80-10056

NASA CR- 160445

DEPARTMENT OF MATHEMATICS

UNIVERSITY OF HOUSTON HOUSTON, TEXAS

(E80-10056) DEVELOPMENT OF MATHEMATICAL TECHNIQUES FOR THE ANALYSIS OF REMOTE SENSING DATA Final Report, 1 Dec. 1978 - 31 Dec. 1979 (Houston Univ.) 31 p HC A03/MF A01 CSCL 05B G3/43

N80-18508

Unclas 00056

FINAL REPORT
NAS-9-15543

DECEMBER 1,1978-DECEMBER 31,1979

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

HOUSTON, TEXAS 77004

A General Framework for Pixel Classification

The ideas presented here represent an attempt to define a natural set of pixel categories which will be represented in a typical LANDSAT scene and which we hope can be delineated with some success by the use of available spatial/spectral clustering algorithms. The pixel categories and their characteristics are:

P - The set of "pure" pixels; i.e., pixels from within fields. these are characterized by a high degree of local spectral homogeneity; that is, elements of P have adjacent pixels which look spectrally alike.

T_1- The set of "trash" pixels. These pixels do not have homogeneous spatial neighborhoods and are relatively distant, spectrally, from the set P.

B - The set of boundary pixels, pixels at the common boundaries of adjacent fields. Elements of B have spatial neighbors in P and no spatial neighbors in the set T_1.

T_2- All other pixels. These have no pure neighbors or else have neighbors in the class T_1, thus the spatial information is ambiguous. However, elements of T_2 are relatively near, spectrally, to the pure pixels.

Obviously if these four categories can be identified, they require different means of processing to extract estimates of the acreages of the real classes. T_1 will not be processed at all, since there is neither spatial nor spectral evidence that it consists of agriculture. The processing of T_2, if it occurs, will rely almost wholly on spectral measurements, since the spatial information is ambiguous for elements of T_2. B consists of pixels whose spectral response can be properly regarded as a mixture of pure pixel responses and a fairly detailed proposal for handling B follows.

There is considerable doubt about whether a normal mixture model with mixing proportions easily related to "real" class acreage proportions is valid for LANDSAT agricultural data. It seems clearly inappropriate for categories T_1 and B and possibly appropriate for T_2 and P. There are objections to applying it to P. First, the reason for preferring the density estimation approach to the clustering and counting approach, namely that the proportion estimates are unbiased, may be invalid for P because the spectral observations are far from independent. Second, the proportion estimates are meaningless unless the component densities of the mixture are related to real classes. This means the field structure of P must be respected by the estimator. Meeting the first objection requires that the dependence between nearby pixels be somehow modeled. The obvious (but probably not adequate) solution to the second problem is to use the clusters generated in P by a spatial/spectral clustering algorithm which preserves the integrity of fields to initialize the parameters in a maximum likelihood algorithm for the normal mixture distribution. For example, the algorithm AMOEBA, after determining the best clustering of some test data, then assigns whole fields to single clusters by a nearest cluster center classification of the field means. It should be noted that in terms of the assumptions underlying AMOEBA[1], it is senseless to graft the familiar maximum likelihood procedures UHMLE and CLASSY to AMOEBA in exactly the naive way just suggested. Indeed, they are based on the wrong likelihood function for the kind of partitioned sample we are considering with P.

In processing B, the boundary pixels, we suggest that the following procedure should be considered. We assume that the set P of pure pixels has been classified, so that a class label $i(r) \in \{1,\ldots,m\}$ is assigned to each pixel $r \in P$.

Given a pixel r in the scene, let $x(r)$ denote its vector of spectral measurements. For $r \in B$ let

$$P(r) = \{s \in P \mid s \text{ is a neighbor of } r\}.$$

and

$$C(r) = \{i(s) \mid s \in P(r)\}.$$

Thus $C(r)$ is the set of class labels of pure spatial neighbors of r. Let C be a set of ≤ 4 (or ≤ 8) of the class labels $\{1, \ldots, m\}$. Define

