Reprinted from

Symposium on Machine Processing of Remotely Sensed Data

June 3 - 5, 1975

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
75CH1009-0 -C

ACREAGE ESTIMATION, FEATURE SELECTION, AND SIGNATURE EXTENSION DEPENDENT UPON THE MAXIMUM LIKELIHOOD DECISION RULE

by J.A. QUIREIN

LACIE VERIFICATION DEPARTMENT
LOCKHEED ELECTRONICS COMPANY
AEROSPACE SYSTEMS DIVISION
16811 EL CAMINO REAL
HOUSTON, TEXAS 77058

M.C. TRICHEL

EARTH OBSERVATIONS DIVISION
SCIENCE AND APPLICATIONS DIRECTORATE
National Aeronautics and Space Administration
LYNDON B. JOHNSON SPACE CENTER
Houston, Texas

Prepared for

SYMPOSIUM ON MACHINE CLASSIFICATION OF REMOTELY SENSED DATA, LABORATORY FOR APPLICATIONS OF REMOTE SENSING, PURDUE UNIVERSITY, WEST LAFAYETTE, INDIANA

JUNE 3, 1975

ACREAGE ESTIMATION, FEATURE SELECTION, AND SIGNATURE EXTENSION DEPENDENT UPON THE MAXIMUM LIKELIHOOD DECISION RULE

John A. Quirein, M. C. Trichel

Lockheed Electronics Company, Inc.,
Aerospace Systems Division, Houston, Texas; and
Lyndon B. Johnson Space Center, National Aeronautics
and Space Administration, Houston, Texas

I. ABSTRACT

The maximum likelihood decision rule and estimation of the resulting m-class probability of misclassification are discussed. A bound on the variance of a proposed unbiased estimator of the m-class probability of error is derived. The problem of estimating the a priori probabilities for two classes is covered. When the estimator is counting the proportion of classified samples assigned to each class, a bound on the error of the estimate is derived. The problem of m-class feature selection using the Bhattacharyya distance is also addressed. The particular case in which each class density is assumed to be a mixture of multivariate normal densities is considered in detail. In conclusion, the extension of spectral signatures in space and time is also discussed.

II. THE MAXIMUM LIKELIHOOD DECISION RULE AND ESTIMATION OF THE RESULTING m-CLASS PROBABILITY OF MISCLASSIFICATION

The m-class probability of misclassification can be estimated using unlabeled test samples and labeled training samples. All prior probabilities are assumed to be known, and labeled training data are assumed to be available to construct an estimate for each class density function. The class prior probabilities and estimated density functions are used to obtain an estimate of the conditional risk of misclassification at each point of the unlabeled test samples. The average of the risk estimates is shown to be the probability of misclassification. An expression for the variance of the estimate resulting from finite test sample size is also derived.

A similar approach has been made (Fukunaga, Kessel, 1973; Minter, Thadani, to be published). However, in this paper, the results of Fukunaga and Kessel's paper are extended from two to m-classes. In addition, it is shown in a later section of this paper how the m-class estimate for the probability of misclassification can be used in feature selection.

Let $\overline{X}$ be a random n-dimensional measurement vector belonging to one of the m-classes $\omega_1, \omega_2, \cdots, \omega_m$. Let q_i be the prior probability of the *ith* class and $p_i(X)$ be the probability density function of the random vector $\overline{X}$ evaluated at X , and belonging to the class ω_i . Let the mixture density be given by

$$p(X) = \sum_{i=1}^{m} q_i p_i(X) \tag{1}$$

Let $[R_1, R_2, \cdots, R_m]$ be any partition of the underlying n-dimensional vector space such that any vector X is classified into ω_i only if X belongs to R_i. Define the characteristic function

