly to the distribution with density p_{F^*} with respect to μ. Also, by the strong law of large numbers, $\int \log h \, dH_n \to \int p_{F^*} \log h \, d\mu$, $\int \log p_{F^*} \, dH_n \to \int p_{F^*} \log p_{F^*} \, d\mu$, and $\int_Z dH_n \to \int_Z p_{F^*} \, d\mu$. The remainder of the proof is restricted to the event of probability 1 where these limits hold; we show that they imply $\hat{F} \to_w F^*$.

Assume there is a subsequence of $\{\hat{F}\}$ and a subdistribution function F such that $\hat{F} \to F$ along this subsequence at all continuity points of F. (Notice that subsequences do exist with these convergence properties, by the Helly selection theorem.) In the following all limits over n are assumed to be taken along this subsequence.

We first show $p_F(y) > 0$ for every y. According to condition 3, $p_F(y)$ can be 0 for some y only if $\int_\Omega dF = 0$, that is, if $\int_\Omega d\hat{F} \to 0$. But if this were true,

$$\frac{1}{n} l_n(\hat{F}) \le \int_E \log h(y) \, dH_n(y) + \log\left(\int_\Omega d\hat{F}\right) \int_Z dH_n$$

would imply $l_n(\hat{F})/n \to -\infty$, which leads to a contradiction, since

$$\frac{1}{n} l_n(\hat{F}) \ge \frac{1}{n} l_n(F^*) \to \int p_{F^*} \log p_{F^*} \, d\mu \in (-\infty, \infty). \tag{4}$$

Let C be compact. With $p_F > 0$, Lemma 2 implies that $\log p_{\hat{F}} \to \log p_F$ uniformly on C, and hence $|\int_C \log p_{\hat{F}} \, dH_n - \int_C \log p_F \, dH_n| \le \int_C |\log p_{\hat{F}} - \log p_F| \, dH_n \to 0$. Since $I_C \log p_F$ is bounded and upper semicontinuous (I_C is the indicator function of C), we have

$$\limsup_n \int_C \log p_F \, dH_n \le \int_C p_{F^*} \log p_F \, d\mu,$$

by the weak convergence of H_n. Therefore, $\limsup_n \int_C \log p_{\hat{F}} \, dH_n \le \int_C p_{F^*} \log p_F \, d\mu$, and this proves, using $p_{\hat{F}} \le h$, that

$$\limsup_n \int \log(p_{\hat{F}}/h) \, dH_n \le \int_C p_{F^*} \log(p_F/h) \, d\mu$$

for every compact C. Now consider a sequence of compact sets which increases to E. Taking limits along this sequence, Fatou's lemma gives

$$\limsup_{C \uparrow E} \int_C p_{F^*} \log(p_F/h) \, d\mu \le \int p_{F^*} \log(p_F/h) \, d\mu.$$

To summarize, we have shown

$$\limsup_n l_n(\hat{F})/n = \limsup_n \int \log p_{\hat{F}} \, dH_n \le \int p_{F^*} \log p_F \, d\mu.$$

But, by (4), $\liminf_n l_n(\hat{F})/n \ge \int p_{F^*} \log p_{F^*} \, d\mu$. Hence, $\int p_{F^*} \log p_F \, d\mu \ge \int p_{F^*} \log p_{F^*} \, d\mu$, that is, $K(F^*, F) \le 0$, which implies $F = F^*$, by Lemma 1.

Since $F = F^*$ holds for any convergent subsequence of $\{\hat{F}\}$, the result follows. □

The proof of consistency applies to any estimator that achieves a likelihood value at least as large as some constant times the maximum value of the likelihood [see also Pfanzagl (1988)], but this fact is not useful for the study of the estimator obtained by maximum-penalized-likelihood methods.

3.2. *Approximation of mixing distributions.* In order to discuss the convergence of the constrained maximum-likelihood estimator in general, we must produce a candidate limit distribution. If the true mixing distribution has three components but we are estimating two, what are we estimating? This question leads us to consider the error in approximation of F^* by a distribution with a specified number of components, that is,

$$K(F^*, \mathscr{F}_m) = \inf_{F \in \mathscr{F}_m} K(F^*, F),$$

where $\mathscr{F}_m$ is the set of subdistribution functions on Θ with m or fewer components. The following result shows the minimum divergence is attained by a distribution function and is strictly decreasing in m for $m \leq m^*$, where m^* is the number of components of F^*.

