NOTICE

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.
UNIVERSITY OF HOUSTON
HOUSTON, TEXAS

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
Unclas
63/43 04099

HOUSTON, TEXAS 77004
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
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 $X_1, X_2, \ldots$ be independent random vectors in $\mathbb{R}^n$ whose common density is one of a parametric family $(q(x|\theta))_{\theta \in \Theta}$, where $\Theta$ is a subset of $\mathbb{R}^v$. Suppose that instead of the $X_i$ we observe only certain subvectors $B_1 X_1, B_2 X_2, \ldots$, where $\{B_i\}$ is a given sequence of $n \times n$ 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 \( \theta \) 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 \( \{f(x|\tau)\}_{\tau \in \Gamma} \), and let \( \alpha_1, ..., \alpha_m \) be the unknown positive probabilities corresponding to a discrete mixing distribution supported on \( \{\tau_1, ..., \tau_m\} \). The number \( m \) is known. Under what conditions will there be a unique consistent MLE of the parameter \( \theta = (\alpha_1, ..., \alpha_{m-1}, \tau_1, ..., \tau_m) \) 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 \( X_i = (X_{i1}, ..., X_{it_i}) \) of known size from an unknown component density \( f(x|\tau_i) \)? In this example the parameter \( \theta \) is only locally identifiable. 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 [12] 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 $E_{\theta}$ will denote expectation with respect to a distribution determined by a parameter $\theta$ and $D_u, D^2_{uv}$ etc. will denote differentiation or partial differentiation with respect to scalar or vector variables $u, v$. For a scalar valued function, $V_u$ 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 $\sigma$, the symmetric $k$-linear form $f(\eta_1, \ldots, \eta_k) = \sum_{j=1}^{k} \langle \sigma_j \rangle$ will be denoted by $\langle \sigma \rangle^{\cdot, \cdot, \cdot}^k$. Thus, for example, we may write the covariance of a statistic $S$ as $\text{Cov}_T(S) = E_T(\langle S - E_T(S) \rangle \cdot \cdot^2)$. The largest and smallest eigenvalues of a symmetric positive definite operator $A$ will be denoted respectively by $\rho(A)$ and $\omega(A)$. 
2. A General Consistency Theorem. Let $\Theta$ be an open subset of $\mathbb{R}^v$ 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 $\sigma$-finite measure $\lambda_r^\theta$ on $\mathbb{R}^{N_r}$. Let $\theta^0 \in \Theta$ and let $X_1, \ldots, X_p, \ldots$ be a sequence of independent random vectors with $X_r$ having density $q_r(\cdot|\theta^0)$. For $\theta \in \Theta$ define

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

Theorem 1: Suppose

(i) \[ \int_{\mathbb{R}^{N_r}} D^2_{\theta} q_r(x|\theta^0) \, d\lambda_r(x) = 0 \]

(ii) \[ \int_{\mathbb{R}^{N_r}} D^2_{\theta} q_r(x|\theta^0) \, d\lambda_r(x) = 0 \]

and that there is a constant $M$, functions $f_r$, a neighborhood $\Omega$ of $\theta^0$ and $\lambda_r$-null sets $A_r$ in $\mathbb{R}^{N_r}$ such that for all $r$, $\theta \in \Omega$, $x \notin A_r$.

(iii) \[ |D^3_{\theta_i, \theta_j, \theta_k} \log q_r(x|\theta^0)| \leq f_r(x) \quad i, j, k = 1, \ldots, v \]

(iv) \[ E_{\theta^0}(f_r(X_r))^2 \leq M \]

(v) \[ E_{\theta^0}(D_{\theta} \log q_r(X_r|\theta^0))^4 \leq M \quad i = 1, \ldots, v \]

(vi) \[ E_{\theta^0}(\frac{1}{q_r(X_r|\theta^0)^2} (D^2_{\theta_i, \theta_j} q_r(X_r|\theta^0))^2) \leq M \quad i, j = 1, \ldots, v \]

and

(vii) there exists $c > 0$ such that \[ \frac{1}{\sum_{r=1}^{p}} J_r(\theta^0) \geq c I_v \]

for sufficiently large $p$, where $J_r(\theta^0) = E_{\theta^0}(D_{\theta} \log q_r(X_r|\theta^0) v^T_{\theta} \log q_r(X_r|\theta^0))$, $I_v$ is the identity on $\mathbb{R}^v$, and the ordering is the usual one on symmetric operators. Then there is a neighborhood $\Omega^0$ of $\theta^0$ such that with probability 1 there is an integer $p_1$ such that for $p \geq p_1$ there is a unique solution $\theta^0$ in $\Omega^0$ of the likelihood equation.
$D_0L_p(\theta) = 0$. Furthermore, $\theta^p \to \theta^0$ as $p \to \infty$ and $\theta^p$ is a maximum likelihood estimate. The consistent estimator $\theta^p$ 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 $S_p(\theta) = \frac{1}{p} \sum_{r=1}^{p} D_0 \log q_r(x_r|\theta)$. By (i) $E_{\theta_0}(S_p(\theta^0)) = 0$ and by

