DISCRIMINANT FUNCTIONS AND THEIR MISCLASSIFICATION ERRORS

I WAYAN MANGKU

Department of Mathematics,
Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University
Jl. Meranti, Kampus IPB Darmaga, Bogor, 16680 Indonesia

Abstract. This paper is a survey study on discriminant functions and their misclassification errors. Here we consider three groups of discriminant functions, namely discriminant functions for respectively multivariate normal variables, multivariate binary variables, and a mixture of multivariate binary and normal variables. Finally we derive their misclassification errors.
Key words: Discriminant analysis, discriminant function, misclassification error, multivariate normal variable, multivariate binary variable, mixture of multivariate binary and normal variables.

1. Introduction

This paper is concerned with discriminant functions and their misclassification errors. This field of study is known as discriminant analysis, one of the important techniques in multivariate analysis.

In discriminant analysis, given the existence of various groups of individuals, we want to find (i) the best way of exhibiting the difference between groups (discrimination problem), and (ii) a rule for allocating new individuals (observations) into one of the existing groups (classification problem). In our research work, the interest is in the classification problem. To solve this problem, a classification rule needs to be constructed. A number of classification rules have been established in the literature. The choice of the most appropriate rule depends on the type of variables in the data; and the best classification rule is the one that leads to the smallest probability of misclassification.

Based on the type of explanatory variables, the field of discriminant analysis can be grouped into three main categories, namely, (i) discriminant analysis with continuous explanatory variables only, (ii) discriminant analysis with discrete explanatory variables only, and (iii) discriminant analysis with a mixture of discrete and continuous explanatory variables. In this paper, we consider the special cases of the

above three categories, namely (a) discriminant analysis with multivariate normal variables only as a special case of (i); (b) discriminant analysis with multivariate binary variables only as a special case of (ii); and (c) discriminant analysis with a mixture of multivariate binary and normal variables as a special case of (iii).

In this paper, we restrict our study to discriminant analysis problems involving only two groups or populations. These groups are denoted by Π_1 and Π_2. Suppose $\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_p)^T$ is a p-dimensional vector of random variables associated with any individual. We assume that $\underline{\mathbf{X}}$ has different probability distributions in Π_1 and Π_2. Let $\underline{\mathbf{x}}$ be the observed value of $\underline{\mathbf{X}}$ (for an arbitrary individual), $f_1(\underline{\mathbf{x}})$ be the probability density of $\underline{\mathbf{X}}$ in Π_1, and $f_2(\underline{\mathbf{x}})$ be the probability density of $\underline{\mathbf{X}}$ in Π_2. Then the simplest intuitive classification decision is: classify $\underline{\mathbf{x}}$ into Π_1 if it has greater probability of coming from Π_1, or classify $\underline{\mathbf{x}}$ into Π_2 if it has greater probability of coming from Π_2, or classify $\underline{\mathbf{x}}$ arbitrarily into Π_1 or Π_2 if these probabilities are equal.

In real situations it is reasonable to consider some important factors such as prior probabilities of observing individuals from the two populations and the cost due to misclassifications. Let q_1 and q_2 be the prior probabilities that $\underline{\mathbf{x}}$ comes from Π_1 and Π_2 respectively ($q_1 + q_2 = 1$). Also let $c(2|1)$ be the cost due to misclassifying $\underline{\mathbf{x}}$ into Π_2 when it actually belongs to Π_1, and $c(1|2)$ be the cost due to the misclassification of $\underline{\mathbf{x}}$ into Π_1 when it belongs to Π_2. Here, the decision is: classify $\underline{\mathbf{x}}$ into Π_1 if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) > q_2c(1|2)/(q_1c(2|1))$, or classify $\underline{\mathbf{x}}$ into Π_2 if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) < q_2c(1|2)/(q_1c(2|1))$, or classify $\underline{\mathbf{x}}$ arbitrarily into Π_1 or Π_2 if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) = q_2c(1|2)/(q_1c(2|1))$.

In our study we only consider the case with equal prior probabilities and equal cost due to misclassifications (ie. $q_1 = q_2$ and $c(1|2) = c(2|1)$). In other words the decision is as follows: classify $\underline{\mathbf{x}}$ into Π_1 if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) > 1$, classify $\underline{\mathbf{x}}$ into Π_2 if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) < 1$, and classify $\underline{\mathbf{x}}$ arbitrarily into Π_1 or Π_2 if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) = 1$.

