INFORMATION IN REMOTELY SENSED DATA FOR ESTIMATING PROPORTION IN MIXTURE DENSITIES

by

Virgil R. Marco, Jr.
Program in Mathematical Science
University of Texas at Dallas
Richardson, Texas 75080

Report #23

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas
Contract NAS-9-14680-11S

November 1980

INFORMATION IN REMOTELY SENSED DATA FOR ESTIMATING PROPORTION IN MIXTURE DENSITIES[1]

Virgil R. Marco, Jr., and Patrick L. Odell
University of Texas at Dallas
Box 688, Richardson, Texas, 75080

I. INTRODUCTION

Data taken remotely by multichannel sensors on a near earth satellite can be modeled as a collection of multivariate data points. In the application [1] that motivates this paper each $p\times 1$ data vector represents a measure of reflectance from (1.1) acre location on the surface of the earth. Each of the p elements of the data vector is a reflectance measure at a preassigned wave length of light. Conceptually, each crop class defines a set of reflectance measures that can be modeled by a multivariate unimodel probability density function unique for each crop class.

Let there be m-crop classes and let the p.d.f.

$$P_i(x) = p_i(x;\mu_i,\Sigma_i) \quad i = 1,\ldots,m \tag{1.1}$$

denote the distribution of the random data vector X given that the measurements were made on the i^{th} crop class, Π_i, $i = 1,\ldots,m$. Also let the multivariate mixture p.d.f.

---

[1]This research was supported in part by the National Aeronautics and Space Agency, Johnson Space Center under Contract NAS 9-14689-95.

$$p(x) = \sum_{i=1}^{m} \alpha_i p_i(x) \qquad (1.2)$$

such that $\alpha_i \quad 0 \quad i = 1,2,\ldots,m$ and $\sum_{i=1}^{m} \alpha_i = 1$ denote the distribution of the multivariate observations given that the data is unlabeleld, that is modeled by $p(x)$ in (1.2).

Definition 1. A random sample is said to be unlabeled if the random vectors are selected from a population defined by (1.2).

Definition 2. A random sample of unlabeled data is said to be classified data if, according to some classification rule $R = (R_1,R_2,\ldots,R_m)$, each vector in the sample is assigned to one of the (crop) classes $\Pi_1, \Pi_2,\ldots,\Pi_m$.

Definition 3. A random sample of unlabeled data is said to be verified data if each vector is classified as being from the true subclass Π_i for some $i = 1,2,\ldots,$ or with probability one.

Verified data is classified data in which there is zero probability of misclassification.

Definition 4. A random sample is said to be labeled if it is selected from a single class Π_i and the identity of i^{th} population is known.

The difference between verified and labeled data is that the verified data must be labeled a posteriori while the labeled data is labeled prior to taking the sample. In both types of samples, one knows with certainty the label of the population from which the samples came.

The purpose is to estimate the vector or proportions $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_p)^T$ which defines the function $p(x)$ in (1.2). If α_i

denotes the proportion of vectors in the mixture from class Π_i and N the total number of vectors in the region, then

$$\hat{A}_i = (1.1)\ N\ \hat{\alpha}_i \tag{1.3}$$

is an estimate of acreage of crop class Π_i, as a function of an estimate of the proportion $\hat{\alpha}_i$ and α_i. Hence, our interest is to estimate well.

Three different types of data are available for estimating the elements of α arise naturally in the application involving remote sensing from space. They all are maximum likelihood estimators for α using

(a) unlabeled data,

(b) classified data, or

(c) verified data, respectively.

The cost of acquiring unlabeled data is less than the cost of acquiring classified data which is in turn less than the cost of acquiring verified data. The computation of sample size allocations when samples from more than one type of data are available arises naturally. In the case of sample design one can control the type of data to be selected and the optimal mix of sampling can be accomplished. It is important to note that one <u>always</u> has available a random sample of unlabeled data; hence if C_u denotes the cost per unit of taking unlabeled data then

$$C_v = C_u + c_v = K_v C_u$$

$$C_c = C_u + c_c = K_c C_u$$

are the per unit cost where C_v and C_u are the costs of classifying and verifying in unlabeled data point respectively. The values K_v and K_c are multiplicative constants that give in addition to an additive model a second multiplicative representation of the costs.

One would expect $C_u < C_c < C_v$ in most space science applications. It is important to note that in the space application unlabeled data is available as basic for two of the three methodologies for estimating α, and except for missing data that the totality of unlabeled data is also available. The cost of machine processing every vector is a realistic limiting factor for unlabeled and classified data while the cost of resources to visit each location for verification is the major limiting factor for obtaining verified data.

However, it is not intuitively clear which type of data contains greatest amount of information for estimating α for a fixed sample size. The purpose of this paper is to compute and order with respect to magnitude the information content of the three types of data, and discuss the implications of that ordering for the space application.

The term information content of the data is defined as the inverse of the Cramer-Rao matrix lower bound for unbiased estimators for α. This is the matrix form of Fisher's Information.

II. INFORMATION CONTENT OF VARIOUS TYPES OF DATA

2.1 Fisher's Information: Let X denote a random observation from a multivariate (p-variate) population whose p.d.f. is defined by (1.2). If we denote the parameter vector by $\alpha = (\alpha_1,\ldots,\alpha_{m-1})^T$ then by the usual theory (Cramer [2], Rao [3]) the $(m-1 \times 1)$ random vector

$$S = \frac{\partial \ell np(x)}{\partial \alpha} \qquad (2.1.1)$$

is such that

$$E[S] = \phi$$

and

$$E[S\ S^T] = -\ E\left[\frac{\partial^2 \ell np(x)}{\partial\alpha\ \partial\alpha}\right] = -\ E\left\{\frac{\partial^2 \ell np(x)}{\partial\alpha_i\ \partial\alpha_j}\right\} \overset{def}{=} \Lambda(\alpha) \qquad (2.1.2)$$

where $\Lambda(\alpha)$ denotes Fisher's information for α contained in the sample X.

If $X_1,\ldots,X_n$ denote a random sample from a multivariate population whose p.d.f. is defined by (1.2), then the Fisher's information for α contained in this sample can be shown to be

$$E[S\ S^T] = n\ \Lambda(\alpha)\ . \qquad (2.1.3)$$

Furthermore, $\Lambda^{-1}(\alpha)$ is the Cramer-Rao lower covariance matrix bound for unbiased estimators of the vector α. That is, if $\hat{\alpha}$ is any <u>unbiased</u> estimator for α, then the covariance matrix $\Lambda(\hat{\alpha})$ will never be less than $\Lambda^{-1}(\alpha)$. Note that if A and B are two positive

definite matrices of the same size and $A - B$ is positive semi-definite then we say B is less or equal to (when $A - B = \Phi$) than A.

