Conference on

Machine Processing of

Remotely Sensed Data

October 16 - 18, 1973

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

FEATURE SELECTION VIA AN UPPER BOUND (TO ANY DEGREE TIGHTNESS) ON PROBABILITY OF MISCLASSIFICATION

Cecil R. Hallum

Loyola University, New Orleans, Louisiana

I. ABSTRACT

Currently, many techniques exist for feature selection purposes which are related but, unfortunately, in an indeterminable way to the probability of misclassification. In this paper a procedure is presented which yields an upper bound (to any degree tightness) on the probability of misclassification in sample Gaussian maximum likelihood classification between each pair of categories in p-dimensional space. The technique permits features to be selected so that the optimal q ($q \leq p$) features have the property that no other subset of q features yield a smaller value to the upper bound on the probability of misclassification. A computer-assessible transformation is utilized which permits a multiple integral over the misclassification region in p-dimensional space to be approximated, to any degree of accuracy, by the product of p iterated integrals, each over univariate space, and each of which may be obtained by a simple table-look-up procedure. Quite often, transformations are used without consideration of loss of information; however, the one utilized in this procedure results in no loss of information and leaves the standard likelihood ratio invariant in value.

II. INTRODUCTION

There seems to be general agreement among remote sensing community personnel, and classification theorists in general, that the cost function which has applicable meaning to the user scientist is expected cost of misclassification. Any procedure, then, which is optimal for feature selection purposes should be one which minimizes this cost function. A popular (and realistic) assumption in almost all remote sensing applications is that the costs of misclassifying an individual from category C_i as being from category C_j, $i \neq j = 1,2,...,m$, are equal, in which case the expected cost of misclassification reduces to probability of misclassification. Currently, many techniques exist for feature selection purposes (eigenvalue/eigenvector techniques (Chien and Fu, 1968) including factor analysis and principal components, standard regression techniques, Wilk's scatter technique (Wilks, 1962), the divergence criterion (Marill and Green, 1963), Sammon's nonlinear mapping (Sammon, 1970), LaMotte-Hocking regression techniques (LaMotte and Hocking, 1970), the Battacharyya distance measure (Kailath, 1967), the Matusita distance (Matusita, 1966), the concept of equivocation (Babu, 1972), and others, including trial-and-error approaches with sets of features with varied composition) which are related but, unfortunately, in an indeterminable way to the probability of misclassification. In this paper a procedure is presented which yields an upper bound (to any degree desired tightness) on the probability of misclassification in sample Gaussian maximum likelihood classification between each pair of categories in p-dimensional space. The technique permits features to be selected between pairs of categories so that the optimal q ($q \leq p$) features have the property that no other subset of q features yields a smaller value to the upper bound on the probability of misclassification. A simple, computer-assessible transformation is utilized which permits a multiple integral over the misclassification region in p-dimensional space to be approximated, to any degree of accuracy, by the product of p iterated integrals, each over univariate space, and each of which may be obtained by a simple table-look-up procedure. The utility of this transformation in the classification processing phase of pattern recognition recently appeared in the literature (Minter and Hallum, 1972) and, unlike many transformations which are used without consideration of loss of information, the one utilized herein results in no loss of information and leaves the standard likelihood ratio invariant in value.

III. TRANSFORMATION TO AN EQUIVALENT PATTERN RECOGNITION SPACE

Let C_i and C_j be two categories ($i \neq j$) which have the Gaussian distribution with respective densities

$$P(X|C_i) = \frac{1}{(2\pi)^{p/2}|\hat{M}_i|^{1/2}}\exp\{-1/2(X - \hat{U}_i)^T\hat{M}_i^{-1}(X - \hat{U}_i)\} \tag{1}$$

and

$$P(X|C_j) = \frac{1}{(2\pi)^{p/2}|\hat{M}_j|^{1/2}}\exp\{-1/2(X - \hat{U}_j)^T\hat{M}_j^{-1}(X - \hat{U}_j)\} \tag{2}$$

where $\hat{U}_i$ and $\hat{U}_j$ are the respective sample means, $\hat{M}_i$ and $\hat{M}_j$ are the sample variance-covariance matrices. Transforming to a new space by the nonsingular transformation Q, where Q is the matrix with the property

$$Q^T\hat{M}_iQ = D \;; \quad Q^T\hat{M}_jQ = I$$

(D is a diagonal matrix and I the identity matrix), the likelihood ratio (assuming the classification scheme is Bayes with equal a priori class probabilities and with cost matrix elements $c_{ij} = c - c\delta_{ij}$, where c is a constant and δ_{ij} is the Kronecker delta) of $P(X|C_i)$ to $P(X|C_j)$ takes the form

$$\begin{aligned}
\frac{P(X|C_i)}{P(X|C_j)} &= \frac{\frac{1}{(2\pi)^{p/2}|\hat{M}_i|^{1/2}}\exp\{-1/2(X - \hat{U}_i)^T\hat{M}_i^{-1}(X - \hat{U}_i)}{\frac{1}{(2\pi)^{p/2}|\hat{M}_j|^{1/2}}\exp\{-1/2(X - \hat{U}_j)^T\hat{M}_j^{-1}(X - \hat{U}_j)\}} \\
&= \frac{\frac{1}{(2\pi)^{p/2}|Q^T\hat{M}_iQ|^{1/2}}\exp\{-1/2(X - \hat{U}_i)^T(Q^{T^{-1}}Q^T\hat{M}_iQQ^{-1})^{-1}(X - \hat{U}_i)\}}{\frac{1}{(2\pi)^{p/2}|Q^T\hat{M}_jQ|^{1/2}}\exp\{-1/2(X - \hat{U}_j)^T(Q^{T^{-1}}Q^T\hat{M}_jQQ^{-1})^{-1}(X - \hat{U}_j)\}} \\
&= \frac{\frac{1}{(2\pi)^{p/2}|D|^{1/2}}\exp\{-1/2(Q^TX - Q^T\hat{U}_i)^TD^{-1}(Q^TX - Q^T\hat{U}_i)\}}{\frac{1}{(2\pi)^{p/2}}\exp\{-1/2(Q^TX - Q^T\hat{U}_j)^T(Q^TX - Q^T\hat{U}_j)\}}
\end{aligned} \tag{3}$$

