Elimination of continuous variates used in classification and discrimination when both binary and continuous variables are present
by Dennis Patrick Brady

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Philosphy in Mathematics
Montana State University

Abstract:
This thesis investigates the elimination of continuous variables used, in classification or discrimination when the vector data consists of mixed binary, and continuous variables. The location model which is used associates a unique pattern of the binary variables with a multinomial cell and assumes a conditional multivariate normal distribution for the continuous variables in each cell.

Two types of elimination are examined — removal of continuous variables within a cell and global removal of continuous variables over all cells. For each type in the case of parameters known, the minimum expected loss for the remaining variables forms the criterion for determining which variables are to be eliminated. With parameters unknown, an estimate of the actual expected loss is used.

With cell-wise elimination, a method is presented which identifies the variables to be eliminated without necessarily calculating the expected loss or the estimated actual expected loss. For global elimination, two methods are considered. The first determines the exact variables to be eliminated whereas the second method considers four techniques which are computationally simple but estimate those variables to be removed. In the latter case, a study is undertaken to assess the performance of the technique in determining the continuous variables to be eliminated.

ELIMINATION OF CONTINUOUS VARIATES USED IN CLASSIFICATION AND DISCRIMINATION WHEN BOTH BINARY AND CONTINUOUS VARIABLES ARE PRESENT

by

DENNIS PATRICK BRADY

A thesis submitted in partial fulfillment
of the requirements for the degree

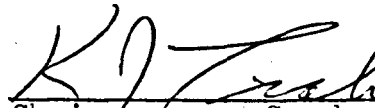
of

DOCTOR OF PHILOSOPHY

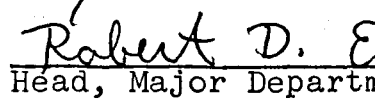
in

Mathematics

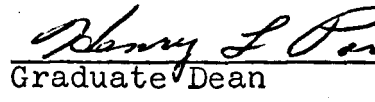
Approved:



Chairperson, Graduate Committee



Head, Major Department



Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

November, 1976

ACKNOWLEDGMENT

The author wishes to express his gratitude to the chairman of his graduate committee, Dr. Kenneth J. Tiahrt, for his able guidance and patience in the preparation of this thesis.

Appreciation is also extended to Professors Robert D. Engle, Martin A. Hamilton, Rodney T. Hansen, Richard E. Lund and Harry Townes for serving on his graduate committee.

Finally, appreciation is extended to my wife, Kerry, for her encouragement and understanding during my involvement with this thesis; and to Mrs. Diane Stovall, who typed this manuscript so efficiently.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

This thesis investigates the elimination of continuous variables used in classification or discrimination when the vector data consists of mixed binary and continuous variables. The location model which is used associates a unique pattern of the binary variables with a multinomial cell and assumes a conditional multivariate normal distribution for the continuous variables in each cell.

Two types of elimination are examined -- removal of continuous variables within a cell and global removal of continuous variables over all cells. For each type in the case of parameters known, the minimum expected loss for the remaining variables forms the criterion for determining which variables are to be eliminated. With parameters unknown, an estimate of the actual expected loss is used.

With cell-wise elimination, a method is presented which identifies the variables to be eliminated without necessarily calculating the expected loss or the estimated actual expected loss. For global elimination, two methods are considered. The first determines the exact variables to be eliminated whereas the second method considers four techniques which are computationally simple but estimate those variables to be removed. In the latter case, a study is undertaken to assess the performance of the techniques in determining the continuous variables to be eliminated.

CHAPTER I

INTRODUCTION

Vector data composed of both binary and continuous variables is often encountered by research workers in medicine, education, engineering, biology, and psychology. However, discrimination and classification of such data has been given only sparse attention until the last few years. Methods have been highly developed for data composed exclusively of either binary variables or continuous variables. However, because of the more complicated computational aspects associated with the mixed-variables, many investigators either converted the continuous variables and proceeded with discrete methods or treated the binary variables as if they were of the continuous type and then used continuous variable techniques. Linhart [30] is usually credited as being the first to propose the former; that is, he attempted to convert continuous variables to binary variables. Two years later, in 1961, Cochran and Hopkins [9] followed up Linhart's idea and determined optimal procedures for the conversion. Many instances are found in the literature where the latter occurs; namely, the discrete nature of a binary variable is not exploited in the discrimination process. As a result, various

authors have studied the adequacy of such continuous variable techniques on binary variables. Gilbert [19] concluded that the performance of a continuous variable technique was satisfactory on a vector of Bernouilli variables. However, an expanded study in 1973 by Moore [32] concluded otherwise. A suitable discrimination technique for mixed-variables was still unknown.

With the advance of computer sophistication, the computational difficulties of mixed-variables were quickly overcome and in 1974 Krzanowski [26] published a unified work on discrimination and classification using both binary and continuous variables. The main results of his efforts are summarized in an article appearing in 1975 in the Journal of the American Statistical Association [27]. However, the literature for data of this type focuses very little on one of the more important considerations; namely, the selection or elimination of the variables used in discrimination or classification. Quite often the investigator may be able to easily obtain the values of the binary variables whereas the cost or effort of measurement of a continuous variable may be enormous compared to that of a binary variable. For example, in educational research the binary variables are usually of the demographic type --

sex, marital status, etc., -- or yes/no responses; and the continuous variables are scores from various tests administered to ascertain certain skills or knowledge of a subject. Similarly, in medical research several measurements may be taken on an individual. The binary variables are often of the demographic, or yes/no, or presence/absence type. The continuous variables are sometimes laboratory readings obtained from sophisticated and costly to operate devices that measure percentages or amounts of certain body elements. Thus both of the above applications illustrate the need for techniques to eliminate one or more continuous variables in the presence of mixed-variable data. Furthermore, it is quite natural to attempt to reduce the analysis to a less complex problem and therefore not complicate the situation with unnecessary continuous variables. The objective of this thesis is to investigate the elimination of a continuous variable in data of the mixed-type and to derive techniques to accomplish such elimination. The expected loss from classification forms the criterion for elimination. Although variable elimination has not been investigated in the framework of mixed-variable data, some workers have treated the problem of variable elimination or selection in various other areas. See, for example, Cochran [8],

Farver [15], Hills [23], Hollingsworth [24], Myers [33], and Weiner [43].

The Model

Various models have been proposed to describe the mixed-variable data. The three most common models employed are: the dichotomized model, in [21,39,40]; the logistic model, in [11,12,13]; and the location model, in [1,7,35]. The dichotomized model assumes an underlying multivariate distribution; and then by "dichotomizing" some of the normal variables, this conversion results in those variables taking on binary values of 0 and 1. The logistic model assumes the continuous variables are marginally multivariate normal whereas the binary variables have a conditional logistic probability distribution whose argument is linear in the continuous variables. The location model assumes a conditional multivariate normal distribution for the continuous variables and that the binary variables form a multinomial distribution. Krzanowski [26] investigated the adequacy of these three models with respect to classification and discrimination of mixed-variable data. Since the dichotomized model and the logistic model are almost identical over a suitable range of values, only a comparison of

the logistic and location models was undertaken. Due to computational problems of a discriminant function associated with the logistic model, Krzanowski concluded that the most suitable model was the location model. Accordingly this model is employed in this work and a full description follows.