$$\delta_{R_i}(X) = 1, \ X \text{ in } R_i$$

$$\delta_{R_i}(X) = 0, \ X \text{ not in } R_i \tag{2}$$

and the functions

$$\tilde{p}(X) = \sum_{i=1}^{m} q_i p_i(X) \ \delta_{R_i}(X) \tag{3}$$

$$r(X) = \frac{\tilde{p}(X)}{p(X)} \tag{4}$$

By definition, the probability of misclassification R for m-classes is given by

$$R = \sum_{i=1}^{m} \int_{R_i} \left[p(X) - q_i p_i(X)\right] dX$$

$$= 1 - \sum_{i=1}^{m} \int_{R_i} q_i p_i(X) \ dX \tag{5}$$

It follows from the definitions of $\delta_{R_i}(X)$, $\tilde{p}(X)$, and $r(X)$ that Equation (5) may be written as

$$R = 1 - \sum_{i=1}^{m} \int q_i p_i(X) \ \delta_{R_i}(X) \ dX$$

$$= 1 - \int \tilde{p}(X) \ dX$$

$$= 1 - \int \left[\frac{\tilde{p}(X)}{p(X)}\right] p(X) \ dX$$

$$= 1 - \int r(X) p(X) \ dX \tag{6}$$

where in the above, the region of integration is the entire measure space. Thus, the probability of misclassification is just the complement with respect to 1 of the expectation of $r(X)$ with respect to the random vector $\overline{X}$, so that

$$R = 1 - E[r(\overline{X})] \tag{7}$$

It is important to note that the probability of misclassification R can be estimated by the sample mean of $r(X_i)$ for N_t test samples as

$$\hat{R} = 1 - \frac{1}{N_t} \sum_{i=1}^{N_t} r\left(\overline{X}_i\right) \geq 0 \tag{8}$$

where $\hat{R}$ is a random variable, the $\overline{X}_i$'s are drawn from the mixture density $p(X)$ and the class identities of the $\overline{X}_i$ are not needed. Note that $\hat{R}$ is minimized in accordance with Bayes rule if R_i, $i = 1, \cdots, m$, is defined such that

$$X_j \text{ exist in } R_i \text{ if and only if } q_i p_i\left(X_j\right) \geq q_k p_k\left(X_j\right) \qquad k=1,\cdots,m \tag{9}$$

Equation (9) is the classical maximum likelihood decision rule (Anderson, 1958) so that both R and the estimate $\hat{R}$ of R are minimized by the same likelihood decision rule. Using a partition defined by (9), $\hat{R}$ is then an estimate of the Bayes error. Let X belong to R_i . Then

$$r(X) = \frac{\tilde{p}(X)}{p(X)} = \frac{q_i p_i(X)}{p(X)}$$

is the posterior probability; i.e., the conditional probability of a measurement X belonging to class ω_i , so that

$$\frac{1}{N_t} \sum_{i=1}^{N_t} r(X_i)$$

represents the average conditional probability for the unlabeled but classified test samples, and it follows from Equation (8) that the estimate of the probability of misclassification is minimized by maximizing the average conditional probability. Also, since $1 - r(X_i)$ is the conditional risk of misclassification, it follows from Equation (8) that the average of the risk estimates is the probability of misclassification.

Since the $\overline{X}_i$'s are independent and identically distributed random vectors, the $r(\overline{X}_i)$'s are independent and identically distributed random variables. Therefore, from Equation (7), the estimate of Equation (8) is unbiased as

$$E[\hat{R}] = 1 - \frac{1}{N_t} \sum_{i=1}^{N_t} E\left[r\left(\overline{X}_i\right)\right] = 1 - E[r(\overline{X})] = R \tag{10}$$

We now derive an expression for the variance of $\hat{R}$ for the particular case of a Bayes partition as defined by Equation (9). Since

$$0 \leq 1 - r(\overline{X}) \leq 1 - \frac{1}{m} \tag{11}$$

it follows

$$[1 - r(\overline{X})]^2 \leq \left[1 - \frac{1}{m}\right] [1 - r(\overline{X})] \tag{12}$$

and

$$E\left\{[1 - r(\overline{X})]^2\right\} \leq \left(1 - \frac{1}{m}\right) R \tag{13}$$