$$= \frac{\frac{1}{(2\pi)^{p/2}|D|^{1/2}}\exp\{-1/2(Z - V)^T D^{-1}(Z - V)\}}{\frac{1}{(2\pi)^{p/2}}\exp\{-1/2(Z - W)^T(Z - W)\}} = \frac{f(Z|C_i)}{f(Z|C_j)}$$

$$= \prod_{k=1}^{p} \frac{p_{ki}(Z_k)}{p_{kj}(Z_k)}$$

where $p_{ki}(Z_k)$ and $p_{kj}(Z_k)$ are the univariate normal densities

$$p_{ki}(Z_k) = \frac{1}{(2\pi)^{1/2} d_k}\exp\{-1/2(Z_k - V_k)^2/d_k^2\} \; ; \tag{4}$$

$$p_{kj}(Z_k) = \frac{1}{(2\pi)^{1/2}}\exp\{-1/2(Z_k - W_k)^2\} \; . \tag{5}$$

In the above, d_k^2 is the $k^{\underline{th}}$ diagonal element of D; Z_k, V_k, and W_k are the $k^{\underline{th}}$ coordinates of the vectors $Z = (Z_1, Z_2, \ldots, Z_p)^T = Q^T X$, $V = (V_1, V_2, \ldots, V_p)^T = Q^T \hat{U}_i$, and $W = (W_1, W_2, \ldots, W_p)^T = Q^T \hat{U}_j$, respectively. The utility of the transformation Q in classification processing has already appeared (Minter and Hallum, 1972) in the literature, and it needs be stressed that Q needs be obtained only once for each pair of categories for a given classification situation (such as a flight-line for classification). Moreover, observe that the invariance of the likelihood ratio (note equations (3)) guarantees that pattern classification in the original and the transformed space is equivalent.

To obtain Q (see Minter and Hallum, 1972), let $Q = P^{-1}O$ where P is the Cholesky (Ratishauser, 1966) factorization of M_j, i.e. $M_j = P^T P$ where P is upper triangular, and O is the orthogonal transformation which diagonalizes the symmetric matrix

$$P^{T^{-1}} \hat{M}_i P^{-1}$$

Routines for obtaining P and O have been implemented for some time and are available via the <u>IBM Scientific Subroutine Package</u> (IBM Corp., 1970) and the <u>Catalog of Mathematical Routines</u> at NASA/MSC in Houston, Texas.

IV. A FIRST DEGREE APPROXIMATION TO THE PROBABILITY OF MISCLASSIFICATION

In both the transformed and the untransformed space, the boundary of the misclassification region between categories C_i and C_j (defined by those points satisfying $P(X|C_i)/P(X|C_j) = f(Z|C_i)/f(Z|C_j) = 1$) is always a hyperellipsoid or a hyper-hyperboloid. In the transformed space, this boundary is specified by the quadratic equation

$$\sum_{k=1}^{p}\left(\frac{Z_k - V_k}{d_k}\right)^2 - \sum_{k=1}^{p}(Z_k - W_k)^2 = \ln \prod_{k=1}^{p} d_k^2 \tag{6}$$

which is the equation of a central conic whose center is located at the point with $k^{\underline{th}}$ coordinate $(d_k^2 W_k - V_k)/(d_k^2 - 1)$ and whose principal axes are parallel to the coordinate axes. In the forthcoming development, the center will be utilized to obtain the smallest hyper-box, with sides parallel to the coordinate axes in the transformed space, which encloses the region of misclassification. Integration, then, over the region of misclassification is approximated to "first degree approximation" (which is sufficient in a large number of cases) by integrating over this box; moreover, the manner of integration for the total probability of misclassification between C_i and C_j is carried out in such a manner that the resulting value is always greater than or equal to the probability of misclassification.

The technique for "first degree approximation" for the case of a hyperellipsoidal region of misclassification is illustrated in Fig. 1 and 2 for two dimensions. The dashed line in Fig. 1 and 2 denotes the equi-probable line between C_i and C_j (specified by equation (6) with p = 2) and is the boundary of the region of misclassification. Figure 2 indicates what we would see in Fig. 1 by looking down on the Z_1, Z_2-plane from the y-direction. From Fig. 1 and 2, the probability of misclassification is given by

$$P = \iint_R f(Z|C_i)\,dZ_1 dZ_2 + \iint_{\tilde{R}} f(Z|C_j)\,dZ_1 dZ_2 \tag{7}$$

where $\tilde{R}$ denotes the region outside of R. In general, the determination of the function to be integrated over R is accomplished by selecting the one having minimum value at the center point C_R of R, which is simply the center of the conic specified by equation (6); the remaining function is integrated over $\tilde{R}$. As is well-known, in p-dimensional space the integration in (7) becomes extremely difficult, if not impossible , utilizing existing techniques. However, the points a_{11}, a_{21}, a_{12}, a_{22} are easily obtained; in general, for p-dimensional space, a_{1j} and a_{2j} are obtained by putting all Z_k, with the exception of Z_j, in (6) equal to the corresponding coordinate of the center C_R, and solving the resulting univariate quadratic in Z_j for Z_j. Applying the quadratic formula, the resulting two values of Z_j are denoted by a_{1j} and a_{2j} and are given by

