Suppose that in (3) and (4) we are dealing with an infinite set of features with parameters p_i, q_i. The order of the features is chosen randomly. We may consider them, therefore, as independent identically distributed. For (3) can be written

$$\lim_{N\to\infty} \sum_{i=1}^{N} r_i = \infty.$$

If $\bar{r}_N$ is the average of the first N terms, we have

$$\lim_{N\to\infty} N\bar{r}_N = \infty.$$

This is satisfied if the limit of $\bar{r}_n$ for $n \to \infty$, being the expectation of r, is positive:

$$Er = \lim_{N\to\infty} \bar{r}_N > 0.$$

Equations (3) and (4) are therefore satisfied if

$$\int_0^1\int_0^1 (p - q)(2p - 1)F(p,q)\, dpdq > 0 \qquad (7)$$

$$\int_0^1\int_0^1 (q - p)(2q - 1)F(p,q)\, dpdq > 0 \qquad (8)$$

are satisfied. $F(p,q)$ is the density function of the parameters. It can be easily verified that these conditions are satisfied for $F(p,q)$ uniform. This implies that if an infinite number of parameters p_i, q_i are outside the regions defined by (5), (3), and (4) can still be satisfied. The significance of the regions introduced by Chandrasekaran and Jain in satisfying conditions (3) and (4) is therefore not clear to us.

Acknowledgment

The author wishes to express his gratitude to Prof. I. T. Young of M.I.T. for the careful reading and discussing of the manuscript.

References

[1] B. Chandrasekaran and K. Jain, "Independence, measurement complexity and classification performance," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, pp. 240–244, Mar. 1975.
[2] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55–63, Jan. 1968.
[3] K. Abend, T. J. Harley, Jr., B. Chandrasekaran, and G. F. Hughes, "Comments on 'The mean accuracy of statistical pattern recognizers'," *IEEE Trans. Inform. Theory* (corresp.), vol. IT-15, pp. 420–423, May 1969.
[4] B. Chandrasekaran, "Independence of measurements and the mean recognition accuracy," *IEEE Trans. Inform. Theory*, vol. 17, pp. 452–456, July 1971.

Dimensionality and Classification Performance with Independent Coordinates

JOHN W. VAN NESS, MEMBER, IEEE

***Abstract*—In an effort to provide better intuition about the effects of increasing the dimension of the observation vector in discriminant analysis, the case where the coordinates of this vector are independent is studied. Some examples and theorems are given which clarify the issue as to when certain results already appearing in the literature are or are not true. A counterexample is also given showing that the divergence J is not useful in bounding the probability of misclassification for Bayes-type decision rules. Some general results are obtained on the benefit of adding more independent dimensions with fixed training data size.**

Manuscript received October 14, 1976; revised January 13, 1977. This work was supported by NSF Grants MPS74-06945 and MCS76-06519.
The author is with Programs in Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080.

I. Introduction

A very important problem frequently arising in classification is determining how many dimensions to use. There are, of course, stepwise procedures for variable selection and various jackknife or reclassification techniques available for determining which set of variables to use. All of these procedures suffer from one or more serious defects: they are only crude approximations, they consume enormous amounts of computer time, or they are based on assumptions which are too restrictive, etc. Thus much work still needs to be done on the general effects of dimension in classification performance. Some results in this direction can be found in such papers as [1], [2], [6], [9], [10], [12], [14], [16], and [19]. There are also strategies to alleviate the high dimension problem such as assuming all variables are independent or assuming groups of variables are independent (see discussions in Fisher *et al.* [7] and Fisher and Kronmal [8]).

A recent paper by Chandrasekaran and Jain [4] is concerned with the effect of high dimensions when the variables are independent, i.e., when the coordinates of the observation vectors are independent random variables. They discuss asymptotic results as the dimension N tends to infinity while the number of training data vectors remains fixed and finite. This is an interesting problem which provides some additional intuition to the user. Unfortunately, the authors do not provide a precise set of sufficient conditions for their results to be valid and the results are not valid in general. This correspondence clarifies this issue and takes up some additional questions involving independent coordinate variables and large dimensions. A counterexample is given which illustrates the fact that the divergence J is not useful for bounding the probability of misclassification when using Bayes-type decision rules.

