General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

Errata

to

Rice University ICSA Report #275-02.-014

by

R. J. P. de Figueiredo

entitled

"Optimal Linear and Nonlinear Feature Extraction Based on the Minimization of the Increased Risk of Misclassification"

(June, 1974)

Research supported in part by NASA contract NAS 9-12776

Errata

(a) Equation (16) should read

$$\Omega_i(A) = \left\{ y \in E^m(A) : P_i f_Y(y \mid H^i, A) > P_j f_Y(y \mid H^j, A) \right\}$$

(b) Equation (54) should read

$$\left.\begin{matrix} r_1^{ij} \\ r_2^{ij} \end{matrix}\right\} = (R^i - R^j) \left\{ R^i \bar{y}^j - R^j \bar{y}^i \pm \right.$$

$$\pm \left[R^i R^j \left((\bar{y}^i - \bar{y}^j)^2 + \right.\right.$$

$$\left.\left. + (R^i - R^j) \log \left(\frac{R^j}{R^i} \frac{P_j^2}{P_i^2} \right) \right]^{\frac{1}{2}} \right\}$$

$$\text{if } R^i \neq R^j$$

or

$$r_1^{ij} = r_2^{ij} = \frac{1}{2}(\bar{y}^i + \bar{y}^j)$$

$$\text{if } R^i = R^j$$

(c) Equation (55) should read

$$P_i f_Y(y \mid H^i, A) = P_j f_Y(y \mid H^j, A),$$

$$i = 1, \ldots, M, \quad j \neq i.$$

Optimal Linear and Nonlinear Feature Extraction Based on the Minimization of the Increased Risk of Misclassification

by
Rui J.P. de Figueiredo
Dept. of Electrical Engineering
and Dept. of Mathematical Sciences
Rice University

ABSTRACT

We consider the problem of determining an optimal not necessarily linear transformation A from a real n-dimensional measure space E^n, in which the raw data to be classified into M ($M \geq 2$) pattern classes appear, to a "feature space" E^m of a prescribed dimension $m<n$ in which classification is made. The Bayes risk in the transformed space E^m, called the "increased risk of misclassification", depends on A and hence will be denoted by $Q_m(A)$. We assume that A belongs to a given class $\mathcal{X}$ of transformations from $E^n \rightarrow E^m$, each member of $\mathcal{X}$ being a prescribed function of a vector parameter $a= (a_1, \ldots\ldots, a_k)$ characterizing the member. So, given an appropriate class $\mathcal{X}$, we select the optimal $\hat{A}$ by minimizing $Q_m(A)$ over all $A \in \mathcal{X}$. Necessary and sufficient conditions for the existence of such an $\hat{A}$ are given, and an iterative algorithm for the determination of $\hat{A}$ is presented. Finally the results obtained are particularized for the case in which the statistics of the data are Gaussian.

Institute for Computer Services & Applications
Rice University
Houston, Texas 77001

June, 1974

Research supported in part under NASA contract NAS-9-12776

OPTIMAL LINEAR AND NONLINEAR FEATURE EXTRACTION BASED ON THE MINIMIZATION OF THE INCREASED RISK OF MISCLASSIFICATION*

Rui J. P. de Figueiredo**
Department of Electrical Engineering
and
Department of Mathematical Sciences
Rice University, Houston, Texas 77001

We consider the problem of determining an optimal not necessarily linear transformation A from a real n-dimensional measure space E^n, in which the raw data to be classified into M ($M \geq 2$) pattern classes appear, to a "feature space" E^m of a prescribed dimension $m < n$, in which classification is to be made. The Bayes risk in the transformed space E^m, called the "increased risk of misclassification", depends on A and hence will be denoted by $Q_m(A)$. We assume that A belongs to a given class χ of transformations from $E^n \to E^m$, each member of χ being a prescribed function of a vector parameter $a = (a_1,\ldots,a_k)$ characterizing the member. For example, if χ is the class of linear transformations, then members of χ are constant m x n matrices, the components of the vector parameter a characterizing a given matrix A consisting of the mn elements of that matrix. So, given an appropriate class χ, we select the optimal $\hat{A}$ by minimizing $Q_m(A)$ over all $A \in \chi$. Necessary and sufficient conditions for the existence of such an $\hat{A}$ are given, and an iterative algorithm for the determination of $\hat{A}$ is presented. Finally, the results obtained are particularized for the case in which the statistics of the data are Gaussian.

1. Introduction

Suppose that a data vector $x = \mathrm{col}(x_1,\ldots,x_n)$, belonging to the real n-dimensional Euclidian space E^n, is to be classified as pertaining to one of the M pattern classes $H^1,\ldots,H^M$. Then x may be considered to be a realization of a random vector $X = \mathrm{col}(X_1,\ldots,X_n)$. We will assume that X_i, $i=1,\ldots,n$, are continuous random variables possessing well defined probability density functions.

For $j=1,\ldots,M$, let P_j denote the prior probability for the pattern class H^j, and $f_X(./H^j)$ the probability density function*** for X conditioned on the class H^j (called the likelihood function for the class H^j). Note that, endowed with these probabilities and likelihood functions, E^n becomes a measure space.

We will assume that P_j, $j=1,\ldots,M$, are known and $f_X(./H^j)$, $j=1,\ldots,M$, can be learned from available training sets. The functions $f_X(./H^j)$, together with their first and second partial derivatives with respect to the components of x, will be assumed to be continuous and integrable on E^n.

Given an integer m, such that $1 \leq m < n$, let A be a function belonging to a given class χ of functions from E^n to $E^m(A)$. Here the m-dimensional Euclidian space E^m is shown to be a function of A because the measure on E^m (introduced by the prior probabilities and the likelihood functions in E^m) is dependent on the transformation A.

In order to formulate the optimal feature extraction problem, we need to be given one more entity, namely a criterion functional, whose value corresponding to a given A will be denoted by

$$Q(A;\ P_1,\ldots,P_M;\ f_X(./H^1),\ldots,f_X(./H^M)), \qquad (1a)$$

which, when the other arguments are clear from the context, will be written simply as

$$Q(A)\ . \qquad (1b)$$

Then the optimal feature extraction problem may be stated precisely as follows:

Problem 1: Given P_j and $f_X(./H^j)$, $j=1,\ldots,M$, a class χ, and a criterion functional Q, all defined as above, find $\hat{A}$ which minimizes* $Q(A)$ over all $A \in \chi$.

In the existing literature (see for example [1] through [6] and the references therein), solutions to Problem 1 have been obtained assuming Gaussian statistics, using classes of linear transformations, and based on criterion functionals Q that are probabilistic distances, such as the divergence, the Bhattacharyia distance, and the Matusita distance. In general, such distances lead to solutions that are at best suboptimal, that is, these solutions minimize a bound on the risk of misclassification rather than the risk of misclassification itself.