$$a_{1j} = C_{Rj} - \sqrt{T_j} \quad ; \quad a_{2j} = C_{Rj} + \sqrt{T_j} \quad , \; j = 1,2,\ldots,p \tag{8}$$

where

$$C_{Rj} = \frac{d_j^2 W_j - V_j}{d_j^2 - 1} \tag{9}$$

is the $j^{\underline{th}}$ coordinate (j = 1,2,...,p) of the center of the conic specified in equation (6) and

$$T_j = C_{Rj}^2 - \frac{d_j^2}{1 - d_j^2}\left\{- \frac{C_{Rj}^2}{d_j^2} + \sum_{k=1}^{p} [C_{Rk}^2 + V_k^2 - d_k^2(W_k^2 + \ln d_k^2)]/d_k^2\right\}. \tag{10}$$

An upper bound approximation to the first integral in (7) is given by integrating over the box B, i.e.

$$\iint_R f(Z|C_i)dZ_1dZ_2 = \iint_R p_{1i}(Z_1)p_{2i}(Z_2)dZ_1dZ_2$$

$$\leq \int_{a_{11}}^{a_{21}} p_{1i}(Z_1)dZ_1 \int_{a_{12}}^{a_{22}} p_{2i}(Z_2)dZ_2$$

$$= \left[F\left(\frac{a_{21} - V_1}{d_1}\right) - F\left(\frac{a_{11} - V_1}{d_1}\right)\right]\left[F\left(\frac{a_{22} - V_2}{d_2}\right) - F\left(\frac{a_{12} - V_2}{d_2}\right)\right] \tag{11}$$

where F denotes the univariate, cumulative Gaussian distribution function with mean 0 and variance 1 and whose value may be obtained at the indicated points by referring to a normal distribution table (which could easily be stored for quick reference) or by utilizing one of the several efficient techniques which are available for estimating tail probabilities under the univariate Gaussian distribution curve. Note also that the right hand side of (11) exceeds the true value of the multiple integral over R by precisely the amount

$$\sum_{k=1}^{4} \iint_{R_k} f(Z|C_i)dZ_1dZ_2 \ . \tag{12}$$

However, a portion of this excess is eliminated in determining the total probability of misclassification between categories C_i and C_j by noting that the second integral in (7) may be obtained as follows:

$$\iint_{\bar{R}} f(Z|C_j)dZ_1dZ_2 = 1 - \iint_R f(Z|C_j)dZ_1dZ_2 \ . \tag{13}$$

Using the similar approximation as in (11),

$$\iint_R f(Z|C_j)dZ_1dZ_2 \leq \int_{a_{11}}^{a_{21}} p_{1j}(Z_1)dZ_1 \int_{a_{12}}^{a_{22}} p_{2j}(Z_2)dZ_2$$

$$= \left[F(a_{21} - W_1) - F(a_{11} - W_1)\right]\left[F(a_{22} - W_2) - F(a_{12} - W_2)\right] \tag{14}$$

and the right-hand side of (14) exceeds the true value by precisely the amount

$$\sum_{k=1}^{4} \iint_{R_k} f(Z|C_j)dZ_1dZ_2 \ . \tag{15}$$

However, $f(Z|C_i) \geq f(Z|C_j)$ for all Z over the regions R_1, R_2, R_3, and R_4, therefore the upper bound estimate to "first degree approximation" of the probability of misclassification is given by

$$\iint_B f(Z|C_i)dZ_1dZ_2 + 1 - \iint_B f(Z|C_j)dZ_1dZ_2$$

$$= \left(\iint_R f(Z|C_i)dZ_1dZ_2 + 1 - \iint_R f(Z|C_j)dZ_1dZ_2\right) + \varepsilon$$

$$= P + \varepsilon \tag{16}$$

where ε is the non-negative quantity which represents the over-estimate of the probability, P, of misclassification and is given by

$$\varepsilon = \sum_{k=1}^{4} \iint_{R_k} (f(Z|C_i) - f(Z|C_j))dZ_1dZ_2 \ . \tag{17}$$

The preceding discussion was for the case of a hyperellipsoidal region of misclassification, however the same discussion applies verbatim as well to the case of a hyper-hyperboloidal region of misclassification where, for the two-dimensional case, R_1, R_2, R_3, and R_4 are specified in Fig. 3 and $a_{11} = -\infty$, $a_{21} = \infty$. In general, the misclassification region will be hyper-hyperboloidal provided $T_j < 0$ in (10) for at least one value of $j = 1,2,...,p$, in which case, $a_{1j} = -\infty$ and $a_{2j} = \infty$ for each such j.

Utilizing the "first-degree approximation," feature selection is easily implemented by ordering the features as follows: for $m,n \in \{1,2,...,p\}$, feature Z_m is preferable to feature Z_n provided

$$\left[F\left(\frac{a_{2m} - V_m}{d_m}\right) - F\left(\frac{a_{1m} - V_m}{d_m}\right)\right] + 1 - \left[F(a_{2m} - W_m) - F(a_{1m} - W_m)\right]$$

$$< \left[F\left(\frac{a_{2n} - V_n}{d_n}\right) - F\left(\frac{a_{1n} - V_n}{d_n}\right)\right] + 1 - \left[F(a_{2n} - W_n) - F(a_{1n} - W_n)\right] \ . \tag{18}$$