$$B_C = \{r \in B \mid C(r) = C\}.$$

Thus B_C is the set of boundary pixels whose pure spatial neighbors have exactly those classifications listed in C. For acreage estimation, we treat each set B_C separately and then combine the estimates to get an acreage estimate for B. For simplicitly we suppose that $C = \{1,2\}$. The generality of the discussion will be obvious. If $r \in B_C$ then r has pure neighbors in classes 1 and 2 only. (Recall that r may have impure neighbors, but none of them belong to the trash class T_1.) Let

$$P_1(r) = \{s \in P(r) \mid i(s) = 1\}$$

$$P_2(r) = \{s \in P(r) \mid i(s) = 2\}$$

and

$$P_1(B_C) = \bigcup_{r \in B_C} P_1(r)$$

$$P_2(B_C) = \bigcup_{r \in B_C} P_2(r)$$

The following are our assumptions about the spectral measurements of elements of B_C. Let r be an arbitrary element of B_C.

1) For each $s \in P(r)$, a fraction $\beta(s,r)$ of the area of pixel r has the same reflectance properties as s. The spectral response from r can be written as

$$x(r) = \beta_1(r)x_1(r) + \beta_2(r)x_2(r) + \epsilon(r),$$

where

$$\beta_1(r) = \sum_{s \in P_1(r)} \beta(s,r)$$

$$\beta_2(r) = \sum_{s \in P_2(r)} \beta(s,r)$$

$$x_1(r) = \sum_{s \in P_1(r)} \frac{\beta(s,r)}{\beta_1(r)} x(s)$$

$$x_2(r) = \sum_{s \in P_2(r)} \frac{\beta(s,r)}{\beta_2(r)} x(s)$$

and ϵ is an error term whose expectation is 0.

2) β_1 and x_1 are uncorrelated as are β_2 and x_2.

3) $E[x_j(r) \mid r \in B_c] = E[x(s) \mid s \in P_j(B_c)]$ $j = 1,2$.

If assumptions (1) - (3) are valid then

(*) $$E[x(r) \mid r \in B_c] = E[\beta_1(r) \mid r \in B_c]E[x(s) \mid s \in P_1(B_c)] + E[\beta_2(r) \mid r \in B_c]E[x(s) \mid s \in P_2(B_c)]$$

The numbers $E[\beta_j(r) \mid r \in B_c]$ are easily related to the acreages of classes 1 and 2 in B_c.

In practice, we intend to estimate $E[\beta_j(r) \mid r \in B_c]$ $j = 1,2$ as least squares solutions of (*). If any set B_c produces an unacceptably large residual error we take that as an indication that the set B_c defined by the algorithm does not consist of boundary pixels. If many sets B_c produce large residual errors, even after experimenting with the tolerances implicit in the definitions of P, B, T_1, and T_2 then we would tend to believe that the boundary pixel model of assumptions (1) - (3) is wrong.

References

1. Jack Bryant, "On the clustering of multidimensional pictorial data." To appear in Pattern Recognition.

1. Introduction

A possible objection to the use of UHMLE or CLASSY in conjunction with AMOEBA is that both these algorithms ignore the association of pixels in fields. Indeed AMOEBA is based on the explicit assumption that pixels in the same field represent the same real class [1], while the assumptions underlying the maximum likelihood algorithms imply that the classification of a pixel is independent of the classification of other pixels. In this report a statistical model based on normal mixtures is proposed which takes into account the organization of LANDSAT agricultural data into fields which are homogeneous as to crop type. Likelihood equations for the parameters of the model are derived which may be solved iteratively as in UHMLE.

2. The Model

We assume that the data elements (pixel data vectors) are real n-vectors each from one of the statistical populations $\Pi_1, \ldots, \Pi_m$ with n-variate density functions $p(x|\Pi_\ell)$, $\ell = 1, \ldots, m$. We assume that the data is organized into sets (fields) $F_1, \ldots, F_p$, where F_j has N_j data elements which have been previously ordered in some arbitrary fashion so that the data elements in F_j form a nN_j-dimensional vector denoted by $X_j = \begin{matrix} x_{j1} \\ \vdots \\ x_{jNj} \end{matrix}$

