AN ABSTRACT OF THE THESIS OF

David A. Cade for the Master of Arts

in Mathematics presented on July 11, 1978

Title: Multivariate Classification Procedures

Abstract approved: Marion P. Emerson

This thesis presents statistical theory relating to classification. A definition of the multinormal distribution is included so that other theory relative to multinormal populations may be discussed. A major portion of the thesis is devoted to the problem of classifying a random observation into one of several populations.

MULTIVARIATE CLASSIFICATION PROCEDURES

A Thesis

Presented to

the faculty of the Department of Mathematics

Emporia State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

David A. Cade

July 1978

John M. Burger
Approved for the Major Department

Harold E Durst
Approved for the Graduate Council

Table of Contents

Chapter Page

CHAPTER I

INTRODUCTION

This thesis presents statistical theory relating to the problem of assigning an individual into one of several given populations: classification. An individual is defined here as a random observation from one of the several populations. The populations under consideration have multivariate distributions. The reader is expected to have had an introductory course in mathematical statistics. Also, an understanding of matrix properties and minimax actions would be helpful.

After the establishment of several preliminary terms, the multinormal (multivariate normal) distribution will be defined. The second major topic will be the classification of an observation into one of two populations with known joint distribution functions. First, a Bayes procedure - requiring the knowledge of a priori probabilities - will be presented and then the minimax procedure for two populations with continuous distributions. Sometimes the classification problem suggests the existence of a third population. This aspect is considered when a test for a hypothesis concerning multinormal populations is defined. Finally, the Bayes classification procedure for several arbitrary populations is presented. Specific populations under discussion will be multinormal populations with common covariance matrix and multivariate discrete populations where a random variable has a Bernoulli distribution. Estimation is employed at various points throughout the thesis.

CHAPTER II

MULTINORMAL DISTRIBUTION

The following symbols are defined: $\underline{G}$ and $\underline{g}$ represent matrices; $\underline{G}'$ and $\underline{g}'$ represent the transposes of $\underline{G}$ and $\underline{g}$, respectively; and X_i is a random variable with a realization x_i, $i = 1,\ldots,m$. One assumes that X_i is normally distributed with mean μ_i and variance $\sigma_i^2 < \infty$,i.e., X_i is $N(\mu_i,\sigma_i^2)$ for $i = 1,\ldots,m$. The covariance of X_i and X_j is defined by

$$\begin{aligned}\text{cov}(X_i,X_j) &= \sigma_{ij}\\ &= E[(X_i - \mu_i)(X_j - \mu_j)]\\ &= E(X_iX_j) - E(X_i)E(X_j)\end{aligned}$$

for $i,j = 1,\ldots,m$. One may notice that $\sigma_{ij} = \sigma_{ji}$. When $i = j$, $\sigma_{ij} = \sigma_{ii} = \sigma_i^2$. The following matrices are defined: $\underline{X} = (X_1,X_2,\ldots,X_m)'$ is a random vector; $\underline{x} = (x_1,x_2,\ldots,x_m)'$ is an observation vector; $\underline{\mu} = (\mu_1,\mu_2,\ldots,\mu_m)'$ is the mean vector for $\underline{X}$; and $\underline{\Sigma} = [\sigma_{ij}]$ is the covariance matrix of order m.

The multinormal distribution of $\underline{X}$ will now be defined. This multivariate distribution has the joint distribution function

$$F(\underline{x}) = P(\underline{X} \leq \underline{x}).$$

The following statements are presented here without proof:

$E(X_i) = \mu_i$, $i = 1,\ldots,m$, so that $E(\underline{X}) = \underline{\mu}$; [4,p.349]

$\text{var}(X_i) = \sigma_{ii}$ and $\text{cov}(X_i,X_j) = \sigma_{ij}$, $j = 1,\ldots,m$, so that $\underline{\Sigma}$ is the matrix of second-order central moments and second-order product moments; [4,p.349] if $\underline{\Sigma}$ is invertible, then $\underline{X}$ has the joint probability density function $f(\underline{x}) = (2\pi)^{-m/2}(\det \underline{\Sigma})^{-1/2}\exp[-\frac{1}{2}(\underline{x} - \underline{\mu})'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu})]$. [5,p.472]

The symbol $N(\underline{\mu},\underline{\Sigma})$ will denote the multinormal distribution of $\underline{X}$.