Let $\underset{\sim}{w}' = (\underset{\sim}{x}', \underset{\sim}{y}')$ be a mixed-variable multivariate vector. Let $\underset{\sim}{x}' = (x_1, x_2, \ldots, x_q)$ be the vector of q binary variables and let $\underset{\sim}{y}' = (y_1, y_2, \ldots, y_p)$ be the vector of p continuous variables. Thus $\underset{\sim}{w}$ is a column vector with (q+p) components. Following the notation of Krzanowski [26,27], the q binary variables may be represented by a vector $\underset{\sim}{z}' = (z_1, z_2, \ldots, z_k)$ where $k = 2^q$, $\sum_{i=1}^{k} z_i = 1$, $z_i = 0,1$. Thus $\underset{\sim}{z}$ has a multinomial distribution with k cells and the correspondence between a unique pattern of $\underset{\sim}{x}$ and cell m is given by $m = 1 + \sum_{i=1}^{q} x_i 2^{(i-1)}$. For example, if $q = 2$, there are $k = 4$ cells. The patterns of $\underset{\sim}{x}'$ which determine the four cells are (0,0), (1,0), (0,1) and (1,1). Thus cell 3 associated with pattern (0,1) is $3 = 1 + x_1 2^0 + x_2 2^1$ $= 1 + (0)2^0 + (1)2^1$. Thus the probability that an individual falls in cell m is $p_m = \text{pr}\{z_m = 1, z_\nu = 0, \nu \neq m = 1,2,\ldots,k\}$ $= p_m$. Assumed, also, is that the distribution of $\underset{\sim}{y}$ given

cell m is multivariate normal with mean $\underset{\sim}{\mu}^{(m)}$ and covariance matrix $\underset{\sim}{\Sigma}$, i.e.,

$$(1.1) \quad (\underset{\sim}{y} \mid z_m = 1,\ z_\nu = 0,\ \nu \neq m = 1,2,\ldots,k) \sim N(\underset{\sim}{\mu}^{(m)}, \underset{\sim}{\Sigma}).$$

Classification Considerations

Since this work is centered in the area of classification, a general overview of the subject is provided here. The problem arises when an individual is to be classified into one of several categories or populations. The focus of this investigation is restricted to two populations, denoted by π_1 and π_2. Two kinds of "costs" may occur in classification: an individual from π_1 may be misclassified as from π_2; or if the individual is from π_2, he may be misclassified as from π_1. Denote the two types of related costs as $C(2|1)(>0)$ and $C(1|2)(>0)$ respectively. A reasonable classification procedure is one which minimizes the overall cost in some sense.

Suppose that measurements are obtainable on each individual and denoted by $\underset{\sim}{x}$. Let the respective densities of $\underset{\sim}{x}$ for populations π_1 and π_2 be $p_1(\underset{\sim}{x})$ and $p_2(\underset{\sim}{x})$. If $\underset{\sim}{x}$ is p-dimensional then an individual is associated with a point in p-dimensional space. A classification procedure

R partitions this space into two exhaustive regions R_1, R_2 such that individuals associated with points in R_i are classified as from π_i $(i = 1,2)$. Various probabilities of correctly classifying or misclassifying individuals from one of the two populations are induced by a given classification procedure. Denote $P(i \mid j,R)$ as the probability of classifying an individual from π_j as from π_i $(i,j = 1,2)$ where $P(i \mid j, R) = \int_{R_i} p_j(\underset{\sim}{x})d\underset{\sim}{x}$. Therefore the two probabilities of misclassification, also known as the error rates, are given by $P(1|2,R)$ and $P(2|1,R)$.

Next suppose there exists a priori probabilities that an individual belongs to the two populations, π_1 and π_2, denoted by $q_1 (>0)$ and $q_2 (>0)$ respectively. The probability of drawing and misclassifying an individual from π_1 with classification procedure R is $q_1\ P(2|1,R)$ and the analogous version from π_2 is $q_2\ P(1|2,R)$. Thus the expected loss from misclassification using classification procedure R is

$$EL = C(2|1)P(2|1,R)q_1 + C(1|2)P(1|2,R)q_2 \tag{1.2}$$

The classification procedure which minimizes EL is thus the best procedure. A solution is given by Anderson [3, p. 131] where he proves that the best procedure is to associate

R_1 with those points $\underset{\sim}{x}$ for which

$$(1.3) \qquad \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} \geq \frac{C(1|2)q_2}{C(2|1)q_1}$$

and let R_2 be all those points $\underset{\sim}{x}$ in the space which do not belong to R_1.

If $p_1(\underset{\sim}{x})$ and $p_2(\underset{\sim}{x})$ are multivariate normal densities with $\underset{\sim}{x}$ having mean $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ in π_1, π_2 respectively and common covariance $\underset{\sim}{\Sigma}$ then Anderson [3,p. 134] shows that R_1 is that collection of points for which

$$(1.4) \quad \underset{\sim}{x}'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2) - \tfrac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2) \geq \log\left\{\frac{C(1|2)q_2}{C(2|1)q_1}\right\}.$$

If the population parameters are unknown, which is usually the case in most applications, then they must be obtained from samples, one from each population. An estimated classification rule may then be obtained from (1.3) by replacing the parameters with their estimates. For multivariate normal populations, $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$, and $\underset{\sim}{\Sigma}$ may be estimated by $\overline{\underset{\sim}{x}}_1$, $\overline{\underset{\sim}{x}}_2$, and $\underset{\sim}{S}$ respectively in the classification rule (1.4). Thus the left-hand side of (1.4) is now given by the classification statistic
$W = \underset{\sim}{x}'\underset{\sim}{S}^{-1}(\overline{\underset{\sim}{x}}_1-\overline{\underset{\sim}{x}}_2) - \frac{1}{2}(\overline{\underset{\sim}{x}}_1+\overline{\underset{\sim}{x}}_2)'\underset{\sim}{S}^{-1}(\overline{\underset{\sim}{x}}_1-\overline{\underset{\sim}{x}}_2)$, also known as Anderson's classification statistic [2]. The first term

of the right-hand side is Fisher's linear discriminant function (LDF) and is a linear function of the variables $x_1, x_2, \ldots, x_p$. This expression was first characterized by Fisher in 1936 [16].