II. Notation and Bayes Classification

Suppose for convenience that the population Ω under study consists of two disjoint classes C_1 and C_2, i.e., $\Omega = C_1 \cup C_2$. Furthermore, assume that we can measure N variables on any object $w \in \Omega$ which is selected for measurement. If these variables are random, the measurement vectors $X = (X_1, \cdots, X_n)$ have as their coordinates N random variables. Thus the training data for classification would consist of the observed vectors on, say, m_1 objects from C_1 and m_2 objects from C_2. We label these two training data sets D_1 and D_2:

$$D_1 \equiv \{y_1, \cdots, y_{m_1}\} \equiv \left\{ \begin{pmatrix} y_{11} \\ \vdots \\ y_{1N} \end{pmatrix}, \cdots, \begin{pmatrix} y_{m_1,1} \\ \vdots \\ y_{m_1,N} \end{pmatrix} \right\}$$

$$D_2 \equiv \{z_1, \cdots, z_{m_2}\} \equiv \left\{ \begin{pmatrix} z_{11} \\ \vdots \\ z_{1N} \end{pmatrix}, \cdots, \begin{pmatrix} z_{m_2,1} \\ \vdots \\ z_{m_2,N} \end{pmatrix} \right\}$$

where, e.g., $y_1' = \boldsymbol{X}'(w_1) = (X_1(w_1), \cdots, X_N(w_1))$.

It is customary to assume that, if w comes from C_1, then $\boldsymbol{X}$ has a multivariate density f_N, and if w comes from C_2, then $\boldsymbol{X}$ is distributed according to the multivariate density g_N. Unless $f_N \neq g_N$, there is no information in the variables for classification. The object of classification is to take any incoming object w, whose class is unknown, and to correctly decide from which class it came using its observation vector $\boldsymbol{x} = \boldsymbol{X}(w) = (x_1, \cdots, x_n)'$, D_1 and D_2, and any prior information.

There are, of course, many algorithms for making the above classification of the unknown w. One very popular algorithm is the Bayes decision rule. Here it is assumed that w is randomly selected from Ω where the probability that w comes from C_1 is

equal to π_1 and that w comes from C_2 is equal to π_2. Here π_1 and π_2 are either known to the person doing the classification or can be estimated by the classifier. Thus we write

$$P\{w \in C_1\} = \pi_1, \qquad P\{w \in C_2\} = \pi_2, \qquad \pi_1 + \pi_2 = 1.$$

If one knows f_N, g_N, π_1, and π_2, the Bayes algorithm says to classify w into that class whose posteriori probability (calculated via Bayes' Theorem) is largest, i.e., if

$$P\{C_1 \mid x\} \equiv \frac{f_N(x)\pi_1}{f_N(x)\pi_1 + g_N(x)\pi_2}$$

is greater than

$$P\{C_2 \mid x\} \equiv \frac{g_N(x)\pi_2}{f_N(x)\pi_1 + g_N(x)\pi_2}$$

then classify w into C_1, otherwise classify w into C_2. In the usual case where f_N and g_N are not known, they are estimated using D_1 and D_2.

All of the above can easily be generalized to more than two classes.

III. A Counterexample

Assume henceforth that all N coordinates of X are independent random variables and, for simplicity, that $\pi_1 = \pi_2 = \frac{1}{2}$. If f_N and g_N are known, then the probability of correct classification for the Bayes algorithm is

$$\begin{aligned} P_c &= \tfrac{1}{2}[P\{f_N(X) \geq g_N(X) \mid w \in C_1\} \\ &\quad + P\{g_N(X) > f_N(X) \mid w \in C_2\}] \qquad (3a) \\ &= \frac{1}{2}\left[P\left\{\ln \frac{f_N(X)}{g_N(X)} \geq 0 \mid w \in C_1\right\}\right. \\ &\quad \left. + P\left\{\ln \frac{g_N(X)}{f_N(X)} > 0 \mid w \in C_2\right\}\right] \qquad (3b) \\ &= 1 - P_e \end{aligned}$$

where P_e is called the probability of misclassification. The form (3b) is convenient here since the coordinates of X are independent and we can write the densities

$$\begin{aligned} f_N(x) &= \prod_{i=1}^{N} f^{(i)}(x_i) \\ g_N(x) &= \prod_{i=1}^{N} g^{(i)}(x_i) \qquad (3c) \end{aligned}$$