After ordering the coordinates in the above manner, denote them by $Z_{(1)}, Z_{(2)},...,Z_{(p)}$; if we want to achieve a probability of misclassification not to exceed a preset value α, we select the first k ordered coordinates such that

$$\prod_{n=1}^{k-1} P_{(n)} > \alpha \tag{19}$$

but

$$\prod_{n=1}^{k} P_{(n)} \leq \alpha \tag{20}$$

where

$$P_{(n)} = \left[F\left(\frac{a_{2(n)} - V_{(n)}}{d_{(n)}}\right) - F\left(\frac{a_{1(n)} - V_{(n)}}{d_{(n)}}\right)\right] +$$

$$1 - \left[F(a_{2(n)} - W_{(n)}) - F(a_{1(n)} - W_{(n)})\right]$$

for $n = 1,2,\ldots,p$. If $\prod_{n=1}^{p} P_{(n)} > \alpha$ then the conclusion is, to the "first-degree approximation," there is reason to doubt that it is possible to classify with this probability of misclassification. We do know, however, that it is possible to distinguish, using the first k ordered features, between categories C_i and C_j with probability of misclassification not to exceed the value of the left side of (20).

V. AN UPPER BOUND (TO ANY DEGREE TIGHTNESS) ON PROBABILITY OF MISCLASSIFICATION

The previous section summarizes the approach for a "first-degree approximation" to the probability of misclassification and provides an upper bound estimate which, in many instances, will be sufficient. The procedure may be generalized further to obtain as tight an upper bound on the probability of misclassification as one chooses. In the following, explicit expressions are given for upper bound estimates, to any degree of desired tightness, for each of the two possible regions of misclassification.

A. A HYPERELLIPSOIDAL REGION OF MISCLASSIFICATION

In general, the region of misclassification in p-dimensional space will be hyperellipsoidal provided

$$T_j > 0 \quad \text{for each } j = 1,2,\ldots,p, \tag{21}$$

(or equivalently, if $d_k^2 - 1$ is of the same sign for all $k = 1,2,\ldots,p$) where T_j is given by equation (10). To aid in following the discussion below, refer to Fig. 4 for the case of two dimensions.

Utilizing the coordinates of the center of the conic of equation (6), increments along a principal axis can be made and, e.g. in two-dimensional space, the points

$$a_{12}^{(1)},\ a_{22}^{(1)},\ a_{12}^{(2)},\ a_{22}^{(2)},\ a_{12}^{(3)},\ a_{22}^{(3)},\ldots,\ a_{12}^{(k)},\ a_{22}^{(k)},\ldots,\ a_{12}^{(N-1)},\ a_{22}^{(N-1)}$$

(e.g., in Fig. 4 N = 6) easily obtained by repeated application of the quadratic formula. In general, if $Z_1, Z_2,\ldots, Z_q$ is a given subset of q features ($q \leq p$) of p-dimensional space, the notation $a_{ij}^{(k_1,\ldots,k_q)}$ will be used to represent the i^{th} of the two j^{th} coordinate values (lower coordinate value if i = 1, upper value if i = 2) of the point on the boundary of the region of misclassification obtained after moving a distance of $k_1\Delta$ from a_{11} to a_{21}, a distance of $k_2\Delta$ from a_{12} to a_{22}, ..., and a distance of $k_q\Delta$ from a_{1q} to a_{2q}. The increment length Δ might, for example, be $\Delta = \max_n |a_{2n} - a_{1n}|/N$. Requiring N to be an <u>even</u> integer in the following will simplify the problem of making certain the region over which integration is carried out completely encloses the region of misclassification. For a particular $j \in \{1,2,\ldots,q\}$, the two values (i.e. for i = 1,2) of $a_{ij}^{(k_1,\ldots,k_q)}$ are obtained by replacing the Z_k's in equation (6), with the exception of the $j^{\underline{th}}$, by the coordinates of the point obtained after moving a distance of $k_1\Delta$ from a_{11} toward a_{12}, a distance of $k_2\Delta$ from a_{12} to a_{22},..., and a distance of $k_q\Delta$ from a_{1q} to a_{2q}, and solving the resulting quadratic in Z_j for Z_j using the quadratic formula. These two values of Z_j denoted by $a_{1j}^{(k_1,\ldots,k_q)}$ and $a_{2j}^{(k_1,\ldots,k_q)}$ are given by

$$a_{1j}^{(k_1,\ldots,k_q)} = C_{Rj} - \sqrt{C_{Rj}^2 - \frac{V_j^2 - d_j^2(W_j^2 + R_j^{(k_1,\ldots,k_q)})}{1 - d_j^2}} \tag{22}$$

$$a_{2j}^{(k_1,\ldots,k_q)} = C_{Rj} + \sqrt{C_{Rj}^2 - \frac{V_j^2 - d_j^2(W_j^2 + R_j^{(k_1,\ldots,k_q)})}{1 - d_j^2}} \tag{23}$$

where

$$R_j^{(k_1,\ldots,k_q)} = \ln \prod_{k=1}^{q} d_k^2 - \sum_{i\neq j}^{q}\left[\left(\frac{a_{1i} + k_i \quad - V_i}{d_i}\right)^2 + (a_{1i} + k_i \quad - W_i)^2\right]$$
$$- \left[\left(\frac{C_{Rj} - V_j}{d_j}\right)^2 - (C_{Rj} - W_j)^2\right]. \tag{24}$$