CHAPTER II

CELL-WISE ELIMINATION OF A CONTINUOUS VARIABLE: PARAMETERS KNOWN

Preliminary Considerations

This chapter considers the consequences of the removal of a continuous variable from a given cell. That is, one continuous variable is removed from each of the k cells, and different variables may be removed from different cells. The purpose for studying cell-wise removal of a continuous variable is that the binary variables' values may be obtained first, determining a given cell. Within the determined cell, it is possible to eliminate any one of the p continuous variables, thus leaving p-1 continuous variables to determine the classification rule and the expected loss.

An application of such a procedure is in the area of educational testing. The measurement of the binary variables is first obtained on an individual. This determines a cell which in turn designates p-1 tests out of p tests available for the individual to take. The two populations are success (π_1) and failure (π_2) of the individuals in a program in which they are applying for admission. It is desired that the p-1 tests administered to the individual

discriminate as well as or better than any other set of p-1 tests out of the p tests available that might be taken.

As an example, suppose mixed-variable data consists of five continuous variables and two binary variables. Thus $p=5$, $q=2$, and $k=2^q=4$. An elimination process which eliminates continuous variables y_1, y_4, y_5, and y_1 in cell 1 through cell 4, respectively, is denoted by $\underset{\sim}{EP}' = (1\ 4\ 5\ 1)$. To avoid confusion, $\underset{\sim}{EP}$ and $\underset{\sim}{EP}'$ will refer to the same elimination process and both are used only to distinguish between a column vector version and a row vector form of the same elimination process.

The following definition is a result of the preceding discussion and example.

Definition 2-1. A cell-wise elimination process, denoted $\underset{\sim}{EP}$, may be represented by a $k \times 1$ vector with the mth component value, denoted ep(m), an integer between 1 and p, inclusively, designating the continuous variable to be eliminated in the mth cell.

The location model is now generalized for dealing with two populations, and the optimal classification rule and its associated expected loss is derived, both for mixed-variable data in general and also when an $\underset{\sim}{EP}$ is involved.

Classification and the Expected Loss

The formulation of the location model, given in (1.1), is now extended to the case of two populations in order to deal with classification and elimination problems. Again following the notation of Krzanowski [26,27], assume that $\underset{\sim}{y}$ has a multivariate normal distribution with mean $\underset{\sim}{\mu}_i^{(m)}$ in cell m of population π_i (i = 1,2; m = 1,2,...,k), and with common covariance matrix $\underset{\sim}{\Sigma}$ in all cells for both populations, i.e.,

$$(2.1)\quad (\underset{\sim}{y}|z_m=1,\ z_\nu=0,\ \nu\neq m=1,2,\ldots,k) \sim N(\underset{\sim}{\mu}_i^{(m)},\underset{\sim}{\Sigma}) \text{ in } \pi_i (i=1,2).$$

Also assume that the probability of obtaining an observation in cell m is p_{im} for population π_i (i=1,2;m=1,2,...,k).

Assuming all population parameters to be known, the optimal classification procedure is given in Theorem 2.1 and the related error rates and the expected loss follow. These results are obtained in a straightforward manner by Krzanowski [26,Ch. 6], using the general theory of classification as presented by Anderson [3,Ch. 6]. The results are given in detail since virtually all aspects of continuous variable elimination considered in this thesis are linked to a classification rule and its expected loss.

Theorem 2.1: If the distributions of π_i are given by (2.1) (i=1,2), the best classification rule is to classify $(\underset{\sim}{x}',\underset{\sim}{y}')$ to π_1 if $1 + \sum_{i=1}^{q} x_i 2^{(i-1)} = m$ and $L_m \geq \log(f_m)$, and otherwise to π_2; where

$$(2.2) \quad L_m = \underset{\sim}{y}'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1^{(m)}-\underset{\sim}{\mu}_2^{(m)}) - \tfrac{1}{2}(\underset{\sim}{\mu}_1^{(m)}+\underset{\sim}{\mu}_2^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1^{(m)}-\underset{\sim}{\mu}_2^{(m)})$$

and

$$(2.3) \qquad f_m = \frac{p_{2m}C(1|2)q_2}{p_{1m}C(2|1)q_1} .$$

Proof: The best classification rule, as derived by Anderson [3,p. 131] and given by (1.3) in chapter 1, states that $\underset{\sim}{w}' = (\underset{\sim}{x}',\underset{\sim}{y}')$ is classified to π_1 if $\frac{p_1(\underset{\sim}{w})}{p_2(\underset{\sim}{w})} \geq \frac{C(1|2)q_2}{C(2|1)q_1}$, and otherwise to π_2; where $p_i(\underset{\sim}{w}) = p_i(\underset{\sim}{x},\underset{\sim}{y})$ is the joint density of $\underset{\sim}{x}$ and $\underset{\sim}{y}$ in population π_i (i=1,2).

The location model implies that $p_i(\underset{\sim}{w}) = p_i(\underset{\sim}{x})p_i(\underset{\sim}{y}|\underset{\sim}{x})$ (i=1,2). Conditioning first on $\underset{\sim}{x}$, suppose that this results in the occurance of cell m; i.e., $1 + \sum_{i=1}^{q} x_i 2^{(i-1)} = m$. Next,

$$(2.4) \qquad p_i(\underset{\sim}{w}) = p_i(z_m,\underset{\sim}{y}) = p_{im}p_i(\underset{\sim}{y}|z_m)$$

where $p_i(\underset{\sim}{x}|z_m)$ is the multivariate normal density

$$(2.5)\quad (2\pi)^{-\frac{1}{2}p}\,|\underset{\sim}{\Sigma}|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_i^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_i^{(m)})\}\quad (i=1,2),$$

with covariance $\underset{\sim}{\Sigma}$ assumed to be a positive definite matrix.

Thus,

$$(2.6)\quad \frac{p_1(\underset{\sim}{x})}{p_2(\underset{\sim}{x})} = \frac{p_{1m}}{p_{2m}}\exp\{-\tfrac{1}{2}[(\underset{\sim}{x}-\underset{\sim}{\mu}_1^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1^{(m)})-(\underset{\sim}{x}-\underset{\sim}{\mu}_2^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2^{(m)})]\}.$$

The left hand side, LHS, of (1.3) is now replaced by the right hand side, RHS, of (2.6). Then, moving the cell probabilities, p_{1m} and p_{2m}, to the RHS of (1.3) and taking the logarithm of both sides since this function is monotonic increasing, the classification rule is to classify $\underset{\sim}{x}$ to π_1 if $1 + \sum_{i=1}^{q} x_i 2^{(i-1)} = m$ and

$$(2.7)\quad -\tfrac{1}{2}[(\underset{\sim}{x}-\underset{\sim}{\mu}_1^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1^{(m)}) - (\underset{\sim}{x}-\underset{\sim}{\mu}_2^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2^{(m)})] \geq \log\left\{\frac{p_{2m}C(1|2)q_2}{p_{1m}C(2|1)q_1}\right\},$$

and otherwise to π_2.