From (1.2) it follows that

$$p(x) = \sum_{j=1}^{m-1} \alpha_j p_j(x) + \left(1 - \sum_{j=1}^{m-1} \alpha_j\right) p_m(x) \qquad (2.1.4a)$$

$$= \sum_{j=1}^{m-1} \alpha_j [p_j(x)-p_m(x)] + p_m(x) . \qquad (2.1.4b)$$

It follows from (2.1.1) that

$$S_j = \frac{p_j(x) - p_m(x)}{\sum_{j=1}^{m} \alpha_j p_j(x)}$$

$$= \frac{p_j(x)-p_m(x)}{p(x)} \qquad (2.1.5)$$

and

$$\frac{\partial S_j}{\partial \alpha_K} = - \frac{[p_j(x)-p_m(x)][p_k(x)-p_m(x)]}{[p(x)]^2} . \qquad (2.1.6)$$

Therefore, the information for α is given by

$$\Lambda(\alpha) \overset{\text{def}}{=} \left\{ -E\left[\frac{\partial S_j}{\partial \alpha_K} \right] \right\}_{(m-1)\times(m-1)} . \qquad (2.1.7)$$

Fisher's information can be seen as the information contained in a random variable X about the parameter α. This should be interpreted

as the extent to which, on the average, the accuracy of estimating the unknown parameter α can be increased as a result of the observed value x of the random variable X.

In the ensuing sections of this paper, information for α contained in unlabeled, classified and verified data, defined earlier will be ordered.

Above, information is defined in terms of unbiased estimators.

2.2 Likelihood Function. If $X_1,X_2,\ldots,X_n$ denotes a simple random sample from $p(x)$ defined by (1.2) then the likelihood function is

$$L_u(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i) \tag{2.2.1a}$$

$$= \prod_{\ell=1}^{n}\left[\sum_{j=1}^{m} \alpha_j\, p_j(X_i)\right] \tag{2.2.1b}$$

the likelihood function for unlabeled data.

Let $X_1,X_2,\ldots,X_n$ denote a simple random sample from $p(x)$ which has been classified according to a rule $R = (R_1,R_2,\ldots,R_m)$, then each data vector X_k, $k = 1,2,\ldots,n$ generates through classification new data defined by the random variable $Y_i(X_k)$, $i = 1,2,\ldots,m$, where

$$Y_i(X_k) = 1 \quad \text{if} \quad X_k \in R_i \tag{2.2.2}$$

$$= 0 \quad \text{if} \quad X_k \notin R_k$$

whose joint p.d.f. is for each X_k a multinomial

$$h_{Y_1 \ldots Y_m}(y_1(X_k),\ldots,y_m(X_k)) = \prod_{i=1}^{m} g_i^{y_i(X_k)} \qquad (2.2.3)$$

where

$$g_j = Pr[X_k \in R_j]$$

$$= \int_{R_i} p(x)dx$$

$$= \sum_{j=1}^{m} \alpha_j \int_{R_i} p_j(x)dx$$

$$= \sum_{j=1}^{m} \alpha_j p(i\,|\,j)\ ,$$

the probability of classifying $I(X_k)$ in Π_j .

The likelihood function for classified data follows from (2.2.3), and is

$$L_C = L(Y_1(X_1),\ldots,Y_m(X_1);\ldots;Y_1(X_n),\ldots,Y_m(X_n))$$

$$= \prod_{k=1}^{n} \prod_{i=1}^{m} g_i^{Y_i(X_k)}$$

$$= \prod_{k=1}^{n} \prod_{i=1}^{m} \left[\sum_{j=1}^{m} \alpha_j P(i|j) \right]^{Y_i(X_k)}$$

$$= \prod_{i=1}^{m} \left[\sum_{j=1}^{m} \alpha_j P(i|j) \right]^{N_i} \qquad (2.2.4)$$

where

$$N_i = \sum_{k=1}^{n} Y_i(X_k) \qquad (2.2.5)$$

the number of sample vectors in R_i .

Let $I_1(X_1), I_2(X_2), \ldots, I_n(X_n)$ denote a random sample whose labels are known with probability one, that is, the data has been verified, then

$$T_j(I_k) = 1 \quad \text{if} \quad I_k \in \Pi_j$$
$$= 0 \quad \text{if} \quad I_k \in \Pi_j \tag{2.2.6a}$$

then the p.d.f. of $T = (T_1, \ldots, T_m)^T$ for each I_k is

$$f_{T_1,\ldots,T_m}(t_1,\ldots,t_m) = \prod_{i=1}^{m} [\alpha_i]^{t_i(I_k)} \quad . \tag{2.2.6b}$$

The likelihood function of a verified sample is

$$L_v = L_v(T_1(I_1),\ldots,T_m(I_1);\ldots;T_1(I_m),\ldots,T_m(I_n))$$

$$= \prod_{k=1}^{n} \prod_{i=1}^{m} [\alpha_i]^{T_i(I_k)}$$

$$= \prod_{i=1}^{m} [\alpha_i]^{n_i} \tag{2.2.7}$$

where

$$n_i = \sum_{k=1}^{n} T_i(I_k) \ , \tag{2.2.8}$$

the number of individuals in the sample from Π_i .

2.3 Information for α Contained in Unlabeled Data.

Let the following denote the information for α contained in unlabeled data: $X_1, \ldots, X_n$:

$$\Lambda_u(\alpha) = n\left\{\Lambda^u_{ij}(\alpha)\right\}_{(m-1)\times(m-1)} .$$

Using (2.1.2), (2.2.1b) and synthetic division, it can be shown that for $i = j$

$$\Lambda^u_{ij} = \left(\frac{\alpha_i+\alpha_m}{\alpha_i\alpha_m}\right)\left[1 - (\alpha_i+\alpha_m)B_{im} - \frac{\alpha_m}{(\alpha_i+\alpha_m)}\sum_{\substack{k=1\\k\neq i}}^{m-1}\alpha_k B_{ij} - \frac{\alpha_i}{(\alpha_i+\alpha_m)}\sum_{\substack{j=1\\j\neq i}}^{m=1}\alpha_j B_{jm}\right] \tag{2.3.1a}$$

and for $i \neq j$

$$\Lambda^u_{ij} = \frac{1}{\alpha_m}\left[1 - (\alpha_i+\alpha_m)B_{im} - (\alpha_j+\alpha_m)B_{jm} - \sum_{\substack{k=1\\k\neq i,j}}^{m-1}\alpha_k B_m + \alpha_m B_{ij}\right] \tag{2.3.1b}$$

where

$$0 \leq B_{ij} = \int_{\mathbb{R}^p} \frac{p_i(x)p_j(x)}{p(x)}\,dx \leq 1 \tag{2.3.1c}$$

and $B_{jk} = B_{kj}$, for all $j \neq k$.