Therefore, a bound on the variance of $1 - r(\overline{X})$, $\sigma^2[1 - r(\overline{X})]$, is

$$\begin{aligned} \sigma^2[1 - r(\overline{X})] &= E\left\{[1 - r(\overline{X})]^2\right\} - E^2\left\{[1 - r(\overline{X})]\right\} \\ &= E\left\{[1 - r(\overline{X})]^2\right\} - R^2 \\ &\leq \left(1 - \frac{1}{m}\right) R - R^2 \\ &= R(1 - R) - \frac{1}{m} R \end{aligned} \tag{14}$$

which is the same as the expression derived for two classes except that m now appears in the denominator rather than 2 (Fukunaga, Kessel, 1973). Thus, the variance of $\hat{R}$ is given by

$$\mathrm{var}[\hat{R}] = \frac{\sigma^2[1 - r(\overline{X})]}{N_t} \tag{15}$$

and by Equation (14) satisfies

$$\mathrm{var}[\hat{R}] \le \frac{1}{N_t}\left[R(1 - R) - \frac{1}{m} R\right] \tag{16}$$

The results of this section are summarized below.

Theorem 1: Let $r(X)$ be integrable with respect to the mixture density function $p(X)$. Then the probability of misclassification R is given by

$$R = 1 - \int r(X)p(X)dX = 1 - E[r(\overline{X})] \tag{17}$$

so that the probability of misclassification is the complement with respect to 1 of the expectation of $r(\overline{X})$ with respect to the random vector $\overline{X}$.

For any partition $[R_1,\cdots,R_m]$, $\hat{R}$ is an unbiased estimator of R in that

$$E[\hat{R}] = R \tag{18}$$

Furthermore, if the partition is the Bayes partition defined by Equation 9, then the variance of $\hat{R}$ satisfies

$$\mathrm{var}[\hat{R}] \le \frac{1}{N_t}\left[R(1 - R) - \frac{1}{m} R\right] \tag{19}$$

III. ESTIMATING A PRIORI PROBABILITIES FOR TWO CLASSES

Assume the existence of two classes ω_1 and ω_2 with density functions, $p_1(X)$ and $p_2(X)$, respectively. The functions are not necessarily multivariate normal. For example, $p_1(X)$ could be the density function for wheat and $p_2(X)$ could be the density function for nonwheat, with the mixture density function given by

$$p(X) = q_1p_1(X) + q_2p_2(X) \tag{20}$$

In this case, the partition defined by Equation (9) is equivalent to

$$X \text{ (classified as wheat) exists in } R_1 \text{ if and only if } q_1p_1(X) \ge 1/2\ p(X) \tag{21}$$

Here only training samples are needed for estimating $p_1(X)$, because the mixture density $p(X)$ can be estimated from the unlabeled test samples.

In general, with an arbitrary decision rule independent of the a priori probabilities q_1 and q_2, the space can be partitioned into two regions R_1 and R_2. In this case, an estimate $\hat{q}_1$ of q_1 can be obtained by counting the proportion of test samples classified as

$$\hat{q}_1 = \int_{R_1} p(X)\ dX \tag{22}$$

If the actual q_1 is unknown, it is difficult to estimate the error associated with $\hat{q}_1$. However, it is possible to derive a measure of the error. Let

$$a = \int_{R_1} p_1(X)\ dX \tag{23}$$

$$b = \int_{R_2} p_2(X)\ dX \tag{24}$$

Then it can be algebraically verified that for any partition, $[R_1,R_2]$, the estimate $\hat{q}_1$ (Equation 22) satisfies

$$a q_1 \leq \hat{q}_1 \leq 1 + \left(q_1 - 1\right) b \tag{25}$$

where $0 \leq a \leq 1$ and $0 \leq b \leq 1$. By noting that the probability of misclassification can be written as

$$R = 1 - q_1\, a - q_2\, b \tag{26}$$

it follows that as a and b approach 1, the probability of misclassification approaches 0, and from Equation (25)

$$\hat{q}_1 \longrightarrow q_1 \tag{27}$$

It follows as in the previous section that estimates of a and b can be obtained using unlabeled test samples and are given by

$$\hat{a} = \frac{1}{N_t} \sum_{X \varepsilon R_1} \left[\frac{p_1(X)}{p(X)}\right] \tag{28}$$

$$\hat{b} = \frac{1}{N_t} \sum_{X \varepsilon R_2} \left[\frac{p_2(X)}{p(X)}\right] \tag{29}$$