The LHS of (2.7) can now be rewritten as

$$\underset{\sim}{x}'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1^{(m)}-\underset{\sim}{\mu}_2^{(m)}) - \tfrac{1}{2}(\underset{\sim}{\mu}_1^{(m)}+\underset{\sim}{\mu}_2^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1^{(m)}-\underset{\sim}{\mu}_2^{(m)}) = L_m.$$

This completes the proof. Note that the rule may be applied even if the numerator, or the denominator, of f_m is zero -- caused by p_{2m}, or p_{1m}, respectively being zero. However, if both p_{1m} and p_{2m} are zero, f_m is of indeterminate form. Therefore, the offending cell m is removed, and work is done with only k-1 cells.

The probabilities of misclassification and the expected loss (EL) are now obtained. Let $D_m^2 = (\underset{\sim}{\mu}_1^{(m)}-\underset{\sim}{\mu}_2^{(m)})'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_1^{(m)}-\underset{\sim}{\mu}_2^{(m)})$ (>0), the Mahalanobis distance between populations π_1 and π_2 conditionally on the individual falling in cell m. As is readily shown and obtained by Anderson [3, p. 134], $L_m \sim N(\frac{1}{2}D_m^2, D_m^2)$ when $(\underset{\sim}{x} | z_m=1, z_\nu=0, \nu\neq m=1,2,\ldots,k) \sim N(\underset{\sim}{\mu}_1^{(m)},\underset{\sim}{\Sigma})$; i.e., when π_1 has occurred. Therefore, the conditional probability of misclassifying an individual from π_1 as from π_2 in cell m is

$$P_m(2|1) = pr\{L_m < \log(f_m) \mid \pi_1\} = \Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} \tag{2.8}$$

where $\Phi(x)$ is the cumulative standard normal distribution function. Also, $L_m \sim N(-\frac{1}{2}D_m^2, D_m^2)$ when $(\underset{\sim}{x} \mid z_m=1, z_\nu=0, \nu\neq m=1,2,\ldots,k) \sim N(\underset{\sim}{\mu}_2^{(m)},\underset{\sim}{\Sigma})$; i.e., when

population π_2 has occurred. Thus, the probability of misclassifying an individual from π_2 as from π_1 conditionally in cell m is

$$(2.9) \qquad P_m(1|2) = pr\{L_m \geq \log(f_m) \mid \pi_2\}$$

$$= \Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} .$$

Equations (2.8) and (2.9) will be defined to be the cell-wise error rates. Now the probability of cell m occurring is p_{im} for population π_i (i=1,2; m=1,2,...,k). Hence, the unconditional (overall) error rates are

$$(2.10) \qquad P(2|1) = \sum_{m=1}^{k} p_{1m}P_m(2|1)$$

$$= \sum_{m=1}^{k} p_{1m}\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\}$$

and

$$P(1|2) = \sum_{m=1}^{k} p_{2m}\,P_m(1|2)$$

$$= \sum_{m=1}^{k} p_{2m}\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} .$$

Since the error rates considered in this thesis are always associated with the optimal classification rule, or an estimated optimal classification rule, the reference to a

classification rule "R" is not necessary for either expression in (2.10). Accordingly, it is omitted throughout.

The cell-wise expected loss is given by

$$(2.11)\quad p_{1m}C(2|1)q_1\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} + p_{2m}C(1|2)q_2\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\}$$

$$= p_{1m}C(2|1)q_1P_m(2|1) + p_{2m}C(1|2)q_2P_m(1|2).$$

The (overall) expected loss using the optimal classification is therefore

$$(2.12)\quad EL = C(2|1)q_1P(2|1) + C(1|2)q_2P(1|2)$$

$$= C(2|1)q_1\sum_{m=1}^{k} p_{1m}\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} + C(1|2)q_2\sum_{m=1}^{k} p_{2m}\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\}$$

$$= \sum_{m=1}^{k}\left[p_{1m}C(2|1)q_1\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} + p_{2m}C(1|2)q_2\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\}\right]$$

$$= \sum_{m=1}^{k}\left[p_{1m}C(2|1)q_1P_m(2|1) + p_{2m}C(1|2)q_2P_m(1|2)\right].$$

Consideration of Theorem 2.1, along with the expected loss of (2.12), leads to the realization that the classifying of observations and calculation of the error rates may be accomplished in a cell-wise manner since there is a different rule for each of the cells. Thus, for each cell m, an optimal classification rule can be obtained based on only p-1 continuous variables instead of p. The error rates (cell-wise and overall) and the expected loss (cell-wise and overall) may be derived from the set of p-1 continuous variables selected in each cell as determined by the $\underset{\sim}{EP}$. Notation necessary to obtain these results with an $\underset{\sim}{EP}$ now follows.

Let $\underset{\sim}{\Sigma}_j$ be the covariance matrix for the p-1 continuous variables with y_j omitted. Note that $\underset{\sim}{\Sigma}_j$ is the matrix formed by removing row j and column j from $\underset{\sim}{\Sigma}$. Let $\underset{\sim}{\mu}_{i,j}^{(m)}$ be the column vector of $\underset{\sim}{\mu}_i^{(m)}$ after removing the mean of variable y_j in cell m of π_i. Also let $\underset{\sim}{x}_j$ be the column vector of $\underset{\sim}{x}$ formed by removing variable y_j. Therefore, the conditional distribution of the (p-1) × 1 vector $\underset{\sim}{x}_j$ is multivariate normal with mean $\underset{\sim}{\mu}_{i,j}^{(m)}$ in cell m of population π_i with covariance matrix $\underset{\sim}{\Sigma}_j$, i.e.,

$$(2.13)\quad (\underset{\sim}{x}_j \mid z_m=1,\ z_\nu=0,\ \nu\neq m=1,2,\ldots,k) \sim N(\underset{\sim}{\mu}_{i,j}^{(m)},\underset{\sim}{\Sigma}_j) \text{ in } \pi_i \ (i=1,2).$$

It is noted that $\underset{\sim}{\Sigma}_j$ is a $(p-1) \times (p-1)$ positive definite matrix since $\underset{\sim}{\Sigma}$ is assumed to be a positive definite matrix. For a proof see Anderson [3, p. 337].

The following theorem provides the optimal classification procedure in the presence of an $\underset{\sim}{EP}$.

Theorem 2.2: If populations π_i have distributions given by (2.13)(i=1,2), the best classification rule with an $\underset{\sim}{EP}$ present is to classify $(\underset{\sim}{x}', \underset{\sim}{y}')$ to π_1 if $1 + \sum_{i=1}^{q} x_i 2^{(i-1)} = m$ and $L_{m,ep(m)} \geq \log(f_m)$, and otherwise to π_2; where

$$(2.14)\quad L_{m,ep(m)} = \underset{\sim}{y}'_{ep(m)} \underset{\sim}{\Sigma}^{-1}_{ep(m)} (\underset{\sim}{\mu}^{(m)}_{1,ep(m)} - \underset{\sim}{\mu}^{(m)}_{2,ep(m)}) - \frac{1}{2}(\underset{\sim}{\mu}^{(m)}_{1,ep(m)} + \underset{\sim}{\mu}^{(m)}_{2,ep(m)})' \underset{\sim}{\Sigma}^{-1}_{ep(m)} (\underset{\sim}{\mu}^{(m)}_{1,ep(m)} - \underset{\sim}{\mu}^{(m)}_{2,ep(m)})$$

and f_m is from (2.3).