where $f^{(i)}$ is the density of X_i given that $w \in C_1$ and $g^{(i)}$ is the density of X_i given $w \in C_2$. In [4] it is claimed that, if for almost all x, $f^{(i)}(x) \neq g^{(i)}(x)$, then

$$\begin{aligned} E_{f_N}\left\{\ln \frac{f_N(X)}{g_N(X)}\right\} &= O(N) \\ \sigma_{f_N}{}^2\left\{\ln \frac{f_N(X)}{g_N(X)}\right\} &= O(N) \qquad (3d) \end{aligned}$$

where, e.g., E_f means the expectation with respect to f. This is *not* necessarily true and Chandrasekaran and Jain acknowledge this in [5]. The following example gives insight into the situation.

Example 1

Suppose the sample space for all X_i is the interval [0,1] and $f^{(i)}(x) = 1$ for all $x \in [0,1]$, $i = 1, \cdots, N$. Furthermore, let

$$g^{(i)}(x) = \begin{cases} 1 + d_i, & 0 \leq x \leq \frac{1}{2} \\ 1 - d_i, & \frac{1}{2} < x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where $0 < d_i < 1$, $i = 1, \cdots, N$. Then

$$\begin{aligned} E_{f_N} \ln \frac{f_N(X)}{g_N(X)} &= -\sum_{i=1}^{N} \int_0^1 \ln g^{(i)}(x_i)\, dx_1 \\ &= -\frac{1}{2}\sum_{1}^{N} [\ln (1 + d_i) + \ln (1 - d_i)] \\ &= -\frac{1}{2}\sum_{1}^{N} \ln (1 - d_i{}^2) = \frac{1}{2}\sum_{1}^{N} a_i \end{aligned}$$

where $d_i = (1 - e^{-a_i})^{1/2}$ and therefore $0 < a_i < \infty$. Choose a_i to be any positive convergent series such as $a_i = i^{-2}$; then

$$E_{f_N}\left\{\ln \frac{f_N(X)}{g_N(X)}\right\} = 0(1).$$

Similarly, choose $a_i = i$ and we get $O(N^2)$. Similar examples can be found for any nontrivial range, not just [0,1].

Thus we see from Example 1 that any order (≥ 1) can be achieved in (3d). If we take $f^{(i)} = f$ and $g^{(i)} = g$, $i = 1, \cdots, N$, then we get $O(N)$ as in (3d). Intuitively, one can imagine several possibilities as N increases. The distributions $f^{(i)}$ and $g^{(i)}$ for large i could quickly approach one another (but stay unequal) or diverge from one another or anything in between. Therefore, the conclusions made [4, p. 241] about perfect discrimination are true only under certain circumstances. Additional sufficient conditions for the conclusions are discussed in the next section and in [5].

IV. Asymptotically Perfect Discrimination

What are some sufficient conditions such that under the assumption of Section III, the algorithm will tend to have perfect discrimination ($P_c = 1$) as $N \to \infty$?

First of all, it should be noted that we are dealing with what in information theory is called "the mean information for discriminating f in favor of g," i.e., we can write this information as

$$I(f:g) \equiv E_f \ln \frac{f(X)}{g(X)}$$

(see. e.g. Kullback [13, ch. 1]). Furthermore, the fact that $I(f:g)$ is large does not imply that $I(g:f)$ is large.

The idea of [4] is to say that if $N \to \infty$, $I(f_N:g_N)$ gets large compared with the standard deviation,

$$\sigma(f_N:g_N) \equiv \left[E_{f_N}\left(\ln \frac{f_N(X)}{g_N(X)}\right)^2 - I^2(f_N:g_N)\right]^{1/2}$$

then we could use the argument of Proposition 1 to say

$$P\left\{\ln \frac{f_N(X)}{g_N(X)} > 0\right\}$$

gets close to 1.

Proposition 1: If $I(f_N:g_N)/\sigma(f_N:g_N) \to \infty$, then the probability of correctly classifying an object from C_1 tends to one.