CHAPTER III

CLASSIFICATION INTO ONE OF TWO POPULATIONS WITH KNOWN JOINT DISTRIBUTION FUNCTIONS

A. KNOWN A PRIORI PROBABILITIES

The classification of an observation into one of two arbitrary populations with known joint probability density functions and a priori probabilities will now be considered. A Bayes procedure will be presented.

One lets π_1 and π_2 represent two populations with density functions $f_1(\underline{x})$ and $f_2(\underline{x})$, respectively. One defines the prior probability that an observation comes from π_1 as q_1 and from π_2 as q_2. One assumes that $q_1 + q_2 = 1$.

An observation $\underline{x} = (x_1,x_2,\ldots,x_m)'$ from either $\pi_1$or π_2 may be considered as a point in some m-dimensional sample space R. "We divide this space into two regions. If the observation falls in R_1, we classify it as coming from population π_1, and if it falls in R_2 we classify it as coming from population π_2." [2,p.127] Here are four probabilities associated with classification:

$P(1|1) = \int_{R_1} f_1(\underline{x})d\underline{x}$, where $d\underline{x} = dx_1\ldots dx_m$, is the probability of correctly classifying an observation from π_1;

$P(2|1) = \int_{R_2} f_1(\underline{x})d\underline{x}$ is the probability of misclassifying an observation from π_1 into π_2;

$P(2|2) = \int_{R_2} f_2(\underline{x})d\underline{x}$ is the probability of correctly classifying an observation from π_2;

and $P(1|2) = \int_{R_1} f_2 f(\underline{x})d\underline{x}$ is the probability of misclassifying an observation from π_2 into π_1.

One lets $C(2|1)$ be the cost of misclassifying an observation from π_1 into π_2 and the cost of misclassifying an observation from π_2 into π_1 be $C(1|2)$. The probability that an observation is drawn from π_1 and misclassified is $q_1P(2|1)$. Similarly, the probability that an observation is drawn from π_2 and misclassified is $q_2P(1|2)$. The expected cost of misclassification $C(2|1)q_1P(2|1) + C(1|2)q_2P(1|2)$ is to be minimized. [2,pp.128-131] When the rule of choosing the regions for classification is

$$R_1: \quad \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{C(1|2)q_2}{C(2|1)q_1},$$

$$R_2: \quad \frac{f_1(\underline{x})}{f_2(\underline{x})} < \frac{C(1|2)q_2}{C(2|1)q_1},$$

the expected cost is minimized. [2,p.131] If $P[f_1(\underline{x})/f_2(\underline{x}) = C(1|2)q_2/C(2|1)q_1|\pi_i] = 0$, $i = 1,2$, and if sets of probability zero are excluded from consideration, then this Bayes procedure is unique for continuous distributions. [2,p.131]

When $q_1 = q_2 = \frac{1}{2}$ and $C(1|2) = C(2|1)$, or equality may be assumed, the likelihood-ratio rule can be used for classification. [6,p.234] π_1 is the more likely population for an observation if the density function $f_1(\underline{x})$ is larger than $f_2(\underline{x})$. Similarly, π_2 is more likely when $f_2(\underline{x})$ is larger. This rule states that $\underline{x}$ is classified into π_1 if $\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq 1$; otherwise classify $\underline{x}$ as from π_2. [6,p.234] One may notice that this procedure is a specific case of the Bayes procedure.

If discrete distributions are encountered in the two populations, the joint probability function $f(\underline{x}) = P(\underline{X} = \underline{x})$ should be substituted for the density in the previously mentioned rules of classification. [3,p.153] The probability of misclassifying an observation from π_i into π_j is $P(j|i) = \Sigma_{R_j} f_i(\underline{x})$, $i,j = 1,2$, $i \neq j$.

B. UNKNOWN A PRIORI PROBABILITIES FOR THE CASE OF POPULATIONS WITH CONTINUOUS DISTRIBUTIONS

One may assume now that an observation $\underline{x} = (x_1, x_2, \ldots, x_m)'$ is to be classified into one of two populations with known density functions, the costs of misclassification are known, and the prior probabilities are unknown.