Proof: The proof is very similar to that of Theorem 2.1 and is not repeated here.

The probabilities of misclassification and the expected loss for the optimal classification rule with an $\underset{\sim}{EP}$ involved are now considered. Let $D^2_{m,j} = (\underset{\sim}{\mu}^{(m)}_{1,j} - \underset{\sim}{\mu}^{(m)}_{2,j})' \underset{\sim}{\Sigma}^{-1}_{j} (\underset{\sim}{\mu}^{(m)}_{1,j} - \underset{\sim}{\mu}^{(m)}_{2,j}) (>0)$ be the Mahalanobis distance between π_1 and π_2 conditionally on the individual

falling in cell m and based on p-1 continuous variables with y_j omitted. Derivation of the cell-wise error rates for p-1 continuous variables with y_j omitted in cell m is similar to the methods for finding (2.8) and (2.9). Thus the cell-wise error rates for cell m with y_j removed are

$$(2.15) \qquad P_m(2|1)_j = \Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,j}}{D_{m,j}}\right\} \text{ and}$$

$$P_m(1|2)_j = \Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,j}}{D_{m,j}}\right\} .$$

The unconditional (overall) error rates with a given $\underset{\sim}{EP}$, obtained in a manner analogous to those in (2.10) are

$$(2.16) \quad P(2|1)_{\underset{\sim}{EP}} = \sum_{m=1}^{k} p_{1m} P_m(2|1)_{ep(m)}$$

$$= \sum_{m=1}^{k} p_{1m}\, \Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} \text{ and}$$

$$P(1|2)_{\underset{\sim}{EP}} = \sum_{m=1}^{k} p_{2m}\, P_m(1|2)_{ep(m)}$$

$$= \sum_{m=1}^{k} p_{2m}\, \Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} .$$

The cell-wise expected loss with an $\underset{\sim}{EP}$ is

$$(2.17)\quad p_{1m}C(2|1)q_1\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} + p_{2m}C(1|2)q_2\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\}$$
$$= p_{1m}C(2|1)q_1P_m(2|1)_{ep(m)} + p_{2m}C(1|2)q_2P_m(1|2)_{ep(m)}\,.$$

Thus, the (overall) expected loss using the optimal classification rule of Theorem 2.2 in conjunction with an $\underset{\sim}{EP}$ is

$$(2.18)\quad EL(\underset{\sim}{EP}) = C(2|1)q_1\sum_{m=1}^{k} p_{1m}\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} + C(1|2)q_2\sum_{m=1}^{k} p_{2m}\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\}$$
$$= \sum_{m=1}^{k}\left[p_{1m}C(2|1)q_1\,\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} + p_{2m}C(1|2)q_2\,\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\}\right]$$
$$= \sum_{m=1}^{k}\left[p_{1m}C(2|1)q_1P_m(2|1)_{ep(m)} + p_{2m}C(1|2)q_2P_m(1|2)_{ep(m)}\right].$$

Note that the expected loss, EL($\underset{\sim}{EP}$), is derived from an optimal classification rule no matter what variables are removed as determined by the $\underset{\sim}{EP}$; however, the $\underset{\sim}{EP}$ may not be "optimal" in the sense that a different cell-wise elimination process may have the smallest expected loss. That is, there are p^k different cell-wise elimination processes, each with an associated (overall) expected loss based on the optimal classification rule of Theorem 2.2. Therefore an $\underset{\sim}{EP}$ with minimum (overall) expected loss is sought. This search is initiated by studying the effect on the cell-wise error rates, given in the next section.

The Effect on the Cell-wise Error Rates: $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$

This section considers cell-wise elimination in the case where the continuous variables are mutually independent. Thus, this implies that $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, where $\underset{\sim}{D}(\sigma_j^2)$ is a diagonal matrix whose diagonal elements -- σ_j^2 -- are the variances of the continuous variables y_j, respectively. A detailed study of this situation is of theoretical interest since it gives valuable insight into the error rates associated with an $\underset{\sim}{EP}$, enables the cell-wise Mahalanobis distances, D_m^2 and $D_{m,ep(m)}^2$, to be easily interpreted, and

focuses attention under simple conditions to areas where problems may arise.

In using an $\underset{\sim}{EP}$, an immediate concern is what happens to the error rates or expected loss as a consequence of the cell-wise elimination. A result which is basic to the investigation of this situation and concerns cell-wise error rates is given in the following theorem.

Theorem 2.3: If the distributions of the π_i (i=1,2) are given by (2.1) for Theorem 2.1 or by (2.13) for Theorem 2.2, depending on non-elimination or an $\underset{\sim}{EP}$, respectively, then using the optimal classification rules of Theorems 2.1 and 2.2 with $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, the effect on cell-wise error rates is either a trivial result or a non-trivial result with three cases. The results are

I. If $\delta^2_{m,ep(m)} = 0$,
then $P_m(2|1) = P_m(2|1)_{ep(m)}$,
and $P_m(1|2) = P_m(1|2)_{ep(m)}$.

II. If $\delta^2_{m,ep(m)} > 0$, and
Case I. If $\log(f_m) > (=) \frac{1}{2}D_m D_{m,ep(m)}$
then $P_m(2|1) < (<) P_m(2|1)_{ep(m)}$
and $P_m(1|2) > (=) P_m(1|2)_{ep(m)}$.

Case II. If $-\frac{1}{2}D_m D_{m,ep(m)} < \log(f_m) < \frac{1}{2}D_m D_{m,ep(m)}$
then $P_m(2|1) < P_m(2|1)_{ep(m)}$
and $P_m(1|2) < P_m(1|2)_{ep(m)}$.

Case III. If $\log(f_m) < (=) -\frac{1}{2}D_m D_{m,ep(m)}$
then $P_m(2|1) > (=) P_m(2|1)_{ep(m)}$
and $P_m(1|2) < (<) P_m(1|2)_{ep(m)}$, where

(2.19) $\delta^2_{m,j} = \dfrac{(\mu^{(m)}_{j1}-\mu^{(m)}_{j2})^2}{\sigma^2_j}$ and $\mu^{(m)}_{ji}$ is the mean of variable y_j in cell m of π_i (i=1,2; j=1,2,...,p).

