Ann. Inst. Statist. Math.
29 (1977), Part A, 433-444

ON LINEAR CLASSIFICATION PROCEDURES BETWEEN TWO CATEGORIES WITH KNOWN MEAN VECTORS AND COVARIANCE MATRICES

AKIHIRO NISHI

(Received Nov. 5, 1976)

Summary

This paper is concerned with probabilities (error probabilities), caused by misclassification, of linear classification procedures (linear procedures) between two categories, whose mean vectors and covariance matrices are assumed to be known, while the distribution of each category may well be continuous or discrete. The tightest upper bounds on the largest of two kinds of error probability of each linear procedure and on the expected error probability for any apriori probabilities are obtained. Moreover in some cases of interest, the *optimal* linear procedure (in the sense of attaining the infimum out of all the upper bounds) is given.

1. Introduction

Suppose there are several categories, characterized by their probability distributions, from which every observation comes. In the general statistical classification problem, we must assign one of the categories to the observation according to some predetermined procedure. We want to seek the procedure, in a class of procedures such as a class of linear ones or a class of all ones (admitting randomized ones), that makes the probability of misclassification as small as possible.

We treat, throughout this paper, the case of two categories, whose mean vectors μ_1, μ_2 (not the same) and covariance matrices Σ_1, Σ_2 (nondegenerate) are assumed to be known. We define linear procedures as follows. Let d ($\neq 0$) be a vector and θ a scalar. An observation x is classified into the first category C_1 if $d'x \leq \theta$ and into the second one C_2 if $d'x > \theta$. The function $d'x-\theta$ is called the discriminant function with respect to this classification procedure.

When the distribution of each category is multivariate normal, Anderson and Bahadur [2] have studied the linear procedures in detail. The error probability of the minimax procedure is

$$\Phi^*(\Delta/2)=\int_{\Delta/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\,, \tag{1.1}$$

where $\Delta^2=4\max_{d\neq 0}\{d'(\mu_1-\mu_2)/(\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d})\}^2$ is the generalized Mahalanobis distance. In the case of $\Sigma_1=\Sigma_2$ $(=\Sigma)$ the minimax procedure among all possible procedures (admitting randomized ones) is given by well known linear discriminant function $x'\Sigma^{-1}(\mu_2-\mu_1)-(1/2)(\mu_2+\mu_1)\Sigma^{-1}(\mu_2-\mu_1)$. Furthermore in this case the procedure is Bayes one with respect to equal apriori probabilities, and Δ^2 becomes usual $(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$.

Our main concern is not with the case where we know the form (suitably parametrized) of the distribution for each category but with the case where we are given only first and second moments of the distribution. Let $F=(F_1, F_2)$, where F_i is the distribution function of the category C_i $(i=1, 2)$, $\mathcal{F}=(\mathcal{F}_1, \mathcal{F}_2)$, where $\mathcal{F}_i=\mathcal{F}(\mu_i, \Sigma_i)$ is a class of all distribution functions with specified mean vector μ_i and covariance matrix Σ_i, and $\Pi=(\Pi_1, \Pi_2)$ $(\Pi_i\geqq 0,\ \Pi_1+\Pi_2=1)$, where Π_i is the apriori probability of C_i. Let $\phi(x)=(\phi_1(x), \phi_2(x))$ (Lebesgue measurable, $\phi_i(x)\geqq 0$, $\phi_1(x)+\phi_2(x)\equiv 1$) be a (randomized) procedure such that an observation x is classified into C_i with probability $\phi_i(x)$, $\Phi=\{\phi(x)\}$ be the collection of all such ϕ's. And define by $\Phi^L=\{\phi^L(x)\}$ be the collection of all linear procedures, where $\phi_1^L(x)=\phi_1^{d,\theta}(x)$ is equal to the indicator function $I_{\{d'x\leqq\theta\}}(x)$ of the half space $\{x\,|\,d'x\leqq\theta\}$, and similarly for $\phi_2^L(x)$. When the true distribution of the category C_i is F_i, the (conditional) error probability of classifying an observation, which is actually taken from C_i, into C_{3-i} $(i=1, 2)$ becomes

$$e_i(\phi, F)=\int(1-\phi_i(x))dF_i\,,\qquad i=1, 2 \tag{1.2}$$

and the expected error probability with respect to the apriori probabilities Π becomes

$$e_\Pi(\phi, F)=\Pi_1 e_1(\phi, F)+\Pi_2 e_2(\phi, F)\,. \tag{1.3}$$

In the case of the univariate and equal apriori probabilities (put $\Pi^0=(1/2, 1/2)$), Chernoff [3] showed that $\sup_F\inf_\phi e_{\Pi^0}(\phi, F)=\sup_F\inf_{\phi^L} e_{\Pi^0}(\phi^L, F)=(1/2)[1+\Delta^2/4]^{-1}$, where $\Delta=2|\mu_1-\mu_2|/(\sigma_1+\sigma_2)$, by two excellent methods which, however, seem difficult to be extended to the multivariate case. On the other hand, Isii and Taga [6], from the point of view of mathematical programming, treated the multivariate case in full generality. Lachenbruch et al. [7] studied robustness of the linear and quadratic classification procedures.

