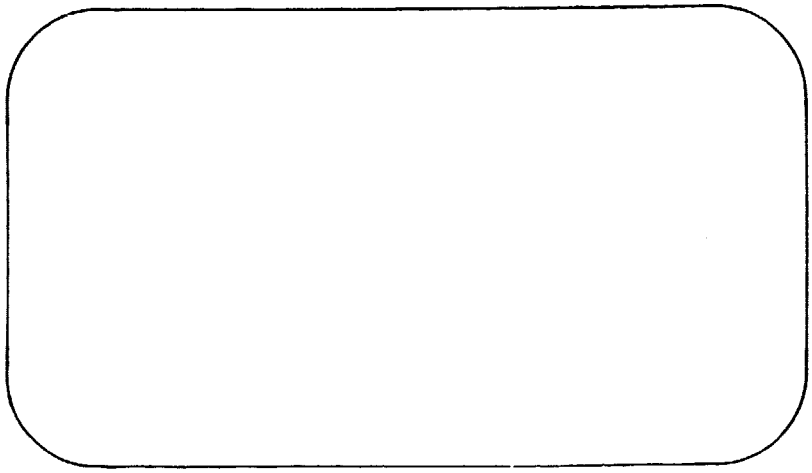
N O T I C E

THIS DOCUMENT HAS BEEN REPRODUCED FROM MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED IN THE INTEREST OF MAKING AVAILABLE AS MUCH INFORMATION AS POSSIBLE

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS



(E81-10099) ON THE EXISTENCE, UNIQUENESS, AND ASYMPTOTIC NORMALITY OF A CONSISTENT SOLUTION OF THE LIKELIHOOD EQUATIONS FOR NONIDENTICALLY DISTRIBUTED OBSERVATIONS: APPLICATIONS TO MISSING DATA PROBLEMS

N81-21416

G3/43 Unclas 00099

HOUSTON, TEXAS 77004

SR-HO-00492
NAS9-14689

AgRISTARS

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

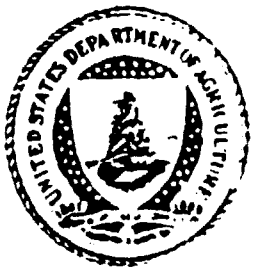
Supporting Research

September 10, 1980

REPORT

ON THE EXISTENCE, UNIQUENESS, AND ASYMPTOTIC NORMALITY OF A CONSISTENT SOLUTION OF THE LIKELIHOOD EQUATIONS FOR NONIDENTICALLY DISTRIBUTED OBSERVATIONS - APPLICATIONS TO MISSING DATA PROBLEMS.

BY: CHARLES PETERS



NASA

SR-HO-00492
NAS9-14689

TECHNICAL REPORT

ON THE EXISTENCE, UNIQUENESS, AND ASYMPOTIC NORMALITY OF A CONSISTENT SOLUTION OF THE LIKELIHOOD EQUATIONS FOR NONIDENTICALLY DISTRIBUTED OBSERVATIONS - APPLICATIONS TO MISSING DATA PROBLEMS

BY

CHARLES PETERS

UNIVERSITY OF HOUSTON
DEPARTMENT OF MATHEMATICS
HOUSTON, TX 77004

SEPTEMBER 10, 1980

1. Introduction

This paper is concerned with the existence, uniqueness, and asymptotic properties of a strongly consistent local maximizer of the likelihood function for a vector parameter in the case of nonidentically distributed samples and without prior assumptions which insure the existence of a global MLE. Well known results pertaining to scalar parameters and i.i.d. samples date back to theorems of Cramér [5] and Huzurbazar [11], while results concerning the consistency of the MLE, under assumptions that insure a unique MLE, may be found in Wald [17], Wolfowitz [19], and LeCam [12]. Somewhat more recently, Silvey [15] has dealt with the asymptotic properties of the MLE without independence. Surprisingly however, a correct proof of the multidimensional version of the combined results of Cramér and Huzurbazar on the existence of a unique consistent solution of the likelihood equations when multiple roots occur did not appear until 1977 in a note by Foutz [10], (see also Tarone & Gruenhage [16], Chanda [3], and Peters and Walker [14,Appendix].) Examples 1 and 2 which follow illustrate the need for a consistency theorem along these lines which relaxes the assumption of identically distributed observations.

