NASA CR-
141478

(NASA-CR-141478) RECURSIVE ESTIMATION OF PRIOR PROBABILITIES USING THE MIXTURE APPROACH (Rice Univ.) 62 p HC $4.25

N75-15387

CSCL 12A

Unclas

G3/65 06655

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

Recursive Estimation of Prior Probabilities Using the Mixture Approach

by

Demetrios Kazakos
ICSA
Rice University

ABSTRACT

In the present work, we consider the problem of estimating the prior probabilities q_k of a mixture of known density functions $f_k(X)$, based on a sequence of N statistically independent observations.

The mixture density is:

$$g(X|Q) = \sum_{k=1}^{M} q_k f_k(X)$$

It is shown that for very mild restrictions on $f_k(X)$, the maximum likelihood estimate of Q is asymptotically efficient.

However, it is difficult to implement. Hence, a recursive algorithm for estimating Q is proposed, analyzed, and optimized.

For the M=2 case, it is possible for the recursive algorithm to achieve the same performance with the Maximum Likelihood one.

For M>2, slightly inferior performance is the price for having a recursive algorithm. However, the loss is computable and tolerable.

Institute for Computer Services & Applications
Rice University
Houston, Texas

September, 1974

Research supported by NASA contract NAS 9-12776

Introduction :

In many pattern classification problems, the probability density function of each class is known accurately, while the prior probabilities of the classes are unknown.

There are instances where the estimation of prior probabilities from unclassified observations is the ultimate purpose of the data processing. This situation occurs in machine processing of remotely sensed Earth Resources data.

The probability density functions of the spectral signatures of the several crops are known, defined in the multidimensional observation space. The objective is the accurate estimation of the proportions of the crops in a given area.

In Section I, the general problem of joint classification of a set of observations and estimation of prior probabilities is formulated. In a related work by the author, [4] the problem of simultaneous optimal classification and recursive estimation of the prior probabilities has been considered. Here, the assumption is that we do not care about the individual classification of each observation, but we are only interested in a good estimate of the prior probabilities.

The method proposed in the present work has the advantages of being recursive in nature, of guaranteed fast convergence of the error variance at a rate that can be computed, achieving the Rao-Cramér lower bound in the two class case.

We are imposing only certain mild constraints to the probability density functions.

I. Likelihood Function

Let $X^N = (X_1 \ldots X_N)$ be a sequence of statistically independent observations.

Each observation $X_i \in E^n$ is distributed according to $f_k(X_i)$, under hypothesis H_k, $k=1, \ldots, M$. The probability density functions $f_k(X)$, $k=1, \ldots, M$ are assumed continuous and positive for every $X \in E^n$.

Let

$$K_i^j = \begin{cases} 1 & \text{if } X_i \in H_j \\ 0 & \text{if } X_i \notin H_j \end{cases}$$

Let

$$K_i = \left(K_i^1 \; K_i^2 \ldots \; K_i^M\right)$$

Then K_i is an M-vector with $M-1$ zeros and a 1 in the j^{th} position if $X_i \in H_j$. Thus K_i indicates the class membership of X_i.

Let

$$K^N = (K_1 \ldots K_N)^T$$

Then K^N is an $N \times M$ matrix, with columns K_i^T. It indicates the class memberships of the observations $(X_1 \ldots X_N)$

Let $\pi = (\pi_1 \ldots \pi_M)^T$ be the vector of prior probabilities of the M classes.

We are interested in determining the conditional likelihood function

$$P\left(X^N, K^N \middle| \pi\right)$$

We have, by the Bayes rule

$$P\left(X^N, K^N \middle| \pi\right) = P\left(X^N \middle| K^N, \pi\right) \; P\left(K^N \middle| \pi\right) =$$

$$= P\left(X^N \middle| K^N\right) \; P\left(K^N \middle| \pi\right)$$

The above conditional probability density functions are :

$$P(X^N \mid K^N) = \prod_{i=1}^{N} \prod_{s=1}^{M} \left[f_s(X_i) \right]^{K_i^s}$$

$$P(X^N \mid \pi) = \prod_{i=1}^{N} \prod_{s=1}^{M} \pi_s^{K_i^s}$$

Substituting, we have :

$$P(X^N, K^N \mid \pi) = \prod_{i=1}^{N} \prod_{s=1}^{M} \left[\pi_s \, f_s \, (X_i) \right]^{K_i^s}$$

In general, both K^N and π may be unknown.

It is interesting to note that the pair (K^N, π) that maximizes $P(X^N, K^N \mid \pi)$ has the following intuitively nice properties.

For known π, the value $K^N = \hat{K}^N$ that maximizes $P(X^N, K^N \mid \pi)$ reduces to the Bayes classifier, i.e.

$$\hat{K}_i^j = \begin{cases} 1 & \text{if } \pi_j \, f_j \, (X_i) = \max_m \pi_m \, f_m \, (X_i) \\ 0 & \text{otherwise} \end{cases}$$

For known K^N, the value $\pi = \hat{\pi}$ that maximizes $P(X^N, K^N \mid \pi)$ is the relative frequency estimate, i.e.

$$\hat{\pi}_s = N^{-1} \sum_{i=1}^{N} K_i^s$$

Hence the estimate

$$(\hat{K}^N, \hat{\pi}) = \arg\max P(X^N, K^N \mid \pi)$$

is intuitively appealing but complicated to realize.

In the present work, we are not interested in estimating K^N. We are only interested in estimating π. If K^N is known, the relative frequency estimate is unbiased :

$$E\,\hat{\pi}_s = N^{-1} \sum_{i=1}^{N} E\,K_i^s = \pi_s$$

The error covariance matrix has elements

$$E\left(\pi_s - \hat{\pi}_s\right)\left(\pi_j - \hat{\pi}_j\right) = \begin{cases} N^{-1}\,\pi_s(1-\pi_s) & \text{for } s=j \\ N^{-1}\,(-\pi_s\,\pi_j) & \text{for } s\neq j \end{cases}$$

Since perfect classification (knowledge of K^N) is an ideal situation for estimating the priors, the above error covariance matrix is a "lower bound" to the achievable error variance in estimating π under unknown K^N.

II. Mixture Approach--2 class case

If we average the conditional p.d.f. $P(X^N, K^N \mid \pi)$ over K^N, the result is :

$$P(X^N \mid \pi) = \sum_{K^N} P(X^N, K^N \mid \pi) =$$

$$= \prod_{m=1}^{N} \left\{ \sum_{s=1}^{M} \pi_s\, f_s(X_m) \right\}$$

We are interested in finding the value of π that will maximize the conditional likelihood function

$$P(X^N \mid \pi)$$

Let

$$g(X \mid \pi) = \sum_{s=1}^{M} \pi_s \, f_s \, (X)$$

The function $g(X \mid \pi)$ is linear in the unknown parameters

$\pi = (\pi_1, \ldots, \pi_M)$

In the present section, we will concentrate on the $M=2$ class case. In this case, the parameter π is one dimensional.

$$g(X \mid \pi) = \pi \, f_1(X) + (1-\pi) \, f_2(X)$$

We make the following assumptions on f_1, f_2:

Assumption 1 :

$f_1(X)$, $f_2(X)$ are continuous and nonzero for all $X \in E^n$

Assumption 2 :

The mixture $g(X \mid \pi)$ is identifiable in the usual sense [5]. That is :

$$\text{if } g(X \mid \pi_1) \equiv g(X \mid \pi_2) \qquad \forall X \in E^n,$$

then $\pi_1 = \pi_2$

Comment :

H has been shown that most of the usual probability density

functions make identifiable mixtures. In [5], there is a list of such p.d.f's.

Because of the convenient form of the function $g(X \mid \pi)$, we are able to use a theorem due to Cramér [6], regarding the behavior of the maximum likelihood estimate, $\hat{q}_N$ where

$$\hat{q}_N = \arg \max_{\pi} P(X^N \mid \pi)$$

In general, the function

$$\ell(X^N, \pi) = \log P(X^N \mid \pi)$$

has a number of local maxima.

The local maxima π_K are solutions of the likelihood equation :

$$\frac{\partial}{\partial \pi} \log P(X^N \mid \pi) = 0$$

The original version of the theorem requires the satisfaction of Conditions 1 - 5, due to Cramér [6].

If Conditions 1 - 5 are satisfied, any solution of the likelihood equation will be a "good" estimate, in a sense to be defined.

For numerical solution of the likelihood equation, it would make things easier if we knew that the likelihood equation has a unique solution. Conditions 6 - 7 due to Perlman [7], guarantee that for large enough N, and with probability 1, we will have a unique solution of the likelihood equation.

The conditions that must be satisfied, are :

Condition 1 :

For almost all $X \epsilon E^n$,

$$\exists \frac{\partial^i}{\partial q^i} \log g(X \mid q) \quad , \quad i=1,2,3 \qquad q \epsilon [0,1]$$

Condition 2 :

$$E \frac{\partial}{\partial q} \log g(X \mid q) \Big|_{q=\pi} = 0$$

where π = true value of the prior probability.

Condition 3 :

$$J(\pi) = E \left(\frac{\partial}{\partial q} \log g(X \mid q) \right)^2 \Bigg|_{q=\pi} < +\infty$$

Condition 4 :

$$E \left. \frac{\partial^2}{\partial q^2} \log g(X \mid q) \right|_{q=\pi} = -J(\pi)$$

Condition 5 :

There exists a function $m(X)$, such that

$$\left| \frac{\partial^3}{\partial q^3} \log g(X \mid q) \right| < m(X), \quad \forall q \in [0,1]$$

and $m(X)$ is finite

Condition 6 :

The Kullback-Leibler information number

$$I(q,\pi) = \int_{E^n} g(X \mid \pi) \log \left[\frac{g(X \mid \pi)}{g(X \mid q)} \right] dx$$

achieves a unique minimum at $q=\pi$.

Condition 7 :

$\frac{\partial}{\partial q} \log g(X \mid q)$ is continuous in q for each $q \in [0,1]$, uniformly in X.