IV. FEATURE SELECTION

Feature selection is the reduction of the dimension n of each observation vector $\overline{X}$ from n to n', where $n' < n$. The dimensionality reduction is obtained by the measurable transformation

$$\overline{Y} = B\overline{X} \tag{30}$$

where $\overline{Y}$ is an n'-dimensional vector. In practice, B is generally a linear transformation, and the space of all such $\overline{Y}$'s is the transformed space. Since B is assumed to be measurable, the Radon-Nikodym theorem (Kullback, 1968) guarantees the existence of density functions $g_i(Y)$ for each of the m-classes, satisfying

$$\int_S g_i(Y)\, dY = \int_{B^{-1}(S)} p_i(X)\, dX \qquad i=1,\cdots,m \tag{31}$$

where S is any measurable set and

$$B^{-1}(S) = \{X | BX \varepsilon S\} \tag{32}$$

PMC_B denotes the minimal probability of misclassification computed in the transformed space for a given transformation B. Similarly, PMC denotes the minimal probability of misclassification computed in the nontransformed space. The first section of this paper shows that if $[S_1,\cdots,S_m]$ is the minimizing partition in the transformed space

$$S_i = \{Y | q_i g_i(Y) \geq q_j g_j(Y) \qquad j=1,\cdots,m\} \tag{33}$$

As shown elsewhere (Quirein, Decell, 1973), $PMC \leq PMC_B$. The two terms are equal when

$$B^{-1}\left(S_i\right) = R_i \qquad i=1,\cdots,m \tag{34}$$

with R_i as defined by Equation (9).

The difference $PMC_B - PMC \geq 0$ can be considered a measure of the average loss of interclass separability resulting from the measurable transformation $\overline{Y} = B\overline{X}$. The objective of feature selection is to find an $n' < n$ and transformation B such that the difference $PMC_B - PMC$ is small. This objective can also be considered as the maintenance of n-dimensional patterns in n'-dimensional space.

In practice, the expressions PMC_B and PMC are difficult to evaluate because the q_i's are generally unknown and because the higher dimensional integrals are difficult to evaluate. However, the approach in the first section of this paper allows the minimizing of an estimate of PMC_B when the q_i's are known and will be discussed briefly since a complete derivation is presented elsewhere (Quirein, Minter, 1974). Let $P\hat{M}C_B$ be an estimate of PMC_B computed in the transformed space, with the variable X_i replaced by BX_i in Equation (8). Assuming differentiability (which for most cases of interest will exist for a given partition), then it is shown elsewhere (Quirein, Minter, 1974), if the estimate $P\hat{M}C_B$ of PMC_B is to be minimized for a partition $[S_1,S_2,\cdots,S_m]$ and matrix B, then B must satisfy the matrix equation

$$\left(\frac{\partial P\hat{M}C_B}{\partial B}\right) = (0) \tag{35}$$

and

$$S_i = \left\{BX_j \middle| q_i g_i\left(BX_j\right) \geq q_k g_k\left(BX_j\right) \quad k=1,\cdots,m\right\} \quad \begin{matrix} \\ i=1,\cdots,m \\ j=1,\cdots,N_t \end{matrix} \tag{36}$$

The expression for the partial derivative $\left(\frac{\partial P\hat{M}C_B}{\partial B}\right)$ is computable and is presented elsewhere (Quirein, Minter, 1974). Because Equations (35) and (36) are numerically complex to satisfy and because the q_i's are generally unknown, a measure ψ of interclass separability has been devised with these properties:

1. The expressions for ψ and ψ_B are easily evaluated. ψ_B is the measure ψ evaluated in the transformed space.
2. $PMC \leq \psi$ and $PMC_B \leq \psi_B$
3. $\psi_B \geq \psi$ with $\psi = \psi_B$ implying $PMC = PMC_B$

To define a measure ψ of interclass separability that satisfies numbers 2 and 3 and sometimes number 1, the Bhattacharyya distance (Kailath, 1967) is used. The interclass Bhattacharyya distance between classes ω_i and ω_k is

$$\psi(i,k) = \int\left[q_i p_i(X) q_k p_k(X)\right]^{1/2} dX \qquad \begin{matrix} i=1,\cdots,m-1 \\ k=i+1,\cdots,m \end{matrix} \tag{37}$$

