DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR-16035

ON THE CONSISTENCY OF
THE MAXIMUM LIKELIHOOD ESTIMATE
OF NORMAL MIXTURE PARAMETERS
FOR A SAMPLE WITH FIELD STRUCTURE

BY
CHARLES PETERS

REPORT 74 ~~SEPT~~ aug 1979

(NASA-CR-160352) ON THE CONSISTENCY OF THE MAXIMUM LIKELYHOOD ESTIMATE OF NORMAL MIXTURE PARAMETERS FOR A SAMPLE WITH FIELD STRUCTURE (Houston Univ.) 18 p HC A02/MF A01 CSCL 12A G3/64

N80-10836

Unclas 45993

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15543

HOUSTON, TEXAS 77004

On the Consistency of the Maximum Likelihood Estimate of Normal Mixture Parameters for a Sample with Field Structure.

by

Charles Peters
Department of Mathematics
University of Houston
Houston, Texas

Report #74
NAS-9-15543 Mod2S
August, 1979

Abstract

General theorems concerning the strong consistency of the MLE of exponential mixture parameters are proved. These theorems imply the strong consistency of the MLE of normal mixture parameters when the data is organized into "fields" each of which is a random sample from one of the component normal distributions

1. Introduction

In [5] a statistical model for LANDSAT agricultural data based on normal mixtures was introduced which admits a specific kind of dependence among the observations, namely their association into fields each representing a single agricultural class. Necessary conditions were derived for a maximum likelihood estimate of the parameters of the model and a numerical procedure for solution of the likelihood equations was suggested. The question of the consistency of the maximum likelihood estimate is complicated by the fact that it is no longer possible to reduce the sample to a set of independent identically distributed variables. The purpose of this note is to establish a general theorem on the existence of a consistent maximum likelihood estimate when the observations are not identically distributed and to show its applicability to the statistical model described in detail below.

We assume that each pixel is identified by a pair (j,k) of positive integers, where the first index j, $1 \le j \le p$, identifies the field containing the pixel and the second index k, $1 \le k \le N_j$, distinguishes it from other pixels in the same field. We suppose that the field structure is predetermined, perhaps as part of a spatial clustering algorithm such as AMOEBA. Let $X_{jk} \in R^n$ be the random vector of spectral measurements from pixel (j,k) and let $\Theta_{jk} \in \{1,\cdots,m\}$ be an unobserved random variable indicating its class index. We assume that the class indices $\Theta_{j1}, \Theta_{j2}, \cdots, \Theta_{jN_j}$ from the jth field are all the same and denote their common value by Θ_j. We further assume that, conditioned on $\Theta_j = \ell$, the measurements $X_{j1}, \cdots, X_{jN_j}$ are independently distributed as $N_n(\cdot, \mu_\ell^o, \Sigma_\ell^o)$, the n-variate normal with unknown mean μ_ℓ^o and unknown covariance Σ_ℓ^o. Let $X_j = (X_{j1}, \cdots, X_{jNj})$. Our final assumptions are that $(X_1, \Theta_1), \cdots, (X_p, \Theta_p)$ are independent and that $\{\Theta_j\}$ are identically distributed with unknown $\alpha_\ell^o = \text{Prob}[\Theta=\ell] > 0$. Under these assumptions, the joint density of all the observations is

$$(1) \qquad p(x_1, \cdots, x_p) = \prod_{j=1}^{p} \sum_{\ell=1}^{m} \alpha_\ell^o \prod_{k=1}^{N_j} N_n(x_{jk}; \mu_\ell^o, \Sigma_\ell^o)$$