In what follows, we propose to solve Problem 1 by choosing the Bayes risk of misclassification, and in particular the probability of misclassification, as the criterion functional to be minimized. While the proposed solution may require more computational effort than the solutions based on probabilistic distances mentioned above, it (the proposed solution) is believed to be of great value for the following two reasons: (1) the feature extraction computation is a "design computation" which is performed off-line and only once and hence the greater computational effort which may be required does not constitute a basic limitation; and (2) the proposed solution would give the maximum possible accuracy in classification achievable in a space of a prescribed dimension m.

2. Feature Extraction Based on the Minimization of the Increased Risk of Misclassification; Problem Formulation

If A is a transformation which sends $x \in E^n$ to $y \in E^m$ we may write

$$y = \mathrm{col}(y_1,\ldots,y_m) = A(x) = \mathrm{col}(A_1(x),\ldots,A_m(x)). \qquad (2)$$

For convenience, we will use the notation

$$\tilde{x} = \mathrm{col}(x_1,\ldots,x_m), \qquad (3)$$

*Supported in part by the NASA Contract NAS-9-12776, the U.S. Army Contract No. DA-31-124-ARO-D-462, and the NSF Grant GK-36375.

**Part of this work was performed while the author held a visiting research professorship at the Mathematics Research Center of the University of Wisconsin at Madison, in the academic year 1972-1973.

***We will denote the function by $f_X(./H^j)$ and its value at x by $f_X(x/H^j)$.

*If the criterion functional involves a probabilistic distance measure to be maximized (rather than minimized), such as the divergence or the Bhattacharyia distance, we define Q to be the negative of such a distance, and minimize Q.

$$\tilde{\tilde{x}} = \mathrm{col}(x_{m+1},\ldots,x_n), \tag{4}$$

and thus express

$$A(x) = A(\tilde{x}, \tilde{\tilde{x}}), \tag{5}$$

$$A_i(x) = A_i(\tilde{x}, \tilde{\tilde{x}}),\ i = 1,\ldots,m. \tag{6}$$

Let us introduce the Jacobian determinant

$$J_A(x) \equiv J_A(\tilde{x}, \tilde{\tilde{x}}) = \begin{vmatrix} \frac{\partial A_1(\tilde{x},\tilde{\tilde{x}})}{\partial x_1} & \cdots & \frac{\partial A_1(\tilde{x},\tilde{\tilde{x}})}{\partial x_m} \\ \cdots & & \\ \frac{\partial A_m(\tilde{x},\tilde{\tilde{x}})}{\partial x_1} & \cdots & \frac{\partial A_m(\tilde{x},\tilde{\tilde{x}})}{\partial x_m} \end{vmatrix}. \tag{7}$$

From now on we will assume that the class χ consists of (not necessarily linear) transformations A from E^n to E^m such that:

(a) The (pure and mixed) second partial derivatives of A(x) with respect to the components of x are continuous;

(b) Except possibly on subsets of E^n where all the likelihood functions vanish, the mapping under (2) of $\tilde{x}$ to y is one-to-one for every $\tilde{\tilde{x}}$; and in particular, $J_A(x) \neq 0$ everywhere, except possibly on the above subsets of E^n.

Under conditions (a) and (b) above, we may, in the region of interest, express the variables $x_1,\ldots,x_m$ in terms of $y_1,\ldots,y_m$, and $x_{m+1},\ldots,x_n$ by inverting (2). Specifically, there is a unique transformation $B: E^n \to E^m$ such that

$$\tilde{x} = B(y,\tilde{\tilde{x}}) \tag{8}$$

for all $\tilde{x}$, $\tilde{\tilde{x}}$, and y satisfying (2) in the region of interest. According to a wellknown procedure[8], the likelihood functions in $E^m(A)$ would then be given by

$$f_Y(y/H^j,A) = \int_{-\infty}^{\infty} dx_n \int_{-\infty}^{\infty} dx_{n-1} \cdots \int_{-\infty}^{\infty} dx_{m+1} \left\{ \frac{f_X(B(y,\tilde{\tilde{x}}),\tilde{\tilde{x}}/H^j)}{|J_A(B(y,\tilde{\tilde{x}}),\tilde{\tilde{x}})|} \right\}. \tag{9}$$

Remark 1: At the expense of complicating our presentation but otherwise adding no difficulty to our formulation, we could have enlarged the class χ of transformations defined above by means of the two weakening conditions: (I) Allow the class χ to include all transformations A for which the vector $\tilde{x}$ consists of any combination of m variables from the set $\{x_1,\ldots,x_n\}$ (rather than only the first m variables from this set) provided conditions (a) and (b), with appropriate ammendments in notation, are satisfied. (II) Weaken condition (b) so that for a given $\tilde{\tilde{x}}$ and y the equation $A(\tilde{x},\tilde{\tilde{x}}) = y$ is permitted to have a finite number of multiple roots, say $\tilde{x}^{(1)} \equiv B^{(1)}(y,\tilde{\tilde{x}}),\ldots,\tilde{x}^{(k)} \equiv B^{(k)}(y,\tilde{\tilde{x}})$. In a standard way, the integrand in (9) would be replaced by

$$\sum_{\ell=1}^{k} \frac{1}{|J_A(B^{(\ell)}(y,\tilde{\tilde{x}}),\tilde{\tilde{x}})|} f_X(B^{(\ell)}(y,\tilde{\tilde{x}}),\tilde{\tilde{x}}/H^j). \tag{10}$$

(End of Remark)

For i,j = 1,...,M, let the nonnegative number c_{ij} represent the cost of classifying a data vector x as arising from H^i when actually it originated from H^j. Again for simplicity in presentation and without loss of generality, we will assume that there is no cost involved in making a correct decision, i.e. that $c_{ii} = 0$, i = 1,...,M .

It is a well known and easily proved fact that, due to the reduction in dimensionality in going from E^n to $E^m(A)$, the Bayes risk of misclassification in $E^m(A)$, denoted by $Q_m(A)$, is greater than that in E^n. For this reason, $Q_m(A)$ will be called the increased risk of misclassification and is expressed by

$$Q_m(A) = \sum_{i=1}^{M} \int_{\Omega_i(A)} \ell_i(y,A)\, dy, \tag{11}$$

where

$$\ell_i(y,A) = \sum_{\substack{j=1 \\ j\neq i}}^{M} c_{ij} P_j f_Y(y/H^j,A),\ i=1,\ldots,M, \tag{12}$$

and $\Omega_i(A)$, i=1,...,M, are decision regions in $E^m(A)$, that is, if $y \in \Omega_i(A)$ one says that it arose from H^i.