Proof: The probability of an error given C_1 is

$$\begin{aligned} &P\left\{\ln \frac{f_N(X)}{g_N(X)} < 0 \mid C_1\right\} \\ &\quad = P\left\{\ln \frac{f_N(X)}{g_N(X)} - I(f_N:g_N) < -I(f_N:g_N) \mid C_1\right\} \\ &\quad \leq P\left\{\left|\ln \frac{f_N(X)}{g_N(X)} - I(f_N:g_N)\right| > I(f_N:g_N) \mid C_1\right\} \\ &\quad \leq \left(\frac{\sigma(f_N:g_N)}{I(f_N:g_N)}\right)^2 \to 0 \qquad \text{as } N \to \infty \end{aligned}$$

by Chebyshev's inequality. Q.E.D.

Proposition 2: Under the assumptions of Section III,

$$I(f_N:g_N) = \sum_1^N I(f^{(i)}:g^{(i)})$$

$$\sigma^2(f_N:g_N) = \sum_1^N \sigma^2(f^{(i)}:g^{(i)})$$

and similarly for $I(g_N:f_N)$ and $\sigma(g_N:f_N)$.

Proof: This is well known and follows immediately from the fact that

$$\ln \frac{f_N(X)}{g_N(X)} = \ln \prod_{i=1}^N \frac{f^{(i)}(X_i)}{g^{(i)}(X_i)} = \sum_1^N \ln \frac{f^{(i)}(X_i)}{g^{(i)}(X_i)}$$

and that $\ln f^{(i)}(X_i) - \ln g^{(i)}(X_i)$ is independent of $\ln f^{(j)}(X_j) - \ln g^{(j)}(X_j)$ for $i \neq j$ under f_N and g_N. Q.E.D.

Corollary 3: Under the assumption of Section III, if $f^{(i)} = f$, $g^{(i)} = g, i = 1, \cdots, N$; and $I(f:g)/\sigma(f:g) \neq 0, I(g:f)/\sigma(f:g) \neq 0$, then one gets perfect discrimination as $N \to \infty$.

The conditions of Corollary 3 are obviously not necessary for asymptotically perfect discrimination. In fact, little in the way of general results involving I/σ can be given that are not immediate from Propositions 1 and 2.

Another approach which is commonly used would be to use bounds on P_c in terms of various types of information measures. There is a large literature on this subject (see, e.g., Toussaint [17], [18] for lists of references) and many reasonable measures of information have been proposed including the divergence (see, e.g., Kullbach [13, ch. 1]),

$$J(f,g) = I(f:g) + I(g:f).$$

The divergence has the advantage that it takes a particularly nice form when the independence assumptions of this correspondence are made and the densities take the form (3c). Two upper bounds for P_c in terms of J are given by Kailath [11],

$$1 - P_c \geq \tfrac{1}{4}e^{(-J/2)} \tag{4a}$$

and by Toussaint [17],

$$1 - P_c \geq \tfrac{1}{2} - \tfrac{1}{2}(1 - 4 \exp [-2H(\pi) - J(\pi_1 f, \pi_2 g])^{1/2} \tag{4b}$$

where

$$H(\pi) = -\pi_1 \ln \pi_1 - \pi_2 \ln \pi_2.$$

Note that if $\pi_1 = \pi_2 = \frac{1}{2}$, then (4b) becomes

$$1 - P_c \geq \tfrac{1}{2} - \tfrac{1}{2}(1 - e^{(-J/2)})^{1/2}. \tag{4c}$$

Unfortunately, no nontrivial lower bounds for P_c in terms of J exist as is shown in the following theorem.

Theorem 1: Assume that a Bayes decision rule is used with $\pi_1 = \pi_2 = \frac{1}{2}$ and f and g known. Then the best upper bound on the probability of error, $P_e \equiv 1 - P_c$, which is a function of J alone and which holds for all f and g is $P_e \leq \frac{1}{2}$.

Proof: First of all, it is well known that $P_e \leq \frac{1}{2}$ for any f and g. We now need to show for arbitrary fixed $J \geq 0$ that we can find an f and g with that J and with the corresponding P_e arbitrarily close to $\frac{1}{2}$. If $J = 0$ we take $f = g$ and hence $P_e = \frac{1}{2}$. Henceforth assume that X can take on only the values 0 and 1. Let $1 > \delta > 0$ be arbitrarily small. If $J = \infty$, choose

$$f(0) = \delta, \qquad f(1) = 1 - \delta$$

$$g(0) = 0, \qquad g(1) = 1$$

then $J = \infty$ and $P_e = (1 - \delta)/2$.