where $x_j = (x_{j1}, \cdots, x_{jNj}) \in R^{nNj}$. This joint density is parametrized by $\{(\alpha_\ell, \mu_\ell, \Sigma_\ell) | \ell=1, \cdots, m\}$ where $\alpha_\ell > 0$; $\sum_{\ell=1}^{m} \alpha_\ell = 1$; $\mu_\ell \in R^n$; and Σ_ℓ is a real nxn positive definite symmetric matrix. For convenience, we let $\psi = \{\alpha_\ell, \mu_\ell, \Sigma_\ell) | \ell = 1, \cdots, m\}$ denote an arbitrary member of the parameter space and ψ^o the true value of the parameter. Thus the likelihood function corresponding to the sample $X_1, \cdots, X_p$ is

$$(2) \qquad L(\psi; X_1, \cdots, X_p) = \prod_{j=1}^{p} \sum_{\ell=1}^{m} \alpha_\ell \prod_{k=1}^{N_j} N_n(X_{jk}; \mu_\ell, \Sigma_\ell) .$$

For $x_j = (x_{j1}, \cdots, x_{jNj}) \in R^{nNj}$ let

$$m_j = m_j(x_j) = \frac{1}{N_j} \sum_{k=1}^{N_j} x_{jk}$$

and

$$S_j = S_j(x_j) = \sum_{k=1}^{N_j} (x_{jk} - m_j)(x_{jk} - m_j)^T$$

be the mean and scatter matrix respectively of the vectors $x_{j1}, \cdots, x_{jNj}$.

(3) $$\prod_{k=1}^{N_j} \mathcal{N}_n(x_{jk};\ \mu_\ell,\ \Sigma_\ell) = (2\Pi)^{-\frac{nNj}{2}} q_j(x_j;\ \mu_\ell,\ \Sigma_\ell)$$

where

(4) $$q_j(x_j;\ \mu_\ell,\ \Sigma_\ell) = |\Sigma_\ell|^{-\frac{Nj}{2}} \exp\{-\tfrac{1}{2}\,\mathrm{tr}\ \Sigma_\ell^{-1}[S_j + N_j(m_j - \mu_\ell)(m_j - \mu_\ell)^T]\}\ .$$

Let

(5) $$q_j(x_j|\psi) = \sum_{\ell=1}^{m} \alpha_\ell q_j(x_j;\ \mu_\ell,\ \Sigma_\ell)\ .$$

By ignoring terms which are independent of the parameters we derive the log likelihood function

(6) $$\ell(\psi) = \sum_{j=1}^{p} \log\ q_j(x_j|\psi)$$

which leads to the following necessary conditions for a local maximum of the likelihood function. Equations (7) - (9) are called the likelihood equations for the present model.

(7) $$\alpha_\ell = \frac{1}{p} \sum_{j=1}^{p} \frac{\alpha_\ell q_j(X_j;\ \mu_\ell,\ \Sigma_\ell)}{q;\ (X;\ |\ \psi)}$$

$$(8) \qquad \mu_\ell = \sum_{j=1}^{p} \frac{N_j q_j(X_j;\, \mu_\ell,\, \Sigma_\ell)}{q_j(X_j|\psi)}\, m_j \Bigg/ \sum_{j=1}^{p} \frac{N_j q_j(X_j;\, \mu_\ell,\, \Sigma_\ell)}{q_j(X_j|\psi)}$$

$$(9) \qquad \Sigma_\ell = \sum_{j=1}^{p} \frac{q_j(X_j;\, \mu_\ell,\, \Sigma_\ell)}{q_j(X_j|\psi)}\, S_j \Bigg/ \sum_{j=1}^{p} \frac{N_j q_j(X_j;\, \mu_\ell,\, \Sigma_\ell)}{q_j(X_j|\psi)} -$$

$$+ \sum_{j=1}^{p} \frac{N_j q_j(X_j;\, \mu_\ell,\, \Sigma_\ell)}{q_j(X_j|\psi)} (m_j-\mu_\ell)(m_j-\mu_\ell)^T \Bigg/ \sum_{j=1}^{p} \frac{N_j q_j(X_j;\, \mu_\ell,\, \Sigma_\ell)}{q_j(X_j|\psi)}$$