The $(N - 1)^{\underline{th}}$ degree upper bound estimate of the probability of misclassification between categories C_i and C_j is then given by

$$\sum_{k_1=1}^{N}\sum_{k_2=1}^{N}\cdots\sum_{k_q=1}^{N}\prod_{n=1}^{q}\int_{A_n}^{B_n} p_{ni}(Z_n)dZ_n \quad +$$

$$1 - \sum_{k_1=1}^{N}\sum_{k_2=1}^{N}\cdots\sum_{k_q=1}^{N}\prod_{n=1}^{q}\int_{A_n}^{B_n} p_{nj}(Z_n)dZ_n$$

$$= \sum_{k_1=1}^{N}\cdots\sum_{k_q=1}^{N}\prod_{n=1}^{q}\left[F\left(\frac{B_n - V_n}{d_n}\right) - F\left(\frac{A_n - V_n}{d_n}\right)\right] +$$

$$1 - \sum_{k_1=1}^{N} \cdots \sum_{k_q=1}^{N} \prod_{n=1}^{q} \left[F(B_n - W_n) - F(A_n - W_n) \right] \tag{25}$$

where, for each $n = 1,2,\ldots,q$,

$$B_n = a_{1n}^{(k_1,\ldots,k_q)} + (k_n + 1)\Delta + \left(a_{1n}^{(k_1,\ldots,k_n+1,\ldots,k_q)} - a_{1n}^{(k_1,\ldots,k_q)}\right) I_n$$

and

$$A_n = a_{1n}^{(k_1,\ldots,k_q)} + k_n\Delta + \left(a_{1n}^{(k_1,\ldots,k_n+1,\ldots,k_q)} - a_{1n}^{(k_1,\ldots,k_q)}\right) I_n$$

where I_n is the indicator function

$$I_n = \begin{cases} 1, & \text{if } k_n > N/2 \\ 0, & \text{if } k_n \leq N/2 \end{cases} .$$

The expression in (25) is the value of the upper bound estimate over a region which encloses the region of misclassification with the property that the enclosing region squeezes down on the hyperellipsoidal region of misclassification as N becomes large.

B. A HYPER-HYPERBOLOIDAL REGION OF MISCLASSIFICATION

The region of misclassification in p-dimensional space will be hyper-hyperboloidal provided $T_j < 0$ (of equation (10)) for one or more $j \in \{1,2,\ldots,p\}$, or equivalently, provided $d_k^2 - 1$ is not of the same sign for all $k = 1,2,\ldots,p$. The procedure for obtaining the upper bound approximation is very similar to that for the hyperellipsoidal region. The primary difference is that of having to integrate over a region in a negative direction from a_0 and a positive direction from a_0' (refer to Fig. 5 for the case of two dimensions). If we let $S = \{j \mid T_j < 0\}$ then the $j^{\underline{th}}$ coordinates of a_0 and a_0', denoted by a_{0j} and a_{0j}', respectively, are given by

$$a_{0j} = \begin{cases} a_{1j}, & \text{if } j \notin S \\ C_{Rj}, & \text{if } j \in S \end{cases} ;$$

$$a_{0j}' = \begin{cases} a_{2j}, & \text{if } j \notin S \\ C_{Rj}, & \text{if } j \in S \end{cases}$$

where a_{1j}, a_{2j}, and C_{Rj} are given by equations (8) and (9). Moreover, for the hyper-hyperboloidal region of misclassification, for each value of $i = 1,2$, the quantity $a_{ij}^{(k_1,\ldots,k_q)'}$ is identical to that of equations (22) and (23) upon replacing C_{Rj} by a_{0j}'; similarly, replacing C_{Rj} by a_{0j} in (22) and (23) and $k_i\Delta$ by $-k_i\Delta$ in expression (24) for $R_j^{(k_1,\ldots,k_q)}$, we obtain $a_{ij}^{(k_1,\ldots,k_q)}$. For larger N and smaller Δ (N and Δ are not necessarily related and N may be even or odd for the hyper-hyperboloidal case), the value of the integration over the enclosing region becomes a tighter upper bound approximation to the probability of misclassification. The explicit expression for this estimate in terms of the cumulative Gaussian distribution function F, with mean 0 and variance 1, is given by

$$P_N = \sum_{k_1=1}^{N} \cdots \sum_{k_q=1}^{N} \left[\prod_{n=1}^{q} \int_{A_n}^{B_n} p_{ni}(Z_n)\,dZ_n + \prod_{n=1}^{q} \int_{A_n'}^{B_n'} p_{ni}(Z_n)\,dZ_n \right] +$$

$$1 - \sum_{k_1=1}^{N} \cdots \sum_{k_q=1}^{N} \left[\prod_{n=1}^{q} \int_{A_n}^{B_n} p_{nj}(Z_n)\,dZ_n + \prod_{n=1}^{q} \int_{A_n'}^{B_n'} p_{nj}(Z_n)\,dZ_n \right]$$

$$= \sum_{k_1=1}^{N} \cdots \sum_{k_q=1}^{N} \left\{ \prod_{n=1}^{q} \left[F\left(\frac{B_n - V_n}{d_n}\right) - F\left(\frac{A_n - V_n}{d_n}\right) \right] + \prod_{n=1}^{q} \left[F\left(\frac{B_n' - V_n}{d_n}\right) - F\left(\frac{A_n' - V_n}{d_n}\right) \right] \right\} +$$

$$1 - \sum_{k_1=1}^{N} \cdots \sum_{k_q=1}^{N} \left\{ \prod_{n=1}^{q} \left[F(B_n - W_n) - F(A_n - W_n) \right] + \prod_{n=1}^{q} \left[F(B_n' - W_n) - F(A_n' - W_n) \right] \right\}$$

where

$$A_n = a_{1n}^{(k_1,\ldots,k_q)} - k_n\Delta\ ;$$

$$B_n = a_{1n}^{(k_1,\ldots,k_q)} - (k_n - 1)\Delta\ ;$$

$$A_n' = a_{1n}^{(k_1,\ldots,k_q)} - (k_n - 1)\Delta\ ;$$

$$B_n' = a_{1n}^{(k_1,\ldots,k_q)} - k_n\Delta\ .$$

VI. CONCLUSIONS

For a given set of q features, the procedure presented herein permits the calculation of as tight an upper bound estimate on the probability of misclassification in Gaussian maximum likelihood classification between each pair of categories as one chooses using these selected features. The original space is transformed to an equivalent pattern recognition space so that all calculations become univariate calculations, including the integrations. An explicit expression is given for as tight an upper bound estimate as one chooses in terms of univariate Gaussian distribution functions which are easily and accurately obtained. The procedure may be utilized to select the optimal q ($q \leq p$) features having the property that no other subset of q features yield a smaller value to the upper bound on the probability of misclassification. In particular, beginning with q = 1, the technique may be applied and q increased, if necessary, until the upper estimate becomes less than some preselected threshold α, where α is the largest value of the probability of misclassification that can be tolerated in a given classification situation. From this set, the q features yielding the smallest value less than α is selected.