Theorem :

Under the regularity Conditions 1-7, the maximum likelihood estimate

$$\hat{P}_N = \arg \max_{q} \prod_{m=1}^{N} g(X_m \mid q)$$

is weakly consistent, i.e.

$$\lim_{N \to \infty} \hat{P}_N = \pi \quad \text{in probability}$$

Furthermore, the estimate $\hat{P}_N$ is asymptotically efficient, i.e., it achieves the Rao-Cramér lower bound:

$$E\left(P_N - \pi\right)^2 \longrightarrow N^{-1} \left[J(\pi)\right]^{-1}$$

Also, with probability 1 there exists an N_o, such that for all $N > N_o$ the likelihood equation has a unique solution in the region $\pi \in [0, 1]$.

——— · ——— · ———

Intuitively speaking, the theorem says that for N "large enough," we will have in $[0,1]$ a unique solution of the likelihood equation. Hence, if N_o is known, we can use an efficient numerical method specifically designed to seek the unique zero of a function.

For the particular problem considered here, we have

$$J(\pi) = \int_{E^n} \left[f_1(X) - f_2(X)\right]^2 \left[\pi f_1(X) + (1-\pi) f_2(X)\right]^{-1} dx$$

In Appendix I, it is shown that Assumption 1 implies that $J(\pi)$ is upper bounded by $[\pi(1-\pi)]^{-1}$.
Hence, for $\pi \neq 0, 1$, $J(\pi)$ is finite. The physical significance of this bound is the following.
The quantity $N^{-1}\pi(1-\pi)$ is the variance of the relative frequency estimate in the case of observations of known classification.
Hence the inequality

$$N^{-1}\left[J(\pi)\right]^{-1} \geq N^{-1}\pi(1-\pi)$$

is natural. It means that the Rao-Cramér lower bound (left hand expression) is higher than the variance of the relative frequency estimate. We have to accept the higher error variance due to the fact that the observed data are unclassified.

In Appendix I, it is also shown that the function

$$A(\pi) = \left[J(\pi)\right]^{-1}$$

is concave in the region $[0,1]$

In such a case, we assume that we know that π lies in an interval $I(\epsilon)$, where

$$I(\epsilon) = \begin{cases} [0,1] & \text{if } J(0) < +\infty,\ J(1) < +\infty \\ [\epsilon,1] & \text{if } J(0) = +\infty,\ J(1) < +\infty \\ [0,1-\epsilon] & \text{if } J(0) < +\infty,\ J(1) = +\infty \\ [\epsilon,1-\epsilon] & \text{if } J(0) = J(1) = +\infty \end{cases}$$

and ϵ is a small positive number. The Conditions 1-7 have to be valid for $\pi \epsilon I(\epsilon)$ in order for the theorem to apply.

In Appendix I, an efficient method for computing $J(\pi)$ in the case of Gaussian densities is demonstrated.

In Appendix II, it is shown that Assumptions 1-2 imply the satisfaction of Conditions 1-7.

Hence, the Maximum Likelihood estimate of π is an efficient method in terms of performance.

The implementation of the estimate requires finding the maximum of the likelihood function, which is an N^{th} degree polynomial. For large N, we cannot afford the computational complexity of the above scheme. Furthermore, the M.L. estimate is non-recursive. We cannot update it efficiently.

We will now consider a recursive estimate of the mixture parameter π. The basic observation is that the value $q = \pi$ minimizes the Kullback-Leibler information number $I(q, \pi)$, and the minimum is unique.

The derivative of $I(q, \pi)$ is :

$$\frac{\partial}{\partial q} I(q, \pi) = -\int_{E^n} g(X \mid \pi) \left[\frac{\partial}{\partial q} \log g(X \mid q) \right] dx =$$

$$= -E \left\{ \frac{\partial}{\partial q} \log g(X \mid q) \;\middle|\; \pi \right\}$$

Hence, the estimate of the gradient of $I(q, \pi)$, for a fixed q and based on one observation X, is :

$$- \frac{\partial}{\partial q} \log g(X \mid q)$$

Motivated by the above observation, we consider the following sequential estimation algorithm :

$$P_{N+1} = P_N + N^{-1} \cdot L(P_N) \; G(X_{N+1}, P_N)$$

where G is the current estimate of the gradient :

$$G(X_{N+1}, q) = \frac{\partial}{\partial q} \log g(X_{N+1} \mid q) =$$

$$= \left[f_1 (X_{N+1}) - f_2 (X_{N+1}) \right] \cdot$$

$$\cdot \left[q \, f_1 (X_{N+1}) + (1-q) \, f_2 (X_{N+1}) \right]$$

and $L(P)$ is a bounded positive function, defined for $P \in [0, 1]$.

$L(P)$ will be chosen later for optimal convergence of the algorithm.

We define the regression function $M(q)$, for $q \in [0, 1]$.

$$M(q) = E \left[L(q) \; G(X, q) \right]$$

$$= L(q) \; F(q)$$

where

$$F(q) = \int_{E^n} \left[f_1(X) - f_2(X)\right] \left[q\, f_1(X) + (1-q)\, f_2(X)\right]^{-1} \cdot$$

$$\cdot \left[\pi f_1(X) + (1-\pi)\, f_2(X)\right] dx$$

The derivative of $F(q)$ is :

$$F'(q) = - \int_{E^n} \left[f_1(X) - f_2(X)\right]^2 \left[q\, f_1(X) + (1-q)\, f_2(X)\right.$$

$$\cdot \left[\pi f_1(X) + (1-\pi)\, f_2(X)\right] dx$$

Hence,

$$F'(q) < 0 \qquad \forall q \in [0,1]$$

Also, we note that

$$F(\pi) = 0$$

$$M(\pi) = 0$$

Therefore, the function $F(q)$ is monotone decreasing in $[0,1]$ and it has a unique zero for $q = \pi$

Let

$$Z(X,q) = G(X,q)\, L(q) + M(q)$$

Obviously, the random variable $Z(X,q)$ has zero mean, conditioned on q

$$E\left[Z(X,q) \mid q\right] = 0$$

To guard against getting an estimate P_{N+1} that is outside of the interval $[a,b]$, I put two reflecting barriers at a and b.

The recursive algorithm then becomes :

$$P'_{N+1} = P_N + N^{-1}\left[Z(X_{N+1}, P_N) - M(P_N)\right]$$

$$P_{N+1} = R(P'_{N+1})$$

The function $R(X)$ truncates to the extreme points of $I(\epsilon)$ any estimate that falls outside.

If $I(\epsilon) = [a, b]$

$$R(X) = \begin{cases} b & \text{if } X \geq b \\ X & \text{if } X \in [a, b] \\ a & \text{if } X \leq a \end{cases}$$

This is standard procedure in algorithms of this type.

For the convergence properties of the above sequential procedure, we now invoke a theorem due to J. Sacks [8]. The conditions of the theorem are expressed for convenience in the notation of the present paper.

They involve the regression function $M(q)$ and the sequence of zero mean, "noisy" observables $\{ Z(X_N, q) \}$.

Condition 1a :

$$M(\pi) = 0$$

and $(q-\pi)\, M(q) < 0$ for all $q \in I(\epsilon)$, $q \neq \pi$

Condition 2a :

For all $q \in I(\epsilon)$ and some positive constant K_1, $|M(q)| \leq K_1 |q - \pi|$, and for every t_1, t_2 such that $0 < t_1 < t_2 < \infty$, $\inf |M(q)| > 0$, where the inf is taken for $t_1 \leq |q-\pi| \leq t_2$, $q \in I(\epsilon)$.

Condition 3a :

For all $q \in I(e)$

$$M(q) = a_1 (q-\pi) + \delta(q,\pi)$$

where $\delta(q,\pi) = 0(|q-\pi|)$ as $|q-\pi| \to 0$

and where $a_1 < 0$.

Condition 4a :

a) $$\sup_{q \in I(e)} E\left[Z^2(X,q) \mid q\right] < \infty$$

b) $$\lim_{q \to \pi} E\left[Z^2(X,q) \mid q\right] = S(\pi)$$

Condition 5a :

(The version of this condition is stronger than necessary, but it is easier to verify for our particular case).

For a fixed value of q , the random variables $\left\{ Z(X_N, q) \right\}_N$ are identically distributed.

Theorem :

(Sacks) Suppose that Conditions 1-5 are satisfied, and assume in addition that $|a_1| > \frac{1}{2}$. Then $N^{\frac{1}{2}} (P_N - \pi)$ is asymptotically normally distributed with mean 0 and variance

$$S(\pi) \left[2 |a_1| - 1 \right]^{-1}$$

In order to satisfy the Conditions 1a - 6a , we constrain the function $L(q)$ to be positive and bounded :

$$0 < C_1 \leq L(q) \leq C_2 < + \infty$$

Then,

$$(q-\pi)\ M(q) = L(q)\ (q-\pi)\ F(q) < 0$$

$$\forall q \neq \pi\ , \qquad q \in I(\epsilon)$$

because the product $(q-\pi)\ F(q)$ is negative for all $q \neq \pi$.

In Appendix III, it is shown that Assumptions 1-2 imply satisfaction of Conditions 1a - 6a .

It is also shown that the constants a_1 and $S(\pi)$ of the theorem are :

$$S(\pi) = L^2(\pi)\ J(\pi)$$

$$a_1 = -\ L(\pi)\ J(\pi) = L(\pi)\ F'(\pi)$$

because :

$$F'(\pi) = -\ J(\pi)$$

We are now able to express the asymptotic error variance of the algorithm in terms of $L(\pi)$, $J(\pi)$ and under the condition

$$2\ |a_1| = L(\pi)\ J(\pi) > 1$$

The variance is :

$$NE\ (P_N - \pi)^2 \longrightarrow J(\pi)\ L^2(\pi)\ \left[\ 2L(\pi)\ J(\pi) - 1\ \right]^{-1}$$

(If the condition $2\ |a_1| > 1$ is not satisfied, Sakrison [ ] has commented that the convergence rate may be slower than N^{-1}).