2. The General Theorem

Let Θ be an open subset of R^2 and let $\psi^o \in \Theta$. Suppose $X_1, X_2, \cdots$, is a sequence of independent random vectors with X_r having N_r-variate density function $q_r(\cdot|\psi^o)$ with respect to some fixed σ-finite measure λ_r on R^{N_r}. Suppose the densitites $q_r(\cdot|\psi)$ are defined for each $\psi \in \Theta$. Given a positive integer p, define a maximum likelihood estimate of ψ^o to be an element $\psi \in \Theta$ which locally maximizes $L_p(\psi) = \sum_{r=1}^{p} \log q_r(X_r|\psi)$. The equation $D_\psi L_p(\psi) = 0$ will be called the likelihood equation, where the symbol D_ψ denotes the Frechet derivative with respect to ψ.

A number of theorems dealing with the consistency of maximum likelihood estimates, under the additional assumption that the X_r's are identically distributed, have been presented in the literature (see for instance Chanda [2], Cramer [4], and Wald [8].) Extending any of these results to the case of nonidentically distributed observations is primarily a matter of finding a convenient set of conditions which insures that a law of large numbers can be invoked at several points in the proofs. The following theorem is such an outgrowth of the proof of strong consistency contained in [6].

Theorem 1: Suppose there is a neighborhood Ω of ψ^o and a λ_r - null sets N_r in R^{Nr} such that for all $\psi \in \Omega$; $x \notin N_r$, $i,j,k = 1,\cdots,2$, $r \in$ (the positive integers) $\frac{\partial q_r(x|\psi)}{\partial \psi_i}$; $\frac{\partial^2 q_r(x|\psi)}{\partial \psi_i \partial \psi_j}$; and $\frac{\partial^3 \log q_r(x|\psi)}{\partial \psi_i \partial \psi_j \partial \psi_k}$ exist and satisfy:

(i)
$$\left|\frac{\partial q_r(x|\psi)}{\partial \psi_i}\right| \le f_{ir}(x)$$

(ii)
$$\left|\frac{\partial^2 q_r(x|\psi)}{\partial \psi_i \partial \psi_j}\right| \le f_{ijr}(x)$$

(iii)
$$\left|\frac{\partial^3 \log q_r(x|\psi)}{\partial \psi_i \partial \psi_j \partial \psi_k}\right| \le f_{ijkr}(x)$$

where f_{ir} and f_{ijr} are λ_r-integrable on R^{N_r} and

(iv) $$E[f_{ijkr}(X_r)^2] = \int_{R^{Nr}} f_{ijkr}(x)^2 q_r(x|\psi_o)d\lambda_r(x) \le M$$

for all $r \in$, where M is a constant. Suppose also that

(v)
$$E\left\{\left[\frac{\partial \log q_r(X_r|\psi^o)}{\partial \psi_i}\right]^4\right\} \le M$$

and

(vi)
$$E\left\{\frac{1}{q_r(X_r|\psi^o)^2}\left(\frac{\partial^2 q_r(X_r|\psi^o)}{\partial \psi_i \partial \psi_j}\right)^2\right\} \le M$$

for all $i,j=1,\cdots,2$ and $r \epsilon$. Finally suppose that $\exists\, \epsilon > 0$ such that

$$\text{(vii)} \quad J_r(\psi^o) = E[\nabla_\psi \log q_r(X_r|\psi^o)\nabla_\psi \log q_r(X_r|\psi^o)^T] \geq \epsilon I_{\nu x \nu}$$

for all $r \epsilon$, where the ordering is the usual one on $\nu x \nu$ symmetric matrices. Then, it is almost surely true that, given a sufficiently small neighborhood of ψ^o; for large p there is a unique solution of the likelihood equation $D_\psi L_p(\psi) = 0$ in that neighborhood. Furthermore, that solution is a maximum likelihood estimate.