Finally, let J be fixed, $0 < J < \infty$. Let $0 < \delta$ be arbitrarily small and small enough so that

$$e^{-J/\delta} < \tfrac{1}{8} \quad \text{and} \quad \delta < \tfrac{1}{8}. \tag{4d}$$

Let $\delta < \varepsilon < 1$ and

$$f(0) = \varepsilon, \qquad f(1) = 1 - \varepsilon$$

$$g(0) = \varepsilon - \delta, \qquad g(1) = 1 - \varepsilon + \delta$$

then $P_e = \frac{1}{2}(1 - \delta)$ and

$$J = \delta \left(\ln \frac{\varepsilon}{\varepsilon - \delta} + \ln \frac{1 - \varepsilon + \delta}{1 - \varepsilon} \right). \tag{4e}$$

If we can solve (4e) for ε such that $\delta < \varepsilon < 1$, then the theorem is proved. We write (4e) as

$$e^J = \left(\frac{\varepsilon}{\varepsilon - \delta} \cdot \frac{1 - \varepsilon + \delta}{1 - \varepsilon} \right)^\delta$$

so that

$$\varepsilon^2 - \varepsilon - \varepsilon\delta - \frac{\delta}{e^{-J/\delta} - 1} = 0. \tag{4f}$$

This has real solutions if

$$(1 + \delta)^2 + 4 \frac{\delta}{e^{-J/\delta} - 1} > 0$$

which in turn will be satisfied if

$$\left| \frac{4\delta\, e^{J/\delta}}{1 - e^{J/\delta}} \right| \leq 1$$

i.e., if

$$\frac{4\delta\, e^{J/\delta}}{e^{J/\delta} - 1} \leq 1$$

or

$$\delta \leq \frac{e^{J/\delta} - 1}{4\, e^{J/\delta}} = \tfrac{1}{4} - \tfrac{1}{4}e^{-J/\delta}$$

which is satisfied by (4d). We now want to check whether the solution of (4f) is between δ and 1. Substituting 1 for ε in the left-hand side of (4f) we get

$$\delta \left(\frac{-1}{1 - e^{J/\delta}} \right) > 0. \tag{4g}$$

Substituting δ we also get (4g). Since the left-hand side of (4f) is a parabola which tends to $+\infty$ as $\varepsilon \to \pm\infty$ and since its values are positive and equal evaluated at δ and 1 and since solutions exist they must be between δ and 1. Q.E.D.

This leaves one with three options: 1) abandon the bounding of P_c approach, 2) work with other forms of information proposed in the literature (which do not take the particularly nice form that J does in the case of independence), or 3) look for some other convenient (in the case of independence) quantity for which suitable bounds on P_c do exist.

V. Unknown Conditionals

The case of greatest practical interest is, of course, when f and g are unknown and must be estimated from training data. We assume that we have training data D_1 and D_2 as defined in Section II; the points in D_1 are m_1 independent identically distributed observations from the distribution f_N and the m_2

points in D_2 are similarly selected from g_N. From this data we form estimates, $\hat{f}_N$ and $\hat{g}_N$, of f_N and g_N. In [4] the assumption is made that $\hat{f}_N$ and $\hat{g}_N$ take the form (3c) implying that one estimates each marginal f_i, $i = 1,\cdots, N$ and g_i, $i = 1,\cdots, N$, separately. This would be likely to happen only if one knew ahead of time he was dealing with independent coordinate variables since most methods of forming f_N and g_N do not lead to the form (3c). The most common nonparametric method of obtaining density estimates is to use a Parzen-type estimate, for example,

$$f_N(x) = \int h(x - y)\, dF_N(y)$$

where h is a "window" which is usually required to be nonnegative and integrable to 1 and F_N is the sample distribution function of D_1. A common choice for h is the Gaussian window (parameter $\sigma > 0$)

$$h_\sigma(x) = (2\pi\sigma^2)^{-n/2} \left\{\exp - \frac{1}{2\sigma^2} x'x\right\}.$$

In this case

$$f_N(x) = (2\pi\sigma^2)^{-n/2} m_1^{-1} \sum_{y_i \in D_1} \exp - \frac{1}{2\sigma^2} \sum_{j=1}^{N} (x_j - y_{ij})^2$$

which does not take the form (3c).