When $B_{ij} = B$,

$$\Lambda_u(\alpha) = n(1-B)\{\Lambda^u_{ij}\} \tag{2.3.2a}$$

where

$$\Lambda^u_{ij} = \frac{\alpha_i+\alpha_m}{\alpha_i\alpha_m} \quad \text{for} \quad i = j \tag{2.3.2b}$$

$$= \frac{1}{\alpha_m} \quad \text{for} \quad i \neq j \ . \tag{2.3.2c}$$

When $m = 3$, the p.d.f. of a random variable X from a mixture population (unlabeled data) is

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) + \alpha_3 p_3(x) \tag{2.3.3a}$$

where

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{2.3.3b}$$

and

$$\alpha_1 \geq 0 \ , \ \alpha_2 \geq 0 \ , \ \alpha_3 \geq 0 \ . \tag{2.3.3c}$$

It follows from (2.3.1a) - (2.3.1c) that the information contained in unlabeled data is given by

$$\Lambda_u(\alpha) = \begin{bmatrix} \Lambda^u_{11} & \Lambda^u_{12} \\ \\ \Lambda^u_{21} & \Lambda^u_{22} \end{bmatrix}$$

where

$$\Lambda^u_{11} = \frac{(1-\alpha_2)}{\alpha_2\alpha_3}\left[1 - \frac{\alpha_2\alpha_3}{1-\alpha_2} B_{12} - (1-\alpha_2)B_{13} - \frac{\alpha_1\alpha_2}{1-\alpha_2} B_{23}\right] \tag{2.3.4a}$$

$$\Lambda^u_{22} = \frac{(1-\alpha_1)}{\alpha_2\alpha_3}\left[1 - \frac{\alpha_1\alpha_3}{1-\alpha_1} B_{12} - \frac{\alpha_1\alpha_2}{1-\alpha_1} B_{13} - (1-\alpha_1)B_{23}\right] \tag{2.3.4b}$$

$$\Lambda_{12}^{u} = \Lambda_{21}^{u} = \frac{1}{\alpha_3}\left[1+\alpha_3 B_{12} - (1-\alpha_2)B_{13} - (1-\alpha_1)B_{23}\right] \quad . \qquad (2.3.4c)$$

Note that one minus (2.3.1c) can be regarded as a distance measure. That is, when the i^{th} and j^{th} populations are "close together" or "far apart" then $(1-B_{ij})$ will be small or large, respectively. In fact, several investigators [3], [5], [6] , have employed a form of (2.3.1c) as a probabilistic distance measure for feature selection. While Cover and Hart [8] have shown that $2\alpha_i\alpha_j B_{ij}$ corresponds to the asymptotic nearest neighbor probability of error, this motivates a possible estimating procedure (see section 4.) using a nearest neighbor procedure.

It is of interest to consider the behavior of B_{ij} in terms of a popular distance measure as the distance between the i^{th} and j^{th} populations diverges. This behavior is described in Lemma 2.3.1.

Lemma 2.3.1: Let the distance measure between the i^{th} and j^{th} populations be given by

$$\Delta_{ij} = \int \left[p_i(x) - p_j(x)\right] \log\left[\frac{p_i(x)}{p_j(x)}\right] dx \quad . \qquad (2.3.5)$$

If $\Delta_{ij} \to \infty$ for all $i \neq j$, then $B_{ij} \to 0$.

Proof: Toussant [4] has shown that

$$0 \leq B_{ij} \leq \frac{1}{2}\left(\frac{\Delta_{ij}}{4}\right)^{-\frac{1}{4}} \quad .$$

Note that as $\Delta_{ij} \to \infty$ then

$$\left(\frac{\Delta_{ij}}{4}\right)^{-\frac{1}{4}} \to 0 \; .$$

Note that (2.3.5) is known as the divergence between two distributions. For normal distributions with equal covariances, (2.3.5) reduces to the well known Malhanabis distance.

The following example can clarify some of the concepts introduced above:

Example 2.3.1:

$$P_1(x) = \begin{cases} x & , 0<x<1 \\ 2-x, & 1<x<2 \\ 0 & , \text{o.w.} \end{cases} \quad , p_2(x) = \begin{cases} x-1, & 1<x<2 \\ 3-x, & 2<x<3 \\ 0 & , \text{o.w.} \end{cases} \quad , p_3(x) = \begin{cases} x-2, & 2<x<3 \\ 4-x, & 3<x<4 \\ 0 & , \text{o.w.} \end{cases}$$

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) + \alpha_3 p_3(x) \quad .$$

Let $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$ then

$$B_{12} = \int_0^4 \frac{p_1(x)p_2(x)}{p(x)}\, dx = \int \frac{(2-x)(x-1)}{\frac{1}{3}(2-x+x-1)}\, dx$$

$$= \int_1^2 3(2-x)(x-1)dx$$

$$= \int_1^2 (3x-2-x^2)dx$$

$$= \frac{1}{2}$$

$$B_{23} = 3\int_2^3 (3-x)(x-2)dx = \frac{1}{2}$$

$$B_{13} = 0 \ .$$

$$\therefore \Lambda(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}) = \begin{bmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{7}{2} \end{bmatrix}$$

To conclude this section, a result that follows from Lemma 2.3.1 is given.

Theorem 2.3.1: Let Δ_{ij} be a distance measure defined by (2.3.5). If $\Delta_{ij} \to \infty$ for all $i \neq j$ then,

$$\Lambda_u(\alpha) \to \Lambda_V(\alpha) = n\{\Lambda^V_{ij}\}$$

where

$$\Lambda^V_{ij} = \begin{cases} \dfrac{\alpha_i + \alpha_m}{\alpha_i \alpha_m} & \text{for } i = j \\ \dfrac{1}{\alpha_m} & \text{for } i \neq j \end{cases} .$$

Proof: Using equations (2.3.1a) - (2.3.1c) and letting $\Delta_{ij} \to \infty$, the Theorem follows from Lemma 2.3.1.