Restricting the procedures to the linear ones, Yau and Lin [8] obtained an upper bound for $\inf_{\phi^L}\sup_F e_\Pi(\phi^L, F)$, where $\sup_F e_\Pi(\phi^L, F)$ seems to represent the deficiency of ϕ^L with respect to Π. Therefore, if ϕ^{*L} attains the infimum, it is considered optimal with respect to Π. In the particular case of Π^0, their upper bound is

$$\inf_{\phi^L}\sup_F e_{\Pi^0}(\phi^L, F)\leqq 2[(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)]^{-1}. \tag{1.4}$$

Under the further assumption of equal covariance matrices, this bound becomes

$$\inf_{\phi^L}\sup_F e_{\Pi^0}(\phi^L, F)\leqq [\Delta^2/4]^{-1}. \tag{1.5}$$

We shall give, applying Isii's theorem [5] on Chebyshev-type inequalities, the tightest upper bound of $\inf_{\phi^L}\sup_F e_{\Pi}(\phi^L, F)$ (Theorem 2.3, though it has a rather complicated form. From Theorems 3.2 and 3.1 that will be given in Section 3, we obtain corresponding to (1.4), (1.5), respectively,

$$\inf_{\phi^L}\sup_F e_{\Pi^0}(\phi^L, F)\leqq [1+\Delta^2/4]^{-1} \tag{1.6}$$

and

$$\inf_{\phi^L}\sup_F e_{\Pi^0}(\phi^L, F)=\min\left\{\frac{1}{2},\ [1+\Delta^2/4]^{-1}\right\}, \tag{1.7}$$

when $\Sigma_1=\Sigma_2$. We can easily show that (1.6) strictly improves (1.4) by using the following facts (i) (ii),

(i) $\sup_{d\neq 0}\left(\dfrac{d'a}{\sqrt{d'Ad}}\right)^2=a'A^{-1}a$ for any positive definite symmetric matrix A and any constant vector a,

(ii) $(\sqrt{\alpha}+\sqrt{\beta})/\sqrt{2}\leqq\sqrt{\alpha+\beta}$ for any scalar $\alpha\geqq 0$, $\beta\geqq 0$. In fact,

$$\begin{aligned}(1.8)\quad & 2[(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)]^{-1}\\ &=2\left[\sup_{d\neq 0}\left(\frac{d'(\mu_1-\mu_2)}{\sqrt{d'(\Sigma_1+\Sigma_2)d}}\right)^2\right]^{-1}\geqq\left[\sup_{d\neq 0}\left(\frac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d}}\right)^2\right]^{-1}\\ &=(\Delta^2/4)^{-1}>[1+\Delta^2/4]^{-1}.\end{aligned}$$

2. Chebyshev-type inequalities

It may be easily examined that Isii's theorem ([5], Th. 3.1., pp. 285–286) can be used for calculating

$$\sup_F e_i(\phi^L, F)=\sup\left\{\int(1-\phi_i^L(x))dF_i \mid F_i\in\mathcal{F}(\mu_i, \Sigma_i)\right\}. \tag{2.1}$$

Thus we have

$$\text{(2.2)} \qquad \sup_F e_i(\phi^L, F) = \inf\left\{\int g(x)dF_i \,\middle|\, \begin{array}{l} g(x) \geqq 1-\phi_i^L(x) \text{ for all } x \\ g(x)\text{: quadratic function} \end{array}\right\}, \qquad i=1, 2.$$

We easily see from the right side of the above equality that we need not pay attention to $g(x)$ if there exists a quadratic function $h(x)$ satisfying $1-\phi_i^L(x) \leqq h(x) \leqq g(x)$ for all x. Since $g(x)$ is quadratic and always non-negative, we can write in the form $g(x)=(x-a)'A(x-a)+\gamma$, where A is a non-negative definite symmetric matrix, a is a vector, and γ is a non-negative scalar.