(v) $S_p(\theta^0) \to 0$ a.s. as $p \to \infty$. Consider the $v \times v$ matrix $D_0S_p(\theta^0)$ whose $i, j$th element is

$$\frac{1}{p} \sum_{r=1}^{p} D_0^2 \log q_r(x_r|\theta) = \frac{1}{p} \sum_{r=1}^{p} \frac{1}{q_r(x_r|\theta^0)} D_0^2 \log q_r(x_r|\theta^0) - \frac{1}{p} \sum_{r=1}^{p} D_0 \log q_r(x_r|\theta) D_0 \log q_r(x_r|\theta^0) \log q_r(x_r|\theta^0).$$

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

$$D_0S_p(\theta^0) + \frac{1}{p} \sum_{r=1}^{p} J_r(\theta^0) \to 0$$

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

$$D_0S_p(\theta^0) \leq -2nI.$$ 

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

$$\frac{1}{p} \sum_{r=1}^{p} \left| \log q_r(x_r|\theta) - D_0^2 \log q_r(x_r|\theta^0) \right|$$

$$\leq \frac{1}{p} \sum_{r=1}^{p} \sum_{k=1}^{v} \left| \theta_k - \theta_k^0 \right| f_r(x_r) + \frac{1}{p} \sum_{r=1}^{p} \sum_{r=1}^{v} \left| \theta_k - \theta_k^0 \right| \int_{0}^{1} \left| D_0^3 \log q_r(x_r|\theta^0 + t(\theta - \theta^0)) \right| dt$$

$$\leq \frac{1}{p} \sum_{r=1}^{p} \sum_{k=1}^{v} \left| \theta_k - \theta_k^0 \right| f_r(x_r)$$
With probability 1, for large $p$

$$\frac{1}{p} \sum_{r=1}^{p} f_r(X_r) \leq 1 + \frac{1}{p} \sum_{r=1}^{p} E_{\theta_0}(f_r(X_r))$$

$$\leq 1 + M^k.$$

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

$$||D_{\theta_0}^p(\theta) - D_{\theta_0}^p(\theta^0)|| \leq \bar{M}||\theta - \theta^0||.$$

Thus there is a convex neighborhood $\Omega^0$ of $\theta^0$ such that

$$D_{\theta_0}^p(\theta) \leq -nI$$

for all $\theta \in \Omega^0, p \geq p_1$. It now follows that for $p \geq p_1$ $S_p$ is one to one on $\Omega^0$ and that the image under $S_p$ of the sphere $\Omega_\delta(\theta^0)$ at $\theta^0$ of small radius $\delta$ contains the sphere $\Omega_{\eta \delta}(S_p(\theta^0))$ at $S_p(\theta^0)$ of radius $n\delta$. Since $0$ is eventually in $\Omega_{\eta \delta}(S_p(\theta^0))$ there is a unique solution of $D_0^p(\theta) = 0$ in $\Omega_\delta(\theta^0)$. Since $D_0^p(\theta)$ is negative definite, this solution is a MLE.

Let $v_p = \frac{1}{p} \sum_{r=1}^{p} v_r(\theta^0)$. The Cramér-Rao lower bound for $p$ observations is verified without difficulty to be $(p \Sigma_p)^{-1}$. By (v), (vi), and Liapounov's Theorem 14, p. 2001, $p^\frac{1}{2} \Sigma_p^{-\frac{1}{2}} S_p(\theta^0)$ is asymptotically distributed as $N_v(0, I)$. Moreover, in a neighborhood of $\theta^0$ we may write

$$S_p(\theta) = S_p(\theta^0) + A(\theta)(\theta - \theta^0)$$

where $A(\theta) = D_0^p S_p(\theta^0)$ as $\theta \to \theta^0$. It follows that with probability 1.