Similarly, in the transformed space

$$\psi_B(i,k) = \int\left[q_i g_i(Y) q_k g_k(Y)\right]^{1/2} dY \qquad \begin{matrix} i=1,\cdots,m-1 \\ k=i+1,\cdots,m \end{matrix} \tag{38}$$

is defined. The separability measure ψ is defined

$$\psi = \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \psi(i,k) \tag{39}$$

and the separability measure ψ_B is

$$\psi_B = \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \psi_B(i,k) \tag{40}$$

The following is proved below:

Theorem 2: $\quad PMC \leq \psi$

$PMC_B \leq \psi_B$

$\psi_B \geq \psi$

with $\psi_B = \psi$ implying $PMC = PMC_B$.

To prove $PMC \leq \psi$, note that

$$\min\left[q_i p_i(X),\ q_k p_k(X)\right] \leq \left[q_i p_i(X)\ q_k p_k(X)\right]^{1/2} \tag{41}$$

and writing the probability of misclassification as

$$PMC = \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \int_{R_i U R_k} \min\left[q_i p_i(X), q_k p_k(X)\right] dX \tag{42}$$

it follows from Equation (41)

$$\begin{aligned} PMC &\leq \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \int_{R_i U R_k} \left[q_i p_i(X)\ q_k p_k(X)\right]^{1/2} dX \\ &\leq \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \int \left[q_i p_i(X)\ q_k p_k(X)\right]^{1/2} dX \\ &= \psi \end{aligned} \tag{43}$$

Similarly, it can be shown $PMC_B \leq \psi_B$. To complete the proof of the theorem, note that

$$\begin{aligned} \psi_B(i,k) &= \int \left[q_i g_i(Y)\ q_k g_k(Y)\right]^{1/2} dY \\ &= \int \left[q_i g_i(Y) / q_k g_k(Y)\right]^{1/2} q_k g_k(Y)\ dY \\ &= \int \left[q_i g_i(BX) / q_k g_k(BX)\right]^{1/2} q_k p_k(X)\ dX \quad \text{(Halmos, 1950)} \\ &= \int \left[q_k g_k(BX) / q_i g_i(BX)\right]^{1/2} q_i p_i(X)\ dX \end{aligned} \tag{44}$$

so that

$$2\psi_B(i,k) - 2\psi(i,k) = \int\Big\{\left[q_i g_i(BX)/q_k g_k(BX)\right]^{1/2} q_k p_k(X)$$

$$+ \left[q_k g_k(BX)/q_i g_i(BX)\right]^{1/2} q_i p_i(X)$$

$$- 2\left[q_i p_i(X)\ q_k p_k(X)\right]^{1/2}\Big\}\, dX$$

$$= \int\Big\{\left[q_i g_i(BX)/q_k g_k(BX)\right]^{1/4}\left[q_k p_k(X)\right]^{1/2}$$

$$- \left[q_k g_k(BX)/q_i g_i(BX)\right]^{1/4}\left[q_i p_i(X)\right]^{1/2}\Big\}^2\, dX$$

$$\geq 0 \qquad (45)$$

which implies that $\psi_B(i,k) \geq \psi(i,k)$ and $\psi_B(i,k) = \psi(i,k)$ if and only if

$$\frac{p_i(X)}{p_k(X)} = \frac{g_i(BX)}{g_k(BX)} \qquad (46)$$

for all X. Thus it follows that $\psi_B \geq \psi$ and the condition $\psi_B = \psi$ implies by Equation (46) that $B^{-1}(S_i) = R_i$, $i = 1,\cdots,m$. In this case, $PMC = PMC_B$ as mentioned in the beginning of this section, completing the proof of the theorem.