Remark: In the proof we make repeated use of the following simple version of the strong law of large numbers (see Chung [3]): Let $Z_1, Z_2, \cdots$ be uncorrelated random variables and suppose the sequence of variances $\{var(Z_i)\}_{i=1}^\infty$ is bounded. Then $\frac{1}{n}\sum_{i=1}^n (Z_i - E(Z_i)) \to 0$ a.s. as $n \to \infty$.

Proof of the theorem: Let $\mathcal{L}_p(\psi) = \frac{1}{p}\sum_{r=1}^p D_\psi \log q_r(X_r|\psi)$. By assumption (i) $E(\mathcal{L}_p(\psi^o)) = 0$ and by assumption (v) and the strong law, $\mathcal{L}_p(\psi^o) \to 0$ a.s. as $p \to \infty$. Now consider the $\nu x \nu$ matrix $D_\psi \mathcal{L}_p(\psi^o)$ whose i jth element is

$$\frac{1}{p}\sum_{r=1}^p \frac{\partial^2 \log q_r(X_r|\psi^o)}{\partial\psi_i\partial\psi_j} = \frac{1}{p}\sum_{r=1}^p \frac{1}{q_r(X_r|\psi^o)}\frac{\partial^2 q_r(X_r|\psi^o)}{\partial\psi_i\partial\psi_j}$$

$$- \frac{1}{p}\sum_{r=1}^p \frac{\partial \log q_r(X_r|\psi^o)}{\partial\psi_i}\frac{\partial \log q_r(X_r|\psi^o)}{\partial\psi_j}$$

By assumption (ii) the expected value of the first term on the right is zero.

Hence, by assumptions (v) and (vi) $D_\psi \mathcal{L}_p(\psi^o) + \frac{1}{p}\sum_{r=1}^{p} J_r(\psi^o) \to 0$ a.s. as $p \to \infty$. It follows that with probability 1, for each η in $0 < \eta < \frac{\epsilon}{2}$ there is a $p_o \in$ so that for $p \geq p_o$

$$D_\psi \mathcal{L}_p(\psi^o) \leq -2\eta I$$

Without loss of generality we can assume Ω is convex.

Thus, for $\psi \in \Omega$,

$$\frac{1}{p}\sum_{r=1}^{p}\left|\frac{\partial^2 \log q_r(X_r|\psi)}{\partial\psi_i\partial\psi_j} - \frac{\partial^2 \log q_r(X_r|\psi^o)}{\partial\psi_i\partial\psi_j}\right|$$

$$\leq \frac{1}{p}\sum_{r=1}^{p}\sum_{k=1}^{\nu}\left|\psi_k - \psi_k^o\right|\int_0^1\left|\frac{\partial^3 \log q_r(X_r|\psi^o + t(\psi - \psi^o))}{\partial\psi_i\partial\psi_j\partial\psi_k}\right| dt$$

$$\leq \frac{1}{p}\sum_{r=1}^{p}\sum_{k=1}^{\nu}\left|\psi_k - \psi_k^o\right| f_{ijkr}(X_r)$$

With probability 1, for large p

$$\frac{1}{p}\sum_{r=1}^{p} f_{ijkr}(X_r) < 1 + \frac{1}{p}\sum_{r=1}^{p} E[f_{ijkr}(X_r)] < 1 + M^{\frac{1}{2}} .$$

by assumption (iv).

It follows that for any particular norms on R^ν and on the $\nu x \nu$ symmetric matrices there is a constant $\overline{M}$ such that with probability 1 there is a $p_1 \in$ such that for all $p \geq p_1$, and $\psi \in \Omega$,

$$||D_\psi \mathcal{L}_p(\psi) - D_\psi \mathcal{L}_p(\psi^o)|| \le \bar{M}||\psi - \psi^o||$$