2. Discriminant functions for multivariate normal variables

A variety of classification rules has been established in the literature. The earliest and most well-known rule is Fisher's (1936) Linear Discriminant Function (LDF). Let $\underline{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})^T$, be the means and Σ_i be the covariance matrices of $\underline{\mathbf{X}}$ in Π_i ($i = 1, 2$). It is often assumed that $\Sigma_1 = \Sigma_2 = \Sigma$. Let $\bar{\underline{\mathbf{x}}}_1, \bar{\underline{\mathbf{x}}}_2, \mathbf{S}_1, \mathbf{S}_2$, and $\mathbf{S}$ be the sample estimates of $\underline{\mu}_1, \underline{\mu}_2, \Sigma_1, \Sigma_2$ and Σ respectively, using independent random samples of size n_1 and n_2 from Π_1 and Π_2. Denote these random samples (also called training samples) by $\underline{\mathbf{t}}_1$ and $\underline{\mathbf{t}}_2$ respectively, and let $\underline{\mathbf{t}} = \{\underline{\mathbf{t}}_1, \underline{\mathbf{t}}_2\}$ be the entire set of training data of $n = n_1 + n_2$ observations. Also let $N_p(\underline{\mu}, \Sigma)$ denotes the p-variate normal distribution with mean $\underline{\mu}$ and covariance matrix Σ. The estimated Fisher's LDF is then given by

$$L(\underline{\mathbf{x}}) = \underline{\mathbf{x}}^T\mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2). \tag{2.1}$$

This LDF was adopted later by Anderson (1951) to obtain a classification statistics $W(\underline{\mathbf{x}})$, given by

$$W(\underline{\mathbf{x}}) = (\underline{\mathbf{x}} - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2))^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2). \tag{2.2}$$

Using this rule, a new individual $\underline{\mathbf{x}}$ will be allocated into Π_1 if $W(\underline{\mathbf{x}}) \geq 0$, otherwise into Π_2.

Anderson's classification rule above is at its best performance when the two populations are multivariate normal with common covariance matrices. However, when the populations are multivariate normal, but with different covariance matrices, the ideal choice is QDF (Quadratic Discriminant Function). The estimated (sample) QDF is defined as:

$$Q(\underline{\mathbf{x}}) = \frac{1}{2}\ln\left[\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}\right] - \frac{1}{2}\left((\underline{\mathbf{x}} - \bar{\underline{\mathbf{x}}}_1)^T \mathbf{S}_1^{-1}(\underline{\mathbf{x}} - \bar{\underline{\mathbf{x}}}_1) - (\underline{\mathbf{x}} - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}_2^{-1}(\underline{\mathbf{x}} - \bar{\underline{\mathbf{x}}}_2)\right) \tag{2.3}$$

and a new individual $\underline{\mathbf{x}}$ will be allocated into Π_1 if $Q(\underline{\mathbf{x}}) \geq 0$, otherwise into Π_2 (see Gilbert (1969), Marks and Dunn (1974), Wald and Kronmal (1977), Randles et al. (1978), Lachenbruch (1979), and Van Ness (1979)). For this situation, the QDF is optimal as it minimizes the overall probability of misclassification (Marks and Dunn, 1974). However, the efficiency of the sample QDF has been shown to decrease with decrease in sample size and the difference between the covariance matrices (Choi, 1986). Friedman (1989) introduced regularized discriminant analysis which can substantially improve the misclassification risk when the population covariance matrices are not close to being equal.

Ashikaga and Chang (1981) evaluated the robustness of the LDF when the distribution of the two populations are characterized by two-component mixed normal distributions with known parameters. They found that similarity in shape is more important than normality for the robustness of the LDF. Their conclusion was that the LDF is rather robust when the two distributions do not markedly deviate from normality and are moderately distant, particularly if they are similar in shape.

Some further investigations and adjustments of the LDF for various different situations when the explanatory variables are continuous, also have been studied by Gessaman and Gessaman (1972), Anderson (1973), Goldstein (1975), McLachlan (1977), Ganesalinggam and McLachlan (1978), Chhikara and McKeon (1984), Critchley and Ford (1984, 1985), Ambergen (1985), Critchley (1985), Murphy and Moran (1986), Davis (1987), Raveh (1989), and Wakaki (1990).

From all the above information, it is clear that the best classification rule for the case where both populations are multivariate normal with common covariance matrices, is the Anderson's statistics $W(\underline{\mathbf{x}})$.

3. Discriminant functions for multivariate binary variables

Now consider the discriminant analysis when the explanatory variables are multivariate binary only. For this case, some authors have proposed special classification rules, including the full multinomial model,

the first and the second-order of Bahadur models, loglinear and logit models, Martin-Bradley models, and some others besides the LDF (Hills (1967), Gilbert (1968), Glick (1973), Moore (1973), Aitchison and Aitken (1976), Goldstein and Dillon (1978), Hall (1981), and Hand (1983)). Because of the large number of procedures available, it becomes considerable interest to determine the relative effectiveness of these competing classification rules. Some comparative studies in this area have been conducted by Gilbert (1968), Moore (1973), Dillon and Goldstein (1978), and Hand (1983).