Define random variables $\{\theta_{jk} \epsilon \{1, \ldots, m\} | j=1, \ldots, p;\ k=1; \ldots, N_j\}$ by $\theta_{jk}=\ell$ if and only if x_{jk} is from Π_ℓ. We assume that all the observations from F_j are from the same class, so that we may write $\theta_{jk} = \theta_j$ for all $j = 1, \ldots, p$; $k, \ell = 1, \ldots, N_j$. Finally, we assume that $(X_1, \theta_1), \ldots, (X_p, \theta_p)$ are independent, that the θ_j's are identically distributed, that $\alpha_\ell = \text{Prob}\,[\theta_j = \ell] > 0$ and that $\sum_{\ell=1}^{m} \alpha_\ell = 1$. Under the stated assumptions, the joint density of $X_1, \ldots, X_p$ is $p(x_1, \ldots, x_p) = \prod_{j=1}^{p} \sum_{\ell=1}^{m} \alpha_\ell\, p\,(x_j)$, where $p_\ell(x_j) = p(x_{j1}, \ldots, x_{jNj} | \theta_j = \ell)$ is the joint density of the elements of F_j given that F_j represents class Π_ℓ.

Let $N = N_1 + \quad + N_p$ and for each ℓ let M_ℓ denote the total number of the N observations X_{jk} which come from class Π_ℓ. The following proposition shows that with reasonable restrictions on the field sized N_j the values of $\{M_\ell : \ell = 1, \ldots, m\}$ can be inferred from a knowledge of the parameters α_ℓ. Thus, acreage estimates of the classes can be derived from estimates of the parameters α_ℓ.

Proposition 1: (a) $E(M_\ell) = \alpha_\ell N$

(b) $\frac{M_\ell}{N} \to \alpha_\ell$ in probability as $p \to \infty$ if and only if

$$\lim_{p\to\infty} \frac{1}{N^2} \sum_{j=1}^{p} N_j^2 = 0.$$

(c) If $\sum_{j=1}^{\infty} N_j^2 \big/ j^2 < \infty$, then $\frac{M_\ell}{N} \to \alpha_\ell$ almost surely.

Proof: (a) Write $M_\ell = \sum_{j=1}^{p} \sum_{k=1}^{N_j} X_{\underline{\ell}}\ (\theta_{jk})$

$$= \sum_{j=1}^{p} N_j\ X_\ell\ (\theta_j)$$

where $\chi_\ell(r) = \begin{cases} 1 & r = \ell \\ 0 & r \neq \ell \end{cases}$

Then $E(M_\ell) = \sum_{j=1}^{p} N_j\, E(\chi_\ell(\theta_j)) = \sum_{j=1}^{p} N_j\, \alpha_\ell = N\alpha_\ell$.

(b) Since $\frac{M_\ell}{N} - \alpha_\ell = \frac{M_\ell}{N} - \left(E\,\frac{M_\ell}{N}\right)$ is bounded, it converges to zero in probability iff var $\frac{M_\ell}{N} \to 0$ as $p \to \infty$. Since the terms $N_j \chi_\ell(\theta_j)$ are independent,

$$\mathrm{var}\left(\frac{M_\ell}{N}\right) = \frac{1}{N^2} \sum_{j=1}^{p} N_j^2\ \mathrm{var}\,(\chi_\ell(\theta_j)) = \frac{1}{N^2} \sum_{j=1}^{p} N_j^2\, \alpha_\ell(1-\alpha_\ell).$$

The conclusion follows.

(c) The assertion follows immediately from Kolmogorov's version of the strong law of large numbers [3].

3. Maximum Likelihood Estimation of the Parameters

In this section we suppose that the class conditional densities $p(x|\Pi_\ell)$ of the data elements X_{jk} are n-variate normal $N(x{:}\mu_\ell,\Sigma_\ell)$ and that $\{X_{jk}{:}\ k=1, \ldots, N_1\}$ are class conditionally independent; i.e., that