Example 1 (Observations with missing components): Let X_1, X_2, ... be independent random vectors in R^n whose common density is one of a parametric family $\{q(x|\theta)\}_{\theta\in\Theta}$, where Θ is a subset of R^ν. Suppose that instead of the X_i we observe only certain subvectors B_1X_1, B_2X_2, ..., where $\{B_i\}$ is a given sequence of $n_i \times n$ matrices obtained by deleting $n - n_i$ rows from the identity. Clearly we can assume that components are missing at random provided that the B_i's are independent of the X_i's. Under what conditions is there a unique

strongly consistent (and asymptotically efficient) local MLE of θ based on the observations B_1X_1, B_2X_2, ...?

A recent paper by Dahiya and Korwar [6] illustrates that even for a bivariate normal sample, with several simplifying restrictions on the sample and on the parameters, the likelihood equation for Example 1 has multiple roots and requires numerical methods for its solution.

Example 2 (Estimating mixture density parameters with sample blocks of varying sizes): Let $f(x|\tau_1)$, $f(x|\tau_2)$, ..., $f(x|\tau_m)$ be unknown but distinct members of a multivariate parametric family $\{f(x|\tau)\}_{\tau\in T}$, and let α_1, ..., α_m be the unknown positive probabilities corresponding to a discrete mixing distribution supported on $\{\tau_1, \ldots, \tau_m\}$. The number m is known. Under what conditions will there be a unique consistent MLE of the parameter $\theta = (\alpha_1, \ldots, \alpha_{m-1}, \tau_1, \ldots, \tau_m)$ describing the mixture density $q(x|\theta) = \sum_{i=1}^{m} \alpha_i f(x|\tau_i)$, based on a sample of the type X_1, X_2, ..., where the X_i are independent and each X_i is itself a random sample $X_i = (X_{i1}, \ldots, X_{iNi})$ of known size from an unknown component density $f(x|\tau_i)$? In this example the parameter θ is only locally idemtifiable. Moreover, it can easily occur that the likelihood function is unbounded [9]; hence, the need for a consistency theorem for local maximizers is especially clear.

The practical importance of Example 2 is indicated by the fact that estimation of mixture density parameters is often proposed as an alternative to the clustering of large amounts of multivariate data [18]. The asymptotic properties of the MLE are of interest because of the prevalence of large sample considerations in judging cluster validity [8], even though it may be difficult to argue for a statistical basis for a given clustering problem. The presentation of the data in blocks of varying size may occur when the primary sampling units are grouped by physical or spatial associations (see [2] and [13] for an

application of this idea in the analysis of pictorial data.)

Finally we remark that the existence and uniqueness of a consistent solution of the likelihood equations bears on the numerical problem of obtaining the estimate. Each of Examples 1 and 2 is a missing data problem (in Example 2 the random variables which indicate the component population of origin are missing); thus, a natural numerical procedure for obtaining a MLE is one derived from the generalized EM procedure of Dempster, Laird, and Rubin [7]. Such a procedure increases the value of the likelihood at each iterative step; however, this is no guarantee of convergence, since the likelihood function may be unbounded. Generally speaking it is possible to show that the Hessian of the log likelihood is negative definite near the consistent solution of the likelihood equations. Thus, the generalized EM procedure is convergent to it given a good enough starting value (see [14] for a thorough discussion of numerical properties in the case of a mixture of multivariate normal distributions.)

Throughout this paper the symbol E_θ will denote expectation with respect to a distribution determined by a parameter θ and D_u, $D^2_{u,v}$ etc. will denote differentiation or partial differentiation with respect to scalar or vector variables u, v. For a scalar valued function, ∇_u will denote the gradient with respect to an inner product which will usually be understood from the context. Given an inner product $\langle\cdot|\cdot\rangle$ and a vector σ, the symmetric k-linear form $f(\eta_1, \ldots, \eta_k) = \prod_{j=1}^{k} \langle\sigma|\eta_i\rangle$ will be denoted by $\langle\sigma|\cdot\rangle^k$. Thus, for example, we may write the covariance of a statistic S as $\mathrm{Cov}_\tau(S) = E_\tau\{\langle S - E_\tau(S)|\cdot\rangle^2\}$. The largest and smallest eigenvalues of a symmetric positive definite operator A will be denoted respectively by $\rho(A)$ and $\sigma(A)$.