Note that (2.3.2a) can be written as

$$\Lambda_u(\alpha) = n(1-B)\Lambda_V(\alpha) . \tag{2.3.6}$$

The information matrix $\Lambda_V(\alpha)$ is the information for α contained in verified data. This is a topic of the next section.

2.4 Information for α Contained in Verified Data

Let $T_i(I_k)$ be defined as in (2.2.6a). It follows from (2.2.7) that

$$\ell n\, L_V = \ell n \left[\prod_{i=1}^{m} \alpha_i^{n_i} \right]$$

$$= \sum_{i=1}^{m} n_i \, \ell n[\alpha_i] \tag{2.4.1}$$

$$= \sum_{i=1}^{m-1} n_i \, \ell n[\alpha_i] + n_m \, \ell n \left[1 - \sum_{j=1}^{m-1} \alpha_j\right] ,$$

since $\sum_{j=1}^{m} \alpha_j = 1$.

From (2.1.1) then $S_j = \frac{\partial \ell n\, L}{\partial \alpha_j}$ it follows that

$$S_V = \frac{\partial \ell n\, L_V}{\partial \alpha} = \{S_j^V\}$$

where

$$S_j^V = \frac{\partial}{\partial \alpha_j}\left[\sum_{i=1}^{m} n_i \ell n\, \alpha_i\right] \tag{2.4.2}$$

$$= \frac{n_j}{\alpha_j} - \frac{n_m}{\alpha_n} \quad , j = 1,\ldots,m-1 .$$

In matrix notation

$$S_V = \Delta_\alpha \bar{n} \tag{2.4.3}$$

where the $(m-1)\times m$ matrix Δ_α is given by

$$\Delta_\alpha = \begin{bmatrix} \frac{1}{\alpha_1} & 0 & 0 & \ldots & 0 & 0 & -\frac{1}{\alpha_m} \\ 1 & \frac{1}{\alpha_2} & 0 & \ldots & 0 & 0 & -\frac{1}{\alpha_m} \\ \cdot & & & & & \cdot & \cdot \\ \cdot & & & & & \cdot & \cdot \\ \cdot & & & & & 0 & \cdot \\ 0 & 0 & 0 & \ldots & 0 & \frac{1}{\alpha_{m-1}} & -\frac{1}{\alpha_m} \end{bmatrix} \tag{2.4.4}$$

and

$$\bar{n} = (n_1,\ldots,n_m)^T .$$

Note that by the Cramer-Rao theory the expected value of S is the zero vector which we will verify directly.

$$E[S_V] = E[\Delta_\alpha \bar{n}]$$

$$= \Delta_\alpha E[\bar{n}]$$

$$= n\, \Delta_\alpha\, \alpha \quad \text{since} \quad n_j \sim \text{multinomial}\ (n,\alpha_j) \ \text{for} \ j = 1,\ldots,m\ .$$

Now,

$$\Delta_\alpha \alpha = \begin{bmatrix} \frac{1}{\alpha_1} & 0 & 0 & \ldots & 0 & -\frac{1}{\alpha_m} \\ 0 & \frac{1}{\alpha_1} & 0 & \ldots & 0 & -\frac{1}{\alpha_m} \\ \vdots & & & & \vdots & \vdots \\ 0 & \ldots & \ldots & \ldots & \frac{1}{\alpha_{m-1}} & -\frac{1}{\alpha_m} \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \Phi \qquad (2.4.5)$$

Thus,

$$E[S_V] = \Phi \quad . \qquad (2.4.6)$$

The information matrix for α when sampling from verified data can now be computed by finding the covariance matrix $V(S_V)$ of S_V using (2.4.3) and (2.4.6), that is,

$$\Lambda_V(\alpha) = V(S)$$

$$= \Delta_\alpha\, V(\bar{n})\, \Delta_\alpha^T \qquad (2.4.7)$$

where $V(\bar{n})$ is the covariance matrix of the $\bar{n} = (n_1,\ldots,n_m)^T$, a multinomial vector variate; that is,

$$V(n) = n[\mathrm{Diag}(\alpha_1,\ldots,\alpha_m) - \alpha\alpha^T] \quad . \tag{2.4.8}$$

From (2.4.7), (2.4.8) and (2.4.5),

$$\begin{aligned}\Lambda_V(\alpha) &= \Delta_\alpha[\mathrm{Diag}(\alpha_1,\ldots,\alpha_m) - \alpha\alpha^T]\,\Delta_\alpha^T \\ &= \Delta_\alpha[\mathrm{Diag}(\alpha_1,\ldots,\alpha_m)]\,\Delta_\alpha^T \quad . \end{aligned} \tag{2.4.9}$$

For exemplary purposes consider the case when $m = 3$, then since

$$\Delta_\alpha = \begin{pmatrix} \frac{1}{\alpha_1} & 0 & -\frac{1}{\alpha_3} \\ 0 & \frac{1}{\alpha_2} & -\frac{1}{\alpha_3} \end{pmatrix} ,$$

$$\Lambda_V(\alpha) = \begin{pmatrix} \frac{1}{\alpha_1} + \frac{1}{\alpha_3} & \frac{1}{\alpha_3} \\ \frac{1}{\alpha_3} & \frac{1}{\alpha_2} + \frac{1}{\alpha_3} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\alpha_1+\alpha_3}{\alpha_1\alpha_3} & \frac{1}{\alpha_3} \\ \frac{1}{\alpha_3} & \frac{\alpha_2+\alpha_3}{\alpha_2\alpha_3} \end{pmatrix} , \tag{2.4.10}$$

Suppose we are given an unlabeled sample

$$X_1,\ldots,X_n \quad .$$

Then we verify this sample generating the sample

$T_1,\ldots,T_n$, where $T_i = (T_{i1},\ldots,T_{im})^T$.

For estimating α_1 should we disregard the unlabeled sample or consider the joint sample (X_i,T_i) , $i = 1,\ldots,n$? The joint p.d.f. of (X_i,T_i) , $i = 1,\ldots,n$ is

$$p(x_i,t_i) = p(x_i|t_i)p(t_i) \ , \ t_i = (t_{i1},\ldots,t_{im})$$

$$= \prod_{j=1}^{m} [p_j(x_i)]^{t_{ij}} \prod_{j=1}^{m} [\alpha_j]^{t_{ij}}$$

$$= \prod_{j=1}^{m} [\alpha_j \ p_j(x_i)]^{t_{ij}} \ . \qquad (2.4.11)$$

To answer the above question consider the following theorem.