Elementary decision theory also tells us that (for a given A) the choice of $\Omega_i(A)$, i=1,...,M, which minimizes $Q_m(A)$ is given by

$$\Omega_i(A) = \{y \in E^m(A): \ell_i(y,A) < \ell_j(y,A), j\neq i\},$$
$$i = 1,\ldots, M, \tag{13}$$

and, in the particular case in which the cost constants are

$$c_{ij} = 1 - \delta_{ij}, i,j=1,\ldots,M, \tag{14}$$

where δ_{ij} = Kronecker delta, (11) becomes the probability of misclassification, (12) and (13) then reducing respectively to

$$\ell_i(y,A) = \sum_{\substack{j=1 \\ j\neq i}}^{M} P_j f_Y(y/H^j,A), \tag{15}$$

and

$$\Omega_i(A) = \{y \in E^m(A): P_i f_Y(y/H^j,A) > P_j f_Y(y/H^j,A), j\neq i\}.$$
$$i = 1,\ldots,M \tag{16}$$

We are thus able to reformulate Problem 1 as follows:

Problem 2: Given P_j and $f_X(./H^j)$, j=1,...,M, the class χ of functions from E^n to $E^m(A)$ satisfying conditions (a) and (b) above, and the criterion functional Q_m defined by (11), find $\hat{A} \in \chi$ which minimizes $Q_m(A)$ over all $A \in \chi$.

In order to simplify our analysis, we will introduce two additional conditions (c) and (d) to be stated below.

(c) Every transformation A belonging to χ is expressible as $A(x) = \varphi(x,a)$, where $a = \mathrm{col}(a_1,\ldots,a_k)$, belonging to a compact subset $\mathcal{A}$ of E^k, is a real parameter vector and φ is a fixed function from E^{n+k} to E^m; in other words, each member of χ is obtained by assigning a different value to the parameter vector a in the argument of the known function $\varphi(x,\cdot)$. We will assume that φ has continuous second partial derivatives with respect to the components of x and a; and that*

$f_Y(y/H^j,a)$, $\partial f_Y(y/H^j,a)/\partial a_p$, and $\partial^2 f(y/H^j,a)/\partial a_p \partial a_q$, p,q=1,...,k, are continuous and integrable in the product spaces spanned respectively by the variables y, y and a_p, and y, a_p and a_q.

Remark 2: The above condition is not too restrictive. For example, the class of all linear transformations from E^n to E^m, whose representation consists of

*In view of the condition just assumed we will from now on replace capital A by small a in the notation appearing in (9) through (16), e.g. $f_Y(y/H^j,a)$ instead of $f_Y(y/H^j,A)$, except when A denotes a matrix.

m × n real constant matrices of bounded norm satisfies (c). In fact, the number k of parameters in this case is simply the total number m × n of entries in any such matrix.

(End of Remark)

If ϕ is a function of y and a let its gradients with respect to these vectors be defined in the usual way:

$$\nabla_y \phi(y,a) = \mathrm{col}\left(\frac{\partial}{\partial y_1}\phi(y,a),\ldots,\frac{\partial}{\partial y_m}\phi(y,a)\right), \tag{17a}$$

$$\nabla_a \phi(y,a) = \mathrm{col}\left(\frac{\partial}{\partial a_1}\phi(y,a),\ldots,\frac{\partial}{\partial a_k}\phi(y,a)\right). \tag{17b}$$

Denote by $\mathcal{B}_{ij}(a)$ the boundary between $\Omega_i(a)$ and $\Omega_j(a)$, that is

$$\mathcal{B}_{ij}(a) = \{y \in E^m(a): \ell_i(y,a) = \ell_j(y,a), \ell_i(y,a) \le \ell_p(y,a), p \ne i,j\}, \tag{18}$$

and call $\mathcal{B}(a)$ the union of $\mathcal{B}_{ij}(a)$, $i=1,\ldots,M, j\ne i$.

In order to avoid singular points in the description of $\mathcal{B}(a)$, we require that

(d) For every nonzero $a \in \mathcal{K}$, and $i=1,\ldots,M$, $j\ne i$,

$$\nabla_y(\ell_i(y,a)-\ell_j(y,a)) \ne \underline{0},\ y \in \mathcal{B}_{ij}. \tag{19}$$

From now on, we will consider, instead of Problem 2,:

Problem 3: Same as Problem 2 with the additional restrictions (c) and (d).

3. Necessary and Sufficient Conditions for an Optimal Transformation

Consider the Hessian matrix*

$$H_a Q_m(a) = \nabla_a \nabla_a^T Q_m(a). \tag{20}$$

We first assert that under the above conditions, $\nabla_a Q_m$ and $H_a Q_m$ can be evaluated by appropriately carrying out the differentiation operations under the integral sign.

Lemma 1. Under the conditions stated,

$$\nabla_a Q_m(a) = \sum_{i=1}^{M} \int_{\Omega_i(a)} dy(\nabla_a \ell_i(y,a)). \tag{21}$$

Proof. Define the function $\ell: E^{k+m} \to E^1$ by

$$\ell(y,a) = \ell_i(y,a),\ y \in \Omega_i(a) \quad \text{or}$$

$$y \in \mathcal{B}_{ij}(a),$$

$$i = 1,\ldots,M, j \ne i. \tag{22}$$

From our conditions, ℓ is continuous on E^{k+m} and, for any given a and $\Omega_i(a)$ the first and second partials of ℓ with respect to the components of a are continuous on $\Omega_i(a)$ and approach continuous limits as y tends to the boundary of $\Omega_i(a)$. On the boundary, the partials have a simple discontinuity. Since $\mathcal{B}(a)$ is of Lebesgue measure zero in $E^m(a)$ we may write

$$Q_m(a) = \sum_{i=1}^{M} \int_{\Omega_i(a)} \ell_i(y,a)\,dy = \int_{E^m} \ell(y,a)dy, \tag{23}$$

where we have purposely dropped the argument of $E^m(a)$ since it is immaterial in this calculation.

For any given integer q, $1 \le q \le k$, let

$$g(a) = \int_{E^m} dy\, \frac{\partial \ell(y,a_1,\ldots,a_{q-1},a_q,a_{q+1},\ldots,a_k)}{\partial a_q}. \tag{24}$$

Since for every $a_1,\ldots,a_{q-1},a_{q+1},\ldots,a_k$, the integrand in (24) is integrable in the product space spanned by the variables a_q and y, it follows, invoking Fubini's theorem in a standard way (with a_{qo} an arbitrary real constant and $\tilde{a}_q$ a variable of integration) that

$$\int_{a_{qo}}^{a_q} d\tilde{a}_q\, g(a_1,\ldots,a_{q-1},\tilde{a}_q,a_{q+1},\ldots,a_k)$$

$$= \int_{a_{qo}}^{a_q} d\tilde{a}_q \int_{E^m} dy\, \frac{\partial \ell(y,a_1,\ldots,a_{q-1},\tilde{a}_q,a_{q+1},\ldots,a_k)}{\partial \tilde{a}_q}$$

$$= \int_{E^m} dy \int_{a_{qo}}^{a_q} d\tilde{a}_q\, \frac{\partial \ell(y,a_1,\ldots,a_{q-1},\tilde{a}_q,a_{q+1},\ldots,a_k)}{\partial \tilde{a}_q}$$

$$= \int_{E^m} dy[\ell(y,a) - \ell(y,a_1,\ldots,a_{q-1},a_{qo},a_{q+1},\ldots,a_k)]$$

$$= Q_m(a) - Q_m(a_1,\ldots,a_{q-1},a_{qo},a_{q+1},\ldots,a_k). \tag{25}$$

But from the leftmost and rightmost members of the equalities (25), we conclude that Q_m is the antiderivative (with respect to a_q) of g or

$$g(a) = \frac{\partial Q_m(a)}{\partial a_q}. \tag{26}$$

(24) and (26) for q=1,...,k, then establish the validity of (21). Q.E.D.