The selected features will be those which best discriminate between categories C_i and C_j. It is likely that for a third category, say C_k, those features which permit best discrimination between categories C_i and C_k will be different from those that best discriminate between C_i and C_j (and C_i and C_k). Therefore, to achieve the degree of classification accuracy that is hoped for in the near future in remote sensing, the selection of pairwise categorical features should be significantly superior to choosing one set of features for all the categories combined.

As a final consideration, if the a priori class probabilities q_i and q_j are known for categories C_i and C_j, respectively, then they may easily be incorporated into the procedure presented herein by simply replacing

$$\ln \prod_{k=1}^{q} d_k^2 \quad \text{by} \quad \ln \frac{q_i}{q_j} \prod_{k=1}^{q} d_k^2$$

throughout. Also, it needs be pointed out that the space in which feature selection is carried out herein is the same space in which classification processing by thresholding (see Minter and Hallum, 1972) is accomplished; consequently, the recommended procedure is to select features utilizing the technique presented herein and then perform classification processing utilizing the Minter-Hallum classification procedure.

VII. REFERENCES

1. Babu, C. Chitti,"On feature extraction in pattern recognition," Proc. Fifth Hawaii Int. Conf. on Syst. Sci., pp. 20-23, Jan., 1972.

2. Chen, Y. T. and Fu, K. S.,"Selection and ordering of feature observations in a pattern recognition system," Information and Control, vol. 12, May, 1968.

3. Kailath, T.,"The divergence and Battacharyya distance measures in signal detection," IEEE Trans. Commun. Technol., vol. COM-15, pp. 52-60, Feb., 1967.

4. LaMotte, L. R. and Hocking, R. R.,"Computational efficiency in the selection of regression coefficients," Texas A & M University, College Station, Texas, 1970.

5. Marill, T. and Green, D. M.,"On the effectiveness of receptors in recognition systems," IEEE Trans. Information Theory, vol. IT-9, pp. 11-17, Jan., 1963.

6. Matusita, K.,"A distance and related statistics in multi-variate analysis," Multi-Variate Analysis, P. R. Krishnaiah, Ed. New York: Academic Press, 1966, pp. 187-200.

7. Minter, T. C. and Hallum, C. R.,"Maximum likelihood classification by thresholding," Technical Report LEC/HASD No. 640-TR-114 under NASA contract NAS9-12200, 1972.

8. Rutishauser, H.,"Algorithmus 1-Lineares Gleichungs-system mit symmetrischer positiv-definiter Bandmatrix nach Cholesky-Computing," Archives for Electronic Computing, vol. 1, iss. 1, pp. 77-78, 1966.

9. Sammon, J. W., Jr.,"Interactive pattern analysis and classification," IEEE Trans. Comput., vol. C-19, pp. 594-616, July, 1970.

10. Wilks, S. S. Mathematical Statistics. New York: Wiley, 1962.

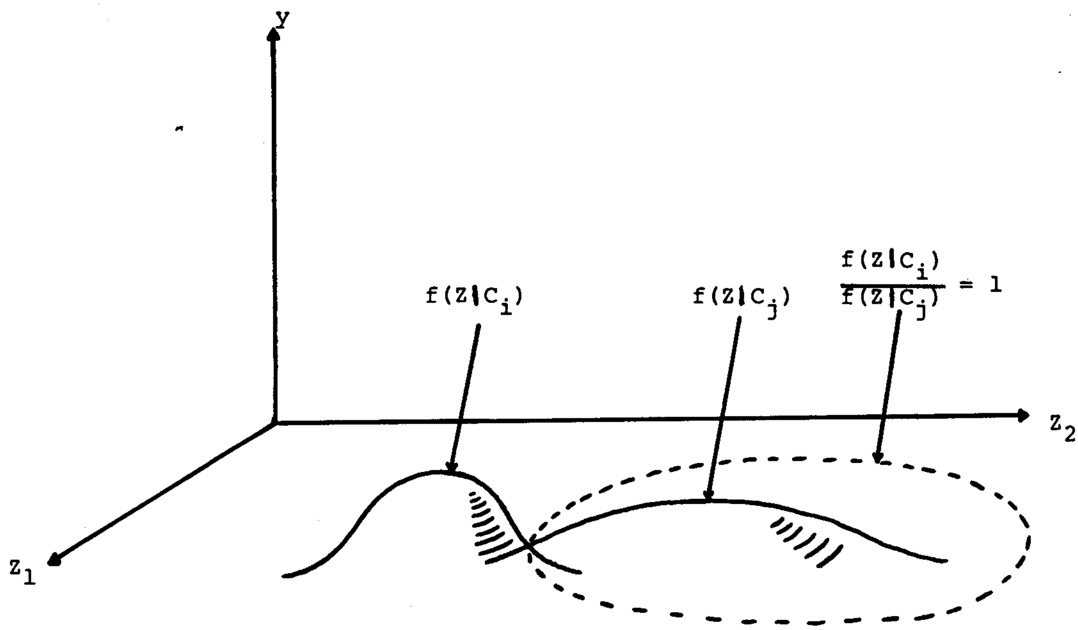

Figure 1. The Misclassification Boundary in Two-dimensional Space Between Categories C_i and C_j.