"A principle that usually leads to a unique procedure is the minimax principle. A procedure is minimax if the maximum expected loss is a minimum. From a conservative point of view, this may be considered an optimum procedure." [2,p.129] "Minimax actions are optimal in the theory of two-person, zero-sum games, in the sense that by using such a strategy, one can be assured of no more loss than the minimum maximum - no matter what the other player (nature, in decision problems) does." [5,p.373]

The expected loss if an observation is incorrectly classified into π_1 is

$$C(1|2)P(1|2), \text{ and } C(2|1)P(2|1)$$

is the expected loss for misclassification into π_2. The minimax procedure is applied to these expected losses, i.e.,

$$C(1|2)P(1|2) = C(2|1)P(2|1), \text{ or}$$

$$C(1|2)\int_{R_1} f_2(\underline{x})d\underline{x} = C(2|1)\int_{R_2} f_1(\underline{x})d\underline{x}.$$

To illustrate this procedure, it is assumed that π_1 is $N(\underline{\mu}_1, \underline{\Sigma})$ and π_2 is $N(\underline{\mu}_2, \underline{\Sigma})$. The ratio of densities is

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} = \frac{(\det \underline{\Sigma})^{1/2} \exp[-\frac{1}{2}(\underline{x} - \underline{\mu}_1)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_1)]}{(\det \underline{\Sigma})^{1/2} \exp[-\frac{1}{2}(\underline{x} - \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_2)]}$$

$$= \exp\{-\frac{1}{2}[(\underline{x} - \underline{\mu}_1)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_1) - (\underline{x} - \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_2)]\}.$$

For some constant c, $\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c$ or $\frac{f_1(\underline{x})}{f_2(\underline{x})} < c$. Taking the natural logarithm of the first of these inequalities,

$$-\frac{1}{2}[(\underline{x} - \underline{\mu}_1)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_1) - (\underline{x} - \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_2)] \geq \log c.$$

The left side of this inequality may be written as

$$\begin{aligned}
&-\frac{1}{2}[\underline{x}'\underline{\Sigma}^{-1}\underline{x} - \underline{x}'\underline{\Sigma}^{-1}\underline{\mu}_1 - \underline{\mu}_1'\underline{\Sigma}^{-1}\underline{x} + \underline{\mu}_1'\underline{\Sigma}^{-1}\underline{\mu}_1 \\
&- \underline{x}'\underline{\Sigma}^{-1}\underline{x} + \underline{x}'\underline{\Sigma}^{-1}\underline{\mu}_2 + \underline{\mu}_2'\underline{\Sigma}^{-1}\underline{x} - \underline{\mu}_2'\underline{\Sigma}^{-1}\underline{\mu}_2] \\
&= \frac{1}{2}\underline{x}'\underline{\Sigma}^{-1}\underline{\mu}_1 + \frac{1}{2}\underline{\mu}_1'\underline{\Sigma}^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_1'\underline{\Sigma}^{-1}\underline{\mu}_1 - \frac{1}{2}\underline{x}'\underline{\Sigma}^{-1}\underline{\mu}_2 \\
&- \frac{1}{2}\underline{\mu}_2'\underline{\Sigma}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_2'\underline{\Sigma}^{-1}\underline{\mu}_2 \\
&= \underline{x}'\underline{\Sigma}^{-1}\underline{\mu}_1 - \frac{1}{2}\underline{\mu}_1'\underline{\Sigma}^{-1}\underline{\mu}_1 - \underline{x}'\underline{\Sigma}^{-1}\underline{\mu}_2 + \frac{1}{2}\underline{\mu}_2'\underline{\Sigma}^{-1}\underline{\mu}_2 \\
&+ \frac{1}{2}\underline{\mu}_1'\underline{\Sigma}^{-1}\underline{\mu}_2 - \frac{1}{2}\underline{\mu}_2'\underline{\Sigma}^{-1}\underline{\mu}_1 \\
&= \underline{x}'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2).
\end{aligned}$$

One defines a new random variable

$$U = \underline{X}'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

and the quantity

$$\alpha = (\underline{\mu}_1 - \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2).$$