Lemma 2. Under the conditions stated,

$$H_a Q_m(a) = \sum_{i=1}^{M} \int_{\Omega_i(a)} dy(\nabla_a \nabla_a^T \ell_i(y,a)) -$$

$$-\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \int_{S_{ij}(a)} ds\, \nabla_a(\ell_i(y,a)-\ell_j(y,a))\nabla_a^T(\ell_i(y,a)-\ell_j(y,a)), \tag{27}$$

where the second set of integrals consists of surface integrals on the subsets of $\mathcal{B}(a)$ corresponding to the boundaries of not more than two regions. Thus $S_{ij}(a)$ is that subset of $\mathcal{B}_{ij}(a)$ which is the common boundary between $\Omega_i(a)$ and $\Omega_j(a)$ only.

Proof. As in the previous proof, we will carry out the integrations over the entire E^m. However our presentation will be simplified using indicator functions for the regions $\Omega_i(a)$.

In fact let the function $u: E^1 \to E^1$ be defined by

$$u(\xi) = \begin{cases} 1, & \xi > 0, \\ 0, & \xi \le 0. \end{cases} \tag{28}$$

Then, (21) may be rewritten as

$$\nabla_a^T Q_m(a) = \sum_{i=1}^{M} \int_{E^m} dy\Big(\Big(\prod_{\substack{j=1 \\ j\ne i}}^{M} u(\ell_j(y,a)-\ell_i(y,a))\Big)\nabla_a^T \ell_i(a,y)\Big). \tag{29}$$

Now, provided we are willing to admit distribution functions, we may transfer further differentiation operations to within the integral sign; that is

*Henceforth the superscript T on a symbol will denote its transpose.

$$\nabla_a \nabla_a^T Q(a) = \sum_{i=1}^{M} \int_{E^m} dy \Big\{ \sum_{\substack{j=1 \\ j\neq i}}^{M} [\delta(\ell_j(y,a)-\ell_i(y,a))]\cdot$$

$$[\nabla_a(\ell_j(y,a)-\ell_i(y,a))]\Big[\prod_{\substack{r=1 \\ r\neq i \\ r\neq j}}^{M} u(\ell_r(y,a)-\ell_i(y,a))\Big]\cdot$$

$$[\nabla_a^T \ell_i(y,a)]$$

$$+ \prod_{\substack{j=1 \\ j=i}}^{M} u(\ell_j(y,a)-\ell_i(y,a)) \nabla_a \nabla_a^T \ell_i(y,a)\Big], \qquad (30)$$

where $\delta(\cdot)$ denotes the delta function.

The last product term in (30) clearly leads to the first integral in (27).

Consider the remaining set of terms, all in square brackets, in (30). The first (delta function) term in square brackets reduces the integration of the set of terms under consideration to a surface integral on $\mathcal{B}_{ij}(a)$. If $y \in \mathcal{B}_{ij}(a)$ is on the boundary of more than two decision regions, say $\Omega_i(a)$, $\Omega_j(a)$, and $\Omega_p(a)$, the product term in the third set of square brackets vanishes because $u(\ell_p(y,a)-\ell_i(y,a)) = 0$ there. Thus points on $\mathcal{B}_{ij}(a)$ common to boundaries of more than two decision regions make no contribution to the surface integrals under discussion. If y is on the boundary common to $\Omega_i(a)$ and $\Omega_j(a)$ only, then the product term in the third set of square brackets is unity there. Thus integration on such points of $\mathcal{B}_{ij}(a)$ leads to

$$\int_{S_{ij}(a)} ds\, \nabla_a(\ell_j(y,a)-\ell_i(y,a))\nabla_a^T \ell_i(y,a). \qquad (31)$$

A similar calculation with the values of i and j interchanged leads to a surface integral of the form

$$\int_{S_{ij}(a)} ds\, \nabla_a(\ell_i(y,a)-\ell_j(y,a))\nabla_a^T \ell_j(y,a)\ . \qquad (32)$$

Thus the total contribution to the surface integral made by $\mathcal{B}_{ij}(a)$ is

$$\int_{S_{ij}(a)} ds\, \nabla_a(\ell_j(y,a)-\ell_i(y,a))\nabla_a^T(\ell_i(y,a)-\ell_j(y,a)), \qquad (33)$$

which establishes the validity of the second (double summation) term in (27). Q.E.D.

Remark 3: Before proceeding any further, it should be pointed out that since $\nabla_a Q_m$ is continuous for almost all a and y, it follows from the Lebesgue dominated convergence theorem[9] that $\nabla_a Q_m$ is continuous for all a.

In order to describe the set $\mathcal{X}$, let us now assume that we are given a function Θ of the variable a, with values in E^q, where $q \leq k$, satisfying:

(e) Θ has continuous second partials and is such that $\|\Theta(a)\| \to \infty$ as $\|a\| \to \infty$.

Following convention, we say that $\hat{a} \in E^k$ is a regular point of Θ if $\nabla_a \Theta^T(\hat{a})$ is of rank q.

We set

$$\mathcal{X} = \{a : \Theta(a) = c = \text{given constant vector}\}. \qquad (34)$$

Necessary and sufficient conditions for the existence of an optimal $\hat{a}$ are formulated in terms of the following two theorems.

Theorem 1. Let conditions (a) through (e) hold and $\mathcal{X}$ be defined as in (34). Then: (i) Problem 3 always has a solution; and (ii) if $\hat{a}$ is such a solution and $\hat{a}$ is a regular point of Θ, then there is a $\hat{\lambda} \in E^q$ such that

$$\sum_{i=1}^{M} \int_{\Omega_i(\hat{a})} dy\, (\nabla_a \ell_i(y,\hat{a})) + \sum_{s=1}^{q} \hat{\lambda}_s \nabla_a \Theta_s(\hat{a}) = \underline{0}. \qquad (35)$$

Proof. $\mathcal{X}$ described by (34) is clearly compact in E^k. Since Q_m is continuous on $\mathcal{X}$, part (i) follows from Weierstrass's Theorem[10]. Part (ii) follows from Lemma 1 and wellknown optimization theory results[11]. Q.E.D.