Thus, there is a convex neighborhood Ω^o of ψ^o such that

$$D_\psi \mathcal{L}_p(\psi) \le -\eta I$$

for all $\psi \in \Omega^o$, $p \ge p_1$. It now follows as in [6] that for $p \ge p$, $\mathcal{L}_p$ is one to one on Ω^o and that the image under $\mathcal{L}_p$ of the sphere $\Omega_\delta(\psi^o)$ at ψ^o of small radius δ contains the sphere $\Omega_{\eta\delta}(\mathcal{L}_p(\psi^o))$ at $\mathcal{L}_p(\psi^o)$ of radius $\eta\delta$. Since 0 is eventually in $\Omega_{\eta\delta}(\mathcal{L}_p(\psi^o))$, there is a unique solution of $\mathcal{L}_p(\psi) = 0$ in $\Omega_\delta(\psi^o)$. Since $D_\psi \mathcal{L}_p(\psi)$ is negative definite, this solution is a maximum likelihood estimate. This concludes the proof.

Theorem 1 shows that by restricting attention to a fixed neighborhood of ψ^o it is possible to speak unambiguously of the unique consistent solution of the likelihood equations or, equivalently, of the unique consistent MLE of ψ^o This terminology will be adopted in the next theorem.

3. Application to Exponential Mixtures

In this section we apply Theorem 1 to a class of mixture models which contains the normal mixture model of Section 1. Referring to the notation of that section, we assume that conditioned on $\Theta_j = \ell$, the random n-vectors $x_{j1},\ldots,x_{jN_j}$ are independent with a common density of exponential type

(1) $$f(x|\tau_\ell) = C(\tau_\ell) \exp \langle \tau_\ell | F(x) \rangle$$

with respect to a dominating σ-finite measure λ where the parameter τ_ℓ is an arbitrary member of an open subset U of a finite dimensional vector space V with inner product $\langle\cdot|\cdot\rangle$. We assume also that C is one to one and that the support of the measure induced on U by F and λ contains an open set. These conditions imply that the parameter τ_ℓ is identifiable [1], and any parametrization of the form (1) satisfying them will be called a <u>canonical representation</u> of the given family of distributions.

The joint density, given $\Theta_j = \ell$, of $X_j = (X_{j1},\ldots,X_{jN_j})$ is also of exponential type; i.e., for $x_j = (x_{j1},\ldots,x_{jN_j})$

$$p_j(x_j|\tau_\ell) = \gamma_j(\tau_\ell) \exp \langle\tau_\ell|G_j(x_j)\rangle \tag{2}$$

where

$$\gamma_j(\tau_\ell) = C(\tau_\ell)^{Nj}$$

$$G_j(x_j) = \sum_{k=1}^{Nj} F(x_{jk})$$

and the representation (2) is also canonical.

Some useful facts about exponential families are collected in the following lemma. For proofs see Barndorff-Nielsen [1].

<u>Lemma 1</u>: Let (1) be a canonical representation of an exponential family. For each $\tau \in U$ let $\kappa(\tau) = -\ln C(\tau) = \ln \int_{\mathbb{R}^n} \exp \langle\tau|F(x)\rangle\, d\lambda(x)$. Then

(i) for each $\tau \in U$, $F(x)$ has moments of all orders with respect to $f(x|\tau)$;

(ii) $\kappa(\tau)$ has derivatives of all orders with respect to τ, which may be obtained by differentiating under the integral sign. Indeed $D_\tau^k\, \kappa(\tau)$ can be represented as a symmetric k-linear form on V which is a polynomial in the first k moments of F. In particular,

(iii) $D_\tau \kappa(\tau) = \langle E_\tau(F)|\cdot\rangle = \int_n \langle F(x)|\cdot\rangle\, f(x|\tau)\, d\lambda(x)$

and

(iv) $D_\tau^2 \kappa(\tau) = \mathrm{cov}_\tau(F) = \int_n \langle F - E_\tau(F)|\cdot\rangle^2 f(x|\tau)\, d\lambda(x)$, which is positive definite.

(v) $\kappa(\tau)$ is strictly convex on U.