When the distribution of $\underline{X}$ is $N(\underline{\mu}_1,\underline{\Sigma})$, the distribution of U is $N(\frac{1}{2}\alpha,\alpha)$. [2,pp.134-135] Also, U is $N(-\frac{1}{2}\alpha,\alpha)$ when $\underline{X}$ is $N(\underline{\mu}_2,\underline{\Sigma})$. [2,p.135]

The probability of misclassifying an observation $\underline{x}$ into π_2 is

$$\begin{aligned}
P(2|1) &= \int_{-\infty}^{\log c}(2\pi\alpha)^{-1/2}\exp[-\frac{1}{2}(u - \frac{1}{2}\alpha)^2/\alpha]du \\
&= P(U \leq \log c).
\end{aligned}$$

One defines a standard normal random variable $Z = \frac{U - \frac{1}{2}\alpha}{\alpha^{1/2}}$. Then,

$$\begin{aligned} P(2|1) &= P(Z\alpha^{1/2} + \tfrac{1}{2}\alpha \le \log c) \\ &= P[Z \le (\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}] \\ &= \Phi[(\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}], \end{aligned}$$

where $\Phi(z)$ is the distribution function for a standard normal distribution. Similarly,

$$\begin{aligned} P(1|2) &= \int_{\log c}^{\infty} (2\pi\alpha)^{-1/2}\exp[-\tfrac{1}{2}(u + \tfrac{1}{2}\alpha)^2/\alpha]du \\ &= 1 - P(U \le \log c) \end{aligned}$$

is the probability of misclassifying an observation into π_1. If $Z = \dfrac{U + \frac{1}{2}\alpha}{\alpha^{1/2}}$ is standard normal, then

$$\begin{aligned} P(1|2) &= 1 - P(Z\alpha^{1/2} - \tfrac{1}{2}\alpha \le \log c) \\ &= 1 - P[Z \le (\log c + \tfrac{1}{2}\alpha)/\alpha^{1/2}] \\ &= 1 - \Phi[(\log c + \tfrac{1}{2}\alpha)/\alpha^{1/2}] \\ &= \Phi[(-\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}], \end{aligned}$$

since $\Phi(z) + \Phi(-z) = 1$.

The constant log c is chosen so that

$$C(1|2)\Phi[(-\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}] = C(2|1)\Phi[(\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}].$$

The minimax procedure classifies $\underline{x}$ as from π_1 when

$$\underline{x}'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) - \tfrac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \ge \log c;$$

otherwise $\underline{x}$ is classified as from π_2. [2,p.136]

One notices that for known prior probabilities q_1 and q_2, the probabilities of misclassification for the Bayes procedure are

$$P(2|1) = \Phi[(\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}]$$

and

$$P(1|2) = \Phi[(-\log c - \tfrac{1}{2}\alpha)/\alpha^{1/2}],$$

where

$$c = \frac{C(1|2)q_2}{C(2|1)q_1}.$$

CHAPTER IV

A TEST FOR A HYPOTHESIS CONCERNING MULTINORMAL POPULATIONS

One supposes that $\underline{x} = (x_1, x_2, \ldots, x_m)'$ is an observation from $\pi_1 : N(\underline{\mu}_1, \underline{\Sigma})$, $\pi_2 : N(\underline{\mu}_2, \underline{\Sigma})$, or perhaps some unspecified population π_3. Whether or not π_3 exists may be unknown, but one assumes the problem of classification suggests the existence of a third population.

One lets $MP = \{\alpha\underline{\mu}_1 + \beta\underline{\mu}_2 | \alpha + \beta = 1\}$ be the set of all points in some m-dimensional space which lie on the line segment joining $\underline{\mu}_1$ and $\underline{\mu}_2$. Here is a test for the hypothesis that $\underline{x}$ comes from the family of multinormal distributions with mean vectors elements of MP and with common covariance matrix $\underline{\Sigma}$, i.e., $H_0 : \underline{x}$ comes from $N(\alpha\underline{\mu}_1 + \beta\underline{\mu}_2, \underline{\Sigma})$. [7,p.492] For this test, the test statistic