Next, we consider the classification rule LDF for multivariate binary variables. Let $\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_p)^T$ be a p-dimensional vector of Bernoulli variables, each of which can take the value 0 or 1. A particular observed value of $\underline{\mathbf{X}}$, denoted by $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_p)^T$, is called a response pattern. Let $\beta_i(\underline{\mathbf{x}})$ be the probability of observing a response pattern $\underline{\mathbf{x}}$ (of 0's and 1's) in Π_i (i=1,2) and $Z_{ij} = (X_j - p_{ij})/(p_{ij}(1-p_{ij}))^{1/2}$. Here, $p_{ij} = \mathsf{E}_i(X_j)$, denote the probability that the j-th variable takes value 1 in Π_i, and $r_i(jk) = \mathsf{E}(Z_{ij}Z_{ik})$, denote the correlation between the j-th and k-th variables in Π_i. Following Bahadur (1961), the second-order approximation to the multinomial cell probabilities is given by

$$\beta_i(\underline{\mathbf{x}}) = \prod_{j=1}^{p} p_{ij}^{x_j}(1-p_{ij})^{1-x_j} \left\{ 1 + \sum_{j} \sum_{j<k} r_i(jk) Z_{ij} Z_{ik} \right\}. \tag{3.1}$$

This re-parameterization enables the multinomial cell probabilities to be described in terms of the population means p_{ij} and the correlations $r_i(jk)$. Then, in the case of multivariate binary data, the classification rule $W(\underline{\mathbf{x}})$ in equation (2.2) can be re-written as

$$\begin{aligned} W(\underline{\mathbf{x}}) &= \sum_{j=1}^{p}\sum_{k=1}^{p} (\hat{p}_{1j} - \hat{p}_{2j})\, S_{kj} x_k - \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p} (\hat{p}_{1j} - \hat{p}_{2j})\, S_{kj}\, (\hat{p}_{1k} + \hat{p}_{2k}) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} \left[\left(x_k - \frac{1}{2}(\hat{p}_{1k} + \hat{p}_{2k}) \right) S_{kj}\, (\hat{p}_{1j} - \hat{p}_{2j}) \right], \end{aligned} \tag{3.2}$$

where x_k is the k-th element of the new individual $\underline{\mathbf{x}}$, $\hat{p}_{ij}$ is the estimate of the j-th element p_{ij} of the population mean vector $\underline{\mu}_i$, and S_{kj} is the (k,j)-th element of the inverse of the pooled covariance matrix derived from the training data. Note that equation (3.2) is equivalent to the usual expression of $W(\underline{\mathbf{x}})$ given in equation (2.2).

Gilbert (1968) compared the performance of the LDF with three other discriminant functions for the data which consist of dichotomous (binary) variables only. Two of these three discriminant functions used were based on a logistic model and the third one based upon the assumption of mutual independent variables. She concluded that the loss involved from using the LDF as the classification rule as opposed to any one of the other procedures is too small to be of any practical importance. Hence she argued that the simplicity and familiarity of the LDF make its use seem desirable. She also added that as the number

of variables increases, the LDF should remain fairly stable and behave superior to any of the other techniques.

Moore (1973) evaluated the performances of five discrimination procedures for binary data. These procedures were: the LDF, the QDF, the full multinomial procedure, and the procedures based on the first order and the second order of the Bahadur models. For these comparisons, he considered three different cases: (a) uncorrelated variables in both populations, (b) correlated variables only in one population (c) positively correlated variables in both populations. The results indicated that the LDF and the first order of the Bahadur model behave superior to the others for cases (a) and (b). He recommended the second order of the Bahadur model for case (c). It was also shown that if the true log likelihood ratio (l.l.r) for the two populations is plotted against the number of variables having value 1, then in some populations this l.l.r does not increase monotonically. For this case it was said to undergo a "reversal". He noted that the LDF gives good results for the populations without reversals. But, for the populations with reversals, the LDF led to a significantly greater actual error rate than a classification procedure based on the full multinomial model.

Dillon and Goldstein (1978) compared the performance of the LDF, the Martin-Bradley models, and the Bahadur models. In their conclusions, they recommended to use the LDF or the first-order of the Bahadur model when the correlations between the variables are moderate, or when large difference exists between the mean vectors of the two populations. For the cases with small difference between the mean vectors of the two populations or with large absolute correlations between the variables, the first-order of Martin Bradley model was recommended.

Hand (1983) compared the performance of the LDF and the kernel method using several real data sets with multivariate binary variables. This kernel method is a nonparametric technique which is ideally suited to binary data. He found that the apparent error rates of the kernel methods are consistently less than those of the LDF. However, it is already known that the apparent error rate is an unreliable estimate of the future classification performance. When the true error rates were estimated using the leave-one-out method, no significant difference was noted between the two classification rules.

Based on the information so far, some authors have attested the robustness of the LDF for data with binary explanatory variables only, especially for the cases with weakly correlated variables or with highly separated mean vectors of the two populations or for the populations without reversals.