For a fixed value of π , we have in Fig. 1, the variance

$$V = J(\pi)\ L^2(\pi)\ \left[\ 2L(\pi)\ J(\pi) - 1\ \right]^{-1}$$

as a function of $L = L(\pi)$

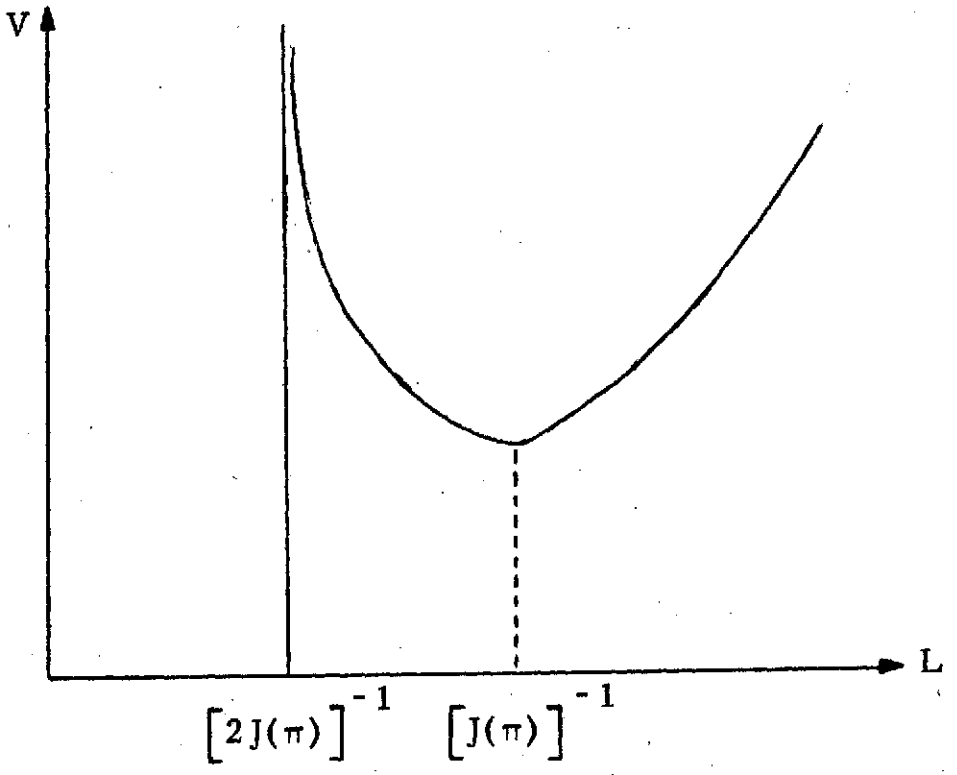

Fig. 1

For $L(\pi) > [2J(\pi)]^{-1}$, the variance V has a global minimum, achievable at

$$L = J^{-1}$$

Hence, we can optimize the nonlinear function L by choosing

$$L(\pi) = [J(\pi)]^{-1}, \quad \pi \in I(e)$$

Substituting the optimum $L(\pi)$ into the variance expression, we find that the resulting minimum asymptotic variance is :

$$E\ (P_N - \pi)^2 \longrightarrow N^{-1}[J(\pi)]^{-1}$$

But this is exactly the Rao-Cramér lower bound, i.e., the sequential procedure is asymptotically efficient.

In other words, if we agree that the mixture approach should be followed, the sequential algorithm presented will perform as well as anything else in estimating π.

The maximum likelihood estimation scheme requires tremendous complexity in order to achieve the Rao-Cramér bound, while the presented sequential scheme is very simple and achieves the same lower bound.

The only difficulty in the implementation, lies in the construction of the nonlinear function $L(\pi)$.

However, it is a one-shot construction, so we can do it off-line. In situations where we have to estimate prior probabilities repeatedly, while the probability density functions remain unchanged, the scheme is increasingly attractive.

In Appendix I, an efficient method for constructing $J(\pi)$ (hence $L(\pi)$) is presented for the case of multivariate Gaussian densities.

III. Mixture Approach : M > 2 Class Case

We now assume that each observation vector $X_K \in E^n$ comes from one of M statistical populations-hypotheses.

Under hypothesis H_m, X_K is distributed according to the p.d.f. $f_m(X_K)$. Let π_m be the prior probability of hypothesis H_m.

We need to estimate only $M-1$ of the prior probabilities (π_m).

Let $\pi = \left[\pi_1 \ \cdots \ \pi_{M-1}\right]^T$ be the vector of true prior probabilities,

and $Q = \left[q_1 \ \cdots \ q_{M-1}\right]^T$ be a vector of arbitrary prior probabilities

Let $g(X \mid Q)$ designate the mixture density :

$$g(X \mid Q) = \sum_{s=1}^{M-1} q_s \, f_s(X) + \left[1 - \sum_{s=1}^{M-1} q_s\right] f_M(X)$$

The likelihood function of a sequence of N independent observations is :

$$\prod_{m=1}^{N} g(X_m \mid Q)$$

We will investigate now the performance of the maximum likelihood estimate of π, based on a sequence on N observations.

The M.L. estimate $\hat{Q}_N$ is determined by the equation :

$$Q_N = \arg \max_{Q \in I_M} \prod_{m=1}^{N} g(X_m \mid Q)$$

where

$$I_M = \{Q;\ Q = (q_1 \cdots q_{M-1}),\ q_s \geq 0,\ s=1,\ldots,M-1,\ \sum_{s=1}^{M-1} q_s \leq 1\}$$

We will make two mild assumptions about the densities $f_m(X)$, similar to the ones for the $M=2$ case.

Assumption 1' :

$f_K(X)$, $K=1,\ldots,M$ are continuous and nonzero for all $X \in E^n$.

Assumption 2' :

The densities $f_K(X)$, $K=1,\ldots,M$ make an identifiable mixture $g(X \mid Q)$.

For assessing the properties of the maximum likelihood estimate, we will use the multidimensional version of the theorem used in Section II. The parameter space now is $M-1$ dimensional.

The Conditions 1′- 5′of the following theorem are due to Cramér, [6] and Conditions 6′- 7′are due to Perlman [7]. The last two Conditions guarantee that for N "large enough," the likelihood equation will have a unique solution in I_M (region of interest).

Condition 1′:

For all $X \in E^n$, the derivatives

$$\frac{\partial^{i+j}}{\partial q_s^i \, \partial q_m^j} \log g(X \mid Q) \quad , \quad s, m = 1, \ldots, M-1$$

exist for all $Q \in I_M$ and $i, j = 1, 2, 3$

Condition 2′:

$$E \frac{\partial}{\partial q_s} \log g(X \mid Q) \Bigg|_{q=\pi} = 0$$

for $s = 1, \ldots, M-1$

where π = true value of the prior probability.

Condition 3′:

$$J_{sK}(\pi) = E \left[\frac{\partial}{\partial q_s} g(X \mid Q) \frac{\partial}{\partial q_K} g(X \mid Q) \right] \Bigg|_{Q=\pi} < \infty$$

for $s, K = 1, \ldots, M-1$

Condition 4′:

$$E \frac{\partial^2}{\partial q_s \, \partial q_K} \log g(X \mid Q) \Bigg|_{Q=\pi} = - J_{sK}(\pi)$$

for $s, K = 1, \ldots, M-1$

Condition 5′ :

There exists a function $m(X)$, such that

$$\left| \frac{\partial^{i+j}}{\partial q_s^i \, \partial q_K^j} \log g(X \mid Q) \right| < m(X) \qquad \forall Q \in I_M$$

for $i, j = 1, 2, 3$, $s, K = 1, \ldots, M-1$

and $m(X)$ is finite, except on a set of probability zero.

Condition 6′ :

The Kullback-Leibler information number

$$I(Q, \pi) = \int_{E^n} g(X \mid \pi) \log \left[\frac{g(X \mid \pi)}{g(X \mid Q)} \right] dx$$

achieves a unique minimum at $Q=\pi$

Condition 7′ :

$\frac{\partial}{\partial q_s} \log g(X \mid Q)$ is continuous at each

$Q \in I_M$, $s = 1, \ldots, M-1$, uniformly in X.

Theorem :

Under the regularity Conditions 1′ - 7′, the maximum likelihood estimate

$$\hat{Q}_N = \arg \max_Q \prod_{m=1}^{N} g(X_m \mid Q)$$

is weakly consistent, i.e.

$$\lim_{N \to \infty} \hat{Q}_N = \pi \quad \text{in probability}$$

Furthermore, the Maximum Likelihood estimate $\hat{Q}_N$ is asymptotically efficient, achieving the Rao-Cramér lower bound.

Also, with probability 1, there exists an N_o, such that for all $N > N_o$, the likelihood equation has a unique solution $\pi^\circ = (\pi_1^\circ \ \ldots \ \pi_{M-1}^\circ)$, in in the region

$$I = \left\{ \pi \ ; \ 0 < \pi_i < 1, \quad i=1,\ldots,M-1, \ \sum_{k=1}^{M-1} \pi_K < 1 \right\}$$

Let

$$R_N(\pi) = E \left(\hat{Q}_N - \pi\right) \left(\hat{Q}_N - \pi\right)^T$$

be the error covariance matrix.

Let $A = (a_1 \ \ldots \ a_{M-1})^T$ be any weighting vector with nonzero norm.

Then the above property stated in the theorem can be expressed as :

$$\lim_{N \to \infty} N \left[A^T R_N^{-1}(\pi) A \right]^{-1} = \left[E \left[A^T \nabla \log g(X \mid \pi) \right]^2 \right]^{-1}$$

——— . ——— . ———

Hence, the maximum likelihood estimator $\hat{Q}_N$ performs better than any estimate.

In Appendix IV, an upper bound to the function $J_{sK}(\pi)$ is found.

The bound is :

$$\left| J_{sK}(\pi) \right| \leq \pi_M^{-1} (\pi_K \ \pi_s)^{-\frac{1}{2}} \left[(\pi_K + \pi_M) (\pi_s + \pi_M) \right]^{3/2}$$

where $\pi_M = 1 - \sum_{K=1}^{M-1} \pi_K$

This bound is finite for

$$\pi_s, \ \pi_K, \ \pi_M \neq 0$$

With arguments similar to those for the $M=2$ case, it can be easily shown that Assumptions $1' - 2'$ imply the satisfaction of Conditions $1' - 7'$. The conclusion is that the maximum likelihood estimate of π "works" for the mixture model.