$$(\underline{x} - \underline{\mu}_1)'\underline{\Sigma}^{-1}(\underline{x} - \underline{\mu}_1) - \frac{[(\underline{x} - \underline{\mu}_1)'\underline{\Sigma}^{-1}(\underline{\mu}_2 - \underline{\mu}_1)]^2}{(\underline{\mu}_2 - \underline{\mu}_1)'\underline{\Sigma}^{-1}(\underline{\mu}_2 - \underline{\mu}_1)}$$

has a chi-square distribution with m - 1 degrees of freedom. [7,p.492] When the value of the test statistic is in or near a specified critical region, $\underline{x}$ possibly belongs to a population which is not $N(\alpha\underline{\mu}_1 + \beta\underline{\mu}_2, \underline{\Sigma})$.

If $\underline{\mu}_1$, $\underline{\mu}_2$, and $\underline{\Sigma}$ are unknown, they may be estimated by using a random sample $\underline{x}_{11}, \ldots, \underline{x}_{1n_1}$ from π_1 and a random sample $\underline{x}_{21}, \ldots, \underline{x}_{2n_2}$ from π_2. The sample sizes, n_1 and n_2, should be made as large as is feasibly possible. When the expected value of an estimator is equal to the parameter being estimated, that estimator is unbiased. [1,p.315] The ith sample mean vector $\underline{\bar{x}}_i = \Sigma_{k=1}^{n}\underline{x}_{ik}/n$ is an estimate for $\underline{\mu}_i$, where

$i = 1$ or 2 and $n = n_1$ or n_2, respectively. [1,p.241] Since

$$\begin{aligned} E(\underline{\bar{X}}_i) &= E(\Sigma^n_{k=1}\underline{X}_{ik}/n) \\ &= \Sigma^n_{k=1}E(\underline{X}_{ik}/n) \\ &= \frac{1}{n}\Sigma^n_{k=1}E(\underline{X}_{ik}) \\ &= \frac{1}{n}(n\underline{\mu}_i) \\ &= \underline{\mu}_i, \end{aligned}$$

$\underline{\bar{x}}_i$ is an unbiased estimate of $\underline{\mu}_i$. The ith sample covariance matrix

$$\underline{S}_i = \frac{1}{n - 1}\Sigma^n_{k=1}(\underline{x}_{ik} - \underline{\bar{x}}_i)(\underline{x}_{ik} - \underline{\bar{x}}_i)'$$

is an unbiased estimate of $\underline{\Sigma}$. [1,p.232] Since $\underline{S}_1$ will not usually be the same as $\underline{S}_2$, the pooled sample covariance matrix

$$\underline{S} = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\underline{S}_1 + (n_2 - 1)\underline{S}_2]$$

is used as an unbiased estimate for $\underline{\Sigma}$. [1,p.241] However, if $\underline{S}_1 = \underline{S}_2$, then $\underline{S}_1 = \underline{S}_2 = \underline{S}$. For large samples, $\underline{\mu}_1$, $\underline{\mu}_2$, and $\underline{\Sigma}$ are replaced by $\underline{\bar{x}}_1$, $\underline{\bar{x}}\ _2$, and $\underline{S}$, respectively. [7,p.492]

CHAPTER V

CLASSIFICATION INTO ONE OF p POPULATIONS

A. BAYES PROCEDURE FOR ARBITRARY POPULATIONS

The Bayes procedure for the classification of an observation $\underline{x} = (x_1, x_2, \ldots, x_m)'$ into one of p populations with known density functions will now be presented. One lets $\pi_1, \ldots, \pi_p$ represent p populations having respective density functions $f_1(\underline{x}), \ldots, f_p(\underline{x})$. The entire sample space is divided into p mutually exclusive regions $R_1, \ldots, R_p$ such that R_i is the region for classification into π_i, $i = 1, \ldots, p$. Also, one lets the cost of misclassifying an observation from π_i into π_j be $C(j|i)$, where the probability of such a misclassification is $P(j|i) = \int_{R_j} f_i(\underline{x})d\underline{x}$, $i,j = 1, \ldots, p$, $i \neq j$. One defines the prior probability that an observation comes from π_i as q_i, $i = 1, \ldots, p$. One assumes that $q_1 + \ldots + q_p = 1$.