LEMMA 2.1. *If $g(x)=(x-a)'A(x-a)+\gamma \geqq I_{\{d'x>\theta\}}(x)$ for all x, then we have the following.*

(i) *There exists a parabolic cylinder function $h(x)$,*

$$g(x) \geqq h(x) \geqq I_{\{d'x>\theta\}}(x) \qquad \text{for all } x,$$

provided that d is contained in the subspace spanned by all eigenvectors for positive eigenvalues of A,

(ii) *$\gamma \geqq 1$, provided that d is not contained in the above subspace.*

PROOF. (i) Normalizing the vector d ($d'd=1$) only for convenience, there exists by the assumption an appropriate orthogonal matrix S, whose first column vector coincides with d, satisfying

$$S'AS = \begin{array}{cc} \begin{array}{cc} r & \end{array} & \\ \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} & \begin{array}{c} r \\ \\ \end{array} \end{array},$$

where $r=\operatorname{rank} A$. Now let us put $x=Sy$, $a=Sp$, $y=\begin{bmatrix} u \\ * \\ * \end{bmatrix}^{r}$, and $p=\begin{bmatrix} q \\ * \\ * \end{bmatrix}^{r}$, then we have $g(x)=(x-a)'A(x-a)+\gamma=(u-q)'B(u-q)+\gamma$. Put further as $u=\begin{bmatrix} u^1 \\ u_2 \end{bmatrix}\begin{array}{l} 1 \\ r-1 \end{array}$, $q=\begin{bmatrix} q^1 \\ q_2 \end{bmatrix}\begin{array}{l} 1 \\ r-1 \end{array}$, $B=\begin{array}{c} \begin{array}{cc} 1 & r-1 \end{array} \\ \begin{bmatrix} b^{11} & b' \\ b & B_{22} \end{bmatrix} \end{array}\begin{array}{l} 1 \\ r-1 \end{array}$ and seek the maximum of $g(x)$ under the constraint $u^1=\eta$ (constant scalar), then we have

$$\text{(2.3)} \qquad \begin{aligned} g(x) &= (u_2-q_2-(\eta-q^1)B_{22}^{-1}b)'B_{22}(u_2-q_2-(\eta-q^1)B_{22}^{-1}b) \\ &\quad +(b^{11}-b'B_{22}^{-1}b)(\eta-q^1)^2+\gamma. \end{aligned}$$

Since B_{22} is positive difinite, $g(x)$ is minimized when $u_2=q_2+(\eta-q^1)B_{22}^{-1}b$ under the constraint $u^1=\eta$, and we have

$$\text{(2.4)} \qquad \begin{aligned} g(x) &\geqq (b^{11}-b'B_{22}^{-1}b)(\eta-q^1)^2+\gamma \\ &= (d'Ad-b'B_{22}^{-1}b)(d'x-d'a)^2+\gamma \geqq I_{\{d'x>\theta\}}(x) \qquad \text{for all } x. \end{aligned}$$

Note $b^{11}-b'B_{22}^{-1}b>0$, because B and B_{22} are positive definite and $|B|=|B_{22}|(b^{11}-b'B_{22}^{-1}b)$. Hence

$$h(x)=(d'Ad-b'B_{22}^{-1}b)(d'(x-a))^2+\gamma \tag{2.5}$$

is a desired parabolic cylinder function.

(ii) This is readily seen to hold by a geometric consideration.

q.e.d.

We shall calculate only $\sup_F e_1(\phi^L, F)$, since $\sup_F e_2(\phi^L, F)$ can be obtained in the same way. Now let us put $\phi^L=\phi^{d,\theta}$, then $1-\phi_1^{d,\theta}(x)=I_{\{d'x>\theta\}}(x)$. We omit the subscript 1 in μ_1 and Σ_1 for simplicity. By an orthogonal transformation $y=S'x$, the family of distribution functions $\mathcal{F}(\mu, \Sigma)$ is mapped onto the family $\mathcal{F}(S'\mu, S'\Sigma S)$. Thus (2.2), after some calculations, becomes

$$\begin{aligned}\sup_F e_1(\phi^{d,\theta}, F)=\inf\{&\mathrm{tr}\,(BS'\Sigma S)+(S'\mu-p)'B(S'\mu-p)+\gamma \mid \\ &(y-p)'B(y-p)+\gamma\geqq I_{\{d'Sy>\theta\}}(y) \text{ for all } y\}\,,\end{aligned} \tag{2.6}$$

where $\mathrm{tr}\,M$ denotes the trace of the matrix M. Considering Lemma 2.1, we may assume that B has a form $B=\begin{bmatrix}\lambda & & 0\\ & 0 & \\ & & \ddots \\ 0 & & & 0\end{bmatrix}$ (λ is a positive scalar) and that the first column vector of S coincides with d. Thus (2.6) becomes

$$\begin{aligned}\sup_F e_1(\phi^{d,\theta}, F)=\inf\{&\lambda d'\Sigma d+\lambda(d'\mu-p^1)^2+\gamma \mid \lambda(y^1-p^1)^2+\gamma \\ &\geqq I_{\{y^1>\theta\}}(y) \text{ for all } y\}\,,\end{aligned} \tag{2.7}$$

where $y'=(y^1, y^2, \cdots)$, $p'=(p^1, p^2, \cdots)$. The infimum of

$$\psi^{d,\theta}(\lambda, p^1, \gamma)\equiv\lambda d'\Sigma d+\lambda(d'\mu-p^1)^2+\gamma \tag{2.8}$$

will now be obtained under the constraint $\lambda(y^1-p^1)^2+\gamma\geqq I_{\{y^1>\theta\}}(y)$ for all y. From Fig. 2.1 it is enough to restrict the functions dominating $I_{\{y^1>\theta\}}$ to the functions satisfying (i) $p^1<\theta$, (ii) $0\leqq\gamma<1$, (iii) $\lambda(\theta-p^1)^2+\gamma=1$. Then $\psi^{d,\theta}$ reduces to