4. Discriminant functions for a mixture of multivariate binary and normal variables

Now consider the discriminant analysis when the data consist of a mixture of multivariate binary and normal variables. In many real situations, multivariate data contain a mixture of discrete and continuous variables. For instance, in soil science one may have continuous laboratory measurements of soil pH, mixed with such categorical attributes as soil colour or texture. In medical research, one may have

continuous measurements of patient temperatures or blood pressures, mixed with binary observations like sex, presence or absence of a certain symptom for each patient. So, multivariate techniques to handle data with mixtures of variables are important. For discriminant analysis in particular, a classification rule using mixtures of explanatory variables needs to be constructed.

Chang and Afifi (1974) used a point-biserial model to obtain a double discriminant function which can be used to classify an observation consists of one dichotomous (binary) and some continuous variables. A comparison of this classification procedure with two other procedures called "the x-continuous procedure" and "the x-out procedure", was conducted by Tu and Han (1982). This double discriminant function was then extended for more general case of a mixture of multivariate binary and continuous variables by Krzanowski (1975). He proposed a likelihood ratio method based on a model called Location Model.

Let $\underline{\mathbf{Z}}^T = (\underline{\mathbf{X}}^T, \underline{\mathbf{Y}}^T)$ be a partitioned vector of random variables, where $\underline{\mathbf{X}}$ consists of q binary variables (say, $X_1, X_2, \ldots, X_q$) and $\underline{\mathbf{Y}}$ consists of p normal variables (say, $Y_1, Y_2, \ldots, Y_p$). Associated with the q binary variables are 2^q multinomial cells and each cell represents a unique response pattern $\underline{\mathbf{x}}$. Let $\underline{\mathbf{x}}_m$ denote the unique response pattern correspond to the multinomial cell m such that

$$m = 1 + \sum_{j=1}^{q} x_j 2^{j-1}$$

for $m = 1, 2, \ldots, 2^q$ where x_j is the observed value of X_j. Under the location model, it is assumed that the continuous variables have a multivariate normal distribution within each cell of the binary variables. Let us assume that the variables vector $\underline{\mathbf{Y}}$ (given $\underline{\mathbf{X}} = \underline{\mathbf{x}}_m$) has a multivariate normal distribution with mean vector $\underline{\mu}_{im}$ and covariance matrix Σ_{im} in cell m and population Π_i ($m = 1, 2, \ldots, 2^q$, i=1,2). It is usually assumed that $\Sigma_{im} = \Sigma$ for all $m = 1, 2, \ldots, 2^q$ and $i = 1, 2$, for simplification. In other words,

$$(\underline{\mathbf{Y}} | \underline{\mathbf{X}} = \underline{\mathbf{x}}_m) \sim N_p(\underline{\mu}_{im}, \Sigma) \tag{4.1}$$

in Π_i. Let $\underline{\mathbf{z}}^T = (\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T)$ be the observed value of $\underline{\mathbf{Z}}^T$ and also let β_{im} be the probability of observing an observation $\underline{\mathbf{z}}$ in cell m for population Π_i. When all the population parameters are known, the optimal classification rule (assuming equal costs and equal prior probabilities) is: given $\underline{\mathbf{x}} = \underline{\mathbf{x}}_m$, allocate $\underline{\mathbf{z}}$ into Π_1 if

$$\left(\underline{\mathbf{y}} - \frac{1}{2}(\underline{\mu}_{1m} + \underline{\mu}_{2m})\right)^T \Sigma^{-1} \left(\underline{\mu}_{1m} - \underline{\mu}_{2m}\right) > \log\left(\frac{\beta_{2m}}{\beta_{1m}}\right), \tag{4.2}$$

otherwise to Π_2. This classification rule leads effectively to a different linear discriminant function for each of the multinomial cells, with cut-off points determined in each cell by the discrete component of the model. Here, we use $\log(\beta_{2m}/\beta_{1m})$ instead of using 0 as the cut-off point for each classification rule corresponding to each multinomial cell m.

In practice, the population parameters are generally unknown, and we have to estimate these parameters using the information from the

training samples. If $\hat{\underline{\mu}}_{1m}, \hat{\underline{\mu}}_{2m}, \mathbf{S}, \hat{\beta}_{1m}$ and $\hat{\beta}_{2m}$ are the estimates of respectively $\underline{\mu}_{1m}, \underline{\mu}_{2m}, \Sigma, \beta_{1m}$ and β_{2m}, then the estimate of the classification rule in (4.2) is then given by the following. Given $\underline{\mathbf{x}} = \underline{\mathbf{x}}_m$, allocate $\underline{\mathbf{z}}$ into Π_1 if

$$\left(\underline{\mathbf{y}} - \frac{1}{2}(\hat{\underline{\mu}}_{1m} + \hat{\underline{\mu}}_{2m})\right)^T \mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right) > \log\left(\frac{\hat{\beta}_{2m}}{\hat{\beta}_{1m}}\right), \tag{4.3}$$

otherwise to Π_2.