Theorem 2.4.1: The amount of information for α contained in the observation (x_i,t_i) is equal to the information for α contained in the observation t_i alone.

Proof: Taking the logs of both sides of the equality in (2.4.11), we see that

$$\ell n \ p \ (x_i,t_i) = \sum_{j=1}^{m} t_{ij} \ \ell n \ p_j(x_i) + \sum_{j=1}^{m} t_{ij} \ \ell n \ \alpha_j \ .$$

Now taking derivative with respect to α_j we have

$$\frac{\partial \ \ell n \ p(x_i,t_i)}{\partial \ \alpha_j} = 0 + \frac{\partial \sum_{j=1}^{m} t_{ij} \ \ell n \ \alpha_j}{\partial \ \alpha_j} = \frac{\partial \ \ell n \ p(t_i)}{\partial \ \alpha_j} \ .$$

Therefore,

$$-E\left[\frac{\partial^2 \ell n p(x_i,t_i)}{\partial \ \alpha_j^2}\right] = - E\left[\frac{\partial^2 \ell n p(t_i)}{\partial \ \alpha_j}\right] \ .$$

Thus, it follows from Theorem 2.4.1 that for estimating α the joint sample (X_i,T_i), $i = 1,\ldots,n$ contains no more information than the sample $T_1,\ldots,T_n$ alone.

2.5 Information for α Contained in Classified Data.

Using the likelihood function given in (2.2.4) for a random sample defined in (2.2.2), it follows that

$$\ell n\, L_c = \sum_{i=1}^{m} N_i\, \ell n\, g_i$$

$$= \sum_{i=1}^{m-1} N_i\, \ell n\, g_i + \left(N - \sum_{i=1}^{m-1} N_i\right) \ell n \left[1 - \sum_{i=1}^{m-1} g_i\right]$$

since

$$\sum_{i=1}^{m} g_i = 1 .$$

Also, from (1.3.6) and $\sum_{i=1}^{m} \alpha_i = 1$ that

$$g_i = \sum_{j=1}^{m-1} \alpha_j \Big[P(i|j) - P(i|m)\Big] + P(i|m) \qquad (2.5.1)$$

and

$$\frac{\partial\, g_i}{\partial\, \alpha_j} = P(i|j) - P(i|m) . \qquad (2.5.2)$$

From (2.1.1) and $S_j^c = \dfrac{\partial\, \ell n\, L_c}{\partial\, \alpha_j}$ it follows that

$$S_j^c = \sum_{i=1}^{m} N_i\, \frac{1}{g_i} \Big[P(i|j) - P(i|m)\Big] \qquad (2.5.3)$$

or in matrix notation

$$S_c = [\Delta_{ij}]^T G^{-1} \bar{N} \tag{2.5.4}$$

where the $(m-1)\times m$ matrix $[\Delta_{ij}]^T$ is defined by its elements

$$\Delta^*_{ij} = P(i|j) - P(i|m) , \tag{2.5.5}$$

$$G = \begin{bmatrix} g_1 & 0 & 0 & \cdots & 0 \\ 0 & g_2 & 0 & \cdots & 0 \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & & \cdot \\ 0 & \cdots & & & g_m \end{bmatrix} , \tag{2.5.6}$$

and

$$N = (N_1, N_2, \ldots, N_m)^T . \tag{2.5.7}$$

Note that by the Cramer-Rao theory the expected value of S_c is the zero vector which we will verify directly.

$$\begin{aligned} E[S_c] &= E[\Delta^*_{ij}]^T G^{-1} \bar{N} \\ &= [\Delta^*_{ij}]^T G^{-1} E[\bar{N}] \\ &= [\Delta^*_{ij}]^T G^{-1}(Ng) , \end{aligned} \tag{2.5.8}$$

where

$$g = (g_1, g_2, \ldots, g_m)^T$$

or

$$g = GJ \tag{2.5.9}$$

where

$$J = (1,1,\ldots,1)^T .$$

It follows from

$$\sum_{i=1}^{m} P(i|j) = 1$$

for $j = 1,2,\ldots,m$ that

$$[\Delta^*_{ij}]J = \Phi \tag{2.5.10}$$

and in turn from (2.5.8) and (2.5.9) that

$$E[S_c] = N[\Delta^*_{ij}]\, G^{-1}\, GJ = \Phi \ . \tag{2.5.11}$$

The covariance matrix $V(S_c)$ of S_c can now be computed using (2.5.4) and (2.5.11), that is

$$V(S_c) = [\Delta^*_{ij}]^T\, G^{-1}\, V(\bar{N})\, G^{-1}[\Delta^*_{ij}] \tag{2.5.12}$$

where $V(\bar{N})$ is the covariance matrix of the $\bar{N} = (N_1,N_2,\ldots,N_m)$, a multinomial vector variate, that is

$$\begin{aligned} V(\bar{N}) &= N[G - GJJ^T G] \\ &= NG(I - JJ^T G) \\ &= N[G - P\alpha\alpha^T P] \end{aligned} \tag{2.5.13}$$

where

$$G = \begin{bmatrix} P_1\alpha & 0 & \cdots & 0 \\ 0 & P_2\alpha & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & P_{m-1}\alpha \end{bmatrix} .$$

From (2.5.10), (2.5.12), and (2.5.13)

$$\Lambda_c(\alpha) = V(S_c) = N[\Delta^*_{ij}]^T G^{-1}[\Delta^*_{ij}] , \qquad (2.5.14)$$

the information for α contained in classified data.

For completeness we state the following theorem.