The implementation of the maximum likelihood estimate of π is numerically difficult. With increasing number of observations, N, the computational complexity of the M.L. estimator increases tremendously.

Motivated by the difficulty in implementation, we will now propose and analyze a recursive estimation procedure.

The intuitive basis is the minization of the functional $I(Q, \pi)$.

$$I(Q, \pi) = E \left\{ \log \left[g(X \mid \pi) \left(g(X \mid Q) \right)^{-1} \right] \middle| \pi \right\}$$

The gradient of I with respect to Q, is:

$$\nabla I(Q, \pi) = E \left\{ \nabla \log \left[g(X \mid \pi) \left(g(X \mid Q) \right)^{-1} \right] \middle| \pi \right\} =$$

$$= - E \left[\nabla \log g(X \mid Q) \middle| \pi \right]$$

Therefore, an estimate of the gradient of $I(Q, \pi)$, based on one observation, X, is the vector

$$\nabla \log g(X \mid Q) = \left[g(X \mid Q) \right]^{-1} \left[f_1(X) - f_M(X), \ldots, \right.$$

$$\left. f_{M-1}(X) - f_M(X) \right]^T$$

This observation motivates the following gradient algorithm for recursive estimation of π.

$$Q_{N+1} = Q_N - (N+1)^{-1} L(Q_N) \nabla \log g(X_{N+1} \mid Q_N)$$

Here, $L(Q)$ is a scalar function of Q, positive and bounded between $[C_1, C_2]$.

$$0 < C_1 \leq L(Q) \leq C_2 < +\infty$$

$L(Q)$ will be adjusted later for optimal convergence of the algorithm. In order to examine the convergence properties of the algorithm, we need to define the regression function $M(Q)$.

$M(Q)$ is an $M-1$ dimensional vector function.

$$M(Q) = E \left\{ L(Q) \nabla \log g(X \mid Q) \;\middle|\; Q \right\}$$

After substitution, we have

$$M(Q) = \left[M_1(Q), \ldots, M_{M-1}(Q) \right]^T$$

where

$$M_K(Q) = - L(Q) \int_{E^n} g(X \mid \pi) \left[g(X \mid Q) \right]^{-1} \cdot$$

$$\cdot \left[f_K(X) - f_M(X) \right] dx$$

$$K = 1, \ldots, M-1$$

We note that

$$M_K(\pi) = 0$$

hence

$$M(\pi) = 0$$

We define the random vector

$$Z(X, Q) = L(Q) \nabla \log g(X \mid Q) - M(Q)$$

we have :

$$E\left(Z(X,Q) \mid Q\right) = 0$$

We will define a region $I_M(A)$ in $M-1$ dimensional Enclidian space. Let $A=(a_1 \ldots a_M)$, where a_i are positive numbers, much smaller than 1. We define the region $I_M(A)$ as follows :

$$I_M(A) = \left\{ Q;\ Q = (q_1 \ \cdots \ q_{M-1}),\ q_K \geq a_K \ , \right.$$

$$\left. K=1, \ldots, M-1 \ , \quad a_M \leq 1 - \sum_{K=1}^{M-1} q_K \right\}$$

We are now ready to apply a multidimensional stochastic approximation theorem due to J. Sacks [ ]. The conditions of the theorem are expressed in terms of the function $M(Q)$ and the random variables $Z(X, Q)$.

Condition 1 :

$M(\pi) = 0$, and for every $e > 0$, $\inf (Q-\pi)^T M(Q) > 0$, where the inf is taken over the region :

$$I_M(A) \cap \left\{ Q;\ e^{-1} > \|Q - \pi\| > e \right\}$$

Condition 2 :

There exists a positive constant K_1, such that, for all $Q \in I_M(A)$,

$$\| M(Q) \| \leq K_1 \| Q - \pi \|$$

Condition 3 :

For all $Q \in I_M(A)$,

$$M(Q) = B(Q - \pi) + \delta(Q, \pi)$$

where B is a positive definite $(M-1) \times (M-1)$ matrix, and

$$\| \delta(Q, \pi) \| = 0 (\| Q - \pi \|) \text{ as } Q - \pi \to 0$$

Condition 4 :

$$\sup_{Q \in I_M(A)} E \left\{ \| Z(X, Q) \|^2 \,\middle|\, Q \right\} < + \infty$$

$$\lim_{Q \to \pi} E \left\{ Z(X, Q) \, Z^T(X, Q) \,\middle|\, Q \right\} = S(\pi)$$

where $S(\pi)$ is a nonnegative definite matrix

Condition 5 :

Conditioned on Q, the sequence of random variables $\{ Z(X_N, Q) \}_N$, is identicaly distributed.

Let $b_1, \ldots, b_{M-1}$ be the eigenvalues of B in decreasing order.

Write $B = P B_1 P^{-1}$, where P = orthogonal matrix and

$B_1 = \text{diag} (b_1 \ldots b_{M-1})$

Let $S_{ij}(\pi) = i, j^{th}$ element of $S(\pi)$

and $S^{*}_{ij}(\pi)$ = i,j^{th} element of

$S^{*}(\pi) = P^{-1} S(\pi) P$

Theorem :

Suppose Conditions 1-5 are satisfied.

Assume, further, that $b_{M-1} > \frac{1}{2}$

Then, $N^{\frac{1}{2}}(Q_N - \pi)$ is asymptotically normal, with mean 0 and covariance matrix PFP^{-1}, where F is the matrix whose $(i,j)^{th}$ element is

$$(b_i + b_j - 1)^{-1} \quad S^{*}_{ij}(\pi)$$

——— . ——— . ———

In Appendix V, it is shown that Assumptions 1-2 imply satisfaction of Conditions 1-5 for the region $Q \in I_M(A)$.

Hence the proposed recursive estimation algorithm will converge to the true value π, and the convergence of the error covariance is of the order N^{-1}.

The reason for achieving high speed of convergence is that the stochastic approximation theorem of Sacks was invoked.

It requires more stringent conditions for convergence than Blum's [9] theorem, for example, and the reward is that a unique zero of the regression function is guaranteed, hence we have speedy convergence.

In order to keep the sequence of estimates $\{Q_N\}$ within the region $I_M(A)$, for convergence purposes, we make a slight modification.

The new computed estimate Q^1_{N+1} is :

$$Q^1_{N+1} = Q_N - (N+1)^{-1} L(Q_N) \nabla \log g(X_{N+1} \mid Q_N)$$

We construct Q_{N+1} from Q^1_{N+1} by truncating to the boundaries the coordinates of Q^1_{N+1} that are outside of $I_M(A)$, so that $Q_{N+1} \in I_M(A)$.

In Appendix V, the error covariance matrix is computed. The result is as follows :

Let $D(\pi)$ be an $(M-1) \times (M-1)$ matrix with elements

$$D_{Ks}(\pi) = \int_{E^n} \left[g(X \mid \pi) \right]^{-1} \left[f_K(X) - f_M(X) \right] \cdot$$

$$\cdot \left[f_s(X) - f_M(X) \right] dx$$

Let $d_1 \geq d_2 \geq \ldots \geq d_{M-1}$ be the eigenvalues of $D(\pi)$.

Let

$$D(\pi) = P^{-1} \text{ diag } (d_1 \ldots d_{M-1}) P$$

where P = orthogonal matrix, consisting of the eigenvectors of $D(\pi)$. Then, using the above theorem, it is found in Appendix V that the asymptotic error covariance matrix is :

$$\lim_{N \to \infty} NE (Q_N - \pi) (Q_N - \pi)^T = PFP^{-1}$$

where

$$F(\pi) = L^2(\pi) \text{ diag } \left[d_1 \left(2L(\pi) d_1 - 1 \right)^{-1}, \ldots, \right.$$

$$\left. , d_{M-1} \left(2L(\pi) d_{M-1} \right)^{-1} \right]$$

The motivation for employing the recursive estimate was to achieve a simpler estimate than the Maximum Likelihood one. It is expected that the convenience of having a recursive estimate will be paid in the form of increased error variance.

The question is, how much performance did we sacrifice ?

Furthermore, it seems at a first glance, that it might be possible to recover some of the incurred loss by cleverly choosing the function $L(\pi)$.

In the case $M=2$, the loss was completely recovered, and the Rao-Cramér bound was achieved with the use of the optimal function $L(\pi)$.

We will compare the performance of the following three estimators of π:

A) Maximum Likelihood Estimator

B) Recursive Estimator

C) Relative Frequency Estimator

Actually, Estimator C can be implemented only when the data are observed noiselessly.

This requirement is equivalent to the densities $f_i(X)$ having disjoint support sets.

Therefore, comparison of Estimator C to the others is only an indication of the loss in performance due to noisy data.

Let

$$R^s(\pi) = \lim_{N \to \infty} NE\,(Q_N - \pi)\,(Q_N - \pi)^T$$

be the asymptotic error covariance of the estimator s.

The supercript s will indicate whether we have the A, B, or C estimator.

Let $A=(a_1 \ldots a_{M-1})$ be an arbitrary weighting vector with nonzero norm.

The magnitude of the quantity

$$\left[A^T \left[R^s(\pi) \right]^{-1} A \right]^{-1}$$

is indicative of the "magnitude" of the error covariance matrix. The error covariance matrix of the recursive estimator satisfies the equation:

$$\left[R^B(\pi) \right]^{-1} = P^{-1} F^{-1} P$$

The Maximum Likelihood estimator achieves the Rao-Cramér lower bound, hence:

$$\left[A^T \left[R^A(\pi) \right]^{-1} A \right]^{-1} = \left[E \left[A^T \nabla \log g(X \mid \pi) \right]^2 \right]^{-1}$$

We have :

$$E \left[A^T \nabla \log g(X \mid \pi) \right]^2 =$$

$$= A^T E \left[\nabla \log g(X \mid \pi) \right] \left[\nabla \log g(X \mid \pi) \right]^T A =$$

$$= A^T D(\pi) A$$

$$= A^T P^{-1} \text{ diag } (d_1 \ldots d_{M-1}) P A$$

$$= A^T P^T \text{ diag } (d_1 \ldots d_{M-1}) P A$$

(because $P^{-1} = P^T$)

The matrix $D(\pi)$ is symmetric.