Theorem 2. Let the conditions as well as $\hat{a}$ and $\hat{\lambda}$ be as described in the preceding theorem. $\hat{a}$ is a local minimum of Q_m on $\mathcal{X}$ if there is an $\epsilon > 0$ such that

$$b^T V^T \Big[H_a Q_m(\hat{a}) + \sum_{s=1}^{q} \hat{\lambda}_s \nabla_a \nabla_a^T \Theta_s(\hat{a})\Big] V b \geq \epsilon \|b\|^2 \qquad (36)$$

for all $b \in E^{k-q}$, where $H_a Q_m(a)$ is as in (27), V is a $k\times(k-q)$ matrix whose columns span the null space of the matrix $(\nabla_a \Theta(\hat{a}))^T$.

Proof. Immediate from Lemma 2 and wellknown results[11] from optimization theory. Q.E.D.

Remark 4: The following is also clear from wellknown facts from optimization theory. Let ξ denote the subset of of $\{1, \ldots, q\}$ such that Θ_i is nonlinear if and only if $i \in \xi$ (ξ may be empty). By enlarging if necessary $\mathcal{X}$, extend it to the set $\mathcal{X}_1$ by replacing in (34) the equality sign by $\leq$ for those Θ_i with $i \in \xi$. If Θ_i, $i \in \xi$, are convex (and hence $\mathcal{X}_1$ is convex) and Q_m is convex on $\mathcal{X}_1$, then $\hat{a}$ satisfying Theorem 2 is a point of global minimum of Q_m on $\mathcal{X}$.

Next we focus our attention on the following important special case.

4. Linear Feature Extraction of Gaussian Features

Let us in fact particularize the results just described to the case in which the statistics of the pattern classes are Gaussian and the class χ of transformations from E^n to E^m is linear. Specifically, we let χ be a compact subset of the real finite-dimensional inner product space $\mathcal{M}$ of real $m\times n$ matrices with the inner product between any two elements A and B of $\mathcal{M}$ defined by

$$\langle A,B\rangle = \mathrm{tr}(AB^T) = \sum_{i=1}^{M}\sum_{j=1}^{M} A_{ij}B_{ij}, \qquad (37)$$

where the abbreviation tr stands for trace.

It is a simple matter to show that (37) is a valid inner product in $\mathcal{M}$.

If g is a real-valued function of a matrix-valued variable A belonging to $\mathcal{M}$, such that at some value of A, say $\tilde{A}$, g is differentiable with respect to the elements of A, one can show from the abstract definition of the gradient[12], that the gradient of g with respect to A, evaluated at $\tilde{A}$, is simply the member $\nabla_A g(\tilde{A})$ of $\mathcal{M}$ whose ij^{th} element is

$$(\nabla_A g(\tilde{A}))_{ij} = \frac{\partial g(\tilde{A})}{\partial A_{ij}}\ . \qquad (38)$$

Formulas for matrix gradients of various types of real-valued functions of matrices have been derived in reference [13]. Recently, Decell and Quirein[6] have used such formulas to express gradients[1-4] of the divergence and Bhattacharyya distance in an easily computable form.

In what follows, we apply the results of Theorem 1 to the Gaussian case by first computing $\nabla_A f_{Y}(y/H^j,\tilde{A})$, at a given $\tilde{A} \in \mathcal{M}$. However, in applying the results of Theorem 2, since the Hessian $H_A f_{Y}(y/H^j,\hat{A})$ is a linear operator that cannot be represented in matrix form without destroying the matrix structure of the element of $\mathcal{M}$ on which it is acting, we obtain instead the

matrix $H_{\hat{A}} f_Y(y/H^j,\tilde{A})B$, for arbitrary $B_{\hat{A}}$ and hence, by integration, $H_{\hat{A}} Q_m(\hat{A})B$. Note that $H_{\hat{A}} Q_m(\hat{A})B$ is all that is needed in connection with (36), where the vector Vb now corresponds to the matrix $B \in \mathcal{M}$.

$H_{\hat{A}} f_Y(y/H^j,\tilde{A})B$ is the Gateau differential[12] $\delta_{\hat{A}} \nabla_A i_Y(y/H^j,A;B)$ of $\nabla_A f_Y(y/H^j,.)$ at $\tilde{A}$ along B and is easily computable from the formula:

$$\delta_A \nabla_A f_Y(y/H^j,\tilde{A};B) =$$

$$\lim_{t\to 0} \frac{1}{t} [\nabla_A f_Y(y/H^j,\tilde{A}+tB) - \nabla_A f_Y(y/H^j,\tilde{A})] = \quad (39)$$

$$[\frac{\partial}{\partial t} \nabla_A f_Y(y/H^j,\tilde{A}+tB)]_{t=0}, \quad t \in E^1. \quad (40)$$

Returning now to our original problem, under the normality hypothesis the probability density functions conditioned on pattern classes, in the transformed space $E^m(A)$ are

$$f_Y(y/H^j,A) = (2\pi)^{-\frac{m}{2}} |R^j|^{-\frac{1}{2}} \exp[-\frac{1}{2}(y-\bar{y}^j)^T (R^j)^{-1}(y-\bar{y}^j)], \quad j=1,\ldots,M, \quad (41)$$

where A is a $m \times n$ matrix (belonging to χ), $|R^j|$ denotes the determinant of R^j, and $\bar{y}^j$ and R^j are the mean and covariance pertaining to H^j in $E^m(A)$. The latter two entities are related to the given mean $\bar{x}^j$ and covariance* $\tilde{R}^j$ associated with H^j in the original space E^n by

$$\bar{y}^j = A\,\bar{x}^j, \quad (42)$$

$$R^j = A\,\tilde{R}^j\,A^T. \quad (43)$$

We will require:

Lemma 3. For $f_Y(y/H^j,A)$, $j=1,\ldots,M$, as in (41),

$$\nabla_A f_Y(y/H^j,A) =$$

$$f_Y(y/H^j,A)[((R^j)^{-1}(y-\bar{y}^j)(y-\bar{y}^j)^T - I)(R^j)^{-1}A\tilde{R}^j + (R^j)^{-1}(y-\bar{y}^j)(\bar{x}^j)^T]. \quad (44)$$

Proof. From (41) we obtain

$$\nabla_A f_Y(y/H^j,A)$$

$$= ((2\pi)^{-\frac{m}{2}} \exp[-\tfrac{1}{2}(y-\bar{y}^j)^T (R^j)^{-1}(y-\bar{y}^j)])([\nabla_A |A\tilde{R}^j A^T|^{-\frac{1}{2}}]$$

$$- \tfrac{1}{2}|R^j|^{-\frac{1}{2}} [\nabla_A \{(y-A\bar{x}^j)^T (A\tilde{R}^j A^T)^{-1}(y-A\bar{x}^j)\}]). \quad (45)$$

