# SUBSPACE METHOD IN PATTERN RECOGNITION*

Satosi Watanabe and Nikhil Pakvasa

University of Hawaii
Honolulu, Hawaii, U.S.A.

## Abstract

The usual algorithms in pattern recognition assume that an n-dimensional domain in the n-dimensional representation space corresponds to each of the classes. The subspace model assumes that an m-dimensional subspace (m<n) passing the origin corresponds to a class. The shortcomings of the original version of the subspace model, called CLAFIC, were eliminated in the new version, called MOSS, that is explained in the present paper. The MOSS version assigns mutually orthogonal subspaces to classes, where the number of classes can be equal to or larger than 2. It also provides a special "reject" subspace. This revised version was made possible by new theoretical and computational improvements. A preliminary experiment promises a large and fruitful field of applications for this model in the future.

## 1. Introduction

According to the vectorial method of pattern recognition, it is customary to assign an n-dimensional region or "zone" in the n-dimensional representation space to each of the classes into which objects are to be classified. Such a zone may be decided either directly by a decision-function derived from the paradigms (class-samples) or indirectly by the condition that the probability of a point belonging to a certain class is larger than the probability of its belonging to any other class.

In contradistinction to the zone model, another model was introduced by one of the present authors (S.W.),[1,2] according to which a subspace of dimensions less than n (usually passing the origin) is assigned to a class. This is a natural outgrowth of the SELFIC method[3] which may be interpreted as an adaptation of the principal component method of statistics to the problem of dimensionality reduction in the vectorial representation. In fact, when we have a collection of vectors belonging to different classes, an application of the SELFIC method reveals that except for small errors, most of the vectors are located within a subspace of dimensionality considerably smaller than n. If this is the case, it is obviously tempting to try to apply the same method to the vectors of each class separately. It is to be readily expected that vectors of different classes will define different subspaces, thus providing the basis for a "subspace model" of classes. The decision of class-affiliation of a new vector may be decided for instance by comparison of the magnitudes of the projections of the vector onto the different class-subspaces. Thus, the method of CLAFIC was born.[1]

The CLAFIC method was programmed and successfully applied to some problems,[1,4] but the method involved two difficulties. (1) In some cases, the intersection of two subspaces receives a large number of dimensions. (2) In some cases, the "angle" between the subspaces becomes so small that the discrimination becomes very difficult. What we present in this paper is, so to speak, a new version of CLAFIC, avoiding the difficulty (1) by properly introducing a "reject" subspace, and the difficulty (2) by orthogonalizing the class-subspaces by the use of a non-orthogonal linear transformation. This new method will be hereafter referred to as MOSS, standing for multiclass orthogonal subspace method. The modifier "multiclass" here is important because in the meantime Fukunaga and Koontz[5] introduced a kind of special coordinate system that allows to represent two classes as subspaces orthogonal to each other. Their method, however, does not allow to accommodate more than two classes. We have published two interim and partial reports[6,7] on the MOSS method in the past, but the present report is more complete and free from certain minor errors involved in the earlier reports.

The CLAFIC method involved a peculiar difficulty of logical nature.[8] If we interpret the implication A→B (predicate A implies predicate B) meaning that the subspace corresponding to A is a subsubspace of the subspace corresponding to B, then we obtain a modular lattice of predicates, which is not necessarily Boolean, i.e., we cannot indiscriminately use the distributive law of logic.[9] This difficulty is not fatal, because the vectors belonging to a class are found more or less in a subspace, but not all the vectors of the subspace correspond to objects of the class. In the revised version MOSS, however, the class-subspaces are all orthogonal, as a result of which the distributive law is reinstated and the usual logic can be used without restriction. In this respect, too, the MOSS is superior to the CLAFIC.

Before going to explain the MOSS method, we shall insert a section introducing mathematical tools that will be used later. To save space, some of the longer proofs will be omitted.

## 2. Mathematical Tools

The projection operator $P$ in the n-dimensional real space is an $n \times n$ real, symmetric, idempotent matrix, i.e.,

$$P^T = P \tag{2.1}$$

and

$$P^2 = P \tag{2.2}$$

These relations are obviously invariant for an orthogonal transformation. As a result, (2.2) means that the eigenvalues of $P$ must be either 0 or 1. The eigenvectors corresponding to eigenvalue 1 form a subspace (passing the origin) of $m$ dimensions, where

$$m = \text{trace } P \tag{2.3}$$

We can take arbitrarily a set of $m$ mutually orthogonal (normalized) vectors in this m-dimensional subspace, which we denote $\xi^{(k)}$, $k=1,2,\ldots,m$.

$$\xi^{(k)T}\,\xi^{(\ell)} = \delta_{k\ell} \tag{2.4}$$

The matrix $P$ can be written as

$$P = \sum_{k=1}^{m} \xi^{(k)}\,\xi^{(k)T} \tag{2.5}$$

* This work was partly supported by Air Force Grant No. AFOSR-72-2259.

The matrix P defines the subspace subtended by $\xi^{(k)}$, k=1,2,...,m, and conversely a subspace defines a projection operator P. Subspace and projection operator correspond one-to-one; consequently we may use synonymously the term subspace and the projection operator corresponding to it.