Suppose nonetheless that one did know enough to use estimates of the form (3c):

$$\hat{f}_N(x) = \prod_1^N \hat{f}^{(i)}(x_i)$$

$$\hat{g}_N(x) = \prod_1^N \hat{g}^{(i)}(x_i) \tag{5a}$$

where to be more exact one should, e.g., write $\hat{f}^{(i)}$ as a function of $m_1 + 1$ variables:

$$\hat{f}^{(i)}(x_i) = \hat{f}^{(i)}(y_1^{(i)},\cdots, y_{m_1}^{(i)}, x_i).$$

Note that in this case the random variable $\hat{f}^{(i)}$ is independent of $\hat{f}^{(j)}$ and $\hat{g}^{(j)}$, $j \neq i$, but not necessarily independent of $\hat{g}^{(i)}$.

The quantity of interest is then

$$P_c = \tfrac{1}{2}[P\{\hat{f}_N(X) \geq \hat{g}_N(X) \mid w \in C_1\} + P\{\hat{g}_N(X) > \hat{f}_N(X) \mid w \in C_2\}. \tag{5b}$$

Transferring to logarithms as in (3b) we define

$$V_N = \sum_1^N [\ln \hat{f}^{(i)}(Y_1^{(i)}, \cdots, Y_{m_1}^{(i)}, X_i) - \ln \hat{g}^{(i)}(Z_1^{(i)}, \cdots, Z_{m_2}^{(i)}, X_i)]$$

then (5b) becomes

$$P_c = \tfrac{1}{2}[P\{V_N \geq 0 \mid w \in C_1\} + P\{V_N < 0 \mid w \in C_2\}]. \tag{5c}$$

Note that V_N is a sum of N independent random variables. Much is known about such sums, perhaps the most comprehensive recent colligation of results on the central limit problem is in Petrov [15]. Given more information about f, g, $\hat{f}$, and $\hat{g}$ one could use those results to give more detailed statements.

In [4, p. 242] it is claimed that a necessary and sufficient condition for P_c to tend to 1 as $N \to \infty$ is for

$$E_f E_D V_N \to \infty \tag{5d'}$$

and

$$E_g E_D V_N \to -\infty \tag{5d''}$$

where E_D indicates the expectation over the training data. It is reasonably clear, however, that (5d) is neither necessary nor sufficient since one cannot say whether or not $P\{V > 0\} = 1$ knowing only that $EV = \infty$. In the appendix to [4], the authors comment that in order for their result to hold, certain quantities ought to be "sufficiently well-behaved" but do not give any set of additional requirements. A counterexample to the sufficiency of (5d) is given below.

Example 2

Let $N(a,b)$ denote the Gaussian density with mean a and variance b and

$$f_i = N(0,\sigma_i^2)$$

$$g_i = N(1,\sigma_i^2), \qquad i = 1,2,\cdots,$$

where it is assumed that we know that the densities are Gaussian and that the variances are σ_i, $i = 1,2,\cdots$ but we do not know the means. Assume that $m_1 = m_2 = 1$, then the natural estimates are

$$\hat{f}^{(i)}(x_i) = \frac{1}{2\Pi\sigma_i} \exp\, [-(1/2\sigma_i^2)(x_i - y_{1i})^2]$$

$$\hat{g}^{(i)}(x_i) = \frac{1}{2\Pi\sigma_i} \exp\, [-(1/2\sigma_i^2)(x_i - z_{1i})^2].$$

Calculating (5d′) and (5d″) we get, respectively,

$$E_f E_D \sum_1^N \frac{1}{2\sigma_i^2} [(X_i - Z_{1i})^2 - (X_i - Y_{1i})^2] = \sum_1^N \frac{1}{2\sigma_i^2}$$

$$E_g E_D \sum_1^N \frac{1}{2\sigma_i^2} [(X_i - Z_{1i})^2 - (X_i - Y_{1i})^2] = -\sum_1^N \frac{1}{2\sigma_i^2}$$

Choose, for example, $\sigma_i^2 = i$, $i = 1,2,\cdots$; then condition (5d) is satisfied.