2. A General Consistency Theorem. Let Θ be an open subset of R^{ν} and for each positive integer r and each $\theta \in \Theta$, let $q_r(\cdot|\theta)$ be an N_r-variate density with respect to some fixed σ-finite measure λ_r on R^{N_r}. Let $\theta^o \in \Theta$ and let $X_1, \ldots, X_p, \ldots$ be a sequence of independent random vectors with X_r having density $q_r(\cdot|\theta^o)$. For $\theta \in \Theta$ define

$$L_p(\theta) = \sum_{r=1}^{p} \log q_r(X_r|\theta)$$

Theorem 1: Suppose

(i) $$\int_{R^{N_r}} D_\theta\, q_r(x|\theta^o)\, d\lambda_r(x) = 0 ,$$

(ii) $$\int_{R^{N_r}} D_\theta^2\, q_r(x|\theta^o)\, d\lambda_r(x) = 0 ,$$

and that there is a constant M, functions f_r, a neighborhood Ω of θ^o and λ_r-null sets A_r in R^{N_r} such that for all r, $\theta \in \Omega, x \notin A_r$,

(iii) $$|D^3_{\theta_i, \theta_j, \theta_k} \log q_r(x|\theta)| \le f_r(x) \qquad i, j, k = 1, \ldots, \nu$$

(iv) $$E_{\theta^o}\{f_r(X_r)^2\} \le M$$

(v) $$E_{\theta^o}\{[D_{\theta_i} \log q_r(X_r|\theta^o)]^4\} \le M \qquad i = 1, \ldots, \nu$$

(vi) $$E_{\theta^o}\left\{\frac{1}{q_r(X_r|\theta^o)^2}\,[D^2_{\theta_i,\theta_j} q_r(X_r|\theta^o)]^2\right\} \le M \qquad i, j = 1, \ldots, \nu$$

and

(vii) there exists $\epsilon > 0$ such that $\frac{1}{p}\sum_{r=1}^{p} J_r(\theta^o) \ge \epsilon I_\nu$ for sufficiently large p,

where $J_r(\theta^o) = E_{\theta^o}\{\nabla_\theta \log q_r(X_r|\theta^o)\, \nabla_\theta^T \log q_r(X_r|\theta^o)\}$, I_ν is the identity on R^{ν}, and the ordering is the usual one on symmetric operators. Then there is a neighborhood Ω^o of θ^o such that with probability 1 there is an integer p_1 such that for $p \ge p_1$ there is a unique solution θ^p in Ω^o of the likelihood equation

$D_\theta L_p(\theta) = 0$. Furthermore, $\theta^p \to \theta^o$ as $p \to \infty$ and θ^p is a maximum likelihood estimate. The consistent estimator θ^p is asymptotically normal and asymptotically efficient.

Proof: In the proof we make repeated use of the following version of the strong law [4, p. 103]: let $Z_1, Z_2, \ldots$ be uncorrelated random variables such that the variances of the Z_i are bounded. Then $\frac{1}{n}\sum_{j=1}^{n}(Z_u - E[Z_j]) \to 0$ a.s. as $n \to \infty$.

Let $S_p(\theta) = \frac{1}{p}\sum_{r=1}^{p} D_\theta \log q_r(X_r|\theta)$. By (i) $E_{\theta^o}\{S_p(\theta^o)\} = 0$ and by (v) $S_p(\theta^o) \to 0$ a.s. as $p \to \infty$. Consider the $\nu \times \nu$ matrix $D_\theta S_p(\theta^o)$ whose i, j^{th} element is

$$\frac{1}{p}\sum_{r=1}^{p} D^2_{\theta_i,\theta_j} \log q_r(X_r|\theta^o) = \frac{1}{p}\sum_{r=1}^{p} \frac{1}{q_r(X_r|\theta^o)} D^2_{\theta_i,\theta_j} q_r(X_r|\theta^o)$$

$$- \frac{1}{p}\sum_{r=1}^{p} D_{\theta_i} \log q_r(X_r|\theta^o) D_{\theta_j} \log q_r(X_r|\theta^o).$$