Hence,

$$D(\pi) = D^T(\pi) = \left(P^T \text{ diag } (d_1 \ldots d_{M-1}) P \right)^T$$

$$D(\pi) = P \text{ diag } (d_1 \ldots d_{M-1}) P^T$$

Using the above observations, we have :

$$\left[A^T \left[R^A(\pi) \right]^{-1} A \right]^{-1} = \left[A^T P^T \text{ diag } (d_1 \ldots d_{M-1}) P A \right]^{-1}$$

$$\left[A^T \left[R^B(\pi) \right]^{-1} A \right]^{-1} = \left[A^T P^T F^{-1} P A \right]^{-1}$$

where

$$F^{-1} = \left[L(\pi)\right]^{-2} \text{ diag } \left[\left(2L(\pi)\ d_1 - 1 \right) d_1^{-1} \ , \ldots ,\right.$$

$$\left. , \ldots , \left(2L(\pi)\ d_{M-1} - 1 \right) d_{M-1}^{-1} \right]$$

We note now that each of the terms of F^{-1} is smaller than the corresponding d_K.

Because, the inequality

$$\left(2L(\pi)\ d_K - 1 \right) d_K^{-1} \left[L(\pi)\right]^{-2} \leq d_K$$

is equivalent to :

$$\left(L(\pi)\ d_K - 1 \right)^2 \geq 0$$

Hence, the conclusion is the following inequality :

$$\left[A^T \left[R^B(\pi)\right]^{-1} A \right]^{-1} \geq \left[A^T \left[R^A(\pi)\right]^{-1} A \right]^{-1} \qquad (a)$$

This inequality is true for any weighting vector A.

It expresses the exact loss in performance, asymptotically speaking, when we use the recursive estimator instead of the Maximum Likelihood one.

In Fig. 2, the magnitude, y_K, of the K^{th} diagonal term of F^{-1} is plotted as a function of L.

$$d_K^{-1} \ L^{-2} \ (2L d_K - 1)$$

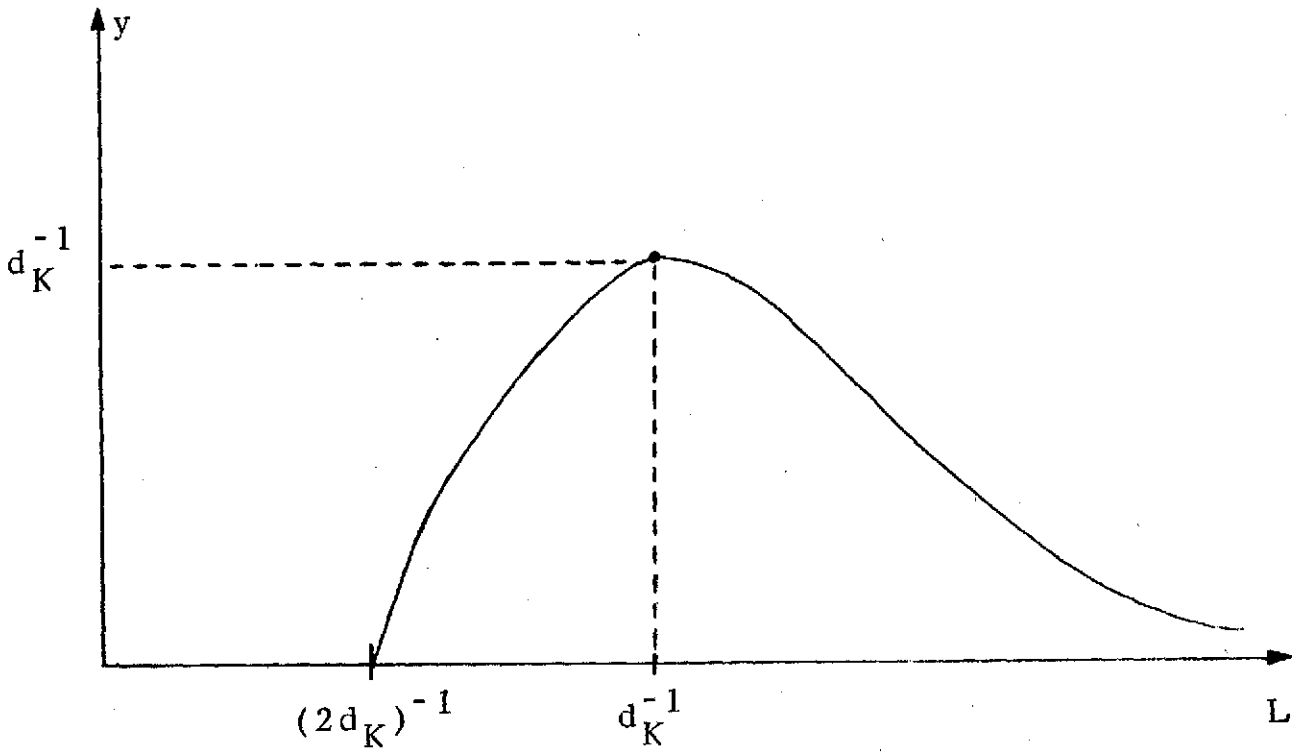

Fig. 2

$y_K(L)$ has a unique global maximum for $L = d_K^{-1}$.

The choice of the function L should be such as to make each y_K as close to its maximum value as possible.

Because then the Rao-Cramér lower bound will be approached as closely as possible.

Obviously, we cannot maximize all y_K simultaneously.

Hence, we choose to maximize their average :

$$T(L) = (M-1)^{-1} \sum_{K=1}^{M-1} y_K(L)$$

We have :

$$T(L) = d^{-1} L^{-2}(2Ld - 1)$$

where

$$d^{-1} = (M-1)^{-1} \sum_{K=1}^{M-1} d_K^{-1}$$

We have :

$$d_1 \geq d \geq d_{M-1}$$

The function $T(L)$ has the same form with $y_K(L)$ if we put $d_K = d$.

Hence, the choice of L that maximizes $T(L)$ is :

$$L_o(\pi) = d^{-1} = (M-1)^{-1} \sum_{K=1}^{M-1} d_K^{-1}$$

Since d_K^{-1} is an eigenvalue of $\left[D(\pi) \right]^{-1}$, we have :

$$L_o(\pi) = (M-1)^{-1} \text{ trace } \left[D(\pi) \right]^{-1}$$

It is much easier to compute $L_o(\pi)$ for each $\pi \in I_M(A)$ by this formula.

If noiseless observations were available, the relative frequency estimate of the prior probabilities would have asymptotic error covariance matrix $R^C(\pi)$, with elements

$$C_{ij} = \pi_j (\delta_{ij} - \pi_i)$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

and $\pi_M = 1 - \sum_{j=1}^{M-1} \pi_j$

The inverse matrix, $\left[R^C(\pi) \right]^{-1}$ has elements

$$h_{ij} = \pi_M^{-1} \left(1 + \pi_j^{-1} \pi_M \delta_{ij} \right)$$

Hence,

$$A^T \left[R^C (\pi) \right]^{-1} A = \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} h_{ij} a_i a_j =$$

$$= \sum_{i=1}^{M-1} a_i^2 \pi_i^{-1} +$$

$$+ \pi_M^{-1} \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} a_i a_j =$$

$$= \sum_{i=1}^{M-1} a_i^2 \pi_i^{-1} + \pi_M^{-1} \left(\sum_{i=1}^{M-1} a_i \right)^2$$

We also have :

$$A^T \left[R^A (\pi) \right]^{-1} A = \int_{E^n} g(X \mid \pi)^{-1} \cdot$$

$$\cdot \left[\sum_{i=1}^{M-1} a_i \left(f_i(X) - f_M(X) \right) \right]^2 dx =$$

$$= \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} a_i \, a_K \, J_{iK}(\pi)$$

For

$$A = (0, \ldots, 0, a_K, 0, \ldots, 0) \qquad , \; a_K \neq 0$$

we have

$$A^T \left[R^A(\pi) \right]^{-1} A = a_K^2 \, J_{KK}(\pi)$$

and

$$A^T \left[R^C(\pi) \right]^{-1} A = a_K^2 \left(\pi_K^{-1} + \pi_M^{-1} \right)$$

Using the result of Appendix IV, we have

$$J_{KK}(\pi) \leq (\pi_K + \pi_M)^3 \, (\pi_K \pi_M)^{-1} \leq (\pi_K^{-1} + \pi_M^{-1})$$

hence, for such A's we have

$$\left[A^T \left[R^A(\pi) \right]^{-1} A \right]^{-1} \geq \left[A^T \left[R^C(\pi) \right]^{-1} A \right]^{-1}$$

I have not been able to prove the above inequality for general A.
I conjecture that it is true in general, because the left side expresses the Rao-Cramér bound on estimating the mixture priors under noisy observations, while the right side expresses the variance of the relative frequency estimate under noiseless (or perfectly classified) observations.
In any case, for a given weight vector A, we can compute both quadratic forms. Their relative sizes will give us a measure of performance loss due to noisy (unclassified) observations in estimating the prior probabilities.

Conclusions

We consider the problem of estimating the mixing prior probabilities when

the probability density functions of a mixture are known.

It was shown that the maximum likelihood estimator is asymptotically efficient, but difficult to implement.

Hence a recursive estimator was proposed and analyzed. Using 2 stochastic approximation theorems due to Sacks, it was possible to show convergence to the true value.

Also, the asymptotic error variance was computed in a closed form. Because of the closed expression, it was possible to see the performance loss due to the use of a recursive algorithm.

For the binary mixture, it was possible to modify the recursive algorithm by means of a memoryless nonlinear transformation, and achieve asymptotical efficiency. For the M ary mixture with $M > 2$, use of a memoryless nonlinear transformation in the recursive algorithm decreased the error covariance, without achieving asymptotic efficiency.