To complete our counterexample we need to show that P_c does not tend to 1 as $N \to \infty$. Look only at

$$P\{V_N \geq 0 \mid w \in C_1\} = P\left\{\sum_1^N \left[\left(\frac{X_i - Z_{1i}}{\sqrt{2\sigma_i}}\right)^2 - \left(\frac{X_i - Y_{1i}}{\sqrt{2\sigma_i}}\right)^2\right] > 0 \mid w \in C_1\right\}. \tag{5e}$$

The above sum is a sum of independent random variables and we will now show that the central limit theorem applies. Let V_{Ni} denote the ith term in the above sum, then as we have stated

$$E(V_{Ni} \mid w \in C_1) = \frac{1}{2\sigma_i^2}.$$

Furthermore, it is easy to calculate that

$$\text{var}\,(V_{Ni} \mid w \in C_1) = 3 + \frac{2}{\sigma_i^2}.$$

Finally, we need only show that

$$E(|V_{Ni} - EV_{Ni}|^3 \mid w \in C_1) < R < \infty. \tag{5f}$$

The random variables

$$V_{Ni}' \equiv \left(\frac{X_i - Z_{1i}}{\sqrt{2}\,\sigma_i}\right)^2$$

and

$$V_{Ni}'' \equiv \left(\frac{X_i - Y_{1i}}{\sqrt{2}\,\sigma_i}\right)^2$$

are squares of $N(1/\sqrt{2}\,\sigma_i,1)$ and $N(0,1)$ random variables, respectively. Since by Minkowski's inequality

$$\left(E\left|V_{Ni}-\frac{1}{2\sigma_i^2}\right|^3\right)^{1/3} \le (EV_{Ni}'^3)^{1/3}+(EV_{Ni}''^3)^{1/3}+\left(\frac{1}{8\sigma_i^6}\right)^{1/3}$$

and $EV_{Ni}'^3$ and $EV_{Ni}''^3$ are sixth moments of these normal random variables we have shown (5f). We can therefore apply Lyapunov form of the central limit theorem (see Breiman [3, p. 186], e.g.).

Write (5e) as

$$P\left\{\frac{V_N-\sum_1^N\frac{1}{2\sigma_i^2}}{\left(3N+2\sum_1^N\sigma_i^{-2}\right)^{1/2}} > -\frac{\sum_1^N\frac{1}{2\sigma_i^2}}{\left(3N+2\sum_1^N\sigma_i^2\right)^{1/2}}\,\middle|\, w\in C_1\right\}.$$

Again let $\sigma_i = 1$, $i = 1,2,\cdots$ and the right-hand side of the above inequality tends to 0 as $N\to\infty$, while the left-hand side tends to a $N(0,1)$ random variable and therefore $P_c \to \frac{1}{2}$.

An interesting example where it pays to ignore training data is given in [4, p. 243, Example B]. This example works because the decision rule does not make use of important information.

VI. Conclusions

The results indicate that perfect discrimination for Bayes-type decision rules may or may not be obtained as the dimension $N\to\infty$ with independent coordinates. This depends on the "degree of separation" provided by the variables being added. It is, of course, not sufficient to look only at the means of $\hat{f}^{(i)}$ and $\hat{g}^{(i)}$ or $\ln \hat{f}^{(i)}$ and $\ln \hat{g}^{(i)}$. The divergence J is also not of much use to indicate perfect separation. To give more detailed results than given here would require more detailed information about $f, g, \hat{f}$, and $\hat{g}$. An interesting open question is to find a convenient information measure for the case of independence which also provides reasonable upper bounds on P_e.