Later, in 1980 Krzanowski generalized his method to incorporate mixtures of continuous and categorical variables. Krzanowski (1977) compared the performance of the LDF and the location model for the case of a mixture of multivariate binary and normal variables, when the assumptions for the location model hold. He found that in some conditions the location model behaves superior to the LDF. In our study, we considered to use the location model for the case with a mixture of multivariate binary and normal variables.

Some further studies in this area have been carried out by Krzanowski (1976,..., 1986), Knoke (1982), Vlachonikolis and Marriot (1982), Wojciechowski (1985), and Vlachonikolis (1990).

5. Misclassification errors

One of the important and interesting problem in discriminant analysis is the evaluation of the performance of a classification rule. This evaluation may focus on either the estimation of the error rate (probability of misclassification) conditional on the correct population assignment, or on the goodness of fit of the estimated posterior (inverse) probabilities to the achieved outcomes conditional on the observed feature vector (Knoke, 1986). In our research, we are interested in the former, the error rate estimation. A bibliography by Toussaint (1974) which was updated recently by McLachlan (1986), shows that interest in techniques of error rate estimation is still strong and that research should be continued.

In fact, there are three types error rates that have been frequently considered for study. They are: (i) the *optimum error rate,* which describes the performance of a classification rule based on known parameters, (ii) the *conditional error rate,* which describes the performance of a classification rule based on parameters estimated by the statistics computed from the training samples, and (iii) the *expected error rate,* which describes the expected performance of a classification rule based on parameters estimated by a randomly chosen training sample.

In practice, the parameters are rarely known, and the expected (or unconditional) error rates depend heavily on the distribution of the discriminant function (in our case, $W(\underline{\mathbf{x}})$) which is very complicated (see for example, Wald (1944), Anderson (1951), Okamoto (1963) and Hills (1966)). Consequently most work associated with error rate have assumed that the sample estimates $\bar{\underline{\mathbf{x}}}_1, \bar{\underline{\mathbf{x}}}_2, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}$ etc. are fixed, leading to the exploration of the *conditional error rate.* Here the word 'conditional' refers to the conditioning of the training samples (and fixing the above sample estimates) from which $W(\underline{\mathbf{x}})$ is constructed. We may

also think of this as either the error rate which would be observed if $W(\underline{\mathbf{x}})$ is applied to an infinite test data, or the probability that the given classifier $W(\underline{\mathbf{x}})$ would incorrectly classify a future observation. It should also be noted that the conditional error rate is the error rate that is important to an experimentor who has already determined the classification rule. This conditional error rate is also referred to as the *actual error rate* or the *true error rate* by many authors. Hence, in our project we concentrate only on the actual error rate and its estimation.

Recall that, only the two-group discriminant analysis problems with equal prior probabilities and equal cost due to misclassifications are considered in our study. Suppose that our classification rule is $W(\underline{\mathbf{x}})$ given by (2.2). Then the actual error rates are given by

$$\begin{aligned} P_1 &= \mathsf{P}(W(\underline{\mathbf{x}}) < 0 \text{ when } \underline{\mathbf{x}} \text{ is from } \Pi_1 | \underline{\mathbf{t}} \text{ fixed}), \\ P_2 &= \mathsf{P}(W(\underline{\mathbf{x}}) \geq 0 \text{ when } \underline{\mathbf{x}} \text{ is from } \Pi_2 | \underline{\mathbf{t}} \text{ fixed}). \end{aligned} \tag{5.1}$$

The overall actual error rate is then defined by

$$AC = \frac{n_1}{n_1+n_2} P_1 + \frac{n_2}{n_1+n_2} P_2. \tag{5.2}$$

Under the assumptions that $\underline{\mathbf{X}} \sim N_p(\underline{\mu}_1, \Sigma)$ on population Π_1 and $\underline{\mathbf{X}} \sim N_p(\underline{\mu}_2, \Sigma)$ on population Π_2, it can easily be shown that

$$P_1 = \Phi \left[\frac{-\left(\underline{\mu}_1 - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2)\right)^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)}{\left((\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)\right)^{1/2}} \right] \tag{5.3}$$

and

$$P_2 = \Phi \left[\frac{\left(\underline{\mu}_2 - \frac{1}{2}(\bar{\underline{\mathbf{x}}}_1 + \bar{\underline{\mathbf{x}}}_2)\right)^T \mathbf{S}^{-1}(\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)}{\left((\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)^T \mathbf{S}^{-1} \Sigma \mathbf{S}^{-1} (\bar{\underline{\mathbf{x}}}_1 - \bar{\underline{\mathbf{x}}}_2)\right)^{1/2}} \right] \tag{5.4}$$

where Φ is the distribution function of a standard normal variate.