Appendix I

The purpose of the present appendix is to show that

$$J(\pi) < [\pi(1-\pi)]^{-1} \quad \text{for} \quad \pi \neq 0, 1$$

and that the function $[J(\pi)]^{-1}$ is concave

for arbitrary densities $f_1(X)$, $f_2(X)$ that are nonzero for all $X \in E^n$. Also a method will be given for computing $J(\pi)$ in the Gaussian case.

Let

$$s = \pi(1-\pi)^{-1}$$

Assume

$$\pi \neq 0, 1$$

$J(\pi)$ can be written :

$$J(\pi) = (1+s) \int_{E^n} \left[1 - f_2(X)\left(f_1(X)\right)^{-1}\right]^2 \cdot$$

$$\cdot \left[s + f_2(X)\left(f_1(X)\right)^{-1}\right]^{-1} f_1(X)dX =$$

$$= (1+s) \int_{E^n} \left\{f_2(X)\left(f_1(X)\right)^{-1} \cdot - (2+s) + (s+1)^2 \cdot \right.$$

$$\left. \cdot \left[s + f_2(X)\left(f_1(X)\right)^{-1}\right]^{-1}\right\} f_1(X)dX$$

Hence

$$(1+s)^{-2} \cdot J(\pi) = -1 + (1+s)s^{-1} \cdot \int_{E^n} f_1(X) \cdot$$

$$\cdot s\left[s + f_2(X)\left(f_1(X)\right)^{-1}\right]^{-1} dX$$

The function $s\left[s+f_2(X)\left(f_1(X)\right)^{-1}\right]^{-1}$ is positive and upper bounded by 1. Hence, we can upper bound $J(\pi)$:

$$J(\pi) \leq (1+s)^2 \cdot s^{-1}$$

or :

$$J(\pi) \leq [\pi(1-\pi)]^{-1}$$

$$[J(\pi)]^{-1} \geq \pi(1-\pi)$$

It is seen that only for $\pi = 0$ or 1 there is a possibility for $J(\pi)$ to be infinite.

A general method will now be given for computing $J(\pi)$ in the case of f_1, f_2 being multivariate Gaussian densities. The approach is an extension of a method in [2] and [4].

Let

$$f_1(X) = N(X, O, R_1)$$

$$f_2(X) = N(X, M_0, R_2)$$

where $M_0 = M_2 - M_1$ = difference of mean vectors.

Let A be the $n \times n$ orthogonal matrix satisfying the relations :

$$AR_1A^T = I$$

$$AR_2A^T = \Lambda$$

where $\Lambda = \text{diag}\,(\lambda_1 \ldots \lambda_n)$

and λ_i are the eigenvalues of R_2 with respect to R_1.

Hence, they satisfy the equation :

$$\left| R_2 - \lambda R_1 \right| = 0$$

Let $M = AM_0 = (m_1 \ldots m_n)^T$

If we make the change of variables

$$Y = AX = (y_1 \ldots y_n)^T$$

the transformed densities are :

$$f_1(Y) = N(Y, O, I)$$

$$f_2(Y) = N(Y, M, \Lambda)$$

It is sufficient to compute the quantity :

$$\int_{E^n} f_1(Y) \cdot \left[s + f_2(Y) \left(f_1(Y) \right)^{-1} \right]^{-1} dY =$$

$$= E\left\{ \left[s + f_2(Y) \left(f_1(Y) \right)^{-1} \right]^{-1} \mid H_1 \right\}$$

Let

$$z = \log \left[f_2(Y) \left(f_1(Y) \right)^{-1} \right]$$

Then

$$z = \frac{1}{2} \sum_{k=1}^{n} y_k^2 - \lambda_k^{-1} (y_k - m_k)^2 - \log \lambda_k$$

The above conditional expectation can be written :

$$E \left\{ \left[s + e^z \right]^{-1} \mid H_1 \right\}$$

Under hypothesis H_1, y_k are Gaussian, zero mean, unit variance independent random variables.

We are now in a position to construct the characteristic function of z under the hypothesis H_1.

Let

$$C(jw) = E\left\{\exp(jwz) \mid H_1\right\}$$

Let

$$a_k = 1 - \lambda_k^{-1}$$

$$b_k = m_k(1-\lambda_k)^{-1}$$

$$h_k = (a_k b_k)^2(1-a_k)^{-1} + \log \lambda_k \qquad k=1, \ldots m$$

Then,

$$C(jw) = \prod_{k=1}^{n} F_k(jw)$$

where

$$F_k(jw) = (1-2a_k jw)^{-\frac{1}{2}} \exp\left[-2(a_k b_k)^2(1-2a_k jw)^{-1} - \right.$$

$$\left. - jwh_k \right]$$

The probability density function $g(z)$ of the random variable z under hypothesis H_1, can be computed from $C(jw)$ by an inverse Fourier transform.

Let

$$q = 3.14159$$

$$g(z) = (2q)^{-1} \int_{-\infty}^{+\infty} C(jw)\exp(-jwz)dw$$

We can finally compute the desired quantity :

$$E\left\{ [s+e^{z}]^{-1} \mid H_1 \right\} = \int_{-\infty}^{+\infty} g(z)\ [s+e^{z}]^{-1}\ dz$$

We will now show that the function $[J(\pi)]^{-1}$ is concave.

This fact was noticed by Boes [1].

The second derivative of $[J(\pi)]^{-1}$ is :

$$\frac{d^2}{d\pi^2}\ [J(\pi)]^{-1} = \Big\{ -2 \int (f_1 - f_2)^2\ g^{-1}\ dx \cdot$$

$$\cdot \int (f_1 - f_2)^4\ g^{-3}\ dx +$$

$$+ 2\Big[\int (f_1 - f_2)^3\ g^{-2}\ dx \Big]^2 \Big\} \Big/$$

$$\Big/ \quad \Big\{ \int (f_1 - f_2)^2\ g^{-1}\ dx \Big\}^3$$

where

$$g = \pi f_1 + (1-\pi) f_2$$

Using Schwarz's inequality, we have :

$$\left\{ \int \left[(f_1 - f_2)\, g^{-1} \right]^3 g \, dx \right\}^2 =$$

$$= \left\{ \int \left[(f_1 - f_2)\, g^{-1} \right] \sqrt{g} \left[(f_1 - f_2)\, g^{-1} \right] \sqrt{g} \, dx \right\}^2 \leq$$

$$\leq \int (f_1 - f_2)^2 \, g^{-2} \, g \, dx \int (f_1 - f_2)^4 \, g^{-4} \, g \, dx =$$

$$= \int (f_1 - f_2)^2 \, g^{-1} \, dx \int (f_1 - f_2)^4 \, g^{-3} \, dx$$

Hence, the numerator of the expression for the second derivative is negative.

Therefore,

$$\frac{d^2}{d\pi^2} \, [J(\pi)]^{-1} < 0 \quad \text{for all} \quad \pi \in [0, 1] \quad \text{and hence} \quad [J(\pi)]^{-1}$$

is concave.

In Fig. 3, we show the shape of $[J(\pi)]^{-1}$ in relation to $\pi(1-\pi)$, which is a lower bound.

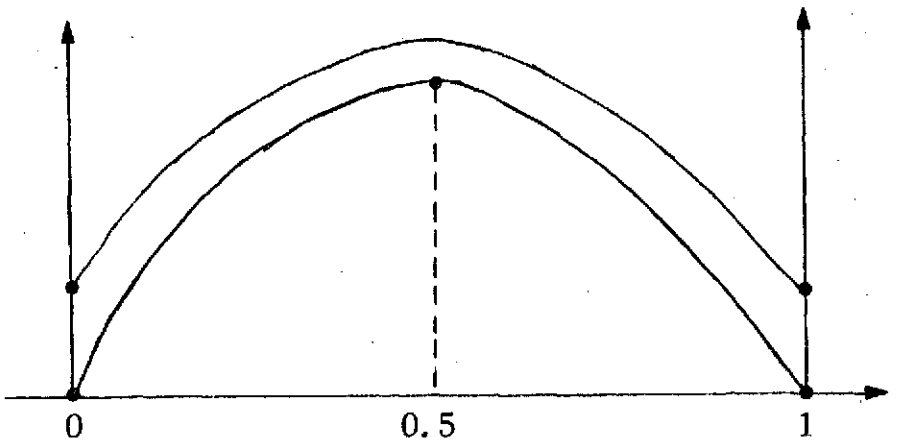

Fig. 3 $\quad [J(\pi)]^{-1} > \pi(1 - \pi)$

Appendix II

We need to check whether conditions 1-7 are satisfied by the class of density functions $f_1(X), f_2(X)$ that satisfy Assumptions 1-2.

The derivatives appearing in Condition 1 are :

$$\frac{\partial^k}{\partial q^k} \log g(X \mid q) = (-1)^{k-1} (k-1)! \left[f_1(X) - f_2(X) \right]^k \cdot \left[q f_1(X) + (1-q) f_2(X) \right]^{-k}$$

for $k = 1, 2, 3$

Using this formula, it is straightforward to check that

$$E \frac{\partial}{\partial q} \log g(X \mid q) \Big|_{q=\pi} = 0$$

$$- E \frac{\partial^2}{\partial q^2} \log g(X \mid q) \Big|_{q=\pi} = E\left(\frac{\partial}{\partial q} \log g(X \mid q) \right)^2 \Big|_{q=\pi} = J(\pi)$$

where

$$J(\pi) = \int_{E^n} \left[f_1(X) - f_2(X) \right]^2 \left[\pi f_1(X) + (1-\pi) f_2(X) \right]^{-1} dX$$

Hence Conditions 1-4 are satisfied.

For Condition 5,

$$\left| \frac{\partial^3}{\partial q^3} \log g(X \mid q) \right| = 2 \left| \frac{f_1(X) - f_2(X)}{q f_1(X) + (1-q) f_2(X)} \right|^3 \leq$$

$$\leq 2 \left| f_1(X) - f_2(X) \right|^3 \left[A(X)\right]^{-3}$$

where $A(X) = \min \left(f_1(X) , f_2(X)\right)$

Since $A(X) > 0 \quad \forall X \in E^n$, and $f_1(X)$, $f_2(X)$ are bounded, Condition 5 is satisfied.