The expected cost of misclassification

$$\Sigma_{i=1}^{p} q_i [\Sigma_{\substack{j=1 \\ j \neq i}}^{p} C(j|i)P(j|i)]$$

is to be minimized. [2,p.142] The first step in the Bayes procedure is to evaluate

$$S_j(\underline{x}) = \Sigma_{\substack{i=1 \\ i \neq j}}^{p} q_i f_i(\underline{x}) C(j|i),$$

$j = 1, \ldots, p$, where $D_j(\underline{x}) = -S_j(\underline{x})$ is called the jth discriminant score. [1,p.246] When the regions of classification, $R_1, \ldots, R_p$, are defined such that $\underline{x}$ is assigned to R_j if $S_j(\underline{x}) = \text{minimum}\{S_1(\underline{x}), \ldots, S_p(\underline{x})\}$, $j = 1, \ldots, p$, the expected cost of misclassification is minimized. [2,p.143] One notices that $\text{minimum}\{S_1(\underline{x}), \ldots, S_p(\underline{x})\} = -\text{maximum}\{D_1(\underline{x}),$

$\ldots, D_p(\underline{x})\}$. If more than one of the discriminant scores have the same maximum value, the selection of a discriminant score for classification purposes among those with maximum value is irrelevant. [2,p.143] When discrete distributions are encountered in the p populations, the joint probability function should be substituted for the density in the discriminant score. [3,pp.155-156] This Bayes procedure may be shown to be unique for continuous distributions under conditions analogous to the case of two populations. [2,pp.143-144]

One assumes now that $f(\underline{x})$ is a joint probability density function or a joint probability function. The appropriate form of the Bayes theorem for the classification problem is

$$P(\pi_i|\underline{x}) = \frac{q_i f_i(\underline{x})}{\Sigma_{k=1}^{p} q_k f_k(\underline{x})},$$

$i = 1,\ldots,p$. [7,p.416] This is the posterior probability of the ith population given that $\underline{x}$ has been observed. One lets $C(j|j) = 0$ for $j = 1,\ldots,p$. If $\underline{x}$ is classified into π_j, $j = 1,\ldots,p$, the expected loss is

$$C(j|1)P(\pi_1|\underline{x}) + \ldots + C(j|p)P(\pi_p|\underline{x}) = \frac{\Sigma_{\substack{i=1\\ i\neq j}}^{p} C(j|i)q_i f_i(\underline{x})}{\Sigma_{k=1}^{p} q_k f_k(\underline{x})}. \quad [2,p.143]$$

Clearly, this expected loss is minimized if the index j is chosen when

$$S_j(\underline{x}) = \Sigma_{\substack{i=1\\ i\neq j}}^{p} C(j|i)q_i f_i(\underline{x})$$

has minimum value, $j = 1,\ldots,p$.

One supposes j is the index such that

$$S_j(\underline{x}) = \text{minimum}\{S_1(\underline{x}),\ldots,S_p(\underline{x})\},$$

$j = 1,\ldots,p$, and the costs of misclassification are equal. One defines $C(k|i)$ as unity for $i,k = 1,\ldots,p$, $i \neq k$. Therefore,

$$\Sigma^{p}_{\substack{i=1\\ i\neq j}} q_i f_i(\underline{x}) < \Sigma^{p}_{\substack{i=1\\ i\neq k}} q_i f_i(\underline{x})$$

for $k = 1,\ldots,p$, $k \neq j$. After subtracting $\Sigma^{p}_{\substack{i=1\\ i\neq j,k}} \; q_i f_i(\underline{x})$ from both sides of the above inequality,

$$q_k f_k(\underline{x}) < q_j f_j(\underline{x}),$$

$k = 1,\ldots,p$, $k\neq j$. [2,p.144] The Bayes procedure classifies $\underline{x}$ as from π_j when

$$q_j f_j(\underline{x})$$

is maximum, $j = 1,\ldots,p$. [2,p.144] One notices that the classification of $\underline{x}$ into π_j when $P(\pi_j|\underline{x})$ has maximum value, $j = 1,\ldots,p$, is an equivalent procedure. [1,p.246]