The eigenvectors corresponding to eigenvalue 0 form a subspace of (n-m) dimensions. If we take arbitrarily (n-m) mutually orthogonal, normalized vectors $\xi^{(k)}$, k=m+1,m+2,...,n, we can form a set of n base vectors $\xi^{(k)}$, k=1,2,...,n. If we express the matrix P using its own eigenvectors, $\xi^{(k)}$, k=1,2,...,n, as the base vectors, the matrix will have first m diagonal elements equal to 1 and all other elements equal to 0. If we apply an orthogonal transformation that leaves invariant the subspace subtended by the first m eigenvectors (and hence the subspace subtended by the last (n-m) eigenvectors, too), this particular form of the matrix will remain unchanged. If we express the matrix P using a set of base vectors other than the $\xi$'s, the matrix will not have this particular form.

If we apply the matrix P from the left on an arbitrary vector x, the result x' will be the projection of x on the subspace defined by P. We can see this by observing

$$x' = Px = \sum_{k=1}^{m} (\xi^{(k)T} x)\, \xi^{(k)} \qquad (2.6)$$

The matrix 0 is obviously a projection operator satisfying (2.1) and (2.2) and corresponds to the subspace of 0 dimension, and the projection of any vector on this subspace is zero. The matrix 1 (identity matrix) is also a projection operator and corresponds to the entire n-dimensional space, and the projection of any vector on this subspace is the original vector itself. The matrix (1-P) obeys also (2.1) and (2.2), meaning that it is a projection operator. It defines the subspace subtended by the eigenvectors of P corresponding to eigenvalue 0. The obvious relation

$$x = Px + (1-P)x \qquad (2.7)$$

means that x is the vectorial sum of the projection of x on the subspace of P and the projection of x on its complementary subspace.

Two subspaces are said to be orthogonal, if any vector of one subspace is orthogonal to all vectors of the other subspace. It is easy to see that this condition is equivalent to the condition that the projection operators $P_1$ and $P_2$ corresponding to them satisfy

$$P_1P_2 = P_2P_1 = 0 \qquad (2.8)$$

which we write also

$$P_1 \perp P_2. \qquad (2.9)$$

Note that $P_1P_2 = 0$ implies that $P_2P_1 = 0$, since $P_2P_1 = P_2^TP_1^T = (P_1P_2)^T = 0^T = 0$. The orthogonality relation is not transitive. A special case of orthogonality is $P \perp (1-P)$:

$$P(1-P) = P-P^2 = P-P = 0 \qquad (2.10)$$

Subspace $P_1$ is said to be "included in" subspace $P_2$ if any vector of $P_1$ is a vector of $P_2$. It is easy to see that this condition is equivalent to

$$P_1P_2 = P_2P_1 = P_1 \qquad (2.11)$$

which we also write

$$P_1 \rightarrow P_2. \qquad (2.12)$$

Note that $P_1P_2 = P_1$ implies $P_2P_1 = P_1$ and vice versa, since $P_2P_1 = P_2^TP_1^T = (P_1P_2)^T = P_1^T = P_1$. The inclusion relation is transitive. If $P_1$ is included in $P_2$, i.e., if (2.11) is the case, the difference $P_2-P_1$ is a projection operator satisfying (2.1) and (2.2). The fact that (1-P) is a projection operator is a special case of this theorem. It should be noted that $0 \rightarrow P \rightarrow 1$ for any P.

Subspaces $P_1$ and $P_2$ are said to be "compatible" if there exists a set of base vectors $\xi^{(k)}$, k=1,2,...,n such that subspace $P_1$ as well as subspace $P_2$ is subtended by some of the $\xi$'s. This condition can be expressed as commutativity of matrices $P_1$ and $P_2$

$$P_1P_2 = P_2P_1 \qquad (2.13)$$

which we also write

$$P_1 \sim P_2 \qquad (2.14)$$

The compatibility relation is not transitive. If (2.12), then (2.14). If $P_1$ and $P_2$ are compatible, the product

$$Q = P_1P_2 = P_1P_2 \qquad (2.15)$$

is easily seen to be a projection operator and

$$Q \rightarrow P_1 \quad \text{and} \quad Q \rightarrow P_2 \qquad (2.16)$$

The three subspaces $P_1$-Q, $P_2$-Q, Q are mutually orthogonal

$$\left.\begin{array}{l} (P_1-Q) \perp (P_2-Q) \\ (P_2-Q) \perp Q \\ Q \perp (P_1-Q) \end{array}\right\} \qquad (2.17)$$

Hence, we can select a single set of orthogonal base vectors such that some of them belong to subspace $P_1$-Q, some others belong to subspace $P_2$-Q, and some others belong to subspace Q. Subspace $P_1$ is subtended by the base vectors of the first group and the base vectors of the third group. Subspace $P_2$ is subtended by the base vectors of the second group and the base vectors of the third group. The remaining base vectors subtend the complementary subspace $1 - (P_1 - Q) - (P_2 - Q) - Q = 1 + Q - P_1 - P_2$. It is easy to show that this expression is a projection operator. Q is the intersection of two compatible subspaces $P_1$ and $P_2$.

The projection operators, being matrices, can be added, subtracted and multiplied, but the result of such an arithmetic operation is not necessarily a projection operator. We have, however, the following rules: (1) If $P_1 \sim P_2$, then $P_1P_2$ ($= P_2P_1$) is a projection operator; (2) If $P_1 \rightarrow P_2$ then $P_2 - P_1$ is a projection operator; (3) If $P_1 \perp P_2$, then $P_1 + P_2$ is a projection operator. The first two have already been mentioned. The third one is easy to prove.

