Reproduced by

CADO

CENTRAL AIR DOCUMENTS OFFICE

WRIGHT-PATTERSON AIR FORCE BASE - DAYTON, OHIO

REEL-C 4983

A.T.I 110634

"NOTICE: When Government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the U.S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto."

UNCLASSIFIED

DISCRIMINATORY ANALYSIS

PERFECT DISCRIMINATION AS THE NUMBER OF VARIABLES INCREASES

William R. Gaffey

PROJECT NUMBER 21-49-004
REPORT NUMBER 5*

* Prepared at the University of California
Under Contract No. 41(128)-31)

USAF SCHOOL OF AVIATION MEDICINE
RANDOLPH FIELD, TEXAS

FEBRUARY 1951

PERFECT DISCRIMINATION AS THE NUMBER OF VARIABLES INCREASES

In applied problems of classification it is generally true that, as the number of measured variables increases, the probability of misclassification becomes smaller.

In some cases this probability will never become zero. One feels intuitively, for example, that in the classical problem of classifying skulls, the probability of misclassifying will remain greater than zero, even though the number of measurements is increased indefinitely. On the other hand, in some problems of botanical classification into species, the separation between the species becomes so great after a few measurements that the probability of misclassification is nil.

Since the obtained measurements are random variables, the problem suggests itself of investigating the conditions on the distributions of these variables which will insure that, as their number is increased, their ability to discriminate becomes perfect.

Obviously, if there are two groups, π_1 and π_2, and the problem is to minimize the probability of misclassifying a member of π_1, a trivial solution is to place every object in π_1. This will not be considered perfect discrimination for π_1, since the total probability of misclassification is not

minimized.

If $x = (x_1, x_2, \cdots, x_s)$ has the density $f_s^{(i)}(x)$ in group π_i with proportion a_i, $i = 1, 2, \cdots, p$, then the partition of the random variable space which minimizes the total probability of misclassification is, for π_i,

$$R_i = \left[x \middle| a_i f_s^{(i)}(x) > a_j f_s^{(j)}(x)\right], \quad j \neq i.$$

The probability of misclassifying a member of π_i in group π_j is then

$$\int_{R_j} f_s^{(i)}(x)\, dx.$$

Therefore

$$\int_{R_j} f_s^{(i)}(x)dx = \int_{\left[x \middle| \log \frac{f_s^{(i)}(x)}{f_s^{(j)}(x)} < \log \frac{a_j}{a_i}\right]} f_s^{(i)}(x)dx = P\left[t < \log \frac{a_j}{a_i} \middle| x \in T_{ij}^{(s)}\right],$$

where

$$t = \begin{cases} \log \dfrac{f_s^{(i)}(x)}{f_s^{(j)}(x)} & \text{if } f_s^{(i)} > 0 \text{ and } f_s^{(j)} > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$T_{ij}^{(s)} = \left[x \middle| f_s^{(i)} > 0 \text{ and } f_s^{(j)} > 0\right].$$

By the Tchebycheff inequality, assuming that the expectations exist,

$$P\left[t < \log \frac{a_i}{a_j} \mid x \epsilon T_{ij}^{(s)}\right] \leq \frac{E(t|x\epsilon T_{ij}^{(s)})^2 - \left[E(t|x\epsilon T_{ij}^{(s)})\right]^2}{\left[E(t|x\epsilon T_{ij}^{(s)}) + \log \frac{a_j}{a_i}\right]^2}.$$

The right side of the inequality approaches zero as $s \longrightarrow \infty$ if and only if

$$I_{ij}^{(s)} = \frac{\left[E(t|x\epsilon T_{ij}^{(s)})\right]^2}{E(t|x\epsilon T_{ij}^{(s)})^2} = \frac{\left[\int_{T_{ij}^{(s)}} \log \frac{f_s^{(i)}(x)}{f_s^{(j)}(x)} \, f_s^{(i)}(x)dx\right]^2}{P\left[x\epsilon T_{ij}^{(s)}\right] \int_{m_{ij}^{(s)}} \left(\log \frac{f_s^{(i)}(x)}{f_s^{(j)}(x)}\right)^2 f_s^{(i)}(x)dx} \longrightarrow 1.$$

Therefore, if $I_{ij}^{(s)} \longrightarrow 1$ as $s \longrightarrow \infty$, the probability of misclassifying a member of π_i into π_j approaches zero. If $I_{ij}^{(s)} \longrightarrow 1$ for all $j \neq i$, then we have asymptotically perfect discrimination.

Example: Multivariate normal distributions. In the special case where $x = (x_1, x_2, \cdots, x_s)$ has multivariate normal density with equal covariance matrices for each π_i, and expectations $\xi^{(i)} = (\xi_1^{(i)}, \xi_2^{(i)}, \cdots, \xi_s^{(i)})$ in π_i, the regions $T_{ij}^{(s)}$ include all values of x, and the requirement that $I_{ij}^{(s)}$ should approach unity is equivalent to requiring that the series

$$\sum_{k=1}^{s} \sum_{m=1}^{s} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})(\xi_m^{(i)} - \xi_m^{(j)})}{\sigma_k \quad \sigma_m}$$

should diverge, where r^{km} is the element of the reciprocal of the correlation matrix R.