(Expressions $\langle\sigma|\cdot\rangle^k$ like that in (iv) are meant to denote k-linear forms; e.g. $\langle\sigma|\cdot\rangle^2$ denotes the bilinear form $b(\eta,\nu) = \langle\sigma|\eta\rangle\langle\sigma|\nu\rangle$.)

We are now ready to apply Theorem 1 to the mixture model

(3) $$q(x|\psi) = \prod_{j=1}^{p} q_j(x_j|\psi)$$

where $$\psi = (\alpha_1,\ldots,\alpha_{m-1},\tau_1,\ldots,\tau_m)$$

$$x = (x_1,\ldots,x_p)$$

(4) $$q_j(x_j|\psi) = \sum_{\ell=1}^{m} \alpha_\ell p_j(x_j|\tau_\ell)$$

$$= p_j(x_j|\tau_m) + \sum_{\ell=1}^{m-1} \alpha_\ell [p_j(x_j|\tau_\ell) - p_j(x_j|\tau_m)]$$

and $p_j(x_j|\tau_\ell)$ has the canonical exponential representation given in (2).

<u>Theorem 2</u>: If the numbers $\{N_j\}$ in the mixture model (3) are bounded, then with probability 1 there is a unique consistent MLE of the parameter ψ^0.

<u>Proof</u>: Using Lemma 1 and writing $\mu_j(\tau_\ell) = E_{\tau_\ell}(G_j)$ the nonzero derivatives of $q_j(x_j|\psi)$ up to order 2 are:

(5) $D_{\alpha_\ell} q_j(x_j|\psi) = p_j(x_j|\tau_\ell) - p_j(x_j|\tau_m)$, $\ell = 1,\ldots,m-1$

(6) $D_{\tau_\ell} q_j(x_j|\psi) = \alpha_\ell p_j(x_j|\tau_\ell) \langle G_j(x_j) - \mu_j(\tau_\ell)|\cdot\rangle$, $\ell = 1,\dots,m$

(7) $D_{\tau_\ell} D_{\alpha_\ell} q_j(x_j|\psi) = p_j(x_j|\tau_\ell) \langle G_j - \mu_j(\tau_\ell)|\cdot\rangle$, $\ell = 1,\dots,m-1$

(8) $D_{\tau_m} D_{\alpha_\ell} q_j(x_j|\psi) = -p_j(x_j|\tau_m) \langle G_j - \mu_j(\tau_m)|\cdot\rangle$, $\ell = 1,\dots,m-1$

(9) $D^2_{\tau_\ell} q_j(x_j|\psi) = \alpha_\ell p_j(x_j|\tau_\ell) \{\langle G_j - \mu_j(\tau_\ell)|\cdot\rangle^2 - \mathrm{cov}_{\tau_\ell}(G_j)\}$,

$\ell = 1,\dots,m$.

Instead of verifying conditions (i) and (ii) of Theorem 1, it is easier to recall that they were needed only in order to conclude that the integrals of the first and second order derivatives of $q_j(x_j|\psi)$ are zero at $\psi = \psi^0$. This is obvious form (5) - (9). Similarly, using Lemma 1 and the boundedness of $\{N_j\}$ the verification of conditions (iii) - (vi) presents no problem more serious than tedium. It remains to verify condition (vii). We may write $J_r(\psi)$ in matrix form as

$$J_r(\psi) = \left[\begin{array}{c|c} I_1 & 0 \\ \hline 0 & N_r^{\frac{1}{2}} I_2 \end{array}\right] E_\psi \left[\begin{array}{c|c} A_r & B_r \\ \hline B_r^* & C_r \end{array}\right] \left[\begin{array}{c|c} I_1 & 0 \\ \hline 0 & N_r^{\frac{1}{2}} I_2 \end{array}\right]$$

where I_1 and I_2 are, respectively, the identity operators on $\mathbb{R}^{m-1}$ and V^m and