The proof of the above theorem shows that the condition $\psi_B = \psi$ is equivalent to

$$\frac{p_i(X)}{p_k(X)} = \frac{g_i(BX)}{g_k(BX)} \qquad \begin{array}{l} i=1,\cdots,m-1 \\ k=i+1,\cdots,m \end{array} \qquad (47)$$

almost everywhere, so that a B , if one exists, can be found satisfying $\psi_B = \psi$ and thus $PMC_B = PMC$ even if the q_i's are unknown.

V. FEATURE SELECTION WHEN EACH CLASS DENSITY IS A MIXTURE OF MULTIVARIATE NORMAL DENSITY FUNCTIONS

Assume that each class density function $p_i(X)$ may be written as

$$q_i p_i(X) = \sum_{k=1}^{j_i} q_{i,k} p_{i,k}(X) \qquad (48)$$

where each $p_{i,k}(X)$ is assumed to be multivariate normal and

$$\sum_{k=1}^{j_i} q_{i,k} = q_i \qquad (49)$$

Defining the separability measure $\tilde{\psi}(i,o)$ between classes ω_i and ω_o as

$$\tilde{\psi}(i,o) = \sum_{k=1}^{j_i}\sum_{\ell=1}^{j_o}\int\left[q_{i,k}p_{i,k}(X)\ q_{o,\ell}p_{o,\ell}(X)\right]^{1/2} dX \qquad (50)$$

then

$$\tilde{\psi} = \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} \tilde{\psi}(i,k) \tag{51}$$

The following is proved below:

<u>Theorem 3</u>:

$$PMC \leq \psi \leq \tilde{\psi}$$

$$PMC_B \leq \psi_B \leq \tilde{\psi}_B$$

$$\tilde{\psi}_B \geq \tilde{\psi}$$

with $\tilde{\psi}_B = \tilde{\psi}$ implying $PMC = PMC_B$.

Before proving Theorem 3, note that any within-class separabilities of the following form need not be evaluated:

$$\int \left[q_{i,k} p_{i,k}(X) \; q_{i,j} p_{i,j}(X) \right]^{1/2} dX \tag{52}$$

This fact is particularly useful in the two-class problems in which each class is assumed to be a mixture of multivariate normal density functions. If $j_1 = 1$ so that the first class has only one subclass, then

$$\tilde{\psi} = \sum_{\ell=1}^{j_2} \int \left[q_{1,1} p_{1,1}(X) q_{2,\ell} p_{2,\ell}(X) \right]^{1/2} dX \tag{53}$$

Moreover, under the assumptions of this section of the paper, expressions such as $\tilde{\psi}$, $\tilde{\psi}_B$, and

$$\left(\frac{\partial \tilde{\psi}_B}{\partial B} \right)$$

are readily evaluated in terms of the means and covariances of each density constituting the mixture (Quirein, Decell, 1973). The proof of the theorem follows:

$$\begin{aligned}
\psi(i,o) &= \int \left[q_i p_i(X) \; q_o p_o(X) \right]^{1/2} dX \\
&\leq \int \sum_{\ell=1}^{j_o} \left[q_i p_i(X) \; q_{o,\ell} p_{o,\ell}(X) \right]^{1/2} dX \\
&\leq \int \sum_{k=1}^{j_i} \sum_{\ell=1}^{j_o} \left[q_{i,k} p_{i,k}(X) \; q_{o,\ell} p_{o,\ell}(X) \right]^{1/2} dX \\
&= \tilde{\psi}(i,o)
\end{aligned} \tag{54}$$

and it immediately follows that

$$PMC \leq \psi \leq \tilde{\psi}$$

and similarly,

$$PMC_B \leq \psi_B \leq \tilde{\psi}_B$$

The inequality $\tilde{\psi}_B \geq \tilde{\psi}$ follows exactly as in the proof of Theorem 2. As in Theorem 2, if $\tilde{\psi}_B = \tilde{\psi}$, then almost everywhere

$$\frac{p_{i,k}(X)}{p_{o,\ell}(X)} = \frac{g_{i,k}(BX)}{g_{o,\ell}(BX)} \qquad \begin{array}{l} k=1,\ldots,j_i \\ \ell=1,\cdots,j_o \\ i=1,\ldots,m-1 \\ o=i+1,\cdots,m \end{array} \tag{55}$$