When the discriminant function consists of multivariate binary variables, the actual error rates can be derived as follows. Let $\underline{\mathbf{x}}_m$ denote a particular response pattern corresponding to the multinomial cell m, and $\beta_{im}(\underline{\mathbf{x}}_m)$ denote the probability of observing a response pattern $\underline{\mathbf{x}}_m$ in population Π_i, i=1,2. Now let $\alpha_m(\underline{\mathbf{x}}_m) = 0, 1/2$, or 1 according as $W(\underline{\mathbf{x}}_m)$ is $>$, $=$, or < 0. The actual error rates P_1 and P_2 are then given by

$$P_1 = \sum_{m=1}^{2^p} \alpha_m(\underline{\mathbf{x}}_m)\beta_{1m}(\underline{\mathbf{x}}_m) \text{ and } P_2 = \sum_{m=1}^{2^p} \left(1 - \alpha_m(\underline{\mathbf{x}}_m)\right) \beta_{2m}(\underline{\mathbf{x}}_m), \tag{5.5}$$

and the overall actual error rate is given by (5.2).

Next we derive the expression of the actual error rate for the case with mixture of multivariate banary and multivariate normal variables. Following the assumption in (4.1), that $(\underline{\mathbf{Y}}|\underline{\mathbf{X}} = \underline{\mathbf{x}}_m) \sim N_p(\underline{\mu}_{1m}, \Sigma)$ in Π_1 and $(\underline{\mathbf{Y}}|\underline{\mathbf{X}} = \underline{\mathbf{x}}_m) \sim N_p(\underline{\mu}_{2m}, \Sigma)$ in Π_2, the actual error rate can be easily derived as follows. Given a fixed training sample, it can be shown that the distribution of $W_m(\underline{\mathbf{y}}|\underline{\mathbf{x}} = \underline{\mathbf{x}}_m)$, when $\underline{\mathbf{z}}^T = (\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T)$

belongs to Π_1, is a univariate normal distribution with mean

$$\left(\underline{\mu}_{1m} - \frac{1}{2}(\hat{\underline{\mu}}_{1m} + \hat{\underline{\mu}}_{2m})\right)^T \mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right) - \log\left(\frac{\hat{\beta}_{2m}}{\hat{\beta}_{1m}}\right),$$

and variance

$$\left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)^T \mathbf{S}^{-1}\Sigma\mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right).$$

As before, $\hat{\underline{\mu}}_{im}, \hat{\beta}_{im}$, for $i = 1, 2$ and $\mathbf{S}$ are respectively the estimate of $\underline{\mu}_{im}, \beta_{im}$, for $i = 1, 2$ and Σ, using the training sample $\underline{\mathbf{t}}$. Then,

$$\begin{aligned} P_1 &= \mathsf{P}\left(W_m(\underline{\mathbf{y}}|\underline{\mathbf{x}} = \underline{\mathbf{x}}_m) < 0 \text{ when } \underline{\mathbf{z}}^T = (\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T) \text{ is from } \Pi_1|\underline{\mathbf{t}} \text{ fixed }\right) \\ &= \mathsf{P}\left[Z < \frac{0 - \left(\underline{\mu}_{1m} - \frac{1}{2}(\hat{\underline{\mu}}_{1m} + \hat{\underline{\mu}}_{2m})\right)^T \mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right) - \log\left(\frac{\hat{\beta}_{2m}}{\hat{\beta}_{1m}}\right)}{\left\{\left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)^T \mathbf{S}^{-1}\Sigma\mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)\right\}^{1/2}}\right] \\ &= \Phi\left[-\frac{\left(\underline{\mu}_{1m} - \frac{1}{2}(\hat{\underline{\mu}}_{1m} + \hat{\underline{\mu}}_{2m})\right)^T \mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right) - \log\left(\frac{\hat{\beta}_{2m}}{\hat{\beta}_{1m}}\right)}{\left\{\left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)^T \mathbf{S}^{-1}\Sigma\mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)\right\}^{1/2}}\right], \quad (5.6) \end{aligned}$$

where $Z \sim N(0, 1)$. Similarly, we can show that

$$\begin{aligned} P_2 &= \mathsf{P}\left(W_m(\underline{\mathbf{y}}|\underline{\mathbf{x}} = \underline{\mathbf{x}}_m) \geq 0 \text{ when } \underline{\mathbf{z}}^T = (\underline{\mathbf{x}}^T, \underline{\mathbf{y}}^T) \text{ is from } \Pi_2|\underline{\mathbf{t}} \text{ fixed }\right) \\ &= \Phi\left[-\frac{\left(\underline{\mu}_{2m} - \frac{1}{2}(\hat{\underline{\mu}}_{1m} + \hat{\underline{\mu}}_{2m})\right)^T \mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right) - \log\left(\frac{\hat{\beta}_{2m}}{\hat{\beta}_{1m}}\right)}{\left\{\left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)^T \mathbf{S}^{-1}\Sigma\mathbf{S}^{-1} \left(\hat{\underline{\mu}}_{1m} - \hat{\underline{\mu}}_{2m}\right)\right\}^{1/2}}\right] \quad (5.7) \end{aligned}$$

As usual, $\Phi(.)$ denotes the standard normal distribution function. The overall probability of misclassification from Π_i is the sum of P_{im}'s, the probabilities of misclassification for each multinomial cell m of Π_i, weighted by the probability of the occurrence of the cell. Hence, the actual error rates P_1 and P_2 are given respectively by

$$P_1 = \sum_{m=1}^{2^q} \beta_{1m} P_{1m} \text{ and } P_2 = \sum_{m=1}^{2^q} \beta_{2m} P_{2m}, \quad (5.8)$$

and the overall actual error rate can be computed using the formula given by (5.2).