$$A_r = \left(\frac{[p_r(X_r|\tau_\ell) - p_r(X_r|\tau_m)][p_r(X_r|\tau_k) - p_r(X_r|\tau_m)]}{q_r(X_r|\psi)^2} \right)_{\ell,k=1,\dots,m-1}$$

$$B_r = \left(\frac{\alpha_k p_r(X_r|\tau_k)[p_r(X_r|\tau_\ell) - p_r(X_r|\tau_m)]}{q_r(X_r|\psi)^2} N_r^{-\frac{1}{2}} \langle G_r - \mu_r(\tau_k)|\cdot\rangle \right)_{\substack{\ell=1,\dots,m-1 \\ k=1,\dots,m}}$$

$$C_r = \left(\frac{\alpha_\ell \alpha_k p_r(X_r|\tau_\ell) p_r(X_r|\tau_k)}{q_r(X_r|\psi)^2} \; N_r^{-1} \, (G_r - \mu_r(\tau_k)) \langle G_r - \mu_r(\tau_\ell) | \cdot \rangle \right)_{k,\ell=1,\ldots,m.}$$

We remark that if $\tau_1,\ldots,\tau_m$ are distinct then as functions of $F \in U$, $e^{\langle\tau_1|F\rangle},\ldots,e^{\langle\tau_m|F\rangle}, e^{\langle\tau_1|F\rangle}F,\ldots,e^{\langle\tau_m|F\rangle}F$ are linearly independent; i.e., if $\lambda_1,\ldots,\lambda_m$ are scalars, $\Lambda_1,\ldots,\Lambda_m \in V$ and $\lambda_1 e^{\langle\tau_1|F\rangle} + \ldots + \lambda_m e^{\langle\tau_m|F\rangle} + e^{\langle\tau_1|F\rangle}\langle F|\Lambda_1\rangle + \ldots + e^{\langle\tau_m|F\rangle}\langle F|\Lambda_m\rangle = 0$ for all $F \in U$, then $\lambda_1 = \ldots = \lambda_m = 0$ and $\Lambda_1 = \ldots = \Lambda_m = 0$. It is easily seen that if $J_r(\psi)$ fails to be positive definite then there is a nontrivial linear combination of these functions which is zero almost surely with respect to the distribution of F. It follows that $J_r(\psi)$ is positive definite for each r. Condition (vii) will be established once it is shown that the smallest eigenvalue of $J_r(\psi)$ is bounded away from zero as $N_r \to \infty$.

Let $\sigma(A)$ denote the smallest eigenvalue of a positive definite operator A. Clearly,

$$\sigma(J_r(\psi)) \geq \sigma\left(E_\psi \left[\begin{array}{c|c} A_r & B_r \\ \hline B_r^* & C_r \end{array} \right] \right)$$

Observe that

$$\frac{p_r(X_r|\tau_\ell)}{p_r(X_r|\tau_k)} = \exp\{-N_r[\kappa(\tau_\ell) - \kappa(\tau_k) - \langle\tau_\ell - \tau_k|\tfrac{1}{N_r}G_r\rangle]\}$$

If X_r is a sample from $f(x|\tau_k)$, then the expression in square brackets converges to

$$\kappa(\tau_\ell) - \kappa(\tau_k) - \langle\tau_\ell - \tau_k | E_{\tau_k}(F)\rangle = \kappa(\tau_\ell) - \kappa(\tau_k) - \kappa'(\tau_k) \cdot (\tau_\ell - \tau_k)$$

which is > 0 by the strict convexity of κ. Hence,

$$\frac{p_r(X_r|\tau_\ell)}{p_r(X_r|\tau_k)} \to 0 \text{ as } N_r \to \infty .$$