Now we can introduce "conjunction" $P_1 \cap P_2$ of two projection operators $P_1$ and $P_2$ with the help of the concept of inclusion ($\rightarrow$). A projection $P_3$ is said to

be conjunction $P_1 \cap P_2$ if $P_3$ satisfies the conditions: (i) $P_3 \to P_1$ and $P_3 \to P_2$, and (ii) if $x \to P_1$ and $x \to P_2$, then $x \to P_3$. With some mathematical manipulation, we can prove[10] that the infinite product of $P_1$ and $P_2$ can be considered to be $P_3$:

$$P_3 = P_1 \cap P_2 = \ldots P_1P_2P_1P_2P_1P_2\cdots \qquad (2.18)$$

It is readily seen the $P_3$ corresponds to subspace consisting of all vectors that belong to both $P_1$ and $P_2$. In other words, $P_3$ is the common subspace, i.e., the intersection, of subspace $P_1$ and subspace $P_2$. It may be noted that if $P_1 \sim P_2$ then $P_1 \cap P_2 = P_1P_2 = P_2P_1$. We defined conjunction with the help of inclusion, but we can conversely define inclusion by conjunction. $P_1 \to P_2$ is equivalent to $P_1 = P_1 \cap P_2$. Special cases are $0 \cap P = 0$ and $1 \cap P = P$ for all P.

The derivation of $(1 - P_1)$ from $P_1$ is called "complementation" and we write

$$\neg P = (1 - P). \qquad (2.19)$$

From this definition follow the three basic laws of complementation:

$$\left.\begin{array}{l} \text{i) } \neg\neg P = P \\ \text{ii) } P_1 \to P_2 \text{ is equivalent to } \neg P_2 \to \neg P_1 \\ \text{iii) If } P \to \neg P \text{ then } P = 0. \end{array}\right\} \qquad (2.20)$$

From our definition, we have $0 = \neg 1$ and $1 = \neg 0$.

The "disjunction" $P_1 \cup P_2$ can now be defined by

$$P_1 \cup P_2 = \neg(\neg P_1 \cap \neg P_2) \qquad (2.21)$$

$$= 1 - [\ldots(1-P_1)(1-P_2)(1-P_1)(1-P_2)\ldots] \qquad (2.22)$$

The disjunction $P_1 \cup P_2$ thus defined is a subspace formed by all the linear combinations of a vector in $P_1$ and a vector in $P_2$. Thus $P_1 \cup P_2$ includes also vectors which are not found in either $P_1$ or $P_2$. Actually, the disjunction can be defined first by conditions similar to the ones used in defining the conjunction, just inverting the direction of arrows, and then prove with the help of (2.20) that the conjunction and disjunction are related by de Morgan's rule (2.21). Then, it is as well justified to define the disjunction directly by de Morgan's rule like (2.21). $P_1 \to P_2$ is equivalent to $P_2 = P_1 \cup P_2$. Special cases are $0 \cup P = P$ and $1 \cup P = 1$ for all P.

The conjunction and disjunction as thus defined obey the following five basic laws:

i) Idempotent law: $P_1 \cap P_1 = P_1$, $P_1 \cup P_1 = P_1$ (2.23)

ii) Commutative law:

$$P_1 \cap P_2 = P_2 \cap P_1, \quad P_1 \cup P_2 = P_2 \cup P_1 \qquad (2.24)$$

iii) Associative law:

$$(P_1 \cap P_2) \cap P_3 = P_1 \cap (P_2 \cap P_3)$$
$$(P_1 \cup P_2) \cup P_3 = P_1 \cup (P_2 \cup P_3) \qquad (2.25)$$

iv) Absorptive law:

$$P_1 \cap (P_1 \cup P_2) = P_1$$
$$P_1 \cup (P_1 \cap P_2) = P_1 \qquad (2.26)$$

v) de Morgan's law:

$$\neg(P_1 \cap P_2) = \neg P_1 \cup \neg P_2$$
$$\neg(P_1 \cup P_2) = \neg P_1 \cap \neg P_2 \qquad (2.27)$$

This shows that the projection operators, including 0 and 1, form a complemented lattice.

It is very important to note that the Distributive laws:

$$(P_1 \cap P_2) \cup P_3 = (P_1 \cup P_3) \cap (P_2 \cup P_3) \qquad (2.28)$$

$$(P_1 \cup P_2) \cap P_3 = (P_1 \cap P_3) \cup (P_2 \cap P_3) \qquad (2.29)$$

do not in general hold. For instance, suppose that $P_1$ and $P_2$ are respectively x-direction and y-direction (each one is a one-dimensional subspace) and $P_3$ is another direction in the x-y-plane. Then $P_1 \cap P_2 = 0$, hence the left hand side of (2.28) is $0 \cup P_3 = P_3$. But, on the right hand side both $(P_1 \cup P_3)$ and $(P_2 \cup P_3)$ are the x-y-plane, as a result the right hand side becomes also the x-y-plane. Hence the distributive law breaks down. The lattice is non-distributive.

We can, however, prove the so-called modular law which states that if

$$P_3 \to P_1 \qquad (2.30)$$

then (2.28) holds. Since (2.30) is equivalent to $P_1 \cup P_3 = P_1$, we can state the theorem as follows: If (2.30) is true then

$$(P_1 \cap P_2) \cup P_3 = P_1 \cap (P_2 \cup P_3) \qquad (2.31)$$

This theorem is also equivalent to the statement that if $P_1 \to P_3$, then (2.24) holds. See Ref. 11 for a proof.