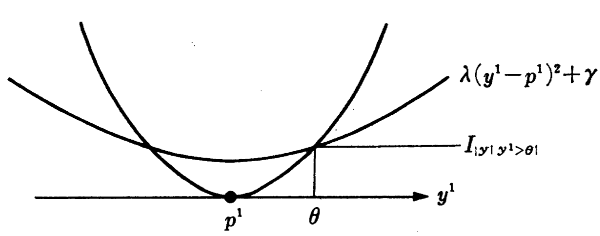

Fig. 2.1

$$\phi^{d,\theta}(\lambda, p^1, \gamma)=(d'\Sigma d+(\theta-d'\mu)^2)\left[\sqrt{\lambda}+\frac{(d'\mu-\theta)\sqrt{1-\gamma}}{d'\Sigma d+(\theta-d'\mu)^2}\right]^2 + \frac{d'\Sigma d+(d'\mu-\theta)^2\gamma}{d'\Sigma d+(\theta-d'\mu)^2}\,. \tag{2.9}$$

Thus we easily have the infimum of $\phi^{d,\theta}$ as

$$\inf\{\phi^{d,\theta}(\lambda, p^1, \gamma)\,|\,\lambda(y^1-p^1)^2+\gamma\geqq I_{\{y^1>\theta\}}(y^1) \text{ for all } y^1\} = \begin{cases} \dfrac{d'\Sigma d}{d'\Sigma d+(\theta-d'\mu)^2}\,, & \text{when } d'\mu<\theta\,. \\ 1\,, & \text{when } d'\mu\geqq\theta\,. \end{cases} \tag{2.10}$$

Summing up those that precede, we obtain the following theorem.

THEOREM 2.1.

$$\text{(i)}\quad \sup_F e_1(\phi^{d,\theta}, F)=\begin{cases} \dfrac{d'\Sigma_1 d}{d'\Sigma_1 d+(\theta-d'\mu_1)^2}\,, & \textit{when } d'\mu_1<\theta \\ 1\,, & \textit{when } d'\mu_1\geqq\theta \end{cases} \tag{2.11}$$

$$\text{(ii)}\quad \sup_F e_2(\phi^{d,\theta}, F)=\begin{cases} \dfrac{d'\Sigma_2 d}{d'\Sigma_2 d+(d'\mu_2-\theta)^2}\,, & \textit{when } d'\mu_2>\theta \\ 1\,, & \textit{when } d'\mu_2\leqq\theta\,. \end{cases} \tag{2.12}$$

By the expression for $\sup e_i(\phi^{d,\theta}, F)$ in the above theorem, it is enough to restrict d to the vectors satisfying $d'\mu_1<d'\mu_2$ for computing $\max_{i=1,2} e_i(\phi^L, F)$, though this might be expected by our definition of $\phi^{d,\theta}$.

THEOREM 2.2.

$$\inf_{\phi^L}\sup_F\max_{i=1,2} e_i(\phi^L, F)=\left(1+\frac{\Delta^2}{4}\right)^{-1}, \tag{2.13}$$

where

$$\Delta^2=4\max_{d\neq 0}\left\{\frac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d}}\right\}^2.$$

Furthermore in the case of $\Sigma_1=\Sigma$ *and* $\Sigma_2=\alpha^2\Sigma$ (α *is a positive scalar*), *we have*

$$\inf_{\phi^L}\sup_F\max_{i=1,2} e_i(\phi^L, F)=\left[1+\frac{1}{(1+\alpha)^2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\right]^{-1}, \tag{2.14}$$

and the $\inf_{\phi^L}$ *on the left side of* (2.14) *is attained only by the linear procedure that classifies an observation* x *into* C_1 *if*

$$x'\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{1+\alpha}(\alpha\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\geqq 0 \tag{2.15}$$

and into C_2 *otherwise.*

PROOF. Setting $\sup_F e_1(\phi^{d,\theta}, F)=\sup_F e_2(\phi^{d,\theta}, F)$ in Theorem 2.1 and solving for θ restricting d to the vectors satisfying $d'\mu_1<d'\mu_2$, we have

$$\theta=(d'\mu_1\sqrt{d'\Sigma_2 d}+d'\mu_2\sqrt{d'\Sigma_1 d})/(\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d})\ . \tag{2.16}$$

Since $\sup_F \max_{i=1,2} e_i(\phi^{d,\theta}, F)=\max_{i=1,2}\sup_F e_i(\phi^{d,\theta}, F)$, we have

$$\inf_\theta \sup_F \max_{i=1,2} e_i(\phi^{d,\theta}, F)=\left[1+\left(\frac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d}}\right)^2\right]^{-1}. \tag{2.17}$$