The above can easily be shown to imply almost everywhere,

$$\frac{p_i(X)}{p_o(X)} = \frac{g_i(BX)}{g_o(BX)} \qquad \begin{array}{l} i=1,\ldots,m-1 \\ o=i+1,\cdots,m \end{array} \tag{56}$$

To see this, consider (when Equation 55 is true)

$$\begin{aligned} \frac{q_i p_i(X)}{q_o p_o(X)} &= \frac{\sum_{k=1}^{j_1} q_{i,k} p_{i,k}(X)}{\sum_{\ell=1}^{j_o} q_{o,\ell} p_{o,\ell}(X)} \\ &= \frac{p_{o,1}(X)}{p_{i,1}(X)} \left[\frac{\sum_{k=1}^{j_i} \frac{q_{i,k} p_{i,k}(X)}{p_{o,1}(X)}}{\sum_{\ell=1}^{j_o} \frac{q_{o,\ell} p_{o,\ell}(X)}{p_{i,1}(X)}} \right] \\ &= \frac{g_{o,1}(BX)}{g_{i,1}(BX)} \left[\frac{\sum_{k=1}^{j_i} \frac{q_{i,k} g_{i,k}(BX)}{g_{o,1}(BX)}}{\sum_{\ell=1}^{j_o} \frac{q_{o,\ell} g_{o,\ell}(BX)}{g_{i,1}(BX)}} \right] \\ &= \frac{\sum_{k=1}^{j_i} q_{i,k} g_{i,k}(BX)}{\sum_{\ell=1}^{j_o} q_{o,\ell} g_{o,\ell}(BX)} \\ &= \frac{q_i g_i(BX)}{q_o g_o(BX)} \end{aligned} \tag{57}$$

This completes the proof of the theorem.

VI. SIGNATURE EXTENSION

In conducting large area crop inventories, spectral signatures obtained from one geographical area, perhaps an 11.1- by 9.2-kilometer (17.9- by 14.8-mile) area may be used to classify another geographical area of a similar size. The two areas are usually separated by approximately 18.5 to 185.2 kilometers (29.8 to 298.0 miles), and the signatures are obtained usually 1 or 2 days apart.

If the random vector variable $\overline{X}$ belongs to area 1 and the random variable $\overline{Y}$ belongs to area 2, the hypothesis states that under certain conditions a linear transformation B and additive vector v can be found satisfying

$$\overline{Y} = B\overline{X} + v \tag{58}$$

If Equation (58) is physically satisfied almost everywhere, then, under certain conditions, known statistics from area 1 can be transformed using B and v to classify area 2. Methods for determining B and a consideration of the conditions to be satisfied are discussed below.

The conditions to be satisfied to find the desired transformation and perform a classification of area 2 using statistics from area 1 fall into two distinct categories — surface conditions and above-surface conditions. The above-surface conditions are affected primarily by the atmosphere, Sun angle, and the sensor. It has been shown (Potter, Shelton, 1974) that a difference in Sun angle only between two areas can be accounted for with a linear transformation. Conceivably, areas with "similar" atmospheric transmission could be delineated by an analyst inspecting satellite imagery.

The surface conditions are primarily affected by surface moisture, soil color, and crop type and stage of maturity. Soil moisture maps could be updated on a real-time basis using meteorological information. Soil color maps could be constructed using historical information. Areas of similar crop type and stage of maturity could be obtained using historical information and updated as needed during the growing season. Below, we consider strategies for satisfying Equation (58) when a solution may exist.

Assume that area 1 consists of m distinct classes

$$\pi_1, \pi_2, \cdots, \pi_m \tag{59}$$

normally distributed with known covariances and means (Ω_j, μ_j). Also, let π_1 be the "class of interest" in that we are only interested in determining the percentage of class π_1 in area 1. The classes $\pi_2, \cdots, \pi_m$ can be assumed to represent the competing crops. Assume that area 2 consists of $m' \leq m$ distinct but generally unidentified classes

$$\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_{m'} \tag{60}$$

normally distributed with known or unknown covariances and means (λ_i, β_i). Only the cases where (λ_i, β_i) are known will be considered. We assume that for each distinct class $\hat{\pi}_i$, the same physical class π_j exists in segment 1. Thus the problem becomes one of associating each of the classes of unknown identity, $\hat{\pi}_i$, in segment 2 with corresponding physical class π_j in segment 1 in some way and then solving Equation (58) such that

$$\begin{aligned} B\Omega_j B^T &= \lambda_i \qquad i=1,\ldots,m' \\ B\mu_j + v &= \beta_i \qquad j=\phi(i) \end{aligned} \tag{61}$$

In general, such an association of classes cannot be made, since, if this were the case, area 2 could be classified using statistics from area 2.