It is important to note that if all the P's at hand are compatible with one another, the distributive law holds for all of them. This is intuitively obvious since in this case there exists a single set of base vectors $\xi^{(k)}$, k=1,2,...,n, such that any P can be determined by a subset of the $\xi$'s. Thus, $P_1 \cap P_2$ is a P which is determined by those $\xi$'s that are commonly included in $P_1$ and $P_2$, and $P_1 \cup P_2$ is a P that is determined by those $\xi$'s that are included in either or both of $P_1$ and $P_2$. Hence, the conjunction and disjunction acquire the usual set-theoretical meaning, and therefore the distributive law has to hold. The lattice becomes distributive or Boolean.

If we assign a proposition to each P, then the lattice can be regarded as the logic of propositions. "$P_1$ is included in $P_2$" is interpreted as "$P_1$ implies $P_2$." The conjunction and disjunction are interpreted as "and" and "or." The complementation is interpreted as negation. The modular lattice is isomorphic with the so-called quantum logic (for the case of finite dimensionality), and the distributive lattice is isomorphic with the usual Boolean logic.

## 3. Class-Subspaces and Their Orthogonalization

Let $x_\alpha^{(k)}$ be an n-dimensional vector representing the $\alpha$-th paradigm (class-sample) of the k-th class, $\alpha=1,2,\ldots,\nu^{(k)}$; k=1,2,...,m. There are m classes,

and there are $\nu^{(k)}$ paradigms (class-samples) in class k. The autocorrelation matrix $G^{(k)}$ for class k is defined by

$$G^{(k)} = \sum_{\alpha=1}^{\nu^{(k)}} x_\alpha^{(k)} x_\alpha^{(k)T} \Big/ \sum_{\alpha=1}^{\nu^{(k)}} x_\alpha^{(k)T} x_\alpha^{(k)} \tag{3.1}$$

which is normalized so that

$$\text{trace } G^{(k)} = 1 \tag{3.2}$$

The orthonormal base vectors peculiar to the k-th class, $\{\psi_j^{(k)}\}$, (j=1,2,...,n), are defined as the eigenvectors of $G^{(k)}$

$$G^{(k)}\psi_j^{(k)} = \lambda_j^{(k)}\psi_j^{(k)} \qquad (j=1,2,\ldots,n) \tag{3.3}$$

where the $\lambda$'s are non-negative, and, due to (3.2), sums up to unity:

$$\sum_{j=1}^{n} \lambda_j^{(k)} = 1. \tag{3.4}$$

We choose the index j so that

$$\lambda_1^{(k)} \geq \lambda_2^{(k)} \geq \cdots \geq \lambda_n^{(k)} \tag{3.5}$$

In most of the examples we encounter in pattern recognition, the eigenvalues $\lambda_j^{(k)}$ decrease very fast with j, which means that the paradigm vectors of class k on the average have very little weight (square of components) along the directions $\psi_j^{(k)}$ for large j. This gives a picture that the vectors of class k are located in a subspace subtended by the $\psi_j^{(k)}$'s with small j's only. If we determine the dimensionality $n^{(k)}$ of the subspace by the condition

$$\sum_{j=1}^{n^{(k)}-1} \lambda_j^{(k)} < \sigma^{(k)} \leq \sum_{j=1}^{n^{(k)}} \lambda_j^{(k)}, \tag{3.6}$$

the subspace so defined accounts on the average for about 100 $\sigma^{(k)}$ % of the weight of the paradigms of class k. We are usually surprised to see that for a relatively large value of $\sigma^{(k)}$ (say, 0.95) the dimension $n^{(k)}$ of the subspace becomes very small compared with n. This fact supports the subspace model of pattern recognition. In practice, we take the same value $\sigma^{(k)}$ for different classes in order to treat different classes on an equal footing. An appropriate value of $\sigma$ in the case of SELFIC (in which we make only one subspace regardless of class affiliation) can be determined more or less by hunch and experience. In our present case, however, we may determine $\sigma$ so that the class discrimination becomes optimal.

The subspace for class k can now be represented by the projection operator

$$P^{(k)} = \sum_{j=1}^{n^{(k)}} \psi_j^{(k)} \psi_j^{(k)T} \tag{3.7}$$

Since the $\psi_j^{(k)}$'s are normalized

$$\text{trace } P^{(k)} = n^{(k)} \tag{3.8}$$

The subspace $P^{(k)}$ consists of those feature variables that have high weights in the collection of paradigms of class k. But, this does not guarantee that the subspaces for two different classes do not contain the same variables. Such common variables have no discriminating power and should be taken out of consideration. We therefore define the "retrenched" subspace $P^{(k)'}$ by subtracting the overlapped subspaces from each subspace $P^{(k)}$.

$$P^{(k)'} = P^{(k)} - \bigcup_{\ell\neq k} (P^{(k)} \cap P^{(\ell)}) \tag{3.9}$$

where $P^{(k)} \cap P^{(\ell)}$ is the "overlap" of $P^{(k)}$ and $P^{(\ell)}$. Since each term $(P^{(k)} \cap P^{(\ell)})$ is a subspace of $P^{(k)}$, the merger of them under $\bigcup_{\ell\neq k}$ is also a subspace of $P^{(k)}$. Hence the difference $P^{(k)'}$ is a projection operator, having the form of $P_1 - P_2$ with $P_2 \to P_1$. Since $P_2 \to P_1$ implies that $P_1$ and $P_2$ are compatible, we can also write $P_2 = P_2 \cap P_1 = P_1P_2$. Hence $P_1 - P_2 = P_1 - P_1P_2 = P_1(1 - P_2) = P_1 \cdot \neg P_2 = P_1 \cap \neg P_2$. Applying this transformation, we can rewrite (3.9) as

$$\begin{aligned} P^{(k)'} &= P^{(k)} \cap \neg \bigcup_{\ell\neq k} (P^{(k)} \cap P^{(\ell)}) \\ &= P^{(k)} \cap \bigcap_{\ell\neq k} (\neg P^{(k)} \cup \neg P^{(\ell)}) \end{aligned} \tag{3.10}$$