By (ii) the expected value of the first term on the right is zero. Hence, by (v) and (vi)

$$D_\theta S_p(\theta^o) + \frac{1}{p}\sum_{r=1}^{p} J_r(\theta^o) \to 0$$

a.s. as $p \to \infty$. Thus, with probability 1, if $0 < \eta < \epsilon/2$ there is $p_o \in N$ so that for $p \geq p_o$

$$D_\theta S_p(\theta^o) \leq -2\eta I .$$

Without loss of generality we can assume Ω is convex. For $\theta \in \Omega$,

$$\frac{1}{p}\sum_{r=1}^{p} |D^2_{\theta_i,\theta_j} \log q_r(X_r|\theta) - D^2_{\theta_i,\theta_j} \log q_r(X_r|\theta^o)|$$

$$\leq \frac{1}{p}\sum_{r=1}^{p}\sum_{k=1}^{\nu} |\theta_k - \theta_k^o| \int_0^1 |D^3_{\theta_i,\theta_j,\theta_k} \log q_r(X_r|\theta^o + t(\theta - \theta^o))|\,dt$$

$$\leq \frac{1}{p}\sum_{r=1}^{p}\sum_{k=1}^{\nu} |\theta_k - \theta_k^o| f_r(X_r)$$

With probability 1, for large p

$$\frac{1}{p}\sum_{r=1}^{p} f_r(X_r) \le 1 + \frac{1}{p}\sum_{r=1}^{p} E_{\theta^o}(f_r(X_r))$$

$$\le 1 + M^{\frac{1}{2}} .$$

It follows that for any particular norms on R^ν and on the symmetric $\nu\times\nu$ matrices there is a constant $\bar{M}$ such with probability 1 there is a positive integer p_1 such that for $p \ge p_1$, $\theta \in \Omega$,

$$||D_\theta S_p(\theta) - D_\theta S_p(\theta^o)|| \le \bar{M}||\theta - \theta^o|| .$$

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

$$D_\theta S_p(\theta) \le -\eta I$$

for all $\theta \in \Omega^o$, $p \ge p_1$. It now follows that for $p \ge p_1$ S_p is one to one on Ω^o and that the image under S_p of the sphere $\Omega_\delta(\theta^o)$ at θ^o of small radius δ contains the sphere $\Omega_{\eta\delta}(S_p(\theta^o))$ at $S_p(\theta^o)$ of radius $\eta\delta$. Since 0 is eventually in $\Omega_{\eta\delta}(S_p(\theta^o))$ there is a unique solution of $D_\theta S_p(\theta) = 0$ in $\Omega_\delta(\theta^o)$. Since $D_\theta S_p(\theta)$ is negative definite, this solution is a MLE.

Let $\Sigma_p = \frac{1}{p}\sum_{r=1}^{p} J_r(\theta^o)$. The Cramér-Rao lower bound for p observations is verified without difficulty to be $(p\,\Sigma_p)^{-1}$. By (v), (vii), and Liapounovs Theorem [4, p. 200], $p^{\frac{1}{2}}\,\Sigma_p^{-\frac{1}{2}}\,S_p(\theta^o)$ is asymptotically distributed as $N_\nu(0, I)$. Moreover, in a neighborhood of θ^o we may write

$$S_p(\theta) = S_p(\theta^o) + A(\theta)(\theta - \theta^o)$$

where $A(\theta) \to D_\theta S_p(\theta^o)$ as $\theta \to \theta^o$. It follows that with probability 1.

$$p^{\frac{1}{2}}\,\Sigma_p^{\frac{1}{2}}\,(\theta^p - \theta^o) = -\,\Sigma_p^{\frac{1}{2}}\,A(\theta^p)^{-1}\,\Sigma_p^{\frac{1}{2}}\,p^{\frac{1}{2}}\,\Sigma_p^{-\frac{1}{2}}\,S_p(\theta^o)$$

for large p. Since $D_\theta S_p(\theta^o) + \Sigma_p \to 0$ and $A(\theta^p) \to D_\theta S_p(\theta^o)$ with probability 1,

the expression $-\Sigma_p^{\frac{1}{2}} A(\theta^p)^{-1}\Sigma_p^{\frac{1}{2}}$ converges almost surely to the identity. Therefore, $p^{\frac{1}{2}}\Sigma_p^{\frac{1}{2}}(\theta^p - \theta^o)$ is asymptotically $N_\nu(0,I)$ and θ^p is asymptotically efficient. This concludes the proof.