If q_i and $P(j|i)$ are unknown, they may be estimated by using p random samples from the combined population of $\pi_1,\ldots,\pi_p$, $i,j = 1,\ldots,p$, $i \neq j$. One lets $\underline{x}_{i1},\ldots,\underline{x}_{in_i}$ be a random sample of size n_i from π_i, $i = 1,\ldots,p$. Also, one lets $N = n_1 + \ldots + n_p$ be the combined sample size. An estimate for q_i is

$$\overline{q}_i = \frac{n_i}{N},\ i = 1,\ldots,p. \quad [1,p.251]$$

After the N observations have been classified using the Bayes procedure, the misclassification probabilities may be estimated. One lets w_{ij} be the number of observations from π_i which were erroneously classified into π_j, $i,j = 1,\ldots,p$, $i \neq j$. A biased estimate for the probability that an observation from π_i is misclassified into π_j is

$$\overline{P}(j|i) = \frac{w_{ij}}{n_i},$$

$i,j = 1,\ldots,p$, $i \neq j$. [1,p.248]

A.1. MULTINORMAL POPULATIONS WITH COMMON COVARIANCE MATRIX

The Bayes procedure - applied to the case of p multinormal populations with common covariance matrix $\underline{\Sigma}$ - will now be presented. One assumes that the parameters and the prior probability of π_i: $N(\underline{\mu}_i, \underline{\Sigma})$, $i = 1,\ldots,p$, are known and the costs of misclassification are equal. For simplicity, one lets $C(j|i) = 1$ for $i,j = 1,\ldots,p$, $i \neq j$.

One supposes j is the index, $j = 1,\ldots,p$, such that

$$q_k f_k(\underline{x}) < q_j f_j(\underline{x})$$

for $k = 1,\ldots,p$, $k \neq j$. This inequality can be written as

$$\frac{f_j(\underline{x})}{f_k(\underline{x})} > \frac{q_k}{q_j}.$$

One defines a new function as

$$\begin{aligned} r_{jk}(\underline{x}) &= \log \frac{f_j(\underline{x})}{f_k(\underline{x})} \\ &= \underline{x}'\underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k) - \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)'\underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k) \\ &= [\underline{x} - \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)]'\underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k). \end{aligned}$$

By a law of logarithms, $r_{jk}(\underline{x}) = -r_{kj}(\underline{x})$. The Bayes classification procedure assigns $\underline{x}$ to region R_j, $j = 1,\ldots,p$, when

$$r_{jk}(\underline{x}) > \log \frac{q_k}{q_j}$$

for $k = 1,\ldots,p$, $k \neq j$. [2,p.147]

If the population parameters are unknown, they may be estimated by using p random samples. One lets $\underline{x}_{i1},\ldots,\underline{x}_{in_i}$ be a sample of size n_i from π_i, $i = 1,\ldots,p$. An estimate for $\underline{\mu}_i$ is

$$\underline{\bar{x}}_i = \Sigma_{k=1}^{n_i} \underline{x}_{ik}/n_i,$$

$i = 1,\ldots,p$. [1,p.247] $\underline{\Sigma}$ may be estimated by the pooled sample covariance matrix

$$\underline{S} = \frac{\Sigma_{i=1}^{p}(n_i - 1)\underline{S}_i}{\Sigma_{i=1}^{p} n_i - p},$$

where $\underline{S}_i$ is the ith sample covariance matrix, $i = 1,\ldots,p$. [1,p.247] In the functions $r_{ij}(\underline{x})$, $\underline{\mu}_i$, $\underline{\mu}_j$, and $\underline{\Sigma}$ are respectively replaced by $\underline{\bar{x}}_i$, $\underline{\bar{x}}_j$, and $\underline{S}$. [2,p.149-150] "Hence, for sufficiently large samples one can use the theory given above." [2,p.150]

One supposes now that $p = 2$ and the probabilities of misclassification must be estimated. An estimate

$$a = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)'\underline{S}^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

for

$$\alpha = (\underline{\mu}_1 - \underline{\mu}_2)'\underline{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

may be used to obtain estimates

$$\bar{P}(2|1) = \Phi[(\log c - \tfrac{1}{2}a)/a^{1/2}]$$

and

$$\bar{P}(1|2) = \Phi[(-\log c - \tfrac{1}{2}a)/a^{1/2}]$$

of the misclassification probabilities. [1,p.244] "It should be noted, however, that these estimators are biased, that is, on the average, the actual probability of misclassification is greater than the estimated one." [1,p.244] One recalls that $c = \dfrac{C(1|2)q_2}{C(2|1)q_1}$.