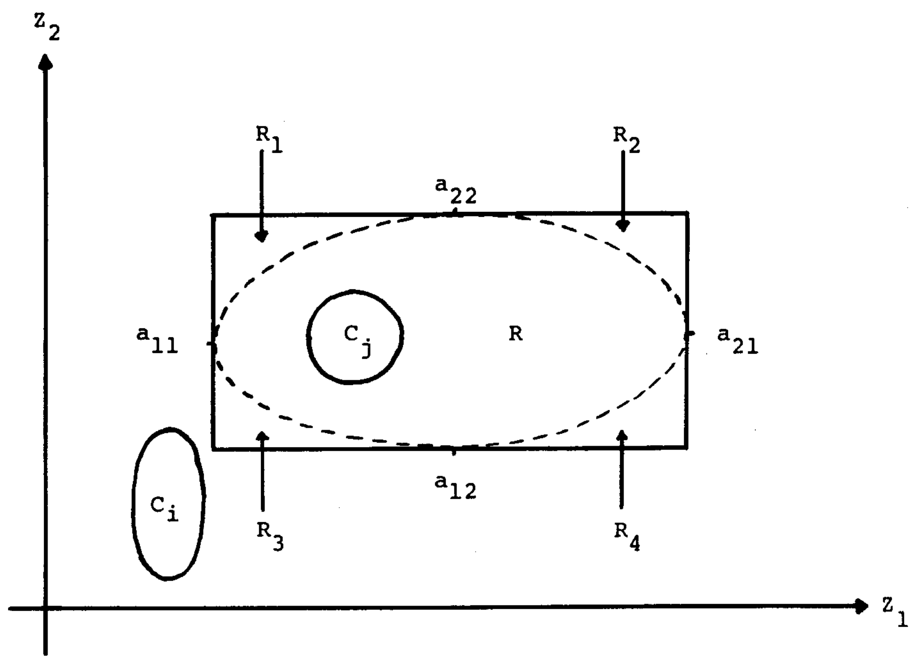

Figure 2. The Box Enclosing the Misclassification Region.

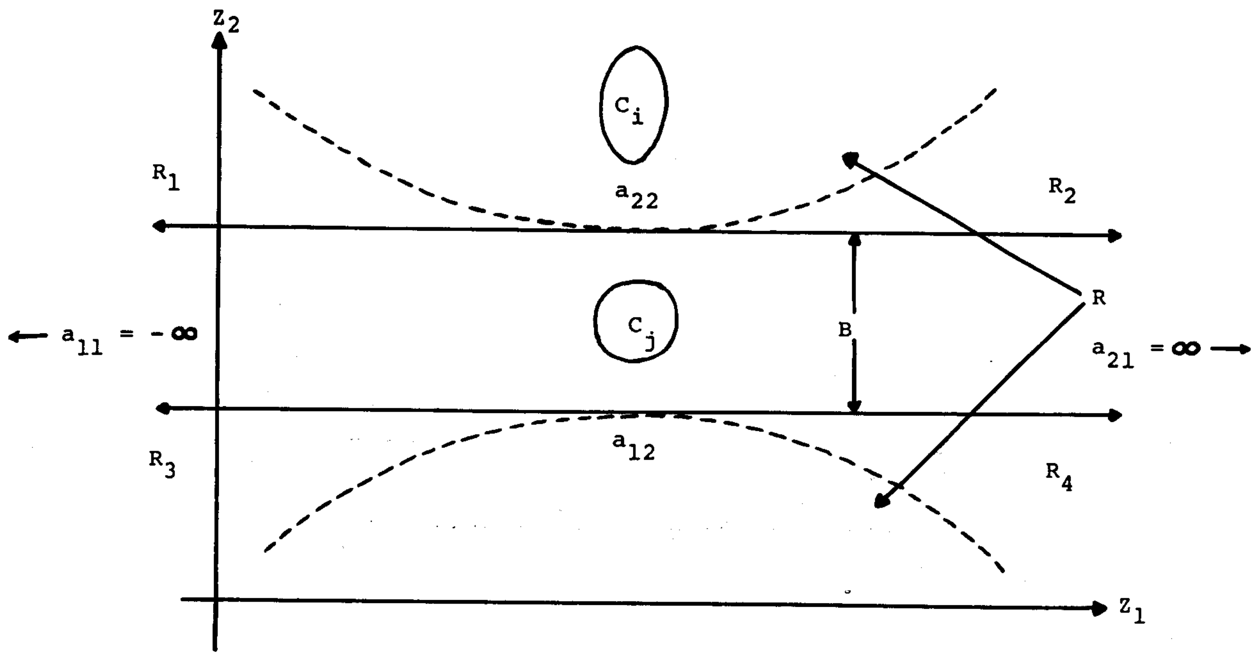

Figure 3. A Hyperboloidal Misclassification Region in Two-dimensional Space Between Categories C_i and C_j.