LEMMA 3. *Assume conditions* 1, 2 *and* 4 *hold. Then, for each* $m \geq 1$, *there is a distribution function* $F_m^* \in \mathscr{F}_m$ *for which* $K(F^*, F_m^*) = K(F^*, \mathscr{F}_m)$. (*In the following,* F_m^* *will denote any such distribution function.*) *If the identifiability property holds, then, for every* $m \geq 1$, $K(F^*, F_{m+1}^*) < K(F^*, F_m^*)$ *if* $F^* \notin \mathscr{F}_m$.

PROOF. (i) Under the vague topology (the topology of convergence of subdistribution functions at continuity points), $\mathscr{F}_m$ is compact (as can be seen from the proof of the Helly selection theorem) and the function $F \to K(F^*, F)$ is lower semicontinuous. But a lower semicontinuous function attains its infimum over a compact set.

(ii) Assume $K(F^*, F_{m+1}^*) = K(F^*, F_m^*)$. Then $K(F^*, F) \geq K(F^*, F_m^*)$ for every $F \in \mathscr{F}_{m+1}$; in particular,

$$\int p_{F^*}(y) \log\left(\frac{(1-\varepsilon)p_{F_m^*}(y) + \varepsilon p(y,\theta)}{p_{F_m^*}(y)}\right) d\mu(y) \leq 0, \qquad \theta \in \Theta, \epsilon > 0.$$

Therefore, using Fatou's lemma,

$$0 \geq \int p_{F^*}(y) \liminf_{\varepsilon \to 0} \frac{1}{\varepsilon} \log\left(\frac{(1-\varepsilon)p_{F_m^*}(y) + \varepsilon p(y,\theta)}{p_{F_m^*}(y)}\right) d\mu(y)$$

$$= \int p_{F^*}(y) \left(\frac{p(y,\theta)}{p_{F_m^*}(y)} - 1\right) d\mu(y).$$

Now, from $\int p_{F^*} p(y, \theta)/p_{F_m^*} \, d\mu \leq 1$ for every θ, we get $\int (p_{F^*})^2/p_{F_m^*} \, d\mu \leq 1$ and

$$K(F^*, F_m^*) = \int p_{F^*} \log(p_{F^*}/p_{F_m^*}) \, d\mu \leq \int p_{F^*}(p_{F^*}/p_{F_m^*} - 1) \, d\mu \leq 0,$$

which implies $F_m^* = F^*$, by Lemma 1. □

The next result shows that the modeling error $K(F^*, F_m^*)$ is negligible for large m.

LEMMA 4. *If conditions* 1, 2, 4 *and* 5 *hold, then* $K(F^*, F_m^*) \to 0$ *as* $m \to \infty$.

PROOF. Fix $y \in E$. For each $m \geq 1$, let there be a finite partition of Θ into intervals (or rectangles, cubes, etc.) $A_1^{(m)}, \ldots, A_{r_m}^{(m)}$. An upper Riemann–Stieltjes sum for the integral $\int_\Theta p(y, \theta) \, dF^*$ based on this partition is $\sum_i \alpha_i^{(m)} p(y, \theta_i^{(m)})$, where $\alpha_i^{(m)} = \int_{A_i^{(m)}} dF^*$ and $\theta_i^{(m)}$ belongs to the closure of $A_i^{(m)}$ and satisfies $p(y, \theta_i^{(m)}) = \sup_{\theta \in A_i^{(m)}} p(y, \theta)$. Therefore, if $F_m = \sum_i \alpha_i^{(m)} \delta_{\theta_i^{(m)}}$, then $p_{F_m} \to p_{F^*}$, and, using $p_{F_m} \geq p_{F^*}$ and conditions 4 and 5, the dominated convergence theorem implies $\int p_{F^*} \log p_{F_m} \, d\mu \to \int p_{F^*} \log p_{F^*} \, d\mu$ as $m \to \infty$. □

3.3. *Consistency of constrained maximum-likelihood estimators.* This section considers the large-sample behavior of the constrained estimator $\hat{F}_m$ defined in Section 1. In order for $\hat{F}_m$ to be consistent, it is necessary that F_m^* be uniquely defined; this means F_m^* uniquely satisfies the following two requirements: $F_m^* \in \mathcal{F}_m$ and $K(F^*, F_m^*) = K(F^*, \mathcal{F}_m)$. This clearly holds if $F^* \in \mathcal{F}_m$ and the identifiability property (3) holds for m-component mixtures, but general statements on this question seem very difficult to obtain. Fortunately this issue is not critical for the study of the maximum-penalized-likelihood estimator. However, the general asymptotic behavior of $\hat{F}_m$ is of interest in its own right and we present the following result.

THEOREM 2. *Assume conditions* 1–5 *hold. With probability* 1, *every limit* F *of* $\{\hat{F}_m\}$ (*i.e.*, $\hat{F}_m \to F$ *at all continuity points of* F) *satisfies* $K(F^*, F) = K(F^*, F_m^*)$. *If* F_m^* *is unique, then* $\hat{F}_m \to_w F_m^*$ *as* $n \to \infty$, *with probability* 1.