Therefore (2.13) follows from the next equality.

$$\inf_{\phi^L} \sup_F \max_{i=1,2} e_i(\phi^L, F)=\inf_{d\neq 0}\left[1+\left(\frac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d}}\right)^2\right]^{-1}. \tag{2.18}$$

If $\Sigma_1=\Sigma$ and $\Sigma_2=\alpha^2\Sigma$, (2.18) becomes

$$\inf_{\phi^L} \sup_F \max_{i=1,2} e_i(\phi^L, F)=\inf_{d\neq 0}\left[1+\frac{1}{(1+\alpha)^2}\left(\frac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma d}}\right)^2\right]^{-1}. \tag{2.19}$$

Obviously the right side of (2.19) is minimized when $\{d'(\mu_1-\mu_2)\}^2/d'\Sigma d$ is maximized. Since $\max_{d\neq 0}[\{d'(\mu_1-\mu_2)\}^2/d'\Sigma d]=(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$ and the maximum is attained only by the vectors proportional to the vector $\Sigma^{-1}(\mu_1-\mu_2)$, we have (2.14). Using the fact that $x'\Sigma^{-1}x+y'\Sigma^{-1}y>x'\Sigma^{-1}y$ for $x'x+y'y>0$, we see that the $\inf_{\phi^L}$ on the left side of (2.14) is attained by $d=\beta\Sigma^{-1}(\mu_2-\mu_1)$, where β is a positive scalar chosen so that $d'd=1$. Hence we have, from (2.16), $\theta=(\beta/(1+\alpha))(\alpha\mu_1+\mu_2)'\Sigma^{-1}(\mu_2-\mu_1)$. Since an observation x is classified into C_1 if and only if $d'x-\theta\leqq 0$, we have thus obtained the linear procedure based on the linear discriminant function (2.15). q.e.d.

Remark. When the distributions of both of the two categories are multivariate normal, the linear procedure determined by some vector d^*, which seems difficult to represent concretely, that attains the $\inf_{d\neq 0}$ of the right side of (2.18), and the corresponding

$$\theta^*=(d^{*\prime}\mu_1\sqrt{d^{*\prime}\Sigma_2 d^*}+d^{*\prime}\mu_2\sqrt{d^{*\prime}\Sigma_1 d^*})/(\sqrt{d^{*\prime}\Sigma_1 d^*}+\sqrt{d^{*\prime}\Sigma_2 d^*}) \tag{2.20}$$

is the minimax procedure in Φ^L (Anderson and Bahadur [2], pp. 427–428).

In the case of $\Sigma_1=\Sigma$ and $\Sigma_2=\alpha^2\Sigma$, the above procedure becomes the linear procedure based on the linear discriminant function (2.15), because we obtain $\theta=(\alpha d'\mu_1+d'\mu_2)/(1+\alpha)$ by solving for θ the equation

$$\begin{aligned}(2.21)\qquad &\int_\theta^\infty \frac{1}{\sqrt{d'\Sigma d}\sqrt{2\pi}}\exp\left(-\frac{(z-d'\mu_1)^2}{2d'\Sigma d}\right)dz\\ &=\int_{-\infty}^\theta \frac{1}{\alpha\sqrt{d'\Sigma d}\sqrt{2\pi}}\exp\left(-\frac{(z-d'\mu_2)^2}{2\alpha^2 d'\Sigma d}\right)dz\end{aligned}$$

with d fixed arbitrarily satisfying $d'\mu_1<d'\mu_2$. Hence the common value, the common conditional error probability, of (2.21) becomes

$$(2.22)\qquad \int_\theta^\infty \frac{1}{\sqrt{d'\Sigma d}\sqrt{2\pi}}\exp\left(-\frac{(z-d'\mu_1)^2}{2d'\Sigma d}\right)dz=\Phi^*\left[\frac{1}{1+\alpha}\frac{d'(\mu_2-\mu_1)}{\sqrt{d'\Sigma d}}\right],$$

whose right side attains its minimum $\Phi^*[(1/(1+\alpha))\sqrt{(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)}]$ when $d=\beta\Sigma^{-1}(\mu_2-\mu_1)$ (β is any positive scalar). Furthermore if $\alpha=1$, then the above procedure reduces to the well known linear procedure that is also Bayes procedure (out of all procedures Φ) with respect to equal apriori probabilities (Anderson [1], Chap. 6).

This theorem seems to justify that we may adopt Δ^2 as the distance between two distributions without any parametrized form and, with not necessarily the same covariance matrices.

Concerning with the expected error probability with respect to the given apriori probabilities Π, we have the following theorem.