For Condition 6, it is known that the Kullback-Leibler information number $I(q,\pi)$ has the following properties :

$$I(q,\pi) = 0 \quad \text{iff } g(X \mid \pi) = g(X \mid q) \quad \forall X \in E^n$$

and $I(q,\pi) > 0$ otherwise.

Because of the identifiability Assumption 2, we can have

$$g(X \mid \pi) = g(X \mid q) \quad \forall X \in E^n \qquad \text{only for } \pi = q$$

Hence, Assumption 2 implies that $I(q,\pi)$ achieves a unique minimum at $q = \pi$, and Condition 6 is satisfied.

The function

$$B(q, X) = \frac{\partial}{\partial q} \log g(X \mid q) =$$

$$= \left[f_1(X) - f_2(X)\right] \left[q\, f_1(X) + \right.$$

$$\left. + (1-q)\, f_2(X)\right]^{-1}$$

is continuous in q for all $q \in [0, 1]$. Furthermore, $B(q, X)$ is bounded, therefore, it is uniformly continuous in q, and Condition 7 is satisfied.

Appendix III

Condition 1a has already been shown to be valid.

For Condition 2a, we have :

$$\left| M(q) \right| \leq C_2 \left| F(q) \right|$$

$F(q)$ and $F'(q)$ will be shown to be bounded.

Let

$$e^z = f_2(X) \left[f_1(X) \right]^{-1}$$

We can write :

$$F(q) = \int_{E^n} f_1(X) \left[\pi + (1-\pi) e^z \right] \left[q + (1-q) e^z \right]^{-1} dX -$$

$$- \int_{E^n} f_2(X) \left[\pi e^{-z} + (1-\pi) \right] \left[q e^{-z} + (1-q) \right]^{-1} dX$$

The second integral has the same form with the first one. If we interchange f_1 and f_2, π and $1-\pi$, q and $1-q$ in the second integral, we get the first one. Hence, it suffices to check the boundedness of the first integral only.

$$\int_{E^n} f_1(X) \left[\pi + (1-\pi) e^z \right] \left[q + (1-q) e^z \right]^{-1} dX =$$

$$= E \left\{ T(z,\pi,q) \mid H_1 \right\}$$

where

$$T(z,\pi,q) = \left[\pi + (1-\pi)\, e^{z} \right] \left[q + (1-q)\, e^{z} \right]^{-1}$$

The derivative of T with respect to z is :

$$\frac{\partial T}{\partial z} = (q-\pi) \left[q + (1-q)\, e^{z} \right]^{-2}$$

Hence T is a monotone function of z.

We have the following bounds :

$$\min\left(\frac{\pi}{q}, \frac{1-\pi}{1-q}\right) \leq T(z,\pi,q) \leq \max\left(\frac{\pi}{q}, \frac{1-\pi}{1-q}\right)$$

Hence $F(q)$ is bounded for $q \neq 0, 1$

The values $F(1), F(0)$ are :

$$f(1) = (1-\pi) \left[1 - J(1) \right]$$

$$F(0) = \pi \left[-1 + J(0) \right]$$

By the definition of the interval $I(\epsilon)$, we see that $F(q)$ is bounded for all q in the interval $I(\epsilon)$.

In a similar manner, it can be shown that $F'(q)$ is bounded for $q \neq 0, 1$.

Hence,

$$\left| M(q) \right| \leq C_2 \left| F'(q) \right| \left| q - \pi \right| \leq C_2 C_3 \left| q - \pi \right|$$

for $q \neq 0, 1$

where $C_3 < +\infty$

The first part of Condition 2a has been satisfied.

The second part is satisfied also, if we observe that $F(q)$ is a strictly monotone function of q.

Because of the boundedness of $F'(q)$, Condition 3a also easily satisfied, with

$$a_1 = M'(\pi) = L(\pi)\ F'(\pi)$$

Also we note that

$$F'(\pi) = -J(\pi)$$

For Condition 4a, we must compute

$$E\left[z^2(X,q) \mid q\right] = E\left[G(X,q)\ L(q) + M(q)\right]^2 =$$

$$= L^2(q)\ E\left[G^2(X,q) \mid q\right] - M^2(q) =$$

$$= L^2(q)\left[-F'(q)\right] - M^2(q)$$

For $q \neq 0, 1$ the above quantity is finite, hence Condition 4a is satisfied. Also, we need to compute the quantity :

$$S(\pi) = \lim_{z \to \pi} E\left[z^2(X,q) \mid q\right] = L^2(\pi)\ J(\pi)$$

Appendix IV

In this appendix, we will seek upper bounds to the integrals $J_{sk}(\pi)$, $s, k = 1, \ldots, M-1$.

$$J_{sk}(\pi) = \int_{E^n} \left(f_s(X) - f_M(X)\right)\left(f_k(X) - f_M(X)\right) \cdot$$

$$\cdot \left[\sum_{m=1}^{M-1} \pi_m f_m(X) + \left(1 - \sum_{m=1}^{M-1} \pi_m\right) \cdot \right.$$

$$\left. \cdot f_M(X)\right]^{-1} dX$$

We will first consider the case $s=k$.

Let

$$\pi_M = 1 - \sum_{j=1}^{M-1} \pi_j$$

$$J_{kk}(\pi) = \int_{E^n} \left(f_k - f_M\right)^2 \left[\pi_k f_k(X) + \pi_M f_M(X) + \right.$$

$$\left. + \sum_{\substack{m=1 \\ m \neq k}}^{M-1} \pi_m f_m(X)\right]^{-1} dX \leq$$

$$\leq \int_{E^n} \left(f_k - f_M\right)^2 \left[\pi_k f_k(X) + \pi_M f_M(X)\right]^{-1} dX$$

Hence,

$$J_{kk}(\pi) \leq \left(\pi_k + \pi_M\right) \int_{E^n} \left(f_k(X) - f_M(X)\right)^2 .$$

$$. \left[\rho(K, M)\ f_k(X) + (1-\rho(K, M)\ f_M(X)\right]^{-1} dX$$

where

$$\rho(K, M) = \pi_k \left[\pi_k + \pi_M\right]^{-1}$$

In Appendix I, an upper bound to this last integral has been found under the condition :

$$\rho(k, M) \neq 0, 1$$

Using this result, we have :

$$J_{kk}(\pi) \leq (\pi_k + \pi_M) \left[\rho(k, M)\ \left(1-\rho(k, M)\right)\right]^{-1}$$

or :

$$J_{kk}(\pi) \leq (\pi_k + \pi_M)^3\ (\pi_k\ \pi_M)^{-1}$$

under the condition :

$$\pi_k\ \pi_M \neq 0$$

Using the Schwarz inequality, we can upper bound $J_{sk}(\pi)$

$$\left[J_{sk}(\pi)\right]^2 = \left\{ \int_{E^n} \left[g(X \mid \pi)\right]^{-\frac{1}{2}} \left[f_s(X) - f_M(X)\right] .\right.$$

$$\left. . \left[g(X|\pi)\right]^{-\frac{1}{2}} \left[f_k(X) - f_M(X)\right] dX \right\}^2 \leq$$

$$\leq \int_{E^n} \left[g(X \mid \pi)\right]^{-1} \left[f_s(X) - f_M(X)\right]^2 dX \cdot$$

$$\cdot \int_{E^n} \left[g(X \mid \pi)\right]^{-1} \left[f_k(X) - f_M(X)\right]^2 dX$$

Hence

$$\left[J_{sk}(\pi)\right]^2 \leq J_{kk}(\pi)\ J_{ss}(\pi)$$

$$\left| J_{sk}(\pi) \right| \leq \left[(\pi_k + \pi_M)\ (\pi_s + \pi_M)\right]^{3/2} \cdot$$

$$\cdot\ (\pi_k\ \pi_s)^{-\frac{1}{2}}\ \pi_M^{-1}$$

This bound is valid for

$$\pi_s,\ \pi_k,\ \pi_M \neq 0$$

As a conclusion, we see that if π lies in the interior of the set I_M, the functions $J_{sk}(\pi)$ are finite.

Hence, the part of condition 3′ related to the finiteness of the above functions, is satisfied.

Appendix V

In the present Appendix, we will check the satisfaction of Conditions 1-5, based on the Assumptions 1-2. For Condition 1, we construct the scalar function

$$A(\lambda) = (Q - \pi)^T M[\pi + \lambda (Q - \pi)]$$

defined for $\lambda \in [0, 1]$

We have

$$A(0) = (Q - \pi)^T M(\pi) = 0$$

$$A(1) = (Q - \pi)^T M(Q)$$

The derivative of $A(\lambda)$ is :

$$A'(\lambda) = \sum_{s=1}^{M-1} (q_s - \pi_s) \frac{\partial}{\partial \lambda} M_s[\pi + \lambda(Q - \pi)]$$

But :

$$M_s[\pi + \lambda(Q - \pi)] =$$

$$= -L(Q) \int_{E^n} g(X \mid \pi) \left[\lambda \sum_{k=1}^{M-1} [f_k(X) - f_M(X)] (q_k - \pi_k) + \sum_{k=1}^{M-1} [f_k(X) - f_M(X)] \pi_k + f_M(X)\right]^{-1} \cdot$$

$$\cdot [f_s(X) - f_M(X)] \, dX$$

Hence :

$$\frac{\partial}{\partial \lambda} M_s[\pi + \lambda(Q - \pi)] = L(Q) .$$

$$\cdot \int_{E^n} g(X \mid \pi) \left[g\left(X \mid \pi + \lambda(Q - \pi)\right) \right]^{-2} \cdot$$

$$\cdot \left[\sum_{k=1}^{M-1} \left[f_k(X) - f_M(X) \right] (q_k - \pi_k) \right] \cdot$$

$$\cdot \left[f_s(X) - f_M(X) \right] dX$$

Substituting, we have the following expression for $A'(\lambda)$:

$$A'(\lambda) = L(Q) \int_{E^n} g(X \mid \pi) \left[g\left(X \mid \pi + \lambda(Q - \pi)\right) \right]^{-2} \cdot$$

$$\cdot \left[\sum_{k=1}^{M-1} (q_k - \pi_k) \left[f_k(X) - f_M(X) \right] \right]^2 dX$$

or, more compactly :

$$A'(\lambda) = L(Q) \int_{E^n} g(X \mid \pi) \left[g\left(X \mid \pi + \lambda(Q - \pi)\right) \right]^{-2} \cdot$$

$$\cdot \left[g(X \mid \pi) - g(X \mid Q) \right]^2 dX$$

We have, therefore :

$$A'(\lambda) \geq 0 \qquad \forall \lambda \in [0, 1]$$

The case $A'(\lambda) = 0$ will occur iff $g(X \mid Q) \equiv g(X \mid \pi) \quad \forall X \in E^n$. But, due to the identifiability assumption of $\{f_i(X)\}$, this would imply $Q = \pi$.