3. Applications.

Suppose that in Example 1 the X_i have a common n variate normal distribution $N_n(\mu, \Sigma)$ and it is desired to estimate μ, Σ by maximum likelihood based on the observed components B_1X_1, B_2X_2, ..., B_pX_p. The likelihood equations for μ and Σ are

$$\sum_{r=1}^{p} B_r^T(B_r \Sigma B_r^T)^{-1} B_r\mu = \sum_{r=1}^{p} B_r^T(B_r\Sigma B_r^T)^{-1}B_rX_r \ . \tag{3.1}$$

and

$$\sum_{r=1}^{p} B_r^T(B_r\Sigma B_r^T)^{-1}B_r = \sum_{r=1}^{p} B_r^T(B_r\Sigma B_r^T)^{-1}B_r(X_r - \mu)(X_r - \mu)^T B_r^T(B_r\Sigma B_r^T)^{-1}B_r \ . \tag{3.2}$$

and have no explicit solution, although for given Σ (3.1) may be solved explicitly for μ provided that the matrix an the left of (3.2) is invertible.

Components i and j are <u>paired</u> in the observation B_rX_r if both the i^{th} and j^{th} columns of B_r contain a 1. Let $\phi(i, j, p)$ denote the relative frequency with which the i^{th} and j^{th} components are paired in the first p observations B_1X_1, ..., B_pX_p, and let $\phi_1(i, j) = \lim_{p \to \infty} \phi(i, j, p)$.

<u>Theorem 2</u>: Let X_1, X_2, ... be independent, identically distributed according to $N_n(\mu, \Sigma)$. If $\phi_1(i, j) > 0$ for all $i, j = 1, \ldots, n$, then there is a unique strongly consistent solution of the likelihood equations (3.1) and (3.2), which has the asymptotic properties given in Theorem 1.

<u>Proof</u>: The only one of conditions (i) - (vii) in Theorem 1 which poses any

difficulty is number (vii). For $\theta = (\mu, \Sigma)$, the information matrix $J_r(\theta)$ corresponding to the density of $B_r X_r$,

$$q_r(\cdot|\theta) = N_{n_r}(B_r\mu, B_r\Sigma B_r^T) ,$$

is

$$J_r(\theta) = \left[\begin{array}{c|c} U_r(\theta) & 0 \\ \hline 0 & U_r(\theta) \otimes U_r(\theta) \end{array}\right] , \tag{3.3}$$

where $U_r(\theta) = B_r^T(B_r\Sigma B_r^T)^{-1}B_r$, and the Kronecker product $U_r(\theta) \otimes U_r(\theta)$ represents the symmetric operator on $n \times n$ real symmetric matrices S (with trace inner product) defined by $U_r(\theta)SU_r(\theta)$. Thus (vii) is satisfied if for each Σ there exists $\epsilon = \epsilon(\Sigma)>0$ such that for all p sufficiently large

$$\frac{1}{p}\sum_{r=1}^{p} Z^T B_r^T(B_r\Sigma B_r^T)^{-1}B_r Z \geq \epsilon Z^T Z \tag{3.4}$$

and

$$\frac{1}{p}\sum_{r=1}^{p} \mathrm{Tr}[B_r^T(B_r\Sigma B_r^T)^{-1}B_r S]^2 \geq \epsilon \mathrm{Tr}S^2 \tag{3.5}$$

for all $Z \in R^n$ and symmetric S. However, (3.5) implies (3.4), as can be seen by taking $S = ZZ^T$. Hence, it suffices to establish (3.5) under the stated hypotheses.

Now,

$$\begin{aligned} &\mathrm{Tr}[B_r^T(B_r\Sigma B_r^T)^{-1}B_r S]^2 \\ &= \mathrm{Tr}[(B_r\Sigma B_r^T)^{-1}(B^r S B_r^T)]^2 \\ &= \mathrm{Tr}[(B_r\Sigma B_r^T)^{-\frac{1}{2}}(B_r S B_r^T)(B_r\Sigma B_r^T)^{-\frac{1}{2}}]^2 \\ &\geq \rho[(B_r\Sigma B_r^T)^{-\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{-\frac{1}{2}}]\ \mathrm{Tr}[B_r S B_r^T]^2 \end{aligned}$$