Proof: Consider the expressions for the cell-wise error rates for non-elimination as given by (2.8) and (2.9); and those for the case of an $\underset{\sim}{EP}$ present as provided in (2.15). Note that $\Phi(x)$ is an increasing function of x on $[-\infty,+\infty]$. Since $\underset{\sim}{\Sigma} = \underset{\sim}{Q}(\sigma^2_j)$,

(2.20) $D^2_m = \sum_{j=1}^{p} \delta^2_{m,j}$, and

$$D^2_{m,ep(m)} = \sum_{\substack{j=1 \\ \neq ep(m)}}^{p} \delta^2_{m,j} = D^2_m - \delta^2_{m,ep(m)}.$$

The proof of the theorem can now deal with the two results.

I. $\delta^2_{m,ep(m)} = 0$.

In this case, $D^2_m = D^2_{m,ep(m)}$. Thus, by inspection of (2.8), (2.9), and (2.15), it follows that $P_m(2|1) = P_m(2|1)_{ep(m)}$

and $P_m(1|2) = P_m(1|2)_{ep(m)}$.

II. $\delta^2_{m,ep(m)} > 0$.

In this situation $D^2_m > D^2_{m,ep(m)}$; hence $D_m > D_{m,ep(m)}$. The three cases of result II of the theorem now follow from the results of A and B below.

A. Suppose $\log(f_m) > (=) \frac{1}{2}D_m D_{m,ep(m)}$.

Then $-\log(f_m)\left\{\frac{D_{m,ep(m)}-D_m}{D_m D_{m,ep(m)}}\right\} > (=) \frac{1}{2}\{D_m - D_{m,ep(m)}\}$ since $D_m - D_{m,ep(m)} > 0$. Next,

$$\frac{-\log(f_m)-\frac{1}{2}D^2_m}{D_m} > (=) \frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}$$ and therefore,

$$\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_m}{D_m}\right\} > (=) \Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\}.$$

Thus, $P_m(1|2) > (=) P_m(1|2)_{ep(m)}$ whenever $\log(f_m) > (=) \frac{1}{2}D_m D_{m,ep(m)}$. Also the results of A imply that $P_m(1|2) < P_m(1|2)_{ep(m)}$ if $\log(f_m) < \frac{1}{2}D_m D_{m,ep(m)}$.

B. Suppose $\log(f_m) < (=) -\frac{1}{2}D_m D_{m,ep(m)}$.

Then $\log(f_m)\left\{\frac{D_{m,ep(m)}-D_m}{D_m D_{m,ep(m)}}\right\} > (=) \frac{1}{2}\{D_m - D_{m,ep(m)}\}$

since $D_m - D_{m,ep(m)} > 0$. Next,

$$\frac{\log(f_m)-\frac{1}{2}D^2_m}{D_m} > (=) \frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}$$ and therefore

$$\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_m^2}{D_m}\right\} > (=)\ \Phi\left\{\frac{\log(f_m)-\frac{1}{2}D_{m,ep(m)}^2}{D_{m,ep(m)}}\right\}.$$

Thus, $P_m(2|1) > (=)\ P_m(2|1)_{ep(m)}$ if $\log(f_m) < (=)\ -\frac{1}{2}D_m D_{m,ep(m)}$. Also, the results of B imply that $P_m(2|1) < P_m(2|1)_{ep(m)}$ whenever $\log(f_m) > -\frac{1}{2}D_m D_{m,ep(m)}$. This completes the proof.

Theorem 2.3 and its proof provide some interesting observations to be made. If $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, then the calculation of $D_{m,j}^2$ can be done quickly once D_m^2 is computed since $D_{m,j}^2 = D_m^2 - \delta_{m,j}^2$ $(j=1,2,\ldots,p)$. Thus, using an $\underset{\sim}{EP}$ does not require any substantial amount of additional computing time to find new cell-wise Mahalanobis distances. It is also seen that eliminating a continuous variable in a cell may actually result in a decrease in one of the cell-wise error rates. However, the elimination cannot decrease both cell-wise error rates and actually may increase both cell-wise error rates. Results concerning the cell-wise expected loss and the overall expected loss will be given later in this chapter.

Thus far, an $\underset{\sim}{EP}$ and non-elimination have been compared in terms of their cell-wise error rates. Since there are actually p^k different $\underset{\sim}{EP}$'s, it is natural to consider mext the interrelationships of cell-wise error rates between two $\underset{\sim}{EP}$'s. Let $\underset{\sim}{EP}$ and $\underset{\sim}{EP}^*$ be two different $k\times 1$ column

vectors. That is, $\underset{\sim}{EP}^*$ must have at least one component value different from $\underset{\sim}{EP}$. The following theorem enables a comparison of $\underset{\sim}{EP}$ and $\underset{\sim}{EP}^*$ to be made on the basis of their cell-wise error rates.

Theorem 2.4: If π_i (i=1,2) is given by (2.13) for both an $\underset{\sim}{EP}$ and an $\underset{\sim}{EP}^*$, then using the optimal classification rules for $\underset{\sim}{EP}$ and $\underset{\sim}{EP}^*$ in Theorem 2.2 and with $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, comparing cell-wise error rates results in either a trivial situation or a non-trivial situation consisting of three cases. Letting $\delta^2_{m,ep(m)} \leq \delta^2_{m,ep^*(m)}$ without loss of generality, the results are

I. If $\delta^2_{m,ep(m)} = \delta^2_{m,ep^*(m)}$

then $P_m(2|1)_{ep(m)} = P_m(2|1)_{ep^*(m)}$

and $P_m(1|2)_{ep(m)} = P_m(1|2)_{ep^*(m)}$.

II. If $\delta^2_{m,ep(m)} < \delta^2_{m,ep^*(m)}$ and

Case I. If $\log(f_m) > (=) \frac{1}{2} D_{m,ep(m)} D_{m,ep^*(m)}$

then $P_m(2|1)_{ep(m)} < (<) P_m(2|1)_{ep^*(m)}$

and $P_m(1|2)_{ep(m)} > (=) P_m(1|2)_{ep^*(m)}$.

Case II. If $-\frac{1}{2} D_{m,ep(m)} D_{m,ep^*(m)} < \log(f_m) < \frac{1}{2} D_{m,ep(m)} D_{m,ep^*(m)}$

then $P_m(2|1)_{ep(m)} < P_m(2|1)_{ep^*(m)}$

and $P_m(1|2)_{ep(m)} < P_m(1|2)_{ep^*(m)}$.

Case III. If $\log(f_m) <(=) -\frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$
then $P_m(2|1)_{ep(m)} >(=) P_m(2|1)_{ep^*(m)}$
and $P_m(1|2)_{ep(m)} <(<) P_m(1|2)_{ep^*(m)}$.