Theorem 2.5.1:

$$\Lambda_c(\alpha) \to \Lambda_v(\alpha) \quad \text{as} \quad P \to I$$

where

$$P = \{P(i|j)\} .$$

Proof: In matrix notation,

$$g = P\alpha .$$

Let $P \to I$, then $g \to \alpha$ and

$$\Delta^*_{ij} \to \Delta^v_{ij} = \begin{cases} 1 & \text{for } i = j \neq m \\ -1 & \text{for } i = m \\ 0 & \text{o.w.} \end{cases}$$

that is,

$$\Delta^*_{ij} \to \begin{pmatrix} I_{m-1} \\ \hline -J^T_{m-1} \end{pmatrix} \quad .$$

Note that (2.4.9) can be written as

$$\Lambda_v(\alpha) = [I_{m-1} \,\vdots\, -J_{m-1}][\mathrm{Diag}(\tfrac{1}{\alpha_1},\ldots,\tfrac{1}{\alpha_m})]\begin{pmatrix} I_{m-1} \\ \hline -J^T_{m-1} \end{pmatrix} \qquad (2.5.15)$$

where I_{m-1} is a $(m-1)\times(m-1)$ identity matrix and

$$-J_{m-1} = \underbrace{(-1,-1,\ldots,-1)}_{m-1}^T \quad .$$

Thus,

$$\Lambda_c(\alpha) = [\Delta^*_{ij}]^T\, G^{-1}[\Delta^*_{ij}] \quad [\Delta^v_{ij}]^T[\mathrm{diag}(\tfrac{1}{\alpha_1},\ldots,\tfrac{1}{\alpha_m})][\Delta^v_{ij}] = \Lambda_v(\alpha)$$

as $P \to I$.

For exemplory purposes consider the case when $m = 2$, then since

$$[\Delta_{ij}]^T = [P(1|1) - P(1|2)\ ,\ P(2|1) - P(2|2)]\ ,$$

$$G = \begin{bmatrix} g_1 & 0 \\ 0 & g_2 \end{bmatrix} \quad ,$$

$$g_1 = 1 - g_2\ ,$$

$$P(1|1) = 1 - P(2|1) \quad \text{and}$$

$$P(2|2) = 1 - P(1|2)\ ,$$

then

$$\Lambda_{11}^{c}(\alpha) = \frac{N[P(1|1)-P(1|2)]^2}{g_1 g_2} \quad . \tag{2.5.16}$$

Suppose further, that if there are no errors in classification, that is,

$$P(1|1) = P(2|2) = 1$$

then

$$g_1 = \alpha_1 \quad \text{and} \quad g_2 = \alpha_2$$

and

$$\Lambda_c(\alpha) = \frac{N}{g_1 g_2} = \frac{N}{\alpha_1 \alpha_2} = \Lambda_v(\alpha) \ .$$

Note that for this case, $\Lambda_c^{-1}(\alpha)$ is the variance of a sufficient statistic $\hat{\alpha}_1 = \frac{N_1}{N}$ for α_1 in a binomial probability density function.

III. THE MAIN RESULT

3.1 The Ordering of the Information for α.

For the two population case $(m=2)$, the information for contained in unlabeled, verified and classified data are given respectively by

$$\Lambda_u(\alpha) = \frac{N(1-B)}{\alpha_1\alpha_2} \quad , \quad \text{where} \quad B = \int_{\mathbb{R}^p} \frac{p_1(x)p_2(x)}{p(x)}\, dx \tag{3.1.1a}$$

$$\Lambda_v(\alpha) = \frac{N}{\alpha_1\alpha_2} \quad , \tag{3.1.1b}$$

and

$$\Lambda_c(\alpha) = \frac{N[P(1|1)-P(1|2)]^2}{g_1 g_2} \quad . \tag{3.1.1c}$$

The similarity of Λ_v, Λ_c and Λ_u is striking and one notes in this case an obvious ordering exists, that is

$$\Lambda_v(\alpha) \geq \Lambda_c(\alpha) \tag{3.1.2a}$$

and

$$\Lambda_v \geq \Lambda_u(\alpha) \quad . \tag{3.1.2b}$$

The inequality (3.1.2a) holds since

$$\Lambda_c(\alpha) = \frac{N}{g_1 g_2}[P(1|1) - P(1|2)]^2$$

$$= \frac{N[P(1|1) - P(1|2)]^2}{[\alpha_1 P(1|1) + \alpha_2 P(1|2)][1-g_1]} \quad .$$

However,

$$g_1 = \alpha_1 P(1|1) + (1-\alpha_1)P(1|2)$$

$$g_2 = 1 - g_1$$

implies

$$g_1 g_2 = \alpha_1(1-\alpha_2)[P(1|1)-P(1|2))^2 + \frac{1}{\alpha_2} P(1|1)[1-P(1|1)]$$

$$+ \frac{1}{\alpha_1} P(1|2)[1-P(1|2)]] \quad .$$

Let

$$R_c = \frac{[P(1|1)-P(1|2)]^2}{\frac{g_1 g_2}{\alpha_1(1-\alpha_1)}} \quad . \qquad (3.1.3)$$

Since $0 \leq R_c \leq 1$, one can conclude for $m = 2$, that

$$\Lambda_c(\alpha) = \frac{N}{\alpha_1(1-\alpha_1)} \; R_c$$

or

$$\Lambda_c(\alpha) \leq \frac{N}{\alpha_1(1-\alpha_1)} = \Lambda_v(\alpha) \quad .$$

From (2.6.1a) and the fact that

$$0 \leq R_u \overset{\text{def}}{=} 1 - B \leq 1 \qquad (3.1.4)$$

implies that (3.1.2b) holds, that is, for $m=2$

$$\Lambda_u(\alpha) \leq \Lambda_v(\alpha) \ .$$

In this section, we will establish the following ordering of the information for α :

$$\Lambda_c(\alpha) \leq \Lambda_u(\alpha) \leq \Lambda_v(\alpha) \ . \qquad (3.1.5)$$

(Note that if A and B are two positive definite matrices of the same size and $A - B$ is positive semi-definite then we say "B is less than A".) This result will be given in a corollary to a Theorem proved by Rao [3].

Note that classified data defined in (2.2.2) is a explicit transformation of the unlabeled data. Knowing this, it follows directly from the following Theorem due to Rao [3] that

$$\Lambda_c(\alpha) \leq \Lambda_u(\alpha) \ .$$

Theorem 3.1.1 (Rao): The matrix $\Lambda_X - \Lambda_T$ is semi-positive definite, where Λ_T is the information matrix in a measureable function T of X .

The ordering between Λ_v with Λ_u and Λ_c is not as straightforward. The ordering (3.1.5) is proved in corollary 3.1.1 which will be proved very similarily to the proof of Theorem 3.1.1 once the following three lemmas are proved.

Suppose one takes an unlabeled sample and then classifies it, then let

$$Z = (X^T, Y(X)) \quad , \quad Y(X) = (Y_1(X), \ldots, Y_m(X))$$

when $Y_j(X) = 1,0$ if $x \in R_j$, $x \notin R_j$ respectively.