The dimension $n^{(k)'}$ of the retrenched subspace is in general smaller than the original subspace.

$$n^{(k)'} = \text{trace } P^{(k)'} \leq n^{(k)} \tag{3.11}$$

These retrenched subspaces contain no common vectors with one another. We can see this point by noticing

$$\begin{aligned} P^{(k)'} \cap P^{(\ell)'} = P^{(k)} \cap P^{(\ell)} \cap \bigcap_{p\neq k} (\neg P^{(k)} \cup \neg P^{(p)}) \\ \cap \bigcap_{q\neq \ell} (\neg P^{(\ell)} \cup \neg P^{(q)}) \end{aligned} \tag{3.12}$$

where there must be a term $(\neg P^{(k)} \cup \neg P^{(\ell)})$ under the conjunctions $\bigcap_{p\neq k}$ and $\bigcap_{q\neq \ell}$ which will cancel the first term $P^{(k)} \cap P^{(\ell)}$, since the conjunction of a projection operator and its complement becomes 0. Hence

$$P^{(k)'} \cap P^{(\ell)'} = \delta_{k\ell} P^{(k)'} \tag{3.13}$$

showing that the $P^{(k)'}$ are projection operators and contain no common vectors if $k\neq\ell$.

The total class-subspace occupied by vectors of all the retrenched class-subspaces is given by $\bigcup_{k=1}^{m} P^{(k)'}$, and the dimension of this total class-subspace is at most n but in general less than n. The subspace orthogonal to the total class-subspace

$$P^{(m+1)'} = \neg \bigcup_{k=1}^{m} P^{(k)'} = 1 - \bigcup_{k=1}^{m} P^{(k)'} \tag{3.14}$$

can be considered as consisting of vectors which are either not included in the original class-subspaces or included in more than one original subspace. Hence,

we consider this as a "reject" subspace and name it the $(m+1)$-st class.

In each subspace $P^{(k)'}$, we can take $n^{(k)'}$ linearly independent vectors. There are in total $\sum_{k=1}^{m} n^{(k)'}$ such vectors, and they are not necessarily mutually linearly independent, even if $\sum_{k=1}^{m} n^{(k)'} \leq n$. However, in actual computation, there seldom happens linear dependence among these vectors. Even if there is linear dependence, a slight modification will make them all linearly independent. If that is the case, the dimension of the total class-subspace $\bigcup_{k=1}^{m} P^{(k)'}$ is the same as the sum of the dimensions of individual class-subspaces. In such a case, the dimension of the reject class will be $n - \sum_{k=1}^{m} n^{(k)'}$. In usual examples, $\sum_{k=1}^{m} n^{(k)'}$ is not larger than $n$, provided $m$ is not excessively large or $\sigma$ is not taken unreasonably high.

The next step is to introduce a linear transformation that will turn the class-subspaces until they become orthogonal to one another. To do this, we take $n^{(k)'}$ orthogonal vectors in $P^{(k)'}$, i.e., we take $n^{(k)'}$ orthogonal degenerate eigenvectors of $P^{(k)'}$ corresponding to eigenvalue unity. If we take all such vectors for all $k$ and include the reject class $P^{(m+1)'}$, we shall get exactly $n$ vectors, which we denote by $\eta^{(a)}$, $a=1,2,\ldots,n$, and the components of $\eta^{(a)}$ in the original coordinate systems by $\eta_\rho^{(a)}$, $\rho=1,2,\ldots,n$. The vectors $\eta^{(a)}$ are not uniquely determined (due to the degeneracy of eigenvalue L), but this ambiguity does not affect the effect of the linear transformation we are going to introduce. The linear transformation L we are now considering is such that the transforms $\eta^{(a)*}$ of $\eta^{(a)}$ by L become all orthogonal, i.e.,

$$\eta^{(a)*T}\eta^{(b)*} = \sum_{\sigma=1}^{n} \eta_\sigma^{(a)*}\eta_\sigma^{(b)} = \delta_{ab} \quad (a,b=1,2,\ldots,n) \tag{3.15}$$

where

$$\eta_\sigma^{(a)*} = \sum_{\rho=1}^{n} L_{\sigma\rho}\,\eta_\rho^{(a)} \tag{3.16}$$

Such an $L$ can be given by the condition

$$\sum_{\rho=1}^{n} L_{\sigma\rho}\,\eta_\rho^{(a)} = \delta_{\sigma a} \qquad \sigma,a=1,2,\ldots,n \tag{3.17}$$

By this transformation (3.17), the vectors $\eta^{(a)}$ belonging to the same subspace $P^{(k)'}$ will remain orthogonal and its magnitude will remain unchanged provided the $\eta^{(a)}$ are normalized originally. As a result, the subspace corresponding to a class remains internally unchanged. Two vectors $\eta^{(a)}$ belonging to two different subspaces $P^{(k)'}$ are made orthogonal to each other by the transformation (3.17).