Therefore,

$$E_\psi\left[\frac{p_r(X_r|\tau_\ell)p_r(X_r|\tau_k)}{q_r(X_r|\psi)^2}\right] = E_{\tau_k}\left[\frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\psi)}\right] .$$

converges to 0 if $\ell \neq k$ and $\frac{1}{\alpha_k}$ if $\ell = k$ as $N_r \to \infty$. Thus,

$$E_\psi[A_r] \to \left(\frac{1}{\alpha_m^2} + \frac{\delta\ell k}{\alpha_k^2}\right) \text{ as } N_r \to \infty .$$

Given that X_r is from $f(x|\tau_k)$, $N_r^{-\frac{1}{2}}(G_r - \mu_r(\tau_k))$ converges in distribution to a normal random variable Z with mean zero and covariance $\mathrm{cov}_{\tau_k}(F)$. Hence,

$$\frac{p_r(X_r|\tau_\ell)}{q_{\ell_r}(X_r|\psi)} N_r^{-\frac{1}{2}}(G_r - \mu_r(\tau_k))$$

converges in distribution to 0 if $\ell \neq k$ and $\frac{1}{\alpha_k}Z$ if $\ell = k$.

Let Λ be any element of V and consider

$$[N_r^{-\frac{1}{2}}\langle G_r - \mu_r(\tau_k)|\Lambda\rangle]^4 = N_r^{-2}\left[\sum_{j=1}^{N_r}\langle F(X_{rj}) - E_{\tau_k}(F)|\Lambda\rangle\right]^4$$

After expanding and taking expectation with respect to τ_k, it will be seen that the only nonvanishing terms are those of the form

$$E_{\tau_k}[\langle F(X_{rj}) - E_{\tau_k}(F)|\Lambda\rangle^2\langle F(X_{r\ell}) - E_{\tau_k}(F)|\Lambda\rangle^2]$$

of which there are $N_r + \binom{N_r}{2} = O(N_r^2)$. Thus

$$E_{\tau_k}[N_r^{-\frac{1}{2}} \langle G_r - \mu_r(\tau_k)|\Lambda\rangle]^4$$

is bounded as $N_r \to \infty$. It follows from a standard theorem on convergence of moments [3 ,p. 95] that

$$E_{\tau_k}\left[\frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\Psi)} N_r^{-\frac{1}{2}} (G_r - \mu_r(\tau_k))\right] \to 0 \quad \text{as} \quad N_r \to \infty.$$

Thus $E_\psi(B_r) \to 0$. Similar reasoning shows that

$$E_\psi(C_r) \to (\delta_{k\ell} \operatorname{cov}_{\tau_k}(F))$$

as $N_r \to \infty$. Therefore $\sigma(J_r(\psi))$ is bounded away from 0 and this concludes the proof.

4. Concluding Remarks.

Clearly the assumption in Theorem 2 that $\{N_r\}$ is bounded can be weakened. In fact, Theorem 1 could be modified in such a way as to show that the MLE of exponential mixture parameters is strongly consistent when $\Sigma N_r^2/r^2 < \infty$.

Redner [7] has shown that when each $N_r = 1$, a certain numerical procedure for obtaining the MLE of exponential mixture parameters is convergent. The generalization to bounded $\{N_r\}$ should not be difficult, and will be addressed in a future report.

References

1. O. Barndorff - Nielsen, Information and Exponential Families in Statistical Theory, Wiley and Sons, New York (1978).

2. K. C. Chanda, A note on the consistency and maxima of the roots of the likelihood equations, Biometrika, 41 (1954), pp. 56-61.

3. K. L. Chung, A Course in Probability Theory, Second Edition, Academic Press, New York (1974).

4. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.

5. C. Peters, A modification of the likelihood equations for normal classes with field structure (preliminary report). Report #73, Department of Mathematics, University of Houston, June 1979.

6. ________ and H. F. Walker, An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. SIAM J. Appl. Math. B., 35, 2, Sept. 1978.

7. R. A. Redner, An iterative procedure for obtaining maximum likelihood estimates in a mixture model (to appear).

8. A. Wald, Note on the consistency of the maximum likelihood estimate, Ann. Math. Stat. 20 (1949), p. 595.