$$P_\ell(x_j) = \prod_{k=1}^{N_j} N(x_{jk};\ \mu_\ell,\ \Sigma_\ell).$$

for $j=1, \ldots, p$. In this case the joint density of $X_1, \ldots, X_p$,

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

is parametrized by $\{(\alpha_\ell, \mu_\ell, \Sigma_\ell) | \ell=1, \ldots, m\}$ where $\alpha_\ell \geq 0$, $\Sigma\, \alpha_\ell = 1$, $\mu_\ell \epsilon R^n$, and Σ_ℓ is a real n x n positive definite symmetric matrix. Whenever a density is evaluated using estimates of its parameters, we denote it, e.g., by $\hat{p}(x_1, \ldots, x_p)$. By a <u>maximum likelihood estimate</u> (MLE) of the parameters $\{(\alpha_\ell, \mu_\ell, \Sigma_\ell)\}$ we mean an element $\{(\hat{\alpha}_\ell, \hat{\mu}_\ell, \hat{\Sigma}_\ell) | \ell=1, \ldots, m\}$ of the parameter set which locally maximizes $\hat{p}(X_1, \ldots, X_p)$. By arguments similar to those used in [2], the following necessary conditions for a MLE are derived.

1) $$\frac{1}{p} \sum_{j=1}^{p} \frac{\hat{p}_\ell(X_j)}{\hat{p}(X_j)} \leq 1 \qquad \text{with equality when } \alpha_\ell > 0$$

2) $$\hat{\mu}_\ell = \sum_{j=1}^{p} \frac{N_j \hat{p}_\ell(X_j)}{\hat{p}(X_j)} \overline{X}_j \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{p}_\ell(X_j)}{\hat{p}(X_j)}$$

3) $$\hat{\Sigma}_\ell = \sum_{j=1}^{p} \frac{\hat{p}_\ell(X_j)}{\hat{p}(X_j)} \sum_{k=1}^{N_j} (X_{jk}-\hat{\mu}_\ell)(X_{jk}\ \hat{\mu}_\ell)^T \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{p}_\ell(X_j)}{\hat{p}(X_j)}$$

In equation (2) $\overline{X}_j = \frac{1}{N_j} \sum_{k=1}^{N_j} X_{jk}$ is the mean of the jth field observations.

By multiplying (1) by $\hat{\alpha}_\ell$ we obtain

4) $$\hat{\alpha}_\ell = \frac{1}{p} \sum_{j=1}^{p} \frac{\alpha_\ell \hat{p}_\ell(X_j)}{\hat{p}(X_j)}$$

which, together with (2) and (3) suggests an iterative procedure for solution of the likelihood equations (2) - (4) analogous to that used in UHMLE [2]. However, the likelihood equations can be considerably simplified by observing that the sequence $(\overline{X}_1, S_1), \ldots, (\overline{X}_p, S_p)$, is a sufficient statistic for the model, where S_j is the sample scatter matrix of the jth field:

$$S_j = \sum_{k=1}^{N_j} (X_{jk} - \overline{X}_j)(X_{jk} - \overline{X}_j)^T.$$

Equation (3) may be rewritten

$$5) \qquad \hat{\Sigma}_\ell = \sum_{j=1}^{p} \frac{\hat{p}_\ell(X_j)}{\hat{p}(X_j)} S_j \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{p}_\ell(X_j)}{\hat{p}(X_j)}$$

$$+ \sum_{j=1}^{p} \frac{N_j \hat{p}_\ell(X_j)}{\hat{p}(X_j)} (\overline{X}_j - \hat{\mu}_\ell)(\overline{X}_j - \hat{\mu}_\ell)^T \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{p}_\ell(X_j)}{\hat{p}(X_j)}$$

The sufficiency of $\{(\overline{X}_j, S_j)\}_{j=1}^{p}$ implies that

$$\frac{\hat{p}_\ell(X_j)}{\hat{p}(X_j)} = \frac{\hat{q}_\ell(\overline{X}_j, S_j)}{\hat{q}(\overline{X}_j, S_j)}$$

where $\hat{q}_\ell(\overline{X}_j, S_j)$ is the estimated joint density of $\overline{X}_j$ and S_j given that F_j represents class ℓ and $\hat{q}(\overline{X}_j, S_j) = \sum_{\ell=1}^{m} \hat{\alpha}_\ell \hat{q}_\ell(\overline{X}_j, S_j)$. The joint density