The term in the first set of square brackets gives

$$\nabla_A |A\tilde{R}^j A^T|^{-\frac{1}{2}} = -\tfrac{1}{2}|R^j|^{-\frac{3}{2}} \nabla_A |R^j|$$

$$= -\tfrac{1}{2}|R^j|^{-\frac{3}{2}} |R^j| \mathrm{tr}[(R^j)^{-1} \nabla_A (A\tilde{R}A^T)]$$

$$= -|R^j|^{-\frac{1}{2}} (R^j)^{-1} A\tilde{R}^j, \quad (46)$$

where here, as in what follows, we use the fact that $\tilde{R}^j$ is symmetric; while the term in the second set of square brackets leads to

$$\nabla_A [(y-A\bar{x}^j)^T (A\tilde{R}^j A^T)^{-1}(y-A\bar{x}^j)]$$

$$= [\nabla_A (y-A\bar{x}^j)^T](R^j)^{-1}(y-\bar{y}^j)$$

$$+ (y-\bar{y}^j)^T (R^j)^{-1}[\nabla_A (A\tilde{R}^j A^T)](R^j)^{-1}(y-\bar{y}^j)$$

$$+ (y-\bar{y}^j)^T (R^j)^{-1}[\nabla_A (y-A\bar{x}^j)]$$

$$= -2(R^j)^{-1}(y-\bar{y}^j)(\bar{x}^j)^T$$

$$- 2(R^j)^{-1}(y-\bar{y}^j)(y-\bar{y}^j)^T (R^j)^{-1} A\tilde{R}^j. \quad (47)$$

Substituting (46) and (47) in (45), (44) is established. Q.E.D.

There is a number of ways one may impose constraints to guarantee compactness of χ. One is to require that

$$\frac{1}{2}\|A\|^2 = \frac{1}{2}\mathrm{tr}(AA^T) = \gamma, \quad A \in \mathcal{M}, \gamma = \text{const.} \quad (48a)$$

Another is to allow only those $A \in \mathcal{M}$ consisting of orthonormal row vectors by requiring that

$$A\,A^T = I, \quad A \in \mathcal{M}. \quad (48b)$$

For simplicity in presentation, we will assume

$$\chi = \{A \in \mathcal{M} : A \text{ satisfies (48a)}\}. \quad (48c)$$

An additional important consideration concerning restrictions on the set χ is spelt out in Remark 7, at the end of this section.

By virtue of the above Lemma, Theorem 1 clearly reduces to:

Theorem 3. Suppose that in Problem 3 the pattern classes Ω_i, $i = 1,\ldots,M$, are Gaussian with means and covariances $\bar{x}^j$ and $\tilde{R}^j$, $j = 1,\ldots,M$, and χ is as in (48c). Then the problem always has a solution. At any such solution $\hat{A}$ it is necessary that the following matrix equation be satisfied

$$\sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} c_{ij} P_j [[(\hat{R}^j)^{-1} \hat{D}^{ij} - I](\hat{R}^j)^{-1} \hat{A}\tilde{R}^j + (\hat{R}^j)^{-1} d^{ij} (\bar{x}^j)^T] + \hat{\lambda}\hat{A} = \underline{0} \quad (49)$$

where I is the identity matrix, $\hat{\lambda}$ is a Lagrange multiplier (to be calculated using (48a),

$$D^{ij} = \int_{\Omega_i(A)} (y-\bar{y}^j)(y-\bar{y}^j)^T f_Y(y/H^j,A)\,dy, \quad (50)$$

$$d^{ij} = \int_{\Omega_i(A)} (y-\bar{y}^j) f_Y(y/H^j,A)\,dy, \quad (51)$$

and the circumflex on a symbol denotes that the corresponding quantity is to be calculated using $\hat{A}$.

Remark 5. The iterative algorithm outlined in the following section requires the left side of (49) to be computed at every iteration using the estimate of $\hat{A}$ obtained in the preceding iteration. At first sight, the need for the evaluation of the multiple integrals in (50) and (51) at each iteration might appear as a serious drawback of this procedure. However, this difficulty can be avoided if we notice that D^{ij} and d^{ij} are in fact proportional to the covariance and mean of the random variable $(Y^j-\bar{y}^j)$ (where Y^j has the density $f_Y(./H^j,A)$) when restricted to the decision region $\Omega_i(A)$. Thus if training sets for the various pattern classes are available, the integrals in question may be replaced by sample averages over appropriate subsets of training sets. Specifically, suppose that, for $i=1,\ldots,M$, pertaining to H^j we have the training set

* We assume $\tilde{R}^j$ nonsingular and $\tilde{R}^j \neq \tilde{R}^p$, $j=1,\ldots,M$, $p\neq j$, the particularization to the special cases when these conditions do not hold being clear.

T^j in $E^m(A)$ consisting of the samples $y^{j1} = A x^{j1},\ldots,y^{jN_j} = A x^{jN_j}$, where $x^{j1},\ldots,x^{jN_j}$ are the given training samples in E^n. Then we have for estimates of D^{ij} and d^{ij}

$$D^{ij} = \frac{1}{N^j} \sum_{y^{jq} \in \Omega_i(A)} (y^{jq}-\bar{y}^j)(y^{jq}-\bar{y}^j)^T, \tag{52}$$

$$d^{ij} = \frac{1}{N^j} \sum_{y^{jq} \in \Omega_i(A)} (y^{jq}-\bar{y}^j) \tag{53}$$

Remark 6. In the particular case in which m=1, $Y = \sum_{j=1}^{m} A_{1j}X_j$ being a scalar random variable, (50) and (51) become single (rather than multiple) integrals and can therefore be easily computed without one having to go through the route described in the preceding remark. Let us assume that the risk is the probability of misclassification and hence (14) holds, then the decision boundaries are real numbers which are chosen, according to (16), from among the roots

$$\left.\begin{matrix} r_1^{ij} \\ r_2^{ij} \end{matrix}\right\} = (R^i-R^j)^{-1}\{R^i\bar{y}^i-R^j\bar{y}^i \pm [R^iR^j\{(\bar{y}^i+\bar{y}^j)^2+(R^i-R^j)\log((\frac{R^j}{R^i})^{\frac{1}{2}}\ \frac{P_j^2}{P_i^2})\}]^{\frac{1}{2}}\},$$

$$i=1,\ldots,M,\ j\neq i, \tag{54}$$

of the equations

$$f_Y(y/H^i,A)=f_Y(y/H^j,A),\ i=1,\ldots,M,\ j\neq i, \tag{55}$$

these densities being described by (41). Suppose that the i^{th} decision region consists of an interval $\alpha_1^i < y < \alpha_2^i$. Then a trivial calculation resulting from the substitution of (41) in (50) and (51) leads to

$$D^{ij}=R^j\{(2\pi)^{-\frac{1}{2}}[\beta_1^{ij}\exp(-\tfrac{1}{2}(\beta_1^{ij})^2)-\beta_2^{ij}\exp(-\tfrac{1}{2}(\beta_2^{ij})^2)] + \mathrm{Erf}(0,1;\beta_2^{ij}) - \mathrm{Erf}(0,1;\beta_1^{ij})\}, \tag{56}$$

$$d^{ij}=(\frac{R^j}{2\pi})^{\frac{1}{2}}[\exp(-\tfrac{1}{2}(\beta_1^{ij})^2)-\exp(-\tfrac{1}{2}(\beta_2^{ij})^2)], \tag{57}$$

where

$$\mathrm{Erf}(0,1;z) = (2\pi)^{-\frac{1}{2}}\int_0^z e^{-\frac{1}{2}s^2}\,ds, \tag{58}$$

$$\beta_1^{ij}=(R^j)^{-\frac{1}{2}}(\alpha_1^i-\bar{y}^j),\ \beta_2^{ij}=(R^j)^{-\frac{1}{2}}(\alpha_2^i-\bar{y}^j). \tag{59}$$