PROOF. As in the proof of Theorem 1 we get that, with probability 1, the limit F of any convergent subsequence of $\{\hat{F}_m\}$ must satisfy

$$\int p_{F^*} \log p_F \, d\mu \geq \int p_{F^*} \log p_{F_m^*} \, d\mu$$

or, equivalently, $K(F^*, F) \leq K(F^*, F_m^*)$. The results follow from this. □

The above consistency theorem enables us to prove the consistency of the maximum-likelihood estimators of the parameters of a finite mixing distribution, when the number of components is known. The following lemma, which is easily proved, states the equivalence of weak convergence of mixing distributions and convergence of the associated parameters in the quotient topology considered by Redner (1981). The quotient topology is defined relative to the equivalence relation under which two sets of parameters are equivalent if they define the same mixing distribution.

LEMMA 5. *Let* $F^* = \sum_{j=1}^m \alpha_j \delta_{\theta_j}$, *where* $\alpha_j > 0$ *for every* j, $\sum_{j=1}^m \alpha_j = 1$, *and the* θ_j *are distinct points of* Θ. *Let* $\{F_k\}$ *be an arbitrary sequence in* $\mathscr{F}_m$, *that is,* $F_k = \sum_{j=1}^m \alpha_j^{(k)} \delta_{\theta_j^{(k)}}$, *where* $\alpha_j^{(k)} \geq 0$, $\theta_j^{(k)} \in \Theta$, $j = 1, \ldots, m$. *Then* $F_k \to_w F^*$ *if and only if* $(\alpha_1^{(k)}, \ldots, \alpha_m^{(k)}, \theta_1^{(k)}, \ldots, \theta_m^{(k)}) \to (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m)$ *in the quotient topology.*

THEOREM 3. *Let* $F^* = \sum_{j=1}^m \alpha_j \delta_{\theta_j}$, *where* $\alpha_j > 0$ *for every* j, $\sum_{j=1}^m \alpha_j = 1$, *and the* θ_j *are distinct points of* Θ. *Assume that the identifiability property* (3) *holds for m-component mixtures and conditions* 1–5 *are satisfied. Let* $\hat{\alpha}_j, \hat{\theta}_j$ *be maximum-likelihood estimators of* α_j, θ_j, $j = 1, \ldots, m$. *Then, with probability* 1, $(\hat{\alpha}_1, \ldots, \hat{\alpha}_m, \hat{\theta}_1, \ldots, \hat{\theta}_m) \to (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m)$ *in the quotient topology, as* $n \to \infty$.

PROOF. In this case $F_m^* = F^*$ is unique (see the paragraph before Theorem 2). Therefore, Theorem 2 gives $\hat{F}_m \to_w F^*$, where $\hat{F}_m = \sum_{j=1}^m \hat{\alpha}_j \delta_{\hat{\theta}_j}$. The conclusion follows by the hypotheses on the parameters, using Lemma 5. □

3.4. *Consistency of maximum-penalized-likelihood estimators.* By combining the results of the previous sections, we can prove $\hat{F}_{\hat{m}_n}$ is consistent, where $\hat{m}_n$ is chosen to maximize a criterion of the form (2); we will need one more preliminary result.

LEMMA 6. *Assume conditions* 1–5 *hold. Then, with probability* 1:

(i) $\lim_{n\to\infty} l_n(\hat{F})/n = \int p_{F^*} \log p_{F^*}\, d\mu$.
(ii) $\lim_{n\to\infty} l_n(\hat{F}_m)/n = \int p_{F^*} \log p_{F_m^*}\, d\mu$, *for every* $m \geq 1$.

PROOF. (i) was proved in the proof of Theorem 1. (ii) is proved in the same way (see the proof of Theorem 2). □

THEOREM 4. *Assume conditions* 1–5 *and the identifiability property* (3) *hold. Let* F^* *have* m^* *components* ($m^* = \infty$ *if* F^* *is not a finite distribution*). *If, for every* $m < m^*$, $a_{m+1,n} \geq a_{mn}$ *for all* n *and* $\limsup_n a_{mn}/n = 0$, *with probability* 1, *then* $\liminf_{n\to\infty} \hat{m}_n \geq m^*$ ($\hat{m}_n \to \infty$ *if* $m^* = \infty$) *and* $\hat{F}_{\hat{m}_n} \to_w F^*$ *as* $n \to \infty$, *with probability* 1.