But,

$$\rho[(B_r\Sigma B_r^T)^{-\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{-\frac{1}{2}}] = 1/\rho[(B_r\Sigma B_r^T)^{\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{\frac{1}{2}}]$$

and

$$\rho[(B_r\Sigma B_r^T)^{\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{\frac{1}{2}}] = \sup_{Tr\Delta^2\le 1} Tr(B_r\Sigma B_r^T)^{\frac{1}{2}}\Delta\,(B_r\Sigma B_r^T)\,\Delta\,(B_r\Sigma B_r^T)^{\frac{1}{2}}$$

$$= \sup_{Tr\Delta^2\le 1} Tr[(B\Sigma B_r^T)\Delta]^2$$

$$= \sup_{Tr\Delta^2\le 1} Tr\Sigma B_r^T\Delta B_r\Sigma B_r^T\Delta B_r$$

$$= \sup_{Tr\Delta^2\le 1} Tr[\Sigma^{\frac{1}{2}}B_r^T\Delta B_r\Sigma^{\frac{1}{2}}]^2$$

$$\le \rho[\Sigma^{\frac{1}{2}}\otimes\Sigma^{\frac{1}{2}}]\sup_{\lambda^2\le 1} Tr[B_r^T\Delta B_r]^2$$

$$= \rho[\Sigma^{\frac{1}{2}}\otimes\Sigma^{\frac{1}{2}}]\ .$$

The last equation follows from $B_rB_r^T = I_{n_r}$. Hence,

$$Tr[B_r^T(B_r\Sigma B_r^T)^{-1}B_rS]^2 \ge \sigma[\Sigma^{-\frac{1}{2}}\otimes\Sigma^{-\frac{1}{2}}]\,Tr[B_rSB_r^T]^2$$

$$= \sigma[\Sigma^{-\frac{1}{2}}\otimes\Sigma^{-\frac{1}{2}}]\,Tr[B_r^TB_rSB_r^TB_r]^2\ .$$

Therefore,

$$\frac{1}{p}\sum_{r=1}^{p} Tr[B_r^T(B_r\Sigma B_r^T)^{-1}B_rS]^2 \ge \sigma[\Sigma^{-\frac{1}{2}}\otimes\Sigma^{-\frac{1}{2}}]\cdot\frac{1}{p}\sum_{r=1}^{p} Tr[B_r^TB_rSB_r^TB_r]^2$$

$$\ge \sigma[\Sigma^{-\frac{1}{2}}\otimes\Sigma^{-\frac{1}{2}}]\sigma[\frac{1}{p}\sum_{r=1}^{p}(B_r^TB_r)\otimes(B_r^TB_r)]TrS^2$$

Since eventually

$$\sigma[\frac{1}{p}\sum_{r=1}^{p}(B_r^TB_r)\otimes(B_r^TB_r)] > \frac{1}{2}\min_{i,j}\phi_1(i,j)\ ,$$

(vii) follows upon taking $\epsilon = \frac{1}{2}\min_{i,j}\phi_1(i,j)\cdot\rho[\Sigma^{\frac{1}{2}}\otimes\Sigma^{\frac{1}{2}}]$. QED.

The second application of Theorem 1 is to the problem outlined in Example 2. We assume that the unknown component densities $f(x|\tau_i)$ are from a regular exponential family (see [1] for definitions) with minimal canonical representation

(3.6) $$f(x|\tau) = C(\tau) \exp \langle \tau | F(x) \rangle \qquad (\tau \in T)$$

with respect to a σ-finite measure λ, where T is an open subset of a finite dimensional space V with inner product $\langle \cdot | \cdot \rangle$. We also assume that for distinct $\tau_1, \ldots, \tau_m$, the functions $e^{\langle \tau_1 | F(x) \rangle}, \ldots, e^{\langle \tau_m | F(x) \rangle}$, together with any components of $F(x)e^{\langle \tau_1 | F(x) \rangle}, \ldots, F(x)e^{\langle \tau_m | F(x) \rangle}$ are linearly independent [1]. The joint density of $X_r = (X_{r1}, \ldots, X_{rN_r})$, given that X_r is a sample from $f(x|\tau_\ell)$ is