(End of Remark)

To verify the conditions of Theorem 2 under the Gaussian hypothesis, we first calculate (omitting the derivation) the term appearing under the first summation in (27):

$$L_i(A,B) = \int_{\Omega_i(A)} dy\ \delta_A \nabla_A^T \ell_i(y,A;B) =$$

$$\sum_{\substack{j=1\\ j\neq i}}^{M} c_{ij}P_j\{$$

$$-2(R^j)^{-1}(B\tilde{R}^jA^T+A\tilde{R}^jB^T)(R^j)^{-1}D^{ij}(R^j)^{-1}A\tilde{R}^j$$

$$+(R^j)^{-1}D^{ij}(R^j)^{-1}B\tilde{R}^j$$

$$-(R^j)^{-1}(B\bar{x}^j(d^{ij})^T+(d^{ij})\bar{x}^jB^T)(R^j)^{-1}A\tilde{R}^j$$

$$-(R^j)^{-1}(B\tilde{R}^jA^T+A(\tilde{R}^j)^TB^T)(R^j)^{-1}d^{ij}(\bar{x}^j)^T$$

$$-(R^j)^{-1}B(\tilde{R}^j+\bar{x}^j(\bar{x}^j)^T)$$

$$+(R^j)^{-1}(B\tilde{R}^jA^T+A\tilde{R}^jB^T)(R^j)^{-1}A\tilde{R}^j$$

$$-\mathrm{tr}[(R^j)^{-1}A\tilde{R}^jB^T][((R^j)^{-1}D^{ij}-I)(R^j)^{-1}A\tilde{R}^j+(R^j)^{-1}d^{ij}(\bar{x}^j)^T]$$

$$+G^{ij}(B)(R^j)^{-1}A\tilde{R}^j$$

$$-\varphi^{ij}(B)(R^j)^{-1}A\tilde{R}^j$$

$$+h^{ij}(B)(\bar{x}^j)^T$$

$$+F^{ij}(B)(R^j)^{-1}A\tilde{R}^j$$

$$-(d^{ij})^T(R^j)^{-1}B\bar{x}^j(R^j)^{-1}A\tilde{R}^j$$

$$+v^{ij}(B)(\bar{x}^j)^T\}, \tag{60}$$

where

$$\varphi^{ij}(B)=\int_{\Omega_i(A)} dy\ f_Y(y/H^j,A)(y-\bar{y}^j)^T(R^j)^{-1}A\tilde{R}^jB^T(y-\bar{y}^j) \tag{61}$$

$$h^{ij}(B)=\int_{\Omega_i(A)} dy\ f_Y(y/H^j,A)(y-\bar{y}^j)^T(R^j)^{-1}\cdot A\tilde{R}^jB^T(y-\bar{y}^j)(R^j)^{-1}(y-\bar{y}^j) \tag{62}$$

$$v^{ij}(B)=\int_{\Omega_i(A)} dy\ f_Y(y/H_j,A)(y-\bar{y}^j)^T(R^j)^{-1}B\bar{x}^j(R^j)^{-1}(y-\bar{y}^j) \tag{63}$$

$$F^{ij}(B)=\int_{\Omega_i(A)} dy\ f_Y(y/H^j,A)(y-\bar{y}^j)^T(R^j)^{-1}B\bar{x}^j\cdot(R^j)^{-1}(y-\bar{y}^j)(y-\bar{y}^j)^T \tag{64}$$

$$G^{ij}(B)=\int_{\Omega_i(A)} dy\ f_Y(y/H^j,A)(y-\bar{y}^j)^T(R^j)^{-1}A\tilde{R}^jB^T(y-\bar{y}^j)\cdot(R^j)^{-1}(y-\bar{y}^j)(y-\bar{y}^j)^T. \tag{65}$$

Since all the above integrals are expectations, they may be computed from the training samples in the same way as (52) and (53).

The surface integral terms in (27) are calculated in a similar fashion. In the special case in which the risk is the probability of misclassification, these terms (taking into account the minus sign that precedes the double summation) reduce to

$$\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} N^{ij}(A), \tag{66}$$

where

$$N^{ij}(A) = \int_{S_{ij}(A)} ds\,[\nabla_A f_Y(y/H^i,A) - \nabla_A f_Y(y/H^j,A)]\,[\nabla_A f_Y(y/H^i,A) - \nabla_A f_Y(y/H^j,A)]^T, \qquad (67)$$

where for $\nabla_A f_Y(y/H^i,A)$, $i = 1,\ldots,M$, we have to use the expression given by the right side of (44). Then quantities similar to those in (61) to (65) result which again can be computed from the training samples.

Theorem 4. If $\hat{A}$ satisfies the conditions of Theorem 3, then satisfaction of (36) is equivalent to the requirement that

$$\mathrm{tr}\,\Big\{\Big[\sum_{i=1}^{M-1}\Big(L_i(\hat{A},B) + \sum_{j=i+1}^{M} N^{ij}(\hat{A})B\Big) + L_M(\hat{A},B) + \hat{\Lambda}B\Big]B^T\Big\} \geq \epsilon\,\mathrm{tr}(BB^T) \qquad (68)$$

for every $B \in \mathcal{R}$ such that the transpose of the p^{th} row of B lies in the null space of the p^{th} column of $\hat{A}^T$, $p = 1, \ldots, m$.

Remark 7: Let ξ denote a set containing m distinct elements from $(1,\ldots,n)$ and denote by A_j the j^{th} column of A. For some given ξ, we may wish to restrict the class χ so that for every $A \in \mathcal{R}$ belonging to χ, the columns A_j, $j \in \xi$ constitute a submatrix of rank m. Since Q_m is invariant with respect to non-singular coordinate transformations in the transformed (feature) space, we may, in the case under consideration, set $(A_j : j \in \xi)$ equal to a permutation P of the columns of the unit $m \times m$ diagonal matrix, leaving the $m(n-m)$ elements in the remaining columns free for the optimization procedure. We would optimize these elements by requiring that the corresponding elements in the matrix equation (49) satisfy (49) (with the remaining elements of $\hat{A}$ set equal to the elements of P).