$$p^\frac{1}{2} \Sigma_p^{-\frac{1}{2}} (\theta^p - \theta^0) = -\Sigma_p^{-\frac{1}{2}} A(\theta^p)^{-1} \Sigma_p^{-\frac{1}{2}} p^\frac{1}{2} \Sigma_p^{-\frac{1}{2}} S_p(\theta^0)$$

for large $p$. Since $D_0^p S_p(\theta^0) + \Sigma_p \to 0$ and $A(\theta^p) \to D_0^p S_p(\theta^0)$ with probability 1,
the expression \(-\frac{1}{p} \sum_{i=1}^{p} A(\theta^P - \theta^0)\) converges almost surely to the identity. Therefore, \(\frac{1}{p} \sum_{i=1}^{p} A(\theta^P - \theta^0)\) is asymptotically \(N(U, I)\) and \(\theta^P\) is asymptotically efficient.
This concludes the proof.

3. Applications.

Suppose that in Example 1 ane \(X_i\) have a common \(n\) variate normal distribution \(N_n(\mu, \Sigma)\) and it is desired to estimate \(\mu, \Sigma\) by maximum likelihood based on the observed components \(B_1 X_1, B_2 X_2, \ldots, B_p X_p\). The likelihood equations for \(\mu\) and \(\Sigma\) are

\[
\sum_{r=1}^{P} B_r^T (B_r \Sigma B_r^T)^{-1} B_r \mu = \sum_{r=1}^{P} B_r^T (B_r \Sigma B_r^T)^{-1} B_r X_r.
\]

and

\[
\sum_{r=1}^{P} B_r^T (B_r \Sigma B_r^T)^{-1} B_r = \sum_{r=1}^{P} B_r^T (B_r \Sigma B_r^T)^{-1} B_r (X_r - \mu)(X_r - \mu)^T B_r^T (B_r \Sigma B_r^T)^{-1} B_r.
\]

and have no explicit solution, although for given \(\Sigma\) (3.1) may be solved explicitly for \(\mu\) provided that the matrix on the right of (3.2) is invertible.

Components \(i\) and \(j\) are paired in the observation \(B_r X_r\) 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 \(B_1 X_1, \ldots, B_p X_p\), and let \(\phi_1(i, j) = \lim_{p \to \infty} \phi(i, j, p)\).

**Theorem 2:** Let \(X_1, X_2, \ldots\) 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.

**Proof:** The only one of conditions (i) - (vii) in Theorem 1 which poses any
difficulty is number (vii). For $\theta = (\mu, \Sigma)$, the information matrix $J_p(\theta)$ corresponding to the density of $B_rX_r$,

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

is

(3.3) $$J_p(\theta) = \left[ \begin{array}{cc} U_p(\theta) & 0 \\ 0 & U_p(\theta) \otimes U_p(\theta) \end{array} \right] ,$$

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

(3.4) $$\frac{1}{p} \sum_{r=1}^{p} \text{Tr} [ T_r(B_r \Sigma B_r^T)^{-1} B_r S ] Z^T Z$$

and

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

for all $Z \in \mathbb{R}^n$ 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,

$$\text{Tr} [ (B_r \Sigma B_r^T)^{-1} B_r S ]^2$$

$$= \text{Tr} [ (B_r \Sigma B_r^T)^{-1} (B_r^T S B_r^T) ]^2$$

$$= \text{Tr} [ (B_r \Sigma B_r^T)^{-1} (B_r S B_r^T) (B_r \Sigma B_r^T)^{-1} ]^2$$

$$> \phi [ (B_r \Sigma B_r^T)^{-1} \otimes (B_r \Sigma B_r^T)^{-1} ] \text{Tr} [ B_r S B_r^T ]^2$$

But,

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

and
\[ \rho ((B_r \Sigma B_r^T)^{1/2} \circ (B_r \Sigma B_r^T)^{1/2}) = \sup_{\text{Tr}^2 s_1} \text{Tr}(B_r \Sigma B_r^T)^{1/2} \Delta (B_r \Sigma B_r^T) \Delta (B_r \Sigma B_r^T)^{1/2} \]

\[ = \sup_{\text{Tr}^2 s_1} \text{Tr}(B_r \Sigma B_r^T)^2 \]

\[ = \sup_{\text{Tr}^2 s_1} \text{Tr}(B_r \Sigma B_r^T \Sigma B_r^T \Delta B_r) \]

\[ = \sup_{\text{Tr}^2 s_1} \text{Tr}(B_r \Sigma B_r^T \Sigma B_r^T \Delta B_r)^2 \]

\[ < \rho(\Sigma^{1/2} \circ \Sigma^{1/2}) \sup_{\text{Tr}^2 s_1} \text{Tr}(B_r \Sigma B_r^T)^2 \]

\[ = \rho(\Sigma^{1/2} \circ \Sigma^{1/2}) . \]