(3.7) $$p_r(x_r|\tau_\ell) = \gamma_r(\tau_\ell) \exp \langle \tau_\ell | G_r(x_r) \rangle$$

where $$x_r = (x_{r1}, \ldots, x_{rN_r})$$

$$\gamma_r(\tau_\ell) = C(\tau_\ell)^{N_r}$$

and

$$G_r(x_r) = \sum_{j=1}^{N_r} F(x_{rj}) \quad .$$

The log-likelihood for the parameter $\theta = (\alpha_1, \ldots, \alpha_{m-1}, \tau_1, \ldots, \tau_m)$ of Example 2, based on the sample $X_1, \ldots, X_p$ is

(3.8) $$L_p(\theta) = \sum_{r=1}^{p} \log q_r(X_r|\theta) \quad ,$$

where

(3.9) $$q_r(X_r|\theta) = \sum_{\ell=1}^{m} \alpha_\ell \, p_r(X_r|\tau_\ell)$$

and $p_r(X_r|\tau_\ell)$ is given by (3.7). The following lemma collects some facts about exponential families which we require. For proofs, see Barndorff-Nielsen [1].

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

Then

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

(ii) $\kappa(\tau)$ has derivatives of all orders which may be obtained by differentiating under the integral sign. $D_\tau^k \kappa(\tau)$ may conveniently be represented as a symmetric k-linear form on V whose coefficients are polynomials in the first k moments of F. In particular,

(iii) $$D_\tau \kappa(\tau) = \langle E_\tau(F) | \cdot \rangle = \int_{R^n} \langle F(x) | \cdot \rangle f(x|\tau) d\lambda(x)$$

and

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

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

We are now ready to establish consistency of the MLE in Example 2.

Theorem 3: If the numbers $\{N_r\}$ are bounded and $L_p(\theta)$ is given by (3.8) then with probability 1 there is a unique consistent solution of $D_\theta L_p(\theta) = 0$ which, moreover, is a MLE of the parameter $\theta^o = (\alpha_1^o, \ldots, \alpha_{m-1}^o, \tau_1^o, \ldots, \tau_m^o)$ and is asymptotically normal and efficient.

Proof: Write $\mu_r(\tau_\ell) = E_{\tau_\ell}(G_r)$; $\mu(\tau_\ell) = E_{\tau_\ell}(F)$. Using Lemma 1, the nonzero derivatives of $q_r(x_r|\theta)$ up to order 2 are:

(3.10) $$D_{\alpha_\ell} q_r(x_r|\theta) = p_r(x_r|\tau_\ell) - p_r(x_r|\tau_m) \quad , \quad 1 \le \ell \le m-1$$

(3.11) $$D_{\tau_\ell} q_r(x_r|\theta) = \alpha_\ell p_r(x_r|\tau_\ell) \langle G_r(x_r) - \mu_r(\tau_\ell) | \cdot \rangle \quad , \quad 1 \le \ell \le m$$

(3.12) $$D^2_{\tau_\ell,\alpha_\ell} q_r(x_r|\theta) = p_r(x_r|\tau_\ell)\langle G_r - \mu_r(\tau_\ell)|\cdot\rangle \ , \quad 1 \le \ell \le m-1$$

(3.13) $$D^2_{\tau_m,\alpha_\ell} q_r(x_r|\theta) = -p_r(x_r|\tau_m)\langle G_r - \mu_r(\tau_m)|\cdot\rangle \ , \quad 1 \le \ell \le m-1$$

(3.14) $$D^2_{\tau_\ell} q_r(x_r|\theta) = \alpha_\ell p_r(x_r|\tau_\ell)\{\langle G_r - \mu_r(\tau_\ell)|\cdot\rangle^2 - \mathrm{cov}_{\tau_\ell}(G_r)\} \ , \quad 1 \le \ell \le m \ .$$

Conditions (i) and (ii) of Theorem 1 follow immediately from (3.10) - (3.14). Similarly, using Lemma 1 and the boundedness of $\{N_r\}$, conditions (iii) - (vi) of Theorem 1 are readily verified. It remain to verify (vii). We may write $J_r(\psi)$ in matrix form as

$$J_r(\theta) = \begin{bmatrix} I_1 & 0 \\ 0 & N_r^{\frac{1}{2}} I_2 \end{bmatrix} E_\theta \begin{bmatrix} A_r & B_r \\ B_r^* & C_r \end{bmatrix} \begin{bmatrix} I_1 & 0 \\ 0 & N_r^{\frac{1}{2}} I_2 \end{bmatrix}$$