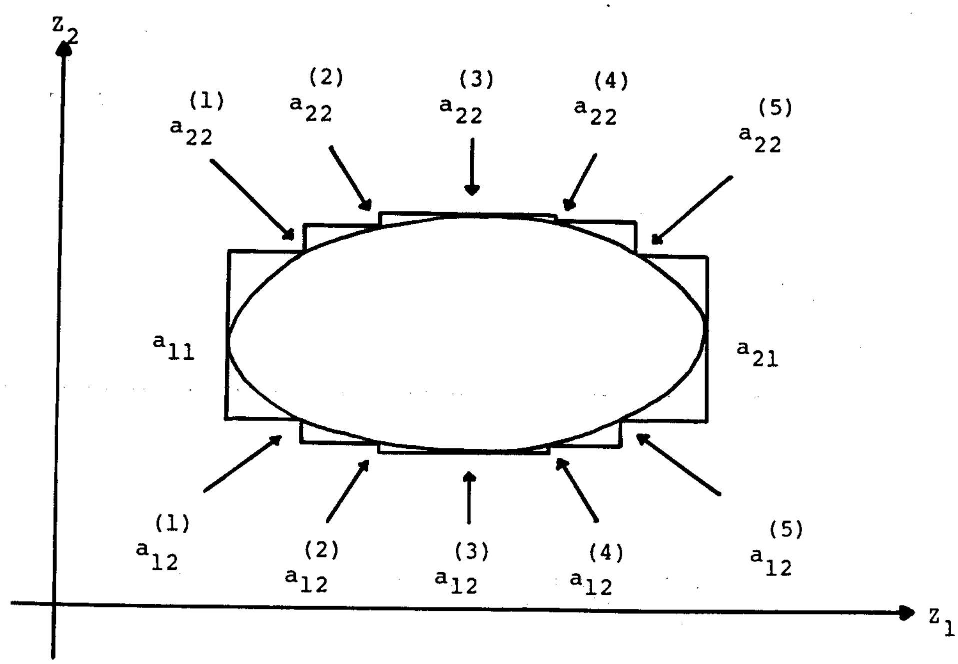

Figure 4. A Tighter Bound on the Misclassification Region

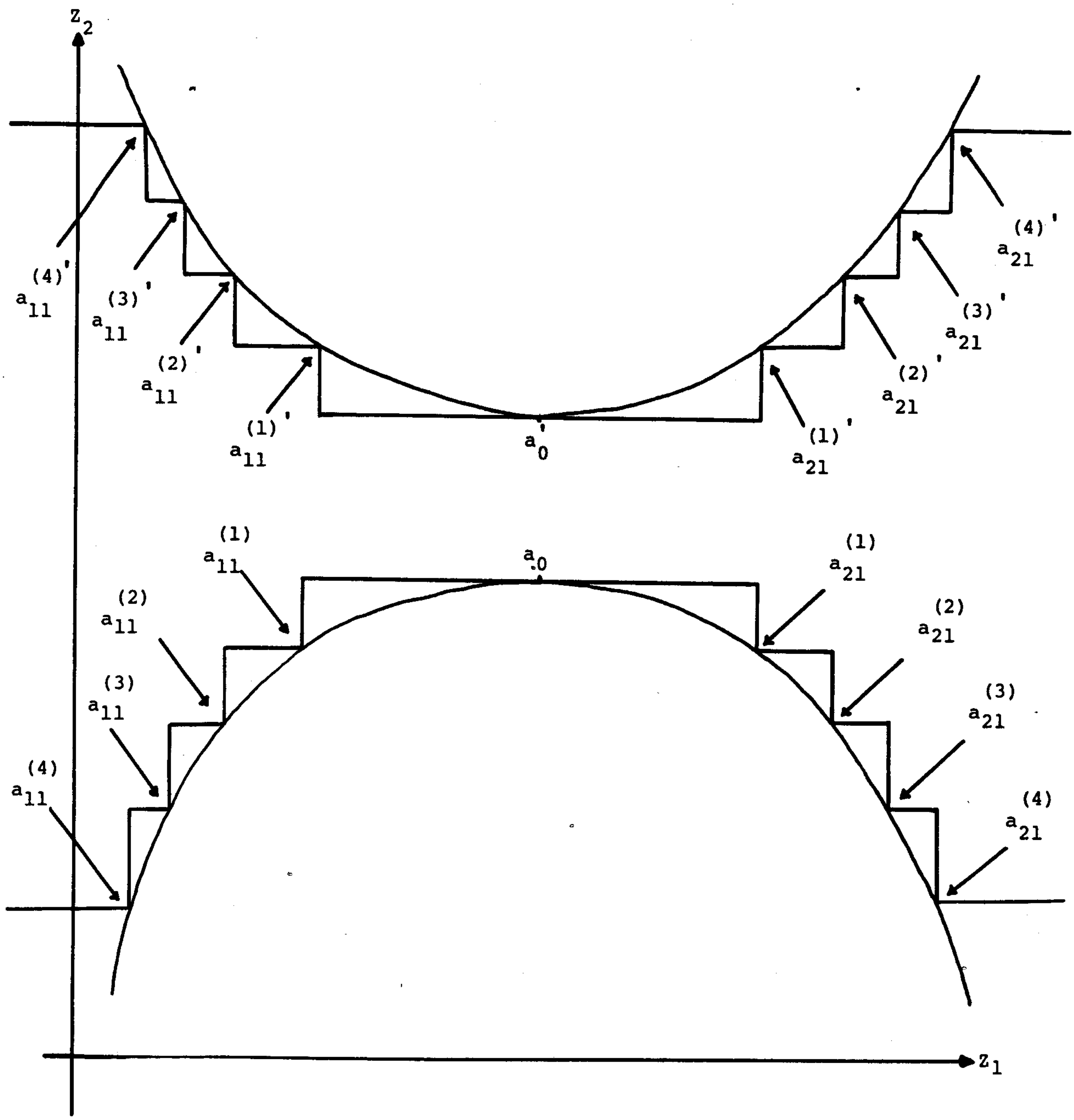

Figure 5. A Tighter Bound on the Hyperboloidal Region of Misclassification