Lemma 3.1.1: The p.d.f. for Z is given by

$$p_Z(z) = \begin{cases} p_X(x) & , \text{ if } X \in R_j \text{ and } y_j = 1 \text{ for some } j = 1,\ldots,m \\ 0 & , \text{ o.w.} \end{cases} \tag{3.1.6}$$

Proof:

$$p_Z(z) = p(x,y)$$

$$= Pr(Y(x) = y | X=x)\ p_X(x)$$

Now (3.1.6) follows from

$$P_r(Y(X) = y \mid X=x) = \begin{cases} 1 & \text{if } X \in R_j \text{ and } y_j = 1 \text{ for some } j = 1,\ldots,m \\ 0 & \text{o.w.} \end{cases}$$

since $P_r(Y_j(x) = 1 \text{ and } Y_k(x) = 1) = 0$ for $j \neq k$.

Recall from Sections 2.3 - 2.5 that

$$S_u = \{S_j^u\} , \tag{3.1.7a}$$

$$S_v = \{S_j^v\} , \tag{3.1.7b}$$

$$S_c = \{S_j^c\} , \tag{3.1.7c}$$

for $j = 1,\ldots,m-1$

where

$$S_j^u = \frac{p_j(x) - p_m(x)}{p(x)} \tag{3.1.8a}$$

$$S_j^v = \frac{T_j}{\alpha_j} - \frac{T_m}{\alpha_m} \tag{3.1.8b}$$

$$S_j^c = \sum_{i=1}^{m} \frac{Y_i}{g_i} \Delta_{ij} \tag{3.1.8c}$$

for $j = 1,\ldots,m-1$.

Furthermore, we know that

$$E\, S_u = E\, S_v = E\, S_C = \Phi \tag{3.1.9}$$

Lemma 3.1.2:

$$\text{(i)} \quad E[S_u|Y=y] = S_c \tag{3.1.10a}$$

$$\text{(ii)} \quad E[S_v|X=x] = S_u \tag{3.1.10b}$$

$$\text{(iii)} \quad E[S_v|Y=y] = S_c \ . \tag{3.1.10c}$$

Proof:

(i) For each $j = 1,\ldots,m-1$, it follows from (3.1.8a) that

$$E[S_j^u|Y=y] = \int \frac{p_j(x)-p_m(x)}{p(x)} \cdot \frac{p(x,y)}{h(y)}\, dx \ .$$

Let

$$Y = y_{(k)} = (0,\ldots,0,1_k,0,\ldots,0)$$

where 1_k indicates that $y_k = 1$. Then it follows from Lemma 3.1.1 that

$$E[S_j^u|Y=y_{(k)}] = \int_{R_k} \frac{p_j(x)-p_m(x)}{p(x)} \frac{p(x)}{g_k} dx =$$

$$= \frac{1}{g_k} [P(k|j)-P(k|m)]$$

$$= \frac{\Delta_{kj}}{g_k} .$$

(Note that $g_k = h(y_{(k)})$.)

Thus, in general we have

$$E[S_j^u|Y=y] = \sum_{k=1}^{m} \frac{y_k}{g_k} \Delta_{kj} = S_j^c , \quad j = 1,2,\ldots,m-1 .$$

(ii) For each $j = 1,\ldots,m-1$, it follows from (3.1.8b) that

$$E[S_j^v|X=x] = \sum_{\{t|p(t|x)>0\}} \frac{t_j}{\alpha_j} - \frac{t_m}{\alpha_m} f(t|x)$$

$$= \frac{f(t_{(j)}|x)}{\alpha_j} - \frac{f(t_{(m)}|x)}{\alpha_m}$$

where

$$t_{(k)} = (0,\ldots,0,1_k,0,\ldots,0) .$$

Note that $f(t|x) = \frac{f(t,x)}{p(x)} = \frac{p(x|t)f(t)}{p(x)}$.

Hence, it follows that

$$E[S_j^v|X = x] = \frac{\alpha_j p_j(x)}{\alpha_j p(x)} - \frac{\alpha_m p_m(x)}{\alpha_m p(x)}$$

$$= \frac{p_j(x)-p_m(x)}{p(x)} = S_j^u \text{ , for } j = 1,\ldots,m-2 .$$

(iii) Suppose $y = y_{(k)}$ for $k = 1,\ldots,m$, then for $j = 1,\ldots,m-1$ it follows from (3.1.8b) that

$$E[S_j^v|Y=y_{(k)}] = \frac{f(t_{(j)}|y_{(k)})}{\alpha_j} - \frac{f(t_{(m)}|y_{(k)})}{\alpha_m} .$$

It can be easily shown as follows:

$$f(t_{(j)}|y_{(t)} = \frac{f(t_{(j)},y_{(k)})}{h(y_{(k)})}$$

$$= \frac{h(y_{(k)}|t_{(j)})f(t_{(j)})}{h(y_{(k)})}$$

$$= \frac{Pr(Y_{(k)}=1|t_{(j)}=1)_j}{g_k}$$

$$= \frac{P(k|j)\alpha_j}{g_k}$$

$$= Q(j|k)$$

Thus,

$$E[S_j^v|Y=y_{(k)}] = \frac{Q(j|k)}{\alpha_j} - \frac{Q(m|k)}{\alpha_m}$$

$$= \frac{\alpha_j P(k|j)}{\alpha_j g_k} - \frac{\alpha_m P(k|m)}{\alpha_m g_k}$$

$$= \frac{1}{g_k}[P(k|j) - P(k|m)]$$

$$= \frac{\Delta_{kj}}{g_k} .$$

In general, we have

$$E(S_j^v|Y=y) = \sum_{i=1}^{m} \frac{y_i \Delta_{ij}}{g_i} = S_j^c , \quad \text{for} \quad j = 1,\ldots,m-1.$$

Lemma 3.1.3: (i) $E(S_c S_u^T) = \Lambda_c$

(ii) $E(S_u S_v^T) = \Lambda_u$

(iii) $E(S_c S_v^T) = \Lambda_c$.

Proof:

(i) $E(S_c S_u^T) = E\{E(S_c S_u^T|Y=y)\}$

$= E\{S_c E(S_u^T|Y=y)\}$.

It follows from Lemma (3.1.2) that

$= E\{S_c S_c^T\} = \Lambda_c$.

(ii) and (iii) are similarly proved.

Corollary 3.1.1:

(i) $\Lambda_u - \Lambda_c = D_1$

(ii) $\Lambda_v - \Lambda_u = D_2$

(iii) $\Lambda_v - \Lambda_c = D_3$

where D_1, D_2 and D_3 are positive semi-definite matrices.