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

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

$$B_r = \left(\frac{\alpha_k p_r(X_r|\tau_k)[p_r(X_r|\tau_\ell) - p_r(X_r|\tau_m)]}{q_r(X_r|\theta)^2} N_r^{-\frac{1}{2}} \langle G_r - \mu_r(\tau_k)|\cdot\rangle \right) \qquad \begin{matrix} \ell = 1, \ldots, m-1 \\ k = 1, \ldots, m \end{matrix}$$

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

The assumptions concerning the linear dependence of the functions $\exp\langle\tau|F(x)\rangle$ and $F(x)\exp\langle\tau|F(x)\rangle$ insure that $J_r(\theta)$ is positive definite for each r. Condition (vii) will be established once it is shown that the smallest eigenvalue of $J_r(\theta)$ is bounded away from zero as $N_r \to \infty$.

Clearly,

$$\sigma(J_r(\theta)) \geq \sigma\left(E_\theta\begin{bmatrix} A_r & B_r \\ B_r^* & C_r \end{bmatrix}\right) .$$

Observe that

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

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 positive by the strict convexity of κ. Hence,

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

Therefore,

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

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

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

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

$$\frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\theta)} N_r^{-1/2}(G_r - \mu_r(\tau_k))$$

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

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

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

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

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

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

is bounded as $N_r \to \infty$. It follows from a standard theorem on convergence of moments [4, p. 95] that

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

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

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

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

4. Concluding Remarks.

Theorem 3 remains true under weake assumptions then the boundedness of the sample sizes N_r, but nothing like the approach embodied in Theorem 1 will work without some restrictions on N_r. Nevertheless, it is far from

intuitively clear that restrictions are needed for the existence of a consistent MLE. Similarly, it seems plausible that the assumption in Theorem 2 that components be paired with nonzero asymptotic frequency might also be weakened. In certain cases, e.g., when a normal mean is to be estimated from data with missing components and the covariance is the identity, the existence of a consistent MLE with desirable asymptotic properties can be shown under weaker hypotheses than those derived from Theorem 1. The condition in Theorem 1 that $\phi_1(i, j) > 0$ for all i and j is nevertheless reasonable since it is equivalent to the condition that the Cramer-Rao lower bound be of the order of $\frac{1}{p}$ as $p \to \infty$.

REFERENCES

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

2. J. Bryant (1979). On the clustering of multidimensional pictorial data, Pattern Recognition, 11(2), 115-125.

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

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

5. H. Cramér (1946). Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J.

6. R. C. Dahiya and R. M. Korwar (1980). Maximum likelihood estimates for a bivariate normal distribution with missing data, Annals of Statistics, 8(3), 687-692.

7. A. D. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. B, 39, 1-38.

8. R. Dubes and A. K. Jain (1979). Validity Studies in clustering methodologies, Pattern Recognition, 11(2), 235-254.

9. R. O. Duda and P. E. Hart (1973). Pattern Classification and Scene Analysis, John Wiley and Sons, New York.

10. V. Foutz (1977). On the unique consistent solution to the likelihood equations, JASA, 72, 357, 147-148.

11. V. S. Huzurbazar (1948). The likelihood equation, consistency and the maxima of the likelihood function, Annals of Eugenics, 14, 3, 185-200.

12. L. LeCam (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, Univ. of California Publ. in Statist., 1, 277-330.

13. C. Peters and F. Kampe (1980). Numerical trials of HISSE, Report no. 75, Department of Mathematics, Univ. of Houston.

14. B. C. Peter, Jr. and H. F. Walker (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions, SIAM J. Appl. Math. B, 35, 2, 362-378.

15. S. D. Silvey (1961). A note on the maximum likelihood in the case of dependent random variables, J. Roy. Statist. Soc. B, 23, 444-452.

16. R. D. Tarone and G. Grueuhage (1975). A note on the uniqueness of roots of the likelihood equations for vector valued parameters, JASA 70, 903-904.

17. A Wald (1949). A note on the consistency of the consistency of the maximum likelihood estemate, Ann. Math. Statist. 20, 595-601.

18. J. H. Wolfe (1970). Pattern clustering by multivariate mixture analysis, Multivariate Behaviorial Research 5, 329-350.

19. J. Wolfowitz (1949). On Wald's proof of the consistency of the maximum likelihood estimate, Ann. Math. Statist. 20, 601-602.