In general a non-orthogonal linear transformation L changes a projection operator $P$ into $LPL^T$ (not $LPL^{-1}$) which are no longer necessarily a projection operator. But, by the special linear transformation L, the $P^{(k)'}$ are transformed into $P^{(k)*}$ which are projection operators and mutually orthogonal. The components of $P^{(k)*}$ are given by

$$P_{\rho\sigma}^{(k)*} = (L\,P^{(k)'}\,L^T)_{\rho\sigma} = \sum_{a\in k} \eta_\rho^{(a)*}\eta_\sigma^{(a)*} \tag{3.18}$$

with the summation with respect to $a$ extends over those vectors included in the subspace $P^{(k)'}$. The product of two such operators becomes

$$(P^{(k)*}\,P^{(\ell)*})_{\rho\sigma} = \sum_{\mu=1}^{n} \sum_{a\in k} \sum_{b\in \ell} \eta_\rho^{(a)*}\eta_\mu^{(a)*}\eta_\mu^{(b)*}\eta_\sigma^{b}$$

$$= \sum_{a\in k} \eta_\rho^{(a)*}\eta_\sigma^{(a)*}\delta_{k\ell} = P^{*(k)}\delta_{k\ell} \tag{3.19}$$

where the condition (3.15) has been used. Eq. (3.19) shows at once that the $P^{(k)*}$ are not only projection operators (idempotency) but are also mutually orthogonal.

From this point on every vector, paradigm or new sample, is transformed by the L-transformation and then we proceed to the next stage of consideration, namely, the problem of discrimination. There remains a certain ambiguity whether the G-matrices should be transformed according to the same formal rule as (3.18) or should be redefined by the same formula as (3.1), only using the vectors with asterisk in both numerator and denominator. In the former case the condition (3.2) will no longer be vigorously true, but the difference should be very little anyway because the paradigm vectors lie more or less entirely in the $P^{(k)}$ space, and the L-transformation will not alter the internal structure of each subspace.

## 4. Decision Procedure

In this section we do not use the asterisk on quantities any longer, but all the quantities, including the new samples, are supposed to have been subjected to the L-transformation.

The quantity in which we are first interested is the probability of an average paradigm vector of class $k$ appearing in subspace $P^{(\ell)}$. This will be obviously given by

$$p(\ell|k) = \text{trace}\,(G^{(k)} \cdot P^{(\ell)}) \,/\, \text{trace}\,G^{(k)} \tag{4.1}$$

which is normalized with respect to subspaces $P^{(\ell)}$, where $\ell$ runs from 1 to m+1:

$$\sum_{\ell=1}^{m+1} p(\ell|k) = 1 \tag{4.2}$$

Obviously $p(k|k)$ will be the largest among $p(\ell|k)$.

The inverse conditional probability that an arbitrary vector in the $P^{(\ell)}$ belongs to class $k$ will be given, with the help of the prior probability $p(k)$ of class $k$, by

$$q(k|\ell) = p(\ell|k)\,p(k) \,/\, \sum_{k=1}^{m} p(\ell|k)\,p(k) \tag{4.3}$$

which is normalized with respect to $m$ classes:

$$\sum_{k=1}^{m} q(k|\ell) = 1 \tag{4.4}$$

Now, when a new object with vector $y$ (of course, after the L-transformation) arrives, we have to find out in which subspace it is located. The probability of vector $y$ being in subspace $P^{(\ell)}$, or the fraction of its weight in subspace $P^{(\ell)}$ is given by

$$w(\ell|y) = y^T P^{(\ell)} y \;/\; y^T y \tag{4.5}$$

with

$$\sum_{y=1}^{m+1} w(\ell|y) = 1 \tag{4.6}$$

Since if $y$ is located in subspace $P^{(\ell)}$, it will belong to class $k$ with probability $q(k|\ell)$, we can conclude that the probability of $y$ belonging to class $k$ is

$$q(k|y) = \sum_{\ell=1}^{m+1} q(k|\ell)\, w(\ell|y) \tag{4.7}$$

with

$$\sum_{k=1}^{m} q(k|y) = 1 \tag{4.8}$$

This quantity $q(k|y)$ thus defined is proportional to $p(k)$, $p(\ell|k)$ and $w(\ell|y)$, which seems to be quite reasonable.

Our decision rule will be to assign class $k_o$ which maximizes $q(k|y)$, i.e.,

$$q(k_o|y) = \operatorname{Max}_k q(k|y) \tag{4.9}$$

The reader may be disturbed by the fact that in this formalism there is no probability $q(m+1|y)$ of $y$ belonging to the reject class. But, this is because we have introduced the reject subspace but not the reject class. The way to utilize the reject subspace for a useful purpose is to compare the probability $w(m+1|y)$ of $y$ being in the reject subspace with the probability $p(m+1|k_o)$ of an average vector of the assigned class being in the reject subspace. If the former is considerably larger than the latter, we have to conclude that the assignment is not well-founded. We can make a criterion of the type

$$w(m+1|y) \leq \theta = \gamma\, p(m+1|k_o) \tag{4.10}$$

where $\theta$ is the threshold for admittance in class $k_o$ and the parameter $\gamma$ is a real number larger than unity. If we decrease the threshold $\theta$, the number of the paradigms of class $k$ being rejected will increase and at the same time the number of the misclassification of paradigms will decrease. We can adjust $\theta$ so that the (weighted) sum of these two undesirable numbers will become minimum.