THEOREM 2.3.

$$\begin{aligned}(2.23)\quad &\inf_{\phi^L}\sup_F e_\Pi(\phi^L, F)\\ &=\min\left[\min(\Pi_1,\Pi_2),\ \inf_d \inf_{d'\mu_1<\theta<d'\mu_2}\left\{\Pi_1\frac{d'\Sigma_1 d}{d'\Sigma_1 d+(\theta-d'\mu_1)^2}\right.\right.\\ &\qquad\left.\left.+\Pi_2\frac{d'\Sigma_2 d}{d'\Sigma_2 d+(d'\mu_2-\theta)}\right\}\right].\end{aligned}$$

PROOF. Since F_1 and F_2 may vary independently in the families of distributions $\mathcal{F}(\mu_1, \Sigma_1)$ and $\mathcal{F}(\mu_2, \Sigma_2)$, respectively, we have

$$(2.24)\qquad \sup_F e_\Pi(\phi^L, F)=\Pi_1\sup_F e_1(\phi^L, F)+\Pi_2\sup_F e_2(\phi^L, F).$$

Hence the result follows immediately from Theorem 2.1. q.e.d.

3. Miscellanea

We consider in this section some special cases of interest. In the case of the same apriori probabilities Π^0, (2.23) of Theorem 2.3 reduces to

$$\inf_{\phi^L}\sup_F e_{\Pi^0}(\phi^L, F) = \min\left[\frac{1}{2}, \inf_d \inf_{d'\mu_1<\theta<d'\mu_2}\left\{\frac{1}{2}\frac{d'\Sigma_1 d}{d'\Sigma_1 d+(\theta-d'\mu_1)^2} + \frac{1}{2}\frac{d'\Sigma_2 d}{d'\Sigma_2 d+(d'\mu_2-\theta)^2}\right\}\right]. \tag{3.1}$$

The above expression, however, seems difficult to be calculated in general. So we will calculate it under the additional assumption that $\Sigma_1=\Sigma_2$ $(=\Sigma)$ in the next theorem.

THEOREM 3.1. *In the case of* $\Sigma_1=\Sigma_2$ $(=\Sigma)$, *we have*

$$\inf_{\phi^L}\sup_F e_{\Pi^0}(\phi^L, F)=\begin{cases}\left[1+\dfrac{\Delta^2}{4}\right]^{-1}, & \text{when } \Delta>2\\[2ex] \dfrac{1}{2}, & \text{when } \Delta\leqq 2,\end{cases} \tag{3.2}$$

where $\Delta^2=(\mu_1-\mu_2)\Sigma^{-1}(\mu_1-\mu_2)$ *(Mahalanobis distance). Furthermore if* $\Delta>2$, *the* $\inf_{\phi^L}$ *on the left side of* (3.2) *is attained only by well known linear procedure that classifies an observation* x *into* C_1 *if* $x'\Sigma^{-1}(\mu_1-\mu_2)-(1/2)(\mu_1+\mu_2)\Sigma^{-1}(\mu_1-\mu_2)\geqq 0$ *and into* C_2 *otherwise.*

PROOF. Let us put

$$\Psi^d(\theta)\equiv\frac{1}{2}\frac{d'\Sigma d}{d'\Sigma d+(\theta-d'\mu_1)}+\frac{1}{2}\frac{d'\Sigma d}{d'\Sigma d+(d'\mu_2-\theta)}, \qquad d'\mu_1<\theta<d'\mu_2. \tag{3.3}$$

By differentiation we have, putting $\delta^2=1/d'\Sigma d$ for simplicity,

$$\frac{d}{d\theta}\Psi^d(\theta)=-2\delta^2\left\{\left(\theta-\frac{d'\mu_1+d'\mu_2}{2}\right)\left(\theta-\frac{d'\mu_1+d'\mu_2}{2}+\varepsilon\right)\cdot\left(\theta-\frac{d'\mu_1+d'\mu_2}{2}-\varepsilon\right)\right\}\Big/ \{1+\delta^2(\theta-d'\mu_1)^2\}^2\{1+\delta^2(d'\mu_2-\theta)^2\}^2, \tag{3.4}$$

where ε is some constant satisfying $0<\varepsilon<(d'\mu_2-d'\mu_1)/2$. Thus we have

$$\min_{d'\mu_1\leqq\theta\leqq d'\mu_2}\Psi^d(\theta) = \begin{cases}\left[1+\dfrac{1}{4}\left(\dfrac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma d}}\right)^2\right]^{-1}, & \text{when } \left|\dfrac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma d}}\right|>\sqrt{2}\\[2ex] \dfrac{1}{2}\left\{1+\left[1+\left(\dfrac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma d}}\right)^2\right]^{-1}\right\}, & \text{when } \left|\dfrac{d'(\mu_1-\mu_2)}{\sqrt{d'\Sigma d}}\right|\leqq\sqrt{2},\end{cases} \tag{3.5}$$

furthermore in the former case in (3.5) $\Psi^d(\theta)$ has its minimum only at

$\theta=(d'\mu_1+d'\mu_2)/2$, and in the latter case at $\theta=d'\mu_1$ or $\theta=d'\mu_2$.