Proof: The proof is similar to that of Theorem 2.3 and uses the expressions of (2.15). Since $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, $D^2_{m,ep(m)} = D^2_m - \delta^2_{m,ep(m)}$ and $D^2_{m,ep^*(m)} = D^2_m - \delta^2_{m,ep^*(m)}$. Assuming $\delta^2_{m,ep(m)} \leq \delta^2_{ep^*(m)}$ implies that $D^2_{m,ep(m)} \geq D^2_{m,ep^*(m)}$, with strict inequality when $\delta^2_{m,ep(m)} < \delta^2_{m,ep^*(m)}$. The proof of the theorem now deals with the two situations.

I. $\delta^2_{m,ep(m)} = \delta^2_{m,ep^*(m)}$

In this situation $D^2_{m,ep(m)} = D^2_{m,ep^*(m)}$; and by inspection of (2.15), it follows that $P_m(2|1)_{ep(m)} = P_m(2|1)_{ep^*(m)}$ and $P_m(1|2)_{ep(m)} = P_m(1|2)_{ep^*(m)}$.

II. $\delta^2_{m,ep(m)} < \delta^2_{m,ep^*(m)}$

In this situation $D^2_{m,ep(m)} > D^2_{m,ep^*(m)}$; thus $D_{m,ep(m)} > D_{m,ep^*(m)}$. The three cases of situation II of the theorem now follow from the results of A and B below.

A. Suppose $\log(f_m) >(=) \frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$.

Then $-\log(f_m)\left\{\frac{D_{m,ep^*(m)} - D_{m,ep(m)}}{D_{m,ep(m)}D_{m,ep^*(m)}}\right\} >(=) \frac{1}{2}\{D_{m,ep(m)} - D_{m,ep^*(m)}\}$

since $D_{m,ep(m)} - D_{m,ep^*(m)} > 0$. Next

$\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}} > (=) \quad \frac{-\log(f_m)-\frac{1}{2}D_{m,ep^*(m)}}{D_{m,ep^*(m)}}$ and therefore, $\Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} > (=) \Phi\left\{\frac{-\log(f_m)-\frac{1}{2}D^2_{m,ep^*(m)}}{D_{m,ep^*(m)}}\right\}$

Thus $P_m(1|2)_{ep(m)} > (=) P_m(1|2)_{ep^*(m)}$ whenever $\log(f_m) > (=) \frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$. Also, the results of A imply that $P_m(1|2)_{ep(m)} < P_m(1|2)_{ep^*(m)}$ if $\log(f_m) < \frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$.

B. Suppose $\log(f_m) < (=) -\frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$.

Then $\log(f_m)\left\{\frac{D_{m,ep^*(m)}-D_{m,ep(m)}}{D_{m,ep(m)}D_{m,ep^*(m)}}\right\} > (=) \frac{1}{2}\{D_{m,ep(m)}-D_{m,ep^*(m)}\}$

since $D_{m,ep(m)}-D_{m,ep^*(m)} > 0$. Next,

$\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}} > (=) \frac{\log(f_m)-\frac{1}{2}D^2_{m,ep^*(m)}}{D_{m,ep^*(m)}}$ and therefore,

$\Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep(m)}}{D_{m,ep(m)}}\right\} > (=) \Phi\left\{\frac{\log(f_m)-\frac{1}{2}D^2_{m,ep^*(m)}}{D_{m,ep^*(m)}}\right\}$. Thus, $P_m(2|1)_{ep(m)} > (=) P_m(2|1)_{ep^*(m)}$ if $\log(f_m) < (=) -\frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$. In addition, the results of A imply that $P_m(2|1)_{ep(m)} < P_m(2|1)_{ep^*(m)}$ if $\log(f_m) > -\frac{1}{2}D_{m,ep(m)}D_{m,ep^*(m)}$. This completes the proof.

The relationship in terms of cell-wise error rates between an EP and a different elimination process, say $\underset{\sim}{EP}^*$, when $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, is thus very similar to the relationship between an $\underset{\sim}{EP}$ and non-elimination. In fact, Theorem 2.3 can be considered a special case of Theorem 2.4 where the situation of non-elimination is considered to be an $\underset{\sim}{EP}$ with $\delta^2_{m,ep(m)} = 0$ $(m=1,2,\ldots,k)$ and the optimal rule of Theorem 2.1 is used. Therefore, in comparing two $\underset{\sim}{EP}$'s in terms of their cell-wise error rates, Theorem 2.4 provides all possibilities; which are seen to depend on the value of $\log(f_m)$ and the cell-wise Mahalanobis distances (or the $\delta^2_{m,ep(m)}$).

The Optimal $\underset{\sim}{EP}$: $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$

Let $O\underset{\sim}{EP}$, an optimal $\underset{\sim}{EP}$, be an $\underset{\sim}{EP}$ such that $EL(O\underset{\sim}{EP}) = \min_{\underset{\sim}{EP}} EL(\underset{\sim}{EP})$. Although the $O\underset{\sim}{EP}$ may not be unique it clearly exists. If $\log(f_m)$ is suitably small, then Theorem 2.4 provides an optimal process in the special case when $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$ which is now given in Corollary 2.1.

Corollary 2.1. If π_i (i=1,2) is given by (2.13) with $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$ and if $-\frac{1}{2} \min_j (D_{m,j}D_{m,i(m)}) < \log(f_m)$ $< \frac{1}{2} \min_j (D_{m,j}D_{m,i(m)})$ for all m where i(m) is such that $\delta^2_{m,i(m)} = \min_j \delta^2_{m,j}$; then using the optimal classification rule in Theorem 2.2, an O$\underset{\sim}{\text{EP}}$ is that $\underset{\sim}{\text{EP}}$ with component value oep(m) = i(m) (m=1,2,...,k).

Proof: Since $-\frac{1}{2} \min_j (D_{m,j}D_{m,i(m)}) < \log(f_m)$ $< \frac{1}{2} \min_j (D_{m,j}D_{m,i(m)})$, Theorem 2.4 implies that $P_m(2|1)_{i(m)} \leq P_m(2|1)_j$ and $P_m(1|2)_{i(m)} \leq P_m(1|2)_j$ for all j using result I and case II of result II. Thus for each cell m, $p_{1m}C(2|1)q_1P_m(2|1)_{i(m)} + p_{2m}C(1|2)q_2P_m(1|2)_{i(m)}$ $\leq p_{1m}C(2|1)q_1P_m(2|1)_j + p_{2m}C(1|2)q_2P_m(1|2)_j$ for all j. Thus EL(O$\underset{\sim}{\text{EP}}$) $\leq$ EL($\underset{\sim}{\text{EP}}$) for all $\underset{\sim}{\text{EP}}$ where oep(m) = i(m) (m=1,2,...,k) and EL($\underset{\sim}{\text{EP}}$) is given by (2.17). This completes the proof.

The above corollary provides an O$\underset{\sim}{\text{EP}}$ only in the special case where $\log(f_m)$ is suitably small. An O$\underset{\sim}{\text{EP}}$ is needed for an arbitrary $\log(f_m)$ when $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_i^2)$. A useful lemma which is necessary in the search for an O$\underset{\sim}{\text{EP}}$ is now given. Note that it does not depend on $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$.