$\hat{q}_\ell(\bar{X}_j, S_j)$ may be expressed as

$$q_\ell(\bar{X}_j, S_j) = N_n\left(\bar{X}_j; \hat{\mu}_\ell, \frac{1}{N_j}\hat{\Sigma}_\ell\right) W_n(S_j; N_j-1, \hat{\Sigma}_\ell)$$

where $N_n(\bar{X}_j; \hat{\mu}_\ell, \frac{1}{N_j}\hat{\Sigma}_\ell)$ is the n-variate normal density of $\bar{X}_j$ and $W_n(S_j; N_{j-1}, \hat{\Sigma}_\ell)$ is the Wishart density of S_j with N_j-1 degrees of freedom [3]. Thus the likelihood equations may be written as

$$6) \qquad \hat{\alpha}_\ell = \frac{1}{p}\sum_{j=1}^{p} \frac{\hat{\alpha}_\ell \hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)}$$

$$7) \qquad \hat{\mu}_\ell = \sum_{j=1}^{p} \frac{N_j \hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)} \bar{X}_j \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)}$$

$$8) \qquad \hat{\Sigma}_\ell = \sum_{j=1}^{p} \frac{\hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)} S_j \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)}$$

$$+ \sum_{j=1}^{p} \frac{N_j \hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)} (\bar{X}_j - \hat{\mu}_\ell)(\bar{X}_j - \hat{\mu}_\ell)^T \Bigg/ \sum_{j=1}^{p} \frac{N_j \hat{q}_\ell(\bar{X}_j, S_j)}{\hat{q}(\bar{X}_j, S_j)}$$

Equations (6) - (8) are to be used as the basis of the iteration procedure. Indeed when each $N_j = 1$ they reduce to the likelihood equations employed in UHMLE.

4. Concluding Remarks.

The questions of the existence of a consistent MLE as $p \to \infty$ and the local convergence of the iterative procedure will be addressed in a future report. We remark that the standard consistency results of Cramer, Chanda, and Wald (see [2] for references) are not directly applicable since the $(\bar{X}_j, S_j)$ are not identically distributed. Numerical results will also be reported at a later date.

REFERENCES

1. J. Bryant, "On the clustering of multidimensional pictorial data", to appear in Pattern Recognition.

2. B. C. Peters, Jr. 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.

3. S. S. Wilks, Mathematical Statistics, Wiley and Sons, New York, 1962.

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 $\alpha_\ell^o = \mathrm{Prob}[\Theta=\ell] > 0$. 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 $x_{j1}, \cdots, x_{jNj}$.

$$(3) \qquad \prod_{k=1}^{N_j} N_n(x_{jk};\ \mu_\ell,\ \Sigma_\ell) = (2\pi)^{-\frac{nNj}{2}}\, q_j(x_j;\ \mu_\ell,\ \Sigma_\ell)$$

where

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

Let

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

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

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

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

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

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

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

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

2. The General Theorem

Let Θ be an open subset of R^2 and let $\psi^0 \in \Theta$. Suppose $X_1, X_2, \cdots$, is a sequence of independent random vectors with X_r having N_r-variate density function $q_r(\cdot|\psi^0)$ 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^0 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) $\dfrac{\partial q_r(x|\psi)}{\partial\psi_i}$; $\dfrac{\partial^2 q_r(x|\psi)}{\partial\psi_i\partial\psi_j}$; and $\dfrac{\partial^3 \log q_r(x|\psi)}{\partial\psi_i\partial\psi_j\partial\psi_k}$ exist and satisfy:

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

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

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

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

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

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

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

and

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

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

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

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 $p_o \epsilon$ so that for $p > p_o$

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

Without loss of generality we can assume Ω is convex.

Thus, for $\psi \epsilon \Omega$,

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

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

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

With probability 1, for large p

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

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 $\bar{M}$ such that with probability 1 there is a $p_1 \epsilon$ such that for all $p \ge p_1$, and $\psi \epsilon \Omega$,

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

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

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

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

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

and

(iv) $D_\tau^2 \kappa(\tau) = \operatorname{cov}_\tau(F) = \int_\Omega \langle F - E_\tau(F)|\cdot\rangle^2\, f(x|\tau)\, d\lambda(x)$, 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 $$\psi = (\alpha_1,\dots,\alpha_{m-1},\tau_1,\dots,\tau_m)$$
$$x = (x_1,\dots,x_p)$$

(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).

Theorem 2: 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.

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

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

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

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

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

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

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 $\psi = \psi^o$. 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

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

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

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

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

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

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

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

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

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

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.