This divergence is also a necessary condition, for now

$$R_j = \left[x \,\middle|\, \sum_{m=1}^{s} \frac{x_m}{\sigma_m} \left[\sum_{k=1}^{s} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})}{\sigma_k} \right] < \log \frac{a_j}{a_i} \right.$$

$$\left. + \frac{1}{2} \sum_{k=1}^{s} \sum_{m=1}^{s} r^{km} \frac{(\xi_k^{(i)} \xi_m^{(i)} - \xi_k^{(j)} \xi_m^{(j)})}{\sigma_k \sigma_m} \right].$$

Making the transformation $x_k = \sigma_k y_k + \xi_k^{(i)}$, $k = 1, 2, \cdots, s$

and then the transformation $y_k = \sum_{j=1}^{s} a_{jk} z_j$, where $\|a_{jk}\| = A$

$R = A'A$, R_j becomes

$$T(R_j) = \left[z \,\middle|\, \sum_{j=1}^{s} z_j \sum_{k=1}^{s} \sum_{m=1}^{s} a_{jm} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})}{\sigma_k} < \log \frac{a_j}{a_i} \right.$$

$$\left. - \frac{1}{2} \sum_{k=1}^{s} \sum_{m=1}^{s} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})(\xi_m^{(i)} - \xi_m^{(j)})}{\sigma_k \sigma_m} \right]$$

and

$$\int_{R_j} f_s^{(i)}(x)\,dx = \int_{R_j} \frac{1}{(\sqrt{2\pi})^s \prod_{k=1}^{s} \sigma_k \sqrt{|R|}} \cdot e^{-\frac{1}{2} \sum_k \sum_m r^{km} \frac{(x_k - \xi_k^{(i)})(x_m - \xi_m^{(i)})}{\sigma_k \sigma_m}} dx_1 \cdots dx_s$$

$$= \frac{1}{(\sqrt{2\pi})^s} \int_{T(R_j)} e^{-\frac{1}{2}\sum_{j=1}^{s} z^2} \, dz_1 \cdots dz_s$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{P} e^{-\frac{x^2}{2}} \, dx,$$

where

$$P = \frac{\log \frac{a_j}{a_i}}{\sqrt{\sum_{k=1}^{s} \sum_{m=1}^{s} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})(\xi_m^{(i)} - \xi_m^{(j)})}{\sigma_k \sigma_m}}}$$

$$= -\sqrt{\sum_{k=1}^{s} \sum_{m=1}^{s} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})(\xi_m^{(i)} - \xi_m^{(j)})}{\sigma_k \sigma_m}}.$$

If the double series does not diverge, the above expression will not approach $-\infty$, and the integral will not approach zero.

A sufficient condition for divergence is that, for $s \geq S$ where S is a finite integer,

$$\text{g. l. b.}_{\substack{1 \leq k \leq s \\ 1 \leq m \leq s}} \left| r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})(\xi_m^{(i)} - \xi_m^{(j)})}{\sigma_k \sigma_m} \right| \geq \epsilon > 0.$$

This condition is satisfied if $x_1, \cdots, x_s$ are independent and if $\left| \frac{\xi_k^{(i)} - \xi_k^{(j)}}{\sigma_k} \right| > \delta > 0$ for all k.

If the variables are not independent, but the lower bound δ for $\left| \frac{\xi_k^{(i)} - \xi_k^{(j)}}{\sigma_k} \right|$ exists, then the sufficient condition becomes, for $s \geq S$,

$$\underset{\substack{1 \leq k \leq s \\ 1 \leq m \leq s}}{\text{g. l. b.}} \left| r^{km} \right| \geq \varepsilon > 0.$$

Since $r^{km} = \frac{|R_{km}|}{|R|}$ where $|R|$ is the determinant of the correlation matrix and $|R_{km}|$ is the determinant of the $(s-1)\times(s-1)$ minor of r_{km}, the sufficient conditions become (1) that the lower bound k exists for $\left| \frac{\xi_k^{(i)} - \xi_k^{(j)}}{\sigma_k} \right|$ and (2) that for all k and m and all $s \geq S$, $|R_{km}| > \delta |R|$ for some $\delta > 0$.

Notice that if $r_{km} = \pm 1$ for some k and m, then the determinant $|R|$ acquires two identical rows or two identical columns and equals zero. Therefore, the $s \times s$ reciprocal matrix is not defined. An $(s-1)\times(s-1)$ reciprocal may be obtained for the reduced matrix R, without the variable with the perfect correlation. In other words, if one new variable correlates perfectly with one of the previous ones, the experimenter has really looked at one of the previous vari-

ables twice, instead of adding a new one, and the series

$$\sum_{k=1}^{s} \sum_{m=1}^{s} r^{km} \frac{(\xi_k^{(i)} - \xi_k^{(j)})(\xi_m^{(i)} - \xi_m^{(j)})}{\sigma_k \sigma_m}$$

acquires no new terms.

(UNPUBLISHED CARD) UNCLASSIFIED

ATI 110 634 (COPIES OBTAINABLE FROM CADO)

UNIVERSITY OF CALIFORNIA, BERKELEY

DISCRIMINATORY ANALYSIS - PERFECT DISCRIMINATION AS THE NUMBER OF VARIABLES INCREASES - PROJECT REPORT

WILLIAM R. GAFFEY FEB 51 7PP

USAF SCHOOL OF AVIATION MEDICINE, RANDOLPH AIR FORCE BASE, TEXAS. (REPORT NO. 5)

STATISTICAL ANALYSIS SCIENCES, GENERAL (33)
MATHEMATICS (3)

UNCLASSIFIED