A possible approach to solving this problem is discussed below. Assume that in each of the areas the same physical class exists and that this class is "easily" identified — with or without ground truth. Such a class is called a "calibration class." Let the calibration class in area 1 be π_j and the corresponding calibration class in area 2 be $\hat{\pi}_i$. Assuming that a B and v satisfying Equation (58) exist, then it suffices to solve the following for B.

$$B\Omega_j B^T = \lambda_i \tag{62}$$

$$B\mu_j + v = \beta_i \tag{63}$$

It is immediately verified that the solution is given by

$$B = \lambda_i^{1/2} \lambda_j^{-1/2} \tag{64}$$

$$v = \beta_i - B\mu_j \tag{65}$$

If a calibration class cannot be obtained, a possible way of obtaining an association between the classes in the two areas is described below. Let

$$\hat{\Omega}_i = B\Omega_i B^T \qquad i=1,\cdots,m \tag{66}$$

$$\hat{\mu}_i = B\mu_i + v \qquad i=1,\cdots,m \tag{67}$$

Let the index i denote the transformed classes from area 1 (Equations 66 and 67) and the index j denote classes from area 2. Using $\psi(i,j)$ to denote the Bhattacharyya distance between classes π_i and $\hat{\pi}_j$ having statistics $(\hat{\Omega}_i,\hat{\mu}_i)$ and (λ_j,β_j), respectively, define

$$\left.\begin{aligned} \gamma_j &= \max_{\substack{i \\ 1\le i\le m}} \psi(i,j) \\ \alpha_j &= i \text{ which maximizes } \psi(i,j) \end{aligned}\right\} j=1,..,m' \tag{68}$$

$$\phi = \sum_{j=1}^{m'} \gamma_j \tag{69}$$

Thus the association is obtained by

$$\max_B \{\phi\} \tag{70}$$

Since for a given association ($\alpha_j = i$ denotes the jth class in area 2 has been associated with the i'th class in area 1), ϕ is a differentiable function of B, the solution could be obtained by iterating Equations 66-70 until the condition

$$\gamma_j > t \qquad j=1,\cdots,m' \tag{71}$$

is satisfied, where t is some predetermined threshold.

VII. REFERENCES

Anderson, T. W.; 1958: An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Inc.

Fukunaga, K.; and Kessel, D. L.; July 1973: "Nonparametric Bayes Error Estimation Using Unclassified Test Samples." IEEE Transactions on Information Theory, Vol. 1, No. 4, pp. 1-19.

Halmos, P. R.; 1950: Measure Theory. New York: D. van Nostrand Company, Inc.

Kailath, T.; February 1967: "The Divergence and Bhattacharyya Distance Measures in Signal Selection." IEEE Transactions on Communication Theory, Vol. 15, No. 1, pp. 52-60.

Kullback, Solomon; 1968: Information Theory and Statistics. Dover Publications.

Minter, T. C.; and Thadani, S.; to be published: Estimation of the Probability of Error Without Ground Truth. Lockheed Memorandum.

Potter, J.; and Shelton, M.; 1974: "Effect of Atmospheric Haze and Sun Angle on Automatic Classification of ERTS-1 Data." Proceedings of the Ninth International Symposium on Remote Sensing of Environment.

Quirein, J. A.; and Decell, H. P.; October 1973: "An Iterative Approach to the Feature Selection Problem." Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue University.

Quirein, J. A.; and Minter, T. C.; November 1974: Feature Selection for M-classes by Minimizing an Estimate of the Probability of Error. LEC 4985.