Lemma 2.1: If π_i (i=1,2) is given by (2.1) for Theorem 2.1 or by (2.13) for Theorem 2.2, depending on non-elimination or an $\underset{\sim}{EP}$, respectively; then using the optimal classification rules of Theorems 2.1 and 2.2 yields a cell-wise expected loss in each cell m which is a decreasing function of the square root of the cell-wise Mahalanobis distance whenever p_{1m} and p_{2m} are simultaneously non-zero.

Proof: Since the form of the cell-wise expected loss is identical for non-elimination and an $\underset{\sim}{EP}$, only the expression for the non-elimination case will be utilized in the proof. Let $G(D_m)$ be the cell-wise expected loss in cell m from (2.12). Thus

$$G(D_m) = p_{1m}C(2|1)q_1P_m(2|1) + p_{2m}C(1|2)q_2P_m(1|2)$$

$$= p_{1m}C(2|1)q_1 \int_{-\infty}^{\left[\frac{\log(f_m)}{D_m} - \frac{D_m}{2}\right]} (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}t^2)dt$$

$$+ p_{2m}C(1|2)q_2 \int_{-\infty}^{\left[\frac{-\log(f_m)}{D_m} - \frac{D_m}{2}\right]} (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}t^2)dt.$$

To obtain the result of the lemma it is sufficient to show that

$G'(D_m) = \frac{d(G(D_m))}{d(D_m)} < 0$ if p_{1m} and p_{2m} are non-zero. Now

$$G'(D_m) = p_{1m}C(2|1)q_1\left\{\frac{-\log(f_m)}{D_m^2} - \frac{1}{2}\right\}(2\pi)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\frac{\log(f_m)}{D_m} - \frac{D_m}{2}\right]^2\right\}$$

$$+ p_{2m}C(2|1)q_1\left\{\frac{\log(f_m)}{D_m^2} - \frac{1}{2}\right\}(2\pi)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\frac{-\log(f_m)}{D_m} - \frac{D_m}{2}\right]^2\right\}.$$

Next,

$$G'(D_m) = p_{1m}C(2|1)q_1\left\{\frac{-\log(f_m)}{D_m^2} - \frac{1}{2}\right\}(2\pi)^{-\frac{1}{2}}$$

$$\times \exp\left\{\frac{-[\log(f_m)]^2}{2D_m^2} + \frac{\log(f_m)}{2} - \frac{D_m^2}{8}\right\}$$

$$+ p_{2m}C(2|1)q_1\left\{\frac{\log(f_m)}{D_m^2} - \frac{1}{2}\right\}(2\pi)^{-\frac{1}{2}}$$

$$\times \exp\left\{\frac{-[\log(f_m)]^2}{2D_m^2} - \frac{\log(f_m)}{2} - \frac{D_m^2}{8}\right\}.$$

Then rearranging terms,

$$G'(D_m) = (2\pi)^{-1/2}\exp\left\{\frac{-[\log(f_m)]^2}{2D_m^2} - \frac{D_m^2}{8}\right\}$$

$$\times \left[p_{1m}C(2|1)q_1\left(\frac{-\log(f_m)}{D_m^2} - \frac{1}{2}\right)\exp\left(\frac{\log(f_m)}{2}\right)\right.$$

$$\left. + p_{2m}C(1|2)q_2\left(\frac{\log(f_m)}{D_m^2} - \frac{1}{2}\right)\exp\left(\frac{-\log(f_m)}{2}\right)\right].$$

However, $\exp\left(\frac{\log(f_m)}{2}\right) = (f_m)^{\frac{1}{2}}$ and $\exp\left(\frac{-\log(f_m)}{2}\right) = (f_m)^{-\frac{1}{2}}$

where $f_m = \frac{p_{2m}C(1|2)q_2}{p_{1m}C(2|1)q_1}$ from (2.3). Thus,

$$G'(D_m) = (2\pi)^{-\frac{1}{2}} \exp\left\{ \frac{-[\log(f_m)]^2}{2D_m^2} - \frac{D_m^2}{8} \right\}$$

$$\times \left[(p_{1m}C(2|1)q_1 p_{2m}C(1|2)q_2)^{\frac{1}{2}} \left(\frac{-\log(f_m)}{D_m^2} - \frac{1}{2} \right) \right.$$

$$\left. + (p_{2m}C(1|2)q_2 p_{1m}C(2|1)q_1)^{\frac{1}{2}} \left(\frac{\log(f_m)}{D_m^2} - \frac{1}{2} \right) \right] .$$

Therefore,

$$G'(D_m) = (2\pi)^{-\frac{1}{2}} \exp\left\{ \frac{-[\log(f_m)]^2}{2D_m^2} - \frac{D_m^2}{8} \right\}$$

$$\times (p_{2m}C(1|2)q_2 p_{1m}C(2|1)q_1)^{\frac{1}{2}}[-1].$$

Hence, $G'(D_m) < 0$ whenever the cell probabilities in cell m are both non-zero. This completes the proof.

Note that if one of the two cell probabilities -- p_{1m} and p_{2m} -- is zero, then the cell-wise expected loss is zero no matter what value is given for the cell-wise Mahalanobis distance in cell m. Thus, one of two probabilities being zero is really the trivial case of the

classification problem and Lemma 2.1 applies to the non-trivial situation.

With Lemma 2.1, a procedure can now be obtained for finding an O$\underset{\sim}{\text{EP}}$ when $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$ and $\log(f_m)$ is arbitrary. The following theorem provides this O$\underset{\sim}{\text{EP}}$.

Theorem 2.5: If π_i (i=1,2) is given by (2.13) with $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$ and if the optimal classification rule of Theorem 2.2 is used, the O$\underset{\sim}{\text{EP}}$ is that $\underset{\sim}{\text{EP}}$ with component values oep(m) where oep(m) is such that $\delta^2_{m,oep(m)} = \min_j \delta^2_{m,j}$ (m=1,2,...,k)

Proof: If the cell probabilities are non-zero in cell m, Lemma 2.1 states that the cell-wise expected loss will be smallest if $D_{m,ep(m)}$ is as large as possible for cell-wise elimination. Since $\underset{\sim}{\Sigma} = \underset{\sim}{D}(\sigma_j^2)$, $D^2_{m,ep(m)} = D^2_m - \delta^2_{m,ep(m)}$ from (2.20); thus the minimum value of $\delta^2_{m,ep(m)}$ implies the maximum value of $D_{m,ep(m)}$; and, hence the smallest cell-wise expected loss in cell m. If p_{1m} or p_{2m} is zero in cell m, the cell-wise expected loss for all ep(m) is zero, including the oep(m). Since EL($\underset{\sim}{\text{EP}}$) is the summation over all cells of each cell's cell-wise expected loss, choosing the component values