An interesting possibility is that we can determine the yet-undetermined parameter, fidelity, $\sigma$ by requiring the maximum discrimination. The discrimination can be measured by the unevenness of $q(k|y)$, since if $q(k|y)$ is distributed more or less evenly over different $k$'s, the decision made by (4.9) will be unreliable. A good measure for this would be the entropy function.

$$S(y) = \sum_{k=1}^{m} q(k|y) \log q(k|y) \tag{4.11}$$

In practice, the best would be to determine the value of $\sigma$ by the requirement that the average entropy $S(y)$ for the given paradigms will become minimum, i.e., the most desirable value $\sigma_o$ of $\sigma$ can be given by

$$S(\sigma_o) = \operatorname{Min}_\sigma S(\sigma) \tag{4.12}$$

where

$$S(\sigma) = \sum_{k=1}^{m} \sum_{a=1}^{\nu(k)} S(x_a^{(k)}) \tag{4.13}$$

Although the expression (4.11) does not show the dependence of $S(y)$ on $\sigma$, but it is obvious that the value of $S(y)$ depends on $\sigma$.

## 5. Preliminary Experimental Test

A computational difficulty arises in the present formalism if utmost care is not taken in calculating the infinite product, $P_3 = P_1 \cap P_2 = \ldots P_1P_2P_1P_2\ldots$, of two projection operators, $P_1$ and $P_2$. First, we make the product $Q = P_1P_2$ which is not a projection operator except in the compatible case. We calculate successively $Q^2$, $Q^4$, $Q^8$,...,$Q^{2^n}$. At the limit $n\to\infty$, $Q^{2^n}$ should become a projection operator $P_3$, but we have to be satisfied with a large value of $n$, at which the two conditions of a projection operator, namely the symmetry and the idempotency, are approximately satisfied. The approximate idempotency, moreover, will guarantee that we have reached almost the limiting value, since it would mean that $Q^{2^n}$ and $Q^{2^{n+1}}$ are approximately the same. Before we began to use the double precision computation, the accumulated errors in multiplication prevented the product to become either symmetric or idempotent.** We have now a program that works beautifully.

An alternative method of calculating $P_3$ would be to repeat the process: $X \to YXY$ and $Y \to XYX$, with starting values: $X = P_1$, $Y = P_2$. Both $X$ and $Y$ will approach $P_3$, but in this case the symmetry, $X^T = X$ and $Y^T = Y$, is automatically satisfied at each stage. To guarantee this, we may add the corrective method: $X \to (X + X^T)/2$, $Y \to (Y + Y^T)/2$ at each stage.

We tested the consistency of our method by analyzing handwritten letters, A, B, and C. They are originally 20 x 20 black-and-white meshes, but we applied the method of crossing number compression[12] to this 400-digit binary information and obtained a 28-digit number, each digit can be occupied by 0, 1, 2 or 3. Thus, our representation space has $n = 28$ dimensions. Each of A, B, and C had 50 paradigms. The actual characters are deliberately deformed and some of them were difficult to decipher even for a human recognizer.

For the fidelity $\sigma$, we took three different values, 0.93, 0.95, and 0.97. The dimensionality of each subspace depends on the $\sigma$-value and the actual values are given in Table 1. The superscript $k = 1,2,3,4$ designate respectively subspace for A, that for B, that for C and the reject subspace. We see, here again, that the dimensionality of each class-subspace is quite small. It is quite clear that as $\sigma$ increases, the dimensionality of the reject subspace decreases.

** The authors would like to thank Dr. Seiji Inatsugu for helpful discussion with regard to the computational difficulties.

| σ \ k | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| .93 | 2 | 4 | 4 | 18 |
| .95 | 3 | 5 | 4 | 16 |
| .97 | 3 | 7 | 6 | 12 |

Table 1

The dimensionality $n^{(k)*}$ of subspace $P^{(k)*}$, where k=1,2,3,4 mean respectively, A, B, C, reject, as dependent on the fidelity σ.

On Tables 2, 3, 4, we give the conditional probability $p(\ell|k)$ that the average vector of class k is found in subspace ℓ, $P^{(\ell)}$. This probability for the reject subspace decreases as σ increases. This probability corresponds to $p(m+1|k)$ of Section 4 and plays an important role in (4.10). Tables 2, 3, 4 correspond respectively to σ = 0.93, 0.95, 0.97.

| k \ ℓ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .8703 | .05331 | .05858 | .01777 |
| 2 | .09889 | .7812 | .09825 | .02160 |
| 3 | .03232 | .08259 | .8621 | .02294 |

Table 2

The values of $p(\ell|k)$ for σ = 0.93

| k \ ℓ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .9099 | .02958 | .04889 | .01155 |
| 2 | .1519 | .6810 | .1549 | .01215 |
| 3 | .03045 | .1391 | .8169 | .0136 |

Table 3

The values of $p(\ell|k)$ for σ = 0.95

| k \ ℓ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .6237 | .1538 | .2174 | .004947 |
| 2 | .1141 | .6683 | .2126 | .004928 |
| 3 | .02454 | .20356 | .76814 | .003481 |

Table 4

The values of $p(\ell|k)$ for σ = 0.97

The next step is to calculate the inverse conditional probability $q(k|\ell)$, which is tabulated in Tables 5, 6, 7, respectively for σ = 0.93, 0.95, 0.97. In Tables 2, 3, 4, the figures add up to unity horizontally whereas in Tables 5, 6, 7, the figures add up to unity vertically. The prior probability p(k) was set equal for all k's.