Obviously both of the right sides of (3.5) are minimized when $\{d'(\mu_1-\mu_2)\}^2/d'\Sigma d$ is maximized. Since $\max_{d\neq 0}[\{d'(\mu_1-\mu_2)\}^2]/d'\Sigma d=(\mu_1-\mu_2)'\cdot\Sigma^{-1}(\mu_1-\mu_2)=\Delta^2$, and the maximum is attained only by the vectors proportional to the vector $\Sigma^{-1}(\mu_1-\mu_2)$, we have

$$\inf_{d}\inf_{d'\mu_1<\theta<d'\mu_2}\Psi^d(\theta)=\begin{cases}\left[1+\dfrac{\Delta^2}{4}\right]^{-1}, & \text{when } \Delta>\sqrt{2}\\[2ex] \dfrac{1}{2}\cdot\dfrac{2+\Delta^2}{1+\Delta^2}, & \text{when } \Delta\leqq\sqrt{2}\,.\end{cases} \tag{3.6}$$

Hence (3.2) follows from (3.1), (3.3) and (3.6). The proof of the latter assertion of the theorem is similar to that of Theorem 2.2. q.e.d.

COROLLARY 3.1. *In the case $\Sigma_1=\Sigma_2$ $(=\Sigma)$, if n observations $x_1, x_2, \cdots, x_n$ are taken independently from one of the two categories, we have*

$$\inf_{\phi^L}\sup_{F} e_{\Pi^0}(\phi^L, F)=\begin{cases}\left[1+n\dfrac{\Delta^2}{4}\right]^{-1}, & \text{when } \Delta>\dfrac{2}{\sqrt{n}}\\[2ex] \dfrac{1}{2}, & \text{when } \Delta\leqq\dfrac{2}{\sqrt{n}}\,.\end{cases} \tag{3.7}$$

Furthermore if $\Delta>2/\sqrt{n}$, the $\inf_{\phi^L}$ on the left side of (3.7) is attained only by the linear procedure that classifies n observations $x_1, x_2, \cdots, x_n$ into C_1 if $\bar{x}'\Sigma^{-1}(\mu_1-\mu_2)-(1/2)(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\geqq 0$ and into C_2 otherwise, where $\bar{x}$ denotes the mean of the observations $x_1, x_2, \cdots, x_n$.

PROOF. If we regard $\begin{bmatrix}\mu_i\\ \mu_i\\ \vdots\\ \mu_i\end{bmatrix}_n$ $(i=1, 2)$ and $\begin{bmatrix}\overbrace{\Sigma \quad\quad 0}^{n}\\ \quad\Sigma\quad \\ \quad\quad\ddots\\ 0 \quad\quad \Sigma\end{bmatrix}_n$ as μ_i $(i=1, 2)$ and Σ in Theorem 3.1, respectively, the proof is the same as that of Theorem 3.1. q.e.d.

Remark. The condition $\Delta>2$ in Theorem 3.1 may be understood geometrically as follows. Let k be the dimentionality of the observation, then the condition $\Delta>2$ is equivalent to

$$\left(\frac{\sqrt{k+2}}{2}\mu_1-\frac{\sqrt{k+2}}{2}\mu_2\right)'\Sigma^{-1}\left(\frac{\sqrt{k+2}}{2}\mu_1-\frac{\sqrt{k+2}}{2}\mu_2\right)>k+2\,. \tag{3.8}$$

Thus the condition $\Delta>2$ means that the ellipsoid of concentration (Cramér [4], p. 300) of $\mathcal{F}((\sqrt{k+2}/2)\mu_i, \Sigma)$ does not contain the vector $(\sqrt{k+2}/2)\mu_{3-i}$ $(i=1, 2)$. The case of $k=2$ is of interest, because the vector $(\sqrt{k+2}/2)\mu_i$ reduces to the mean vector of C_i.

THEOREM 3.2. *With respect to any apriori probabilities Π, we have*

$$\inf_{\phi^L} \sup_F e_\Pi(\phi^L, F) \leqq \left[1+\frac{\Delta^2}{4}\right]^{-1}, \tag{3.9}$$

where $\Delta^2 = 4\cdot\max_{d\neq 0}\{d'(\mu_1-\mu_2)/(\sqrt{d'\Sigma_1 d}+\sqrt{d'\Sigma_2 d})\}^2$. *In the case of* $\Sigma_2=\alpha^2\Sigma_1$ *(α is a positive scalar) and Π^0, the inequality* (3.9) *becomes equality if and only if $\alpha=1$ (i.e., $\Sigma_1=\Sigma_2$) and $\Delta>2$.*