The last equation follows from \( B_r^T B_r = I_{nr} \). Hence,

\[ \text{Tr}(B_r^T (B_r \Sigma B_r^T)^{-1} B_r S)^2 > \rho(\Sigma^{-1/2} \circ \Sigma^{-1/2}) \text{Tr}(B_r \Sigma B_r^T)^2 \]

\[ = \rho(\Sigma^{-1/2} \circ \Sigma^{-1/2}) \text{Tr}(B_r^T \Sigma B_r^T S B_r^T \Sigma B_r)^2 . \]

Therefore,

\[ \frac{1}{p} \sum_{p=1}^{P} \text{Tr}(B_r^T (B_r \Sigma B_r^T)^{-1} B_r S)^2 > \rho(\Sigma^{-1/2} \circ \Sigma^{-1/2}) \cdot \frac{1}{p} \sum_{p=1}^{P} \text{Tr}(B_r^T \Sigma B_r^T S B_r^T \Sigma B_r)^2 \]

\[ > \rho(\Sigma^{-1/2} \circ \Sigma^{-1/2}) \rho(\frac{1}{p} \sum_{p=1}^{P} (B_r^T B_r) \circ (B_r^T B_r) \text{Tr} S^2 . \]

Since eventually

\[ \rho(\frac{1}{p} \sum_{p=1}^{P} (B_r^T B_r) \circ (B_r^T B_r)) > \frac{1}{2} \min_{i,j} \phi_1(i,j) . \]

(vii) follows upon taking \( \cdot \frac{1}{2} \min_{i,j} \phi_1(i,j) \cdot \rho(\Sigma^{1/2} \circ \Sigma^{1/2}) \cdot \text{QED.} \)

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) \[ f(x|\tau) = C(\tau) \exp <\tau|F(x)> \quad (\tau \in \mathcal{T}) \]

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

(3.7) \[ p_r(x_r|\tau_\xi) = \gamma_r(\tau_\xi) \exp <\tau_\xi|G_r(x_r)> \]

where \[ x_r = (x_{r1}, \ldots, x_{rN_r}) \]
\[ \gamma_r(\tau_\xi) = C(\tau_\xi)^{N_r} \]

and \[ G_r(x_r) = \sum_{j=1}^{N_r} F(x_{rj}) \]

The log-likelihood for the parameter $\theta = (\alpha_1, \ldots, \alpha_{m-1}, \lambda_1, \ldots, \lambda_m)$ of Example 2, based on the sample $X_1, \ldots, X_p$ is

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

where \[ q_r(X_r|\theta) = \sum_{\xi=1}^{\ell} a_\xi p_r(X_r|\tau_\xi) \]

and $p_r(X_r|\tau_\xi)$ is given by (3.7). The following lemma collects some facts about exponential families which we require. For proofs, see Barndorff-Nielsen [11].
Lemma 1: Let (1) be a canonical representation of an exponential family. For \( \tau \in T \) let \( \kappa(\tau) = -\ln \mathcal{C}(\tau) = \ln \int \exp<\tau|F(x)>d\lambda(x) \) 

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) \( D_\tau \kappa(\tau) = \int \frac{E_{\tau}(F)}{\cdot} = \int \frac{F(x|\tau)}{\cdot}f(x|\tau)d\lambda(x) \)

and

(iv) \( D_\tau^2 \kappa(\tau) = \text{cov}_\tau(F) = \int \left(F - E_{\tau}(F)\right)\frac{\cdot^2}{\cdot}f(x|\tau)d\lambda(x) \) ; \( D_\tau^2 \kappa(\tau) \) 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 \( \theta = (\alpha_1^0, ..., \alpha_{m-1}^0, \tau_1^0, ..., \tau_m^0) \) and is asymptotically normal and efficient.