From the expressions above, we can see that the arguments in the definition of the actual error rates are still functions of unknown parameters, so these error rates can not be computed directly from the given training data alone. Consequently a procedure for estimating these error rates is needed. A large number of works on estimation of the error rates in discriminant analysis have been reported in literature. For the details we refer to Mangku (1992).

References

[1] Aitchison, J., and Aitken, C.G.G. (1976). "Multivariate Binary Discrimination by the Kernel Method," Biometrika, 63, 413-420.
[2] Ambergen, A.W. (1985). "Interval Estimates for Posterior Probabilities in a Multivariate Normal Classification Model," Journal of Multivariate Analysis, 16, 432-439.
[3] Anderson, T.W. (1951). "Classification by Multivariate Analysis," Psychometrika, 16, 31-50.
[4] Anderson, T.W. (1973). "Asymptotic Expansion of the Distribution of the Studentized Classification Statistic W," Annals of Statistics, 1, 964-972.
[5] Ashikaga, T., and Chang, P.C. (1981). "Robustness of Fisher's Linear Discriminant Function Under Two-Component Mixed Normal Models," Journal of the American Statistical Association, 76, 676-680.
[6] Chang, P.C., and Afifi, A.A. (1974). "Classification Based on Dichotomous and Continuous Variables," Journal of the American Statistical Association, 69, 336-339.
[7] Chhikara, R.S., and McKeon, J. (1984). "Linear Discriminant Analysis with Misallocation in Training Samples," Journal of the American Statistical Association, 79, 899-906.
[8] Choi, S.C. (1986). "Discriminant and Classification: Overview," An International Journal Computers and Mathematics with Applications, 12A, 173-177.
[9] Critchley, F. (1985). "Interval estimation in Discrimination: The Multivariate Normal Equal Covariance Case," Biometrika, 72, 109-116.
[10] Critchley, F., and Ford, I. (1984). "On the Covariance of Two Noncentral F Random Variables and The Variance of the Estimated Linear Discriminant Function," Biometrika, 71, 637-638.
[11] Critchley, F., and Ford, I. (1985). "Interval Estimation in Discrimination: The Multivariate Normal Equal Covariance Case," Biometrika, 72, 109-116.
[12] Davis, A.W. (1987). "Moments of Linear Discriminant Function and an Asymptotic Confidence Interval for the Log Odds Ratio," Biometrika, 74, 829-840.
[13] Dillon, W.R., and Goldstein, M. (1978). "On The Performance of Some Multinomial Classification Rules," Journal of the American Statistical Association, 73, 305-313.
[14] Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problem," Annals of Eugenics, 7, 179-188.
[15] Friedman, J.H. (1989). "Regularized Discriminant Analysis," Journal of the American Statistical Association, 84, 165-175.
[16] Ganesalinggam, S., McLachlan, G.J. (1978). "The Efficiency of a Linear Discriminant Function Based on Unclassified Initial Samples," Biometrika, 65, 658-662.
[17] Gessaman, M.P., and Gessaman, P.H. (1972). "A Comparison of Some Multivariate Discriminant Procedures," Journal of the American Statistical Association, 67, 468-472.
[18] Gilbert, E.S. (1968). "On Discrimination Using Qualitative Variables," Journal of the American Statistical Association, 63, 1399-1412.
[19] Gilbert, E.S. (1969). "The effect of Unequal Variance-Covariance Matrices on Fisher's Linear Discriminant Function," Biometrics, 25, 505-515.
[20] Glick, N. (1973). "Sample-Based Multinomial Classification," Biometrics, 29, 241-256.