The above remark is particularly useful in the verification of the sufficient conditions. For, suppose that to begin with we let all elements of A vary (since we didn't know at the outset which columns of $\hat{A}$ had rank m) and thereby determined $\hat{A}$ by means of an iterative procedure based on (49). Assume that then we find that the first m columns of $\hat{A}$ constitute a non-singular matrix $\hat{A}_1$. Thus let $\hat{A} = (\hat{A}_1, \hat{A}_2)$. We may then replace $\hat{A}$ by $(I, \hat{A}_1^{-1}\hat{A}_2)$ and restrict the verification of the sufficiency conditions with repect to a smaller subset of matrices B than otherwise required (since the only relevant columns of $\hat{A}$ are the last n - m columns).

5. Computational Algorithm

Various iterative methods such as the Newton and gradient methods and their numerous modifications are available for the computation of the optimal feature extraction transformation. We will limit our discussion to the case of linear feature extraction of Gaussian features, our remarks extending trivially to the general non-Gaussian nonlinear case treated in Section 3.

The values A_p and λ_p of the estimates of the optimal matrix $\hat{A}$ and Lagrange multiplier $\hat{\lambda}$ at the p^{th} iteration are given, according to (49) and (48c) by

$$A_p = A_{p-1} - K_p\Big\{\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} c_{ij} P_j\Big[\big[(R^j_{p-1})^{-1} D^{ij}_{p-1} - I\big](R^j_{p-1})^{-1} A_{p-1} + \tilde{R}^j_{p-1} + R^j_{p-1} d^{ij}_{p-1} (\bar{x}^j)^T\Big] + \lambda_{p-1} A_{p-1}\Big\}, \qquad (69a)$$

$$p = 1, \ldots,$$

$$\lambda_p = \lambda_{p-1} - \rho_p\Big(\frac{1}{2}\,\mathrm{tr}(A_{p-1}A_{p-1}^T) - \gamma\Big), \qquad p = 1, \ldots, \qquad (69b)$$

where all the symbols subscripted with p-1 are to be computed with the values A_{p-1} and λ_{p-1} obtained at $(p-1)^{th}$ iteration, K_p and ρ_p are variable matrix and scalar gains determined according to the iterative method selected, the initial estimates A_o and λ_o are set at convenient values, and D^{ij}_{p-1} and d^{ij}_{p-1} are obtained by (56) and (57) if the dimension m of the feature space is one, and by (52) and (53) otherwise. Note that a convenient way of writing (52) and (53) is

$$D^{ij} = \frac{1}{N^j}\sum_{q=1}^{N^j}(y^{jq} - \bar{y}^j)(y^{jq} - \bar{y}^j)^T \Lambda_i(y^{jq}), \qquad (70)$$

$$d^{ij} = \frac{1}{N^j}\sum_{q=1}^{N^j}(y^{jq} - \bar{y}^j)\Lambda_i(y^{jq}), \qquad (71)$$

where

$$\Lambda_i(y) = \prod_{\substack{j=1\\ j\neq i}}^{M} u(\ell_j(y,A) - \ell_i(y,A)) \qquad (72)$$

We have used the Davidon-Fletcher-Powell[11] (D-F-P) method in implementing (69a,b) on the IBM 370/155 computer of the Rice University Institute for Computer Services and Applications (ICSA). The D-F-P procedure requires function evaluations in addition to gradient evaluations. Evaluation of $Q_m(A)$ at the p^{th} iteration may be carried out by the following estimate justified in the same way as (52) and (53):

$$Q_m(A_p) = \sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M}\frac{1}{N^j}\sum_{q=1}^{N^j}\Lambda_i(y^{jq}), \qquad (73)$$

If one knows beforehand that certain features are significant, they may be retained thus reducing the number of parameters of the matrix A to be determined according to Remark 7.

6. Conclusion

General classes of nonlinear and linear transformations A for the reduction of the dimensionality of the classification (feature) space so that, for a prescribed dimension m of this space, the increase of the misclassification risk is minimized, have been investigated. Necessary conditions that must satisfied by the optimal $\hat{A}$ have been presented and sufficient conditions for a local minimum have been indicated. Even though the sufficiency conditions are very complicated, the necessary conditions lend themselves to the formulation of iterative algorithms for the determination of the optimal transformation. In the proposed approach, the multiple integrals which appear at each step of the iteration are replaced by certain sample averages over training sets, a procedure which permits the carrying out of the required computations with reasonable amount of effort even for values of m not too low.

Testing of the proposed method on remotely sensed data provided by the Johnson Space Center Earth Observations Division is in progress at Rice University ICSA computer facilities, and the numerical results obtained will be discussed in a separate paper.

References

[1] Tou, J. T. and Heydorn, R. P., "Some Approaches to Optimum Feature Extraction", in Computers and Information Sciences-II, edited by J. Tou, Academic Press, New York, 1967.

[2] Kailath, T., "The Divergence and Bhattacharyya Distance Measures in Signal Selection", IEEE Trans. on Communication Theory, vol. 15, pp. 52-60, 1967.

[3] Henderson, T. L. and Lainiotis, D., "Application of State Variable Techniques to Feature Extraction", Proc. 1968 Asilomar Conf. on Circuits and Systems, 1968.

[4] Caprihan, A. and de Figueiredo, R. J. P., "On the Extraction of Pattern Features from Continuous Measurements", IEEE Trans. on Systems Science and Cybernetics, vol. SSC-6, pp. 110-115, 1970.

[5] Meisel, W. S., Computer-Oriented Approaches to Pattern Recognition, Academic Press, N. Y., chapter IX, 1972.

[6] Decell, H. P., Jr. and Quirein, J. A., "An Iterative Approach to the Feature Selection Problem", Proceedings of the Purdue University Conference on Machine Processing of Remotely Sensed Data, pp. 3B1-3B12, October 1973.

[7] Apostol, T. M., Mathematical Analysis, Addison-Wiley, Reading, Mass., pp. 193 and 271, 1957.

[8] Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, New York, pp. 201 and 235, 1965.

[9] Natanson, I. P., Theory of Functions of a Real Variable (Translated from the Russian by L. F. Boron), F. Ungar Publ. Co., New York, vol. II, p. 86, 1960, vol. I, p. 161, 1961.

[10] Royden, H. L., Real Variables, MacMillian, New York, p. 140, 1963.

[11] Fiacco, A. V. and McCormick, G. P., Nonlinear Programming: Sequential Unconstrained Minimization Techniques, J. Wiley, New York, Chapter 2, 1968.

[12] Tapia, R. A., "The Differentiation and Integration of Nonlinear Operators", in Nonlinear Functional Analysis and Applications, edited by L. Rall, Academic Press, New York, pp. 45-99, 1971.

[13] Athans, M. and Schweppe, F. C., "Gradient Matrices and Matrix Calculations", MIT Lincoln Lab Tech Note 1965-53, Lexington, Mass., 1965.

[14] Fletcher, R. and Powell, J., "A Rapidly Converging Descent Method for Minimization", British Computer J., vol. 6, pp. 163-168, 1963.

June, 1974