| k \ ℓ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .8689 | .05813 | .05749 | .2852 |
| 2 | .09874 | .8518 | .09642 | .3466 |
| 3 | .03227 | .09005 | .8460 | .3681 |

Table 5

The values of $q(k|\ell)$ for σ = 0.93

| k \ ℓ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .8336 | .03481 | .0479 | .3096 |
| 2 | .1390 | .8014 | .15177 | .3258 |
| 3 | .02788 | .1637 | .8003 | .36454 |

Table 6

The values of $q(k|\ell)$ for σ = 0.95

| k \ ℓ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | .8180 | .1500 | .1814 | .3704 |
| 2 | .1497 | .6515 | .1773 | .3689 |
| 3 | .03220 | .1984 | .6411 | .2606 |

Table 7

The values of $q(k|\ell)$ for σ = 0.97

Finally, we took each paradigm as the y and tested if it would be classified correctly according to our procedure. Tables 8, 9, 10 show the results of our decision procedures for σ = 0.93, 0.95, 0.97 respectively. Since the same vector is used as the "new arrival" as well as one of the 50 paradigms that determine the subspace, this test may be considered as a test of self-consistency only. But, in view of a large variety of input data in each class, the result may be regarded as very satisfactory.

| k | $k_o = 1$ | $k_o = 2$ | $k_o = 3$ | reject |
|---|---|---|---|---|
| 1 | 46 | 0 | 2 | 2 |
| 2 | 1 | 45 | 3 | 1 |
| 3 | 0 | 0 | 47 | 3 |

Table 8

The result of our decision procedure for $\sigma = 0.93$. k is the true class, $k_o$ is the assigned class according to (4.9). Rejection is done according to (4.10).

| k | $k_o = 1$ | $k_o = 2$ | $k_o = 3$ | reject |
|---|---|---|---|---|
| 1 | 49 | 0 | 0 | 1 |
| 2 | 2 | 45 | 3 | 0 |
| 3 | 0 | 1 | 49 | 0 |

Table 9

The result of our decision procedure for $\sigma = 0.95$.

| k | $k_o = 1$ | $k_o = 2$ | $k_o = 3$ | reject |
|---|---|---|---|---|
| 1 | 45 | 1 | 2 | 2 |
| 2 | 4 | 39 | 4 | 3 |
| 3 | 0 | 0 | 49 | 1 |

Table 10

The result of our decision procedure for $\sigma = 0.97$.

We also tried 50 sample vectors which did not belong to either A or B or C, i.e., they are just produced by random numbers. With $\sigma = 0.93$, it classified 92% of them as reject, using the same rule of rejection as in Tables 8, 9, 10.

## 6. Conclusion

It is not claimed that the subspace model is superior to the zone model for all pattern recognition problems. But, it seems justified to believe that for some types of problems the subspace model in the MOSS version as explained in this paper is just as good or even better than the zone model. The only theoretical objection that could be raised against the subspace model in the CLAFIC version was the apparent breakdown of the Boolean logic. The identification of logical implication with inclusion of subspaces is quite natural and unavoidable. Then, the "and" and "or" will follow the definition given in Section 2, and we encounter a strange fact that the distributive law of logic no longer holds in general. This objection should not be taken too seriously, because the vectors of a class are located approximately in a subspace, but not all the vectors of the subspace may correspond to real objects of the class. However, in the case of MOSS version of the subspace method, there cannot be any breakdown of the distributive law. Since the subspaces are all orthogonal, their projection operators are all mutually commutative. Hence, the lattice becomes distributive and the corresponding logic will be the usual one.

## References


1. S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton, and R. Walker, "Evaluation and Selection of Variables in Pattern Recognition," in *Computer and Information Sciences*, vol. 2, (Julius Tou, ed.). New York: Academic Press, 1967, pp. 91-122.

2. S. Watanabe, *Knowing and Guessing*. New York: John Wiley & Sons, 1969.

3. S. Watanabe, "Karhunen-Loève Expansion and Factor Analysis," *Transactions of the Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes*, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1967, pp. 635-660.

4. C. A. Kulikowski, "Pattern Recognition Approach to Medical Diagnosis," in *IEEE Transactions on Systems Science and Cybernetics*, vol. SSC-6, pp. 173-178, July 1970.

5. K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen-Loève Expansion to Feature Selection and Ordering," in *IEEE Transactions on Computers*, vol. C-19, pp. 311-318, April 1970.

6. C. A. Kulikowski and S. Watanabe, "Multiclass Subspace Methods in Pattern Recognition," in *Proceedings of the National Electronics Conference*, vol. XXVI, pp. 468-470, 1970.

7. B. Reiter and S. Watanabe, "Orthogonal Subspaces for Multi-Class Pattern Recognition," in *Proceedings of the Fifth Hawaii International Conference on System Sciences*, (Art Lew, ed.), Western Periodicals Company, pp. 398-400, 1972.

8. S. Watanabe, "Modified Concepts of Logic, Probability and Information Based on Generalized Continuous Characteristic Function," in *Information and Control*, vol. 15, No. 1. New York: Academic Press, 1969, pp. 1-21.

9. S. Watanabe, *Knowing and Guessing*, pp. 449-478.

10. S. Watanabe, *Knowing and Guessing*, p. 495 and p. 563.

11. S. Watanabe, *Knowing and Guessing*, p. 476.

12. S. Watanabe and others, "Evaluation and Selection of Variables in Pattern Recognition," p. 100.