PROOF. (i) First assume F^* is a finite distribution with m^* components. By Lemma 6,

$$\lim_n \frac{l_n(\hat{F}_{m^*}) - l_n(\hat{F}_m)}{n} = \int p_{F^*} \log(p_{F^*}/p_{F_m^*})\, d\mu = K(F^*, F_m^*) \quad \text{for all } m < m^*, \tag{5}$$

with probability 1; the rest of this part of the proof is restricted to this event of probability 1. According to the condition on the rate of growth of a_{mn} and the fact that $K(F^*, F_m^*) > 0$, (5) implies

$$l_n(\hat{F}_{m^*}) - l_n(\hat{F}_m) > a_{m^*,n} - a_{mn} \quad \text{for all } m < m^* \text{ for large } n;$$

therefore, $\liminf_n \hat{m}_n \geq m^*$. By the definition of $\hat{m}_n$,

$$l_n(\hat{F}_{\hat{m}_n}) - a_{\hat{m}_n n} \geq l_n(F^*) - a_{m^* n}.$$

But $\hat{m}_n \geq m^*$, and hence $l_n(\hat{F}_{\hat{m}_n}) \geq l_n(F^*)$, for large n. The argument of the proof of Theorem 1 can now be used to obtain $\hat{F}_{\hat{m}_n} \to_w F^*$.

(ii) Now assume F^* is not a finite distribution. By Lemma 6,

$$\lim_n \frac{l_n(\hat{F}_{m+1}) - l_n(\hat{F}_m)}{n} = \int p_{F^*} \log(p_{F_{m+1}^*}/p_{F_m^*})\, d\mu = K(F^*, F_m^*) - K(F^*, F_{m+1}^*) \quad \text{for all } m \geq 1, \tag{6}$$

on an event of probability 1, which we restrict to for the remainder of the proof. But $K(F^*, F_{m+1}) < K(F^*, F_m^*)$ for every m by Lemma 3, and so (6) implies

$$l_n(\hat{F}_{m+1}) - l_n(\hat{F}_m) > a_{m+1,n} - a_{mn} \quad \text{for all } m \geq 1 \text{ for large } n.$$

This inequality implies that for each $m \geq 1$, for large enough n, $\hat{m}_n > m$ and $l_n(\hat{F}_{\hat{m}_n}) \geq l_n(F_m^*)$. Therefore, in particular, $\hat{m}_n \to \infty$. Let $\hat{F}_{\hat{m}_n}$ converge to F along a subsequence. As in the proof of Theorem 1, $l_n(\hat{F}_{\hat{m}_n}) \geq l_n(F_m^*)$ for large n implies

$$\int p_{F^*} \log p_F\, d\mu \geq \int p_{F^*} \log p_{F_m^*}\, d\mu.$$

But, by Lemma 4, $\int p_{F^*} \log p_{F_m^*}\, d\mu \to \int p_{F^*} \log p_{F^*}\, d\mu$ as $m \to \infty$. Therefore, $F = F^*$ and the proof is complete. □

The results of the above theorem state that the estimator $\hat{m}_n$, in the limit, does not underestimate the number of components of the true mixing distribution, and in case the true number of components is infinite, the estimator is consistent in the sense that it converges to ∞. This property holds also for the number of components in $\hat{F}$ (this result does not appear to have been reported previously although it is a consequence of the consistency of $\hat{F}$). The question

of whether $\hat{m}_n$ is consistent for a finite number of components (possibly under an additional condition on a_{mn}) is worthy of further research.

Acknowledgments. This paper is based on a part of my Ph.D. dissertation in the Department of Statistics at The University of British Columbia. I thank my supervisor, Marty Puterman, and Harry Joe, for helpful discussions and guidance during the completion of this work. The comments of an Associate Editor and three referees led to substantial improvements, in particular to a simplification in the proof of Lemma 4.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.) 267–281. Akademiai Kiado, Budapest.

CHEN, J. (1991). Optimal rate of convergence for finite mixture models. Preprint.

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

HECKMAN, J. and SINGER, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271–320.

HENNA, J. (1985). On estimating the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.* **37** 235–240.

JEWELL, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479–484.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887–906.

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

LINDSAY, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work* (C. Taillie, G. P. Patil and B. A. Baldessari, eds.) **5** 95–110. Reidel, Dordrecht.

LINDSAY, B. G. (1983). The geometry of mixing likelihoods: A general theory. *Ann. Statist.* **11** 86–94.

LINHART, H. and ZUCCHINI, W. (1986). *Model Selection*. Wiley, New York.

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models*. Dekker, New York.

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137–158.

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

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4** 1200–1209.

SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1** 49–58.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

DEPARTMENT OF BIOSTATISTICS, SC-32
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195