PROOF. For any $\phi^L \in \Phi^L$ and any $F \in \mathcal{F}$, we obviously have

$$e_\Pi(\phi^L, F) \leqq \max_{i=1,2} e_i(\phi^L, F) \tag{3.10}$$

for any apriori probabilities Π. Hence (3.9) follows immediately from (3.10) and (2.13) of Theorem 2.2. In the case of $\Sigma_2=\alpha^2\Sigma_1$ (put $\Sigma_1=\Sigma$), we know, from Theorem 2.2, that the $\inf_{\phi^L}$ on the left side of (2.19) is attained when $d^*=\Sigma^{-1}(\mu_2-\mu_1)$ and $\theta^*=(\alpha d^{*\prime}\mu_1+d^{*\prime}\mu_2)/(1+\alpha)$. Let us put similarly to (3.3)

$$\Psi^d(\theta) \equiv \frac{1}{2}\,\frac{d'\Sigma d}{d'\Sigma d+(\theta-d'\mu_1)^2}+\frac{1}{2}\,\frac{\alpha^2 d'\Sigma d}{\alpha^2 d'\Sigma d+(d'\mu_2-\theta)^2}, \quad d'\mu_1<\theta<d'\mu_2\,. \tag{3.11}$$

Now for any $d\neq 0$ we easily obtain by differentiation

$$\frac{d}{d\theta}\Psi^d(\theta)\Big|_{\theta=(\alpha d'\mu_1+d'\mu_2)/(1+\alpha)}=0 \qquad \text{if and only if } \alpha=1\,.$$

Applying this fact to the case of $d^*=\Sigma^{-1}(\mu_2-\mu_1)$, we have the "only if part" of the latter asstertion of the theorem because of $\Psi^{d^*}(\theta^*)=[1+\Delta^2/4]^{-1}$. The "if part" is merely a part of Theorem 3.1. q.e.d.

Remarks. (i) From the theorem of Isii and Taga ([6], Th. 1), we know that

$$\inf_{\phi} \sup_F e_{\Pi^0}(\phi, F) = \frac{1}{2}\left[1+\frac{\Delta^2}{4}\right]^{-1}. \tag{3.12}$$

Thus we have the following inequality from (3.9) and (3.12)

$$\inf_{\phi^L} \sup_F e_{\Pi^0}(\phi^L, F) \leqq 2 \inf_{\phi} \sup_F e_{\Pi^0}(\phi, F)\,, \tag{3.13}$$

where the equality holds when $\Sigma_1=\Sigma_2$ and $\Delta>2$. Moreover since $\inf_\phi \sup_F e_{\Pi^0}(\phi, F) \leqq \inf_\phi \sup_F \max_{i=1,2} e_i(\phi, F)$, and considering (2.13) we have

$$\inf_{\phi} \sup_{F} \max_{i=1,2} e_i(\phi, F) \leqq \inf_{\phi^L} \sup_{F} \max_{i=1,2} e_i(\phi^L, F) \leqq 2 \inf_{\phi} \sup_{F} \max_{i=1,2} e_i(\phi, F) . \tag{3.14}$$

(ii) Considering (2.13) and (3.9) in the case of Π^0, it follows that

$$\inf_{\phi^L} \sup_{F} e_{\Pi^0}(\phi^L, F) \leqq \inf_{\phi^L} \sup_{F} \max_{i=1,2} e_i(\phi^L, F) , \tag{3.15}$$

where the equality holds when $\Sigma_1 = \Sigma_2$ and $\Delta > 2$.

Acknowledgments

The author is greatly indebted to Professors K. Isii of Osaka Univ. and Y. Taga of Chiba Univ., under whose guidance this work has been done, for suggestions and criticism. Also, the author is grateful to Professor Y. Horibe of Shizuoka Univ. for his comments.

SHIZUOKA UNIVERSITY

REFERENCES

[1] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc.

[2] Anderson, T. W. and Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices, *Ann. Math. Statist.*, **33**, 420–431.

[3] Chernoff, H. (1971). A bound on the classification error for discriminating between populations with specified means and variances, *Studi Prob. Statist. Ric. Operat.* 203–211.

[4] Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton Univ. Press.

[5] Isii, K. (1964). Inequalities of the types of Chebyshev and Cramér-Rao and mathematical programming, *Ann. Inst. Statist. Math.*, **16**, 277–293.

[6] Isii, K. and Taga, Y. (1975). Mathematical programming approach to a minimax theorem of statistical discrimination applicable to pattern recognition, *Optimization Techniques IFIP Technical Conference, Novosibirsk July 1974*, Springer-Verlag, 348–352.

[7] Lachenbruch, P. A., Sneeringer, C. and Revo, L. T. (1972). Robustness of the linear and quadratic discriminant function to certain types of non-normality, *Commun. Statist.*, **1**, 39–56.

[8] Yau, S. S. and Lin, T. T. (1968). On the upper bound of the probability of error of a linear pattern classifier for probabilistic pattern classes, *Proc. IEEE.*, **56**, 321–322.