References

[1] K. Abend and T. J. Harley, Jr., "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 420–421, 1969.
[2] D. C. Allais, "The problem of too many measurements in pattern recognition and prediction," *IEEE Int. Conv. Rec.*, vol. 14, pp. 124–130, 1966.
[3] Leo Breiman, *Probability*. Reading, Mass.: Addison-Wesley, 1968.
[4] B. Chandrasekaran and A. K. Jain, "Independence, measurement complexity, and classification performance," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, pp. 240–244, Mar. 1975.
[5] B. Chandrasekaran and A. K. Jain, " 'Independence, measurement complexity, and classification performance': An emendation," *IEEE Trans. Syst., Man, Cybern.*, this issue, pp. 564–566.
[6] O. J. Dunn, "Some expected values for probabilities of current classification in discriminant analysis," *Technometrics*, vol. 13, pp. 345–353, 1971.
[7] L. D. Fisher, R. A. Kronmal, and P. Diehr, "Mathematical aids to medical decision making," manuscript, Dep. of Biostatistics, Univ. of Washington, Seattle, 1973.
[8] L. D. Fisher and R. A. Kronmal, "The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis," manuscript, Dep. of Biostatistics, Univ. of Washington, Seattle, 1974.
[9] D. H. Foley, "Considerations of sample and feature size," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 618–626, 1972.
[10] G. F. Hughes, "On the mean accuracy of statistical pattern recognizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55–63, 1968.
[11] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, Feb. 1967.
[12] E. R. Knezek and T. L. Boullion, "A simulation study of population classification using selected variates," unpublished paper, Mathematics Dep., Texas Tech. Univ., Lubbock, 1973.
[13] S. Kullback, *Information Theory*. New York: Dover, 1968.
[14] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.
[15] V. V. Petrov, *Sums of Independent Random Variables*. New York: Springer-Verlag, 1975.
[16] M. Sorum, "Estimating the expected and the optimal probabilities of misclassification," *Technometrics*, vol. 14, pp. 935–943, 1972.
[17] G. T. Toussaint, "On some measures of information and their application to pattern recognition," in *Proc. Conf. Measures of Information and Their Applications*, Indian Inst. Technol., Bombay, Aug. 1974.
[18] G. T. Toussaint, "Probability of error and equivocation of order α," Rep., School of Computer Science, McGill Univ., Montreal, PQ, Canada, 1975.
[19] J. Van Ness and C. Simpson, "On the effects of dimension in discriminant analysis," *Technometrics*, vol. 18, pp. 175–187, May 1976.

"Independence, Measurement Complexity, and Classification Performance": An Emendation

B. Chandrasekaran, member, ieee, and Anil K. Jain, member, ieee

***Abstract*—Some of the results of our paper, "Independence, Measurement Complexity, and Classification Performance" [1], require clarification and some need to be modified. That task is undertaken.**

The fabric of our paper [1] is marred by one weak thread running through it, and that thread is spun out of an inadequate convergence argument. Van Ness [2] locates the problem clearly. On the other hand, the comments by Duin [3] are a mixture: one mistaken, one a semantic difference, and one true, but rather moot. We shall briefly go through the results of the earlier paper [1] and point out where they need modification. Along the way we shall comment on the points raised in [2].

To recapitulate the notation:

$$f_N(x) \equiv \prod_{i=1}^N f_i(x_i), \qquad g_N(x) \equiv \prod_{i=1}^N g_i(x_i)$$

are the two class-conditional densities; x is the pattern vector, which consists of N independent measurements x_i, $i = 1,\cdots,N$; $\hat{f}$, etc., stand for estimated densities; c_1 and c_2 are the two classes. All summations, unless otherwise stated, are from 1 to N. $E_{x\in c_j}$ will stand for expectation with respect to class c_j, and $E\chi$ for expectation with reference to training data sets χ. In addition, for notational simplicity, we use the following abbreviations in this note:

$$\hat{d}_i \equiv \log \hat{f}_i - \log \hat{g}_i$$

$$D_i^{(j)} \equiv E_{x\in c_j} E\chi \hat{d}_i$$

and

$$V_i^{(j)} \equiv E_{x\in c_j} E\chi[(\hat{d}_i - D_i^{(j)})^2].$$

1) In the case of infinite sample sets, i.e., the case of known $f_N(x)$ and $g_N(x)$, we claim in [1] that if, for all i and almost all x,

$$f_i(x) \neq g_i(x) \tag{1}$$

then

$$E_f\left\{\log\frac{f_N(x)}{g_N(x)}\right\} = O(N)$$

and

$$\sigma_f{}^2\left\{\log\frac{f_N(x)}{g_N(x)}\right\} = O(N)$$

where E_f and $\sigma_f{}^2$ refer to expectation and variance with respect to the f-distribution. This would be true if all the measurement variables are identically distributed. In the general case, however, condition (1) is insufficient to guarantee the $O(N)$ behavior of the

Manuscript received October 18, 1976; revised January 27, 1977. This work was supported by the Air Force Office of Scientific Research under Grant 72-2351.

B. Chandrasekaran is with the Department of Computer and Information Science, Ohio State University, Columbus, OH 43210.
A. K. Jain is with the Department of Computer Science, Michigan State University, East Lansing, MI 48824.