Proof:

(i) Since $ES_c = ES_u = \Phi$, then by definition, the covariance matrix of $S_u - S_c$ is given by

$$E(S_u - S_c)(S_u - S_c)^T . \qquad (3.1.11)$$

Now (3.1.11) can be written as

$$E(S_u S_u^T - S_u S_c^T + S_c S_u^T + S_c S_c^T) = ES_u S_u^T - ES_u S_c^T - ES_c S_c^T + ES_c S_c^T .$$

It follows from Lemma (3.1.3) that

$$E(S_u - S_c)(S_u - S_c)^T = \Lambda_u - \Lambda_c^T - \Lambda_c + \Lambda_c$$

$$= \Lambda_u - \Lambda_c^T$$

$$= \Lambda_u - \Lambda_c \text{ , since } \Lambda_c \text{ is symmetric.}$$

Since by definition, (3.1.11) is positive semi-definite, then $\Lambda_u - \Lambda_c$ is positive semi-definite.

(ii) and (iii) are similarly proved.

Table 4.1. Approximate Values of Λ_u, Λ_B, Λ_{MLE} in a Mixture of Normals

α_1 \ Δ	$\Delta = 1/4$			$\Delta = 1/3$			$\Delta = 1/2$			$\Delta = 1$			$\Delta = 2$			$\Delta = 3$			$\Delta = \infty$
	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_V
.1	.063	.000	.040	.114	.000	.071	.263	.000	.160	1.15	.226	.647	4.56	3.36	2.66	7.98	7.21	5.78	11.1
.2	.063	.000	.040	.111	.000	.071	.250	.010	.158	.967	.435	.619	3.06	2.46	2.24	4.80	4.43	4.11	6.25
.3	.062	.000	.040	.110	.006	.070	.242	.060	.157	.875	.533	.601	2.51	2.08	2.01	3.76	3.50	3.41	4.76
.4	.062	.015	.040	.108	.042	.070	.237	.127	.156	.830	.575	.590	2.27	1.91	1.90	3.33	3.11	3.10	4.17
.5	.062	.040	.040	.108	.070	.070	.236	.156	.156	.816	.587	.587	2.20	1.86	1.85	3.21	3.00	3.00	4.00

$\Delta = (\mu_2 - \mu_1)/\sigma$, $(\sigma = 1)$

Λ_u = Unlabeled Information

Λ_B = Bayes Classified Information

Λ_{MLE} = MLE Classified Information

In Table 4.2 values of information are given for various values of Δ , k and α_1 when $\sigma_2^2 = k\sigma_1^2$ and p(x) is a mixture of two univariate normal p.d.f. The value selected for $\sigma_1^2 = 1$ and $n = 1$.

4.2 Conclusions

The surprising result that classified data has the least information is especially important since current practice in processing remote sensed data is to classify the unlabeled data. It is true that it may be easier to classify than compute the maximum likelihood estimates for α using unlabeled data. Hence classifying the data would be a necessary task. The information in classfied data is nearly equal to but always less than the information in unlabeled data.

Note also, if the expense to verify data is sufficiently small then the unlabeled data taken remotely from sapce is not needed. A random sample of locations on the earth's surface is sufficient to estimate α . The satellite data is of no value except to randomly select sites for verification.

If training data and test data are in reality classified data (that is, unlabeled data classified by photo interpreters) one can and should expect loss of information. However, training data and test data are in fact verified or labeled (ground truth with no classification error) one should expect better results in estimating α .

Table 4.2. Information Λ for Various Types of Data (v,u,c) Versus Values of the Parameters (k,Δ,α_1).

α_1	Type of Data	k = 1			k = 2		
		Δ = 1	2	3	1	2	3
0.1	v	11.11	11.11	11.11	11.11	11.11	11.11
	u	1.15	4.57	7.98	0.60	2.38	5.51
	c	0.65	2.66	5.78	0.47	1.68	3.79
0.3	v	4.76	4.76	4.76	4.76	4.76	4.76
	u	0.88	2.51	3.76	0.62	1.81	3.09
	c	0.60	2.01	3.41	0.48	1.48	2.69
0.5	v	4.00	4.00	4.00	4.00	4.00	4.00
	u	0.82	2.20	3.21	0.68	1.77	2.77
	c	0.59	1.86	3.00	0.61	1.47	2.50

6. REFERENCES

[1] MacDonald, R. B. and Hall, F. G., "Global Crop Forecasting," Science, Vol. 208, May 1980, pp. 670-679.

[2] Cramer, H., Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J., 1946.

[3] Rao, C. R., Linear Statistical Inference and Its Applications, John Wiley and Sons, New York, 1973.

[4] Toussaint, G. T., "On Some Measures of Information and Their Application to Pattern Recognition", Proceedings of Conference on Measures of Information and Their Applications, Indian Institute of Technology, Bombay, India, Aug. 16-18, 1974.

[5] Toussaint, G. T., "On the Divergence Between Two Distributions and the Probability of Misclassification of Several Decision Rules", Proceeding of the Second International Joint Conference on Pattern Recognition, Copenhagen, Aug., 1974.

[6] Trouborst, P. M., and et al., "New Families of Probabilistic Distance Measures", Proceedings of the Second International Joint Conference on Pattern Recognition, Copenhagen, Aug., 1974.

[7] Devijver, P. A., "On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition", IEEE Trans. Inform. Theory, Vol. C-23, No. 1 Jan., 1974, pp. 70-79.

[8] Cover, T. M. and Hart, P.E., "Nearest Neighbor Pattern Classification", IEEE Trans. Inform. Theory, Vol. IT-13, pp. 21-27, Jan., 1967.

Related to our definition of an information function is the concept of an information measure. Starting with the definition of Fisher information Mann (1949) derives three

$$u = \frac{p f_1}{p f_1 + (1-p) f_0}$$

$$\ln u = p f_1 +$$

$$= \ln p + \ln f_1 + \ln (p f_1 + (1-p) f_0)$$

$$\frac{d}{dp} = \frac{1}{p} + \frac{f_1 - f_0}{p f_1 - (1-p) f_0}$$

$$\frac{d^2}{} = \frac{-1}{p^2} + - \frac{(f_1 - f_0)^2}{[p f_1 + (1-p) f_0]^2}$$

$$= \frac{1}{p^2} + \int \frac{(f_1 - f_0)^2}{f}$$