Hence, for $Q \neq \pi$

we have $A'(\lambda) > 0 \qquad \forall \lambda \in [0,1]$.

Therefore,

$$A(1) = (Q - \pi)^T M(Q) > 0 \qquad \forall Q \neq \pi$$

and Condition 1 is satisfied. For Condition 2, we apply the mean value theorem to the scalar function of λ, $M_k[\pi + \lambda(Q - \pi)]$, between the points $\lambda = 0$ and $\lambda = 1$.

$$M_k(Q) = M_k(\pi) + \sum_{s=1}^{M-1} (q_s - \pi_s) \frac{\partial}{\partial q_s} \cdot$$

$$\cdot M_k[\pi + \lambda_k(Q - \pi)]$$

where $\lambda_k \in [0,1]$.

Substituting, we have :

$$M_k(\pi) = L(Q) \sum_{s=1}^{M-1} (q_s - \pi_s) C_{ks}$$

where

$$C_{ks} = \lambda_k \int_{E^n} g(X|\pi)\left[g(X \mid Q_k)\right]^{-2} \left[f_s(X) - f_M(X)\right] \cdot$$
$$\cdot \left[f_k(X) - f_M(X)\right] dX$$

with

$$Q_k = \pi + \lambda_k(Q - \pi) = (P_1 \; P_2 \; \ldots \; P_{M-1})^T$$

Also, let

$$P_M = 1 - \sum_{j=1}^{M-1} P_j$$

Therefore,

$$\left[M_k(Q) \right]^2 = L^2(Q) \left[\sum_{k=1}^{M-1} (q_k - \pi_k) C_{ks} \right]^2 \leq$$

$$\leq L^2(Q) \left[\sum_{s=1}^{M-1} C_{ks}^2 \right] \| Q - \pi \|^2$$

and

$$\| M(Q) \|^2 = \sum_{k=1}^{M-1} \left[M_k(Q) \right]^2 \leq$$

$$\leq L^2(Q) \left[\sum_{k=1}^{M-1} \sum_{s=1}^{M-1} C_{ks}^2 \right] \| Q - \pi \|^2$$

We can bound the quantities C_{ks}, with a method similar to the one used in Appendix IV.

The result is :

$$C_{ks} \leq \tfrac{1}{2} \lambda_k \left[\max \left(P_k^{-1}, P_M^{-1} \right) + \max \left(P_s^{-1}, P_M^{-1} \right) \right]$$

Therefore, for $Q \in I_M(A)$,

and with $K_1 = C_2^2 \sum_{k,s} C_{ks}^2 < +\infty$

we have satisfied Condition 2. For Condition 3, we use the second order mean value theorem for the scalar function of λ,

$M_k[\pi + \lambda(Q - \pi)]$, between the points $[0,1]$

$$M_k(Q) = \sum_{s=1}^{M-1} (q_s - \pi_s) \frac{\partial}{\partial q_s} M_k(\pi) +$$

$$+ \sum_{s=1}^{M-1} \sum_{j=1}^{M-1} (q_s - \pi_s)(q_j - \pi_j) \cdot$$

$$\cdot \frac{\partial^2}{\partial q_s \partial q_j} M_k\left[\pi + \lambda_k(Q - \pi)\right]$$

where $\lambda_k \epsilon [0, 1]$.

Hence, we can write :

$$M(Q) = B(Q - \pi) + (Q - \pi)^T W(Q - \pi)$$

where $B = L(\pi) D$ and D is a $(M-1) \times (M-1)$ matrix with elements D_{ij},

$$D_{ij} = \int_{E^n} \left[g(X \mid \pi)\right]^{-1} \left[f_i(X) - f_M(X)\right] \cdot$$

$$\cdot \left[f_j(X) - f_M(X)\right] dX$$

The matrix W is $(M-1) \times (M-1)$ and has element (s, j) the number :

$$\frac{\partial^2}{\partial q_s \partial q_j} \sum_{k=1}^{M-1} M_k\left[\pi + \lambda_k(Q - \pi)\right]$$

It can be shown, again, that for $Q \epsilon I_M(A)$, the above terms are bounded, with methods similar to those of Appendix I.

Hence

$$\frac{(Q - \pi)^T W (Q - \pi)}{\|Q - \pi\|^2}$$ is upper bounded by a finite number.

Furthermore, let $Y = \left(y_1 \ \ldots \ y_{M-1}\right)^T$ be an arbitrary vector, $\|Y\| \neq 0$.

Then

$$Y^T B Y = L(\pi) \int_{E^n} \left[g(X \mid \pi)\right]^{-1} \cdot \left[\sum_{k=1}^{M-1} y_k \left(f_k(X) - f_M(X)\right)\right]^2 dX$$

$Y^T B Y$ can be zero iff $\sum_{k=1}^{M-1} y_k f_k(X) - f_M(X) \sum_{k=1}^{M-1} y_k = 0$

$\forall X \in E^n$.

The identifiability of the set $\left(f_i(X)\right)$ makes this impossible.

Therefore, B is positive definite. The above facts show that Condition 3 is satisfied.

We must compute

$$E\left[\ \|Z(X,Q)\|^2 \mid Q\ \right] =$$

$$= L^2(Q) \int_{E^n} g(X \mid \pi) \left[g(X \mid Q)\right]^{-2} .$$

$$\cdot \sum_{k=1}^{M-1} \left[f_k(X) - f_M(X) \right]^2 dX - \left\| M(Q) \right\|^2$$

The first integral can be upper bounded in the same manner as C_{kk}.

We have :

$$\int_{E^n} g(X \mid \pi) \left[g(X \mid Q) \right]^{-2} \left[f_k(X) - f_M(X) \right]^2 dX \leq$$

$$\leq \max \left(q_k^{-1} , q_M^{-1} \right)$$

where

$$Q = \left(q_1 \cdots q_{M-1} \right)$$

and

$$q_M = 1 - \sum_{j=1}^{M-1} q_j$$

For

$$q_1 , \ldots , q_{M-1} , q_M > 0 ,$$

each term is bounded.

Hence, for $Q \in I_M(A)$, the expected value of the norm of $Z(X, Q)$ is bounded. The matrix $S(\pi)$ has elements

$$S_{ij}(\pi) = L^2(\pi) \int_{E^n} \left[g(X \mid \pi) \right]^{-1} \left[f_i(X) - f_M(X) \right] \cdot$$

$$\cdot \left[f_j(X) - f_M(X) \right] dX$$

or :

$$S(\pi) = L^2(\pi) \, D(\pi)$$

It has been shown already that $D(\pi)$ is positive definite.

Hence Condition 4 is satisfied. Because of the nature of the algorithm, Condition 5 is easily shown to be satisfied.

For our case, we have :

$$B(\pi) = L(\pi)\ D(\pi)$$

$$S(\pi) = L^2(\pi)\ D(\pi)$$

The matrix $S^*(\pi)$ is :

$$S^*(\pi) = P^{-1}\ S(\pi)\ P = P^{-1}\ L^2(\pi)\ D(\pi)\ P =$$

$$= L(\pi)\ P^{-1}\ B(\pi)\ P =$$

$$= L(\pi)\ \mathrm{diag}(b_1 \ \ldots \ b_{M-1})$$

Let $d_1 \geq d_2 \geq \ldots \geq d_{M-1}$ be the eigenvalues of $D(\pi)$.

Then,

$$b_k = L(\pi)\ d_k$$

and

$$S^*(\pi) = L^2(\pi)\ \mathrm{diag}(d_1 \ldots d_{M-1})$$

The matrix F is, therefore, diagonal :

$$F(\pi) = L^2(\pi)\ \mathrm{diag}\Big[\big(2L(\pi)\ d_1 - 1\big)^{-1}\ d_1\ , \ldots\ ,$$
$$\big(2L(\pi)\ d_{M-1} - 1\big)^{-1}\ d_{M-1}\Big]$$

References

[1] D. C. Boes, "On the Estimation of Mixing Distributions," Ann. Math. Statistics, 1966, p. 177.

[2] K. Fukunaga and T. Krile, "Calculation of Bayes' Recognition Error for Two Multivariate Gaussian Distributions," IEEE Trans. on Comp., March, 1969.

[3] T. Y. Young and G. Coraluppi, "Stochastic Estimation of a Mixture of Normal Density Functions Using Information Criterion," IEEE Trans. on IT, May, 1970.

[4] D. Kazakos, "Optimal Design of an Unsupervised Adaptive Classifier with Unknown Priors," ICSA, Rice University Technical Report, May, 1974. (Will appear as an article.)

[5] S. J. Yakowitz, "Unsupervised Learning and the Identifiability of Finite Mixtures," IEEE Trans. on IT, 1970, (3).

[6] H. Cramér, "Mathematical Methods of Statistics," Princeton University Press, 1946.

[7] M. D. Perlman, "The Limiting Behavior of Multiple Roots of the Likelihood Equation," Department of Statistics, University of Minnesota, Tech. Report 125, July, 1969.

[8] J. Sacks, "Asymptotic Distribution of Stochastic Approximation Procedures," Ann. Math. Statistics, 1958, (2).

[9] J. R. Blum, "Multidimensional Stochastic Approximation Method," Ann. Math. Statistics, 1954, p. 737.

[10] D. Sakrison, "Stochastic Approximation : A Recursive Method for Solving Regression Problems," Advances in Communication Systems, Vol. 2, A. V. Balakrishnan, ed. New York, Academic Press, 1966, pp. 51.