Proof: Write \( u_r(\tau_g) = E_{\tau_g}(G_r) ; u(\tau_g) = E_{\tau_g}(F) \). Using Lemma 1, the nonzero derivatives of \( q_r(x_r|\theta) \) up to order 2 are:

(3.10) \( D_{\alpha_{g,1}} q_r(x_r|\theta) = p_r(x_r|\tau_g) - p_r(x_r|\tau_m) \), \( 1 \leq g \leq m-1 \)

(3.11) \( D_{\tau_g} q_r(x_r|\theta) = \alpha_g p_r(x_r|\tau_g) G_r(x_r) - \mu_r(\tau_g) \cdot \cdot \), \( 1 \leq g \leq m \)
\begin{align}
(3.12) \quad & D_{x_2}^2 q_r(x_r|\theta) = p_r(x_r|\tau_k) < G_r - \mu_r(\tau_k) | \cdot >, \quad 1 \leq k \leq m-1 \\
(3.13) \quad & D_{x_2}^2 q_r(x_r|\theta) = -p_r(x_r|\tau_m) < G_r - \mu_r(\tau_m) | \cdot >, \quad 1 \leq k \leq m-1 \\
(3.14) \quad & D_{x_2}^2 q_r(x_r|\theta) = \alpha_k p_r(x_r|\tau_k) (\cdot)^2 - \text{cov}_{\tau_k}(G_r), \quad 1 \leq k \leq m .
\end{align}

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

\[
J_r(\theta) = \begin{bmatrix}
I_1 & 0 \\
0 & N_r^{m-1}
\end{bmatrix}
E_0 \begin{bmatrix}
A_r & B_r \\
B^*_r & C_r
\end{bmatrix}
\begin{bmatrix}
I_1 & 0 \\
0 & N_r^{m-1}
\end{bmatrix}
\]

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

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

\[
B_r = \frac{\alpha_k p_r(x_r|\tau_k) p_r(x_r|\tau_k)}{q_r(x_r|\theta)^2} N_r^{m-1} < G_r - \mu_r(\tau_k) | \cdot > \quad \ell = 1, \ldots, m-1
\]

\[
C_r = \frac{\alpha_k p_r(x_r|\tau_k) p_r(x_r|\tau_k)}{q_r(x_r|\theta)^2} N_r^{m-1} (G_r - \mu_r(\tau_k) | \cdot ) \quad k, \ell = 1, \ldots, m.
\]

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, 
\[ \alpha(J_\theta(\theta)) \geq \alpha \left( E_\theta \left[ \begin{array}{cc} A_r & B_r \\ B_r^* & C_r \end{array} \right] \right) \].

Observe that 
\[ \frac{p_r(X_r|\tau_k)}{p_r(X_r|\tau_\ell)} = \exp \{-N_r[k(\tau_\ell) - k(\tau_k) - \zeta_k|\frac{1}{N_r} G_r>\} \].

If \( X_r \) is a sample from \( f(x|\tau_k) \), then the expression in square brackets converges to 
\[ k(\tau_\ell) - k(\tau_k) - \zeta_k|E_{\tau_k}(F) > = k(\tau_\ell) - k(\tau_k) - k'(\tau_k) \cdot (\tau_\ell - \tau_k) \]
which is positive by the strict convexity of \( k \). Hence,
\[ \frac{p_r(X_r|\tau_\ell)}{p_r(X_r|\tau_k)} \rightarrow 0 \text{ as } N_r \rightarrow \infty \].

Therefore,
\[ E_\theta \left[ \frac{p_r(X_r|\tau_\ell) p_r(X_r|\tau_k)}{q_r(X_r|\theta)^2} \right] = E_{\tau_k} \left[ \frac{p_r(X_r|\tau_k)}{q_r(X_r|\theta)} \right] \]
converges to 0 if \( \ell \neq k \) and \( \frac{1}{\alpha_k} \) if \( \ell = k \) as \( N_r \rightarrow \infty \). Thus,
\[ E_\theta[A_r] \rightarrow \left( \frac{1}{2} + \frac{\delta_{\ell k}}{\alpha_m} \right) \text{ as } N_r \rightarrow \infty \].

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

Hence,
\[ \frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\theta)} \cdot N^{-\frac{1}{2}}_r(G_r - \mu_r(\tau_k)) \]
converges in distribution to 0 if \( \ell \neq k \) and \( \frac{1}{\alpha_k} Z \) if \( \ell = k \).

Let \( \Lambda \) be any element of \( V \) and consider

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

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

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

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

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

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

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

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

\[
E_\emptyset(C_r) + (\delta_k \text{cov}_{\tau_k}(F))
\]

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

4. **Concluding Remarks.**

Theorem 3 remains true under weaker assumptions than 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