A.2. MULTIVARIATE DISCRETE POPULATIONS WHERE X HAS A BERNOULLI DISTRIBUTION

For an illustration of the Bayes classification procedure when p populations have discrete distributions, one assumes that each random variable in a random vector $\underline{X} = (X_1,\ldots,X_m)'$ has a Bernoulli distribution. X_j attains the values 1 and 0 with respective probabilities

$$P(X_j = 1|\pi_i) = p_{ij} \text{ and } P(X_j = 0|\pi_i) = 1 - p_{ij},$$

$i = 1,\ldots,p$, $j = 1,\ldots,m$. In the ith population, the probability function of X_j is $f_i(x_j) = \begin{cases} p_{ij}^{x_j}(1 - p_{ij})^{1-x_j} & \text{for } x_j = 0,1, \\ 0 & \text{otherwise,} \end{cases}$

$i = 1,\ldots,p$, $j = 1,\ldots,m$.

One supposes $X_1,\ldots,X_m$ are mutually independent and the costs of misclassification are equal. One lets $C(j|i) = 1$, $i,j = 1,\ldots,p$, $i \neq j$. The joint probability function for π_i is

$$f_i(\underline{x}) = \Pi_{j=1}^{m} f_i(x_j),$$

$i = 1,\ldots,p$. [1,p.251] An observation $\underline{x} = (x_1,x_2,\ldots,x_m)'$ is classified as from π_i if

$$P(\pi_i|\underline{x}) = \frac{q_i f_i(\underline{x})}{\Sigma_{k=1}^{p} q_k f_k(\underline{x})}$$

has maximum value for $i = 1,\ldots,p$.

If the Bernoulli distribution mean p_{ij} is unknown, it may be estimated by using a random sample of size n_i from π_i, $i = 1,\ldots,p$, $j = 1,\ldots,m$. One lets n_{ij} be the number of sample points from the ith population which have a 1 in the jth row. An unbiased estimate for p_{ij} is

$$\bar{p}_{ij} = \frac{n_{ij}}{n_i},$$

$i = 1,\ldots,p$, $j = 1,\ldots,m$. [1,p.251] The estimate $\bar{p}_{ij}$ should be substituted for p_{ij} to obtain an estimated posterior probability $\bar{P}(\pi_i|\underline{x})$ of the ith population. [1,p.252]

CHAPTER VI

CONCLUSION

The major objective of this thesis has been to present statistical theory relative to the classification problem. Because the multinormal distribution is one of the more common and important multivariate distributions, a definition of this distribution has been included. An area for further study would be the minimax procedure for several given arbitrary populations.

During an application of classification, an investigator may wish to use some relevant statistical theory that has not been included in this thesis. If two given populations have univariate discrete distributions, one may desire to employ a minimax procedure. [3,p.154-155] Before using the test of a hypothesis concerning multinormal populations, one may want to test for equality of several covariance matrices. [2,p.247-250] Also, one may wish to run some type of test on the multinormal mean vectors. [1,p.230-234] Finally, one may desire to use a minimax classification procedure when three populations with multivariate continuous distributions are involved. [2,p.144-152]

BIBLIOGRAPHY

1. Afifi, A. A. and S. P. Azen. Statistical Analysis A Computer Oriented Approach. New York: Academic, 1972.

2. Anderson, T. W. An Introduction to Multivariate Statistical Analysis. New York: Wiley, 1958.

3. Blackwell, David and M. A. Girshick. Theory of Games and Statistical Decisions. New York: Wiley, 1954.

4. Kendall, Maurice G. and Alan Stuart. The Advanced Theory of Statistics, I. London: Griffin, 1963.

5. Lindgren, Bernard W. Statistical Theory. New York: Macmillan, 1976.

6. Morrison, Donald F. Multivariate Statistical Methods. New York: McGraw-Hill, 1976.

7. Rao, C. Radhakrishna. Linear Statistical Inference and Its Applications. New York: Wiley, 1965.