[21] Goldstein, M. (1975). "Comparison on Some Density Estimate Classification Procedures," Journal of the American Statistical Association, 70, 666-669.
[22] Goldstein, M., and Dillon, W.R. (1978). Discrete Discriminant Analysis. John Wiley & Sons, New York.
[23] Hall, P. (1981). "On Nonparametric Multivariate Binary Discrimination," Biometrika, 68, 287-294.
[24] Hand, D.J. (1983). "A Comparison of Two Methods of Discriminant Analysis Applied to Binary Data," Biometrics, 39, 683-694.
[25] Hills, M. (1966). "Allocation Rules and Their Error Rates," Journal of The Royal Statistical Society, Ser.B, 28, 1-20.
[26] Hills, M. (1967). "Discrimination and Allocation with Discrete Data," Applied Statistics, 16, 237-250.
[27] Knoke, J.D. (1982). "Discriminant Analysis with Discrete and Continuous Variables," Biometrics, 38, 191-200.
[28] Knoke, J.D. (1986). "The Robust Estimation of Classification Error Rate," An International Journal Computers and Mathematics with Applications, 12A, 253-260.
[29] Krzanowski, W.J. (1975). "Discrimination and Classification Using Both Binary and Continuous Variables," Journal of the American Statistical Association, 70, 782-790.
[30] Krzanowski, W.J. (1976). "Canonical Representation of the Location Model for Discriminant or Classification," Journal of the American Statistical Association, 71, 845-848.
[31] Krzanowski, W.J. (1977). "The Performance of Fisher's Linear Discriminant Function Under Non-optimal Condition," Technometrics, 19, 191-200.
[32] Krzanowski, W.J. (1979). "Some Linear Transformation for Mixture of Binary and Continuous Variables, With Particular Reference to Linear Discriminant Analysis," Biometrika, 66, 33-39.
[33] Krzanowski, W.J. (1980). "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," Biometrics, 36, 493-499.
[34] Krzanowski, W.J. (1982). "Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis Testing Approach," Biometrics, 38, 991-1002.
[35] Krzanowski, W.J. (1983a). "Stepwise Location Model Choice in Mixed-Variable Discrimination," Applied Statistics, 32, 260-266.
[36] Krzanowski, W.J. (1983b). "Distance Between Populations Using Mixed Continuous and Categorical Variables," Biometrika, 70, 235-243.
[37] Krzanowski, W.J. (1986). "Multiple Discriminant Analysis in the Presence of Mixed Continuous and Categorical Data," An International Journal Computers and Mathematics with Applications, 12A, 179-185.
[38] Lachenbruch, P.A. (1979). "Note on Initial Misclassification Effects on the Quadratic Discriminant Function," Technometrics, 21, 129-132.
[39] Mangku, I W. (1992). *Error Rate Estimation in Discriminan Analysis: Another Look at Bootstrap and Other Empirical Techniques.* Unpublished Master Thesis, Curtin University of Technology, Perth, Australia.
[40] Marks, S., and Dunn, O.J. (1974). "Discriminant Functions When Covariance Matrices Are Unequal," Journal of the American Statistical Association, 69, 555-559.
[41] McLachlan, G.J. (1977). "Estimation the Linear Discriminant Function from Initial Samples Containing a Small Number of Unclassified Observations," Journal of the American Statistical Association, 72, 403-406.

[42] McLachlan, G.J. (1986). "Error Rate Estimation in Discriminant Analysis: Recent advances," In Advances in Multivariate Statistical Analysis, ed. A. K. Gupta, Dordrecht: D. Reidel, 233-252.
[43] Moore, D.H., II (1973). "Evaluation of Five Discrimination Procedures for Binary Variables," Journal of the American Statistical Association, 68, 389-404.
[44] Murphy, B.J., and Moran, M.A. (1986). "Parametric and Kernel Density Methods in Discriminant Analysis: Another Comparison," An International Journal Computers and Mathematics with Applications, 12A, 197-207.
[45] Okamoto, M. (1963). "An Asymptotic Expansion for The Distribution of The Linear Discriminant Function," Ann. Math. Stat., 34, 1286-1301.
[46] Randles, R.H. et al. (1978). "Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates," Journal of the American Statistical Association, 73, 564-568.
[47] Raveh, A. (1989). "A Nonmetric Approach to Linear Discriminant Analysis," Journal of the American Statistical Association, 84, 176-183.
[48] Toussaint, G.T. (1974). "Bibliography on Estimation of Misclassification," IEEE Transactions on Information Theory, 20, 472-479.
[49] Tu, C.T., and Han, C.P. (1982). "Discriminant Analysis Based on Binary and Continuous Variables," Journal of the American Statistical Association, 77, 447-454.
[50] Van Ness, J.W. (1979). "On the Effects of Dimension in Discriminant Analysis for Unequal Covariance Populations," Technometrics, 21, 119-127.
[51] Vlachonikolis, I.G. (1990). "Predictive Discrimination and Classification With Mixed Binary and Continuous Variables," Biometrika, 77, 657-662.
[52] Vlachonikolis, I.G., and Marriott, F.H.C. (1982). "Discrimination With Mixed Binary and Continuous Data," Applied Statistics, 31, 23-31.
[53] Wakaki, F. (1990). "Comparison of Linear and Quadratic Discriminant Functions," Biometrika, 77, 227-229.
[54] Wald, A. (1944). "On a Statistical Problem Arising in the Classification of an Individual Into One of Two Group," Annals of Mathematical Statistics, 15, 145-169.
[55] Wald, P.W., and Kronmal, R.A. (1977). "Discriminant Function When Covariance Are Unequal and Sample Sizes Are Moderate," Biometrics, 33, 479-484.
[56] Wojciechowski, T. (1985). "The Empirical Bayes Classification Rules for Mixture of Discrete and Continuous Variables," Biometrical Journal, 27, 521-532.