COMPARING OPTIMISM OF ERROR RATE ESTIMATORS IN DISCRIMINANT ANALYSIS BY MONTE CARLO SIMULATION ON MULTIVARIATE NORMAL DATA

I WAYAN MANGKU

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

ABSTRACT. The problem considered in this paper is estimation of the error rate in two-group discriminant analysis. Here, performance of 19 existing error rate estimators are compared and contrasted by mean of Monte Carlo simulations under the ideal condition that both parent populations are multivariate normal with common covariance matrix. The criterion used for comparing those error rate estimators is optimism. Five experimental factors are considered for the simulation, they are the number of variables, the sample size relative to the number of variables, the Mahalanobis squared distance between the two populations, dependency factor among variables, and the degree of variation among the elements of the mean vector of the populations. The result of the simulation shows that there is no estimator performing the best for all situations. However, in general, the estimator $\bar{U}$ proposed by Lachenbruch and Mickey (1968) is the best.

Key words: Discriminant analysis, classification rule, probability of misclassification, actual error rate, Monte Carlo simulation, optimism.

1. INTRODUCTION

The problem in two-groups discriminant analysis considered in this paper is as follows. Given the existence of two groups of individuals, one want to find a classification rule for allocating new individuals (observations) into one of the existing two groups. Corresponding to each classification rule, there is a probability of misclassifications if that classification rule is used to classify new individuals (observations) into one of the two groups. The best classification rule is the one that leads to the smallest probability of misclassifications, which also called error rates.

The error rates that have been frequently considered for study 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.

However, in practice, the parameters are rarely known, and the expected (or unconditional) error rates depend heavily on the distribution of the discriminant function, which is very complicated. Consequently most work associated with error rate have assumed that the samples, which are used to construct the estimated classification rule, are fixed. This leads to the exploration of the *conditional error rate.* Here the word *conditional* refers to the conditioning of the training samples from which the classification rule is constructed. One may also think of this as the probability that the given classification rule 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 this paper we concentrate only on the actual error rate and its estimation.

2. Classification rule

The classification rule used in the current study can be described as follows. Recall that 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 that $\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_p)^T$ 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, that is if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) > 1$; or classify $\underline{\mathbf{x}}$ into Π_2 if it has greater probability of coming from Π_2, that is if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) < 1$; or classify $\underline{\mathbf{x}}$ arbitrarily into Π_1 or Π_2 if these probabilities are equal or if $f_1(\underline{\mathbf{x}})/f_2(\underline{\mathbf{x}}) = 1$.

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. However, in this paper, only the case with equal prior probabilities and equal cost due to misclassifications is considered.

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

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

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. In this paper (2.2) is considered as our classification rule, and sometime the notation $W(\underline{\mathbf{x}}, \underline{\mathbf{t}})$ is used, to give an emphasize that this classification rule is constructed using the training sample $\underline{\mathbf{t}}$, to classify the new individual $\underline{\mathbf{x}}$.

3. Simulation Study Plan

In this comparative study, some existing estimators are compared and contrasted using Monte Carlo simulations. The usefulness of a Monte Carlo assessment is that the population parameters and the true distribution from which the training data are obtained are known, thus the true error rates (in our case the actual error rate) can always be computed. Hence, the estimated error rates can be compared with the true error rate for choosing the best estimator. In this comparative study, behaviour of the 19 estimators are compared and contrasted under ideal conditions that both parent populations are multivariate normal with common covariance matrix. Those 19 estimators are: *Resubstitution* (R) (Smith, 1947), *OS* (Okamoto, 1963), M (McLachlan, 1974), *NS* (Glick, 1978), U (Lachenbruch, 1967), $\bar{U}$ (Lachenbruch and Mickey, 1968), *Jackknife (JK)* (Efron, 1982), *Infinite Seperate Efron (ISE)* (Efron, 1983), *Infinite Mixture Efron (IME)* (Efron, 1983), *Infinite Seperate Chatterjee (ISC)* (Chatterjee and Chatterjee, 1983), *Infinite Mixture Chatterjee (IMC)* (Chatterjee and Chatterjee, 1983), *Finite Seperate Efron (FSE)* (Efron, 1983), *Finite Mixture Efron (FME)* (Efron, 1983), *Finite Seperate Chatterjee (FSC)* (Chatterjee and Chatterjee, 1983), *Finite Mixture Chatterjee (FMC)* (Chatterjee and Chatterjee, 1983), *Infinite Seperate Balanced (ISB)* (Mangku, 2007), *Finite Seperate Balanced (FSB)* (Mangku, 2007), *Infinite Mixture Balanced (IMB)* (Mangku, 2007) and *Finite Mixture Balanced (FMB)* (Mangku, 2007).

The overall error rates (estimated and actual) from these Monte Carlo simulations are used for comparisons. Computer programs written in GAUSS are used in these simulation studies. The criterion used in this comparative study is *optimism*. This criterion is aimed at quantifying the amount of optimism associated with each estimator in estimating the actual error rate. This optimism criterion, denoted by OPT, is the percentage of the number of simulated data sets in which the estimated error rate is less than the corresponding actual error rate. Thus, an estimator with a small value

of OPT is underoptimistic, while a large value indicates that the estimator is overoptimistic. Hence, a good estimator should have a value for OPT in the neighbourhood of 50%.

Without loss generality, it is assumed that mean vectors $\underline{\mu}_1 = \underline{O}$, $\underline{\mu}_2 = \underline{\mu}$ and covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$. We further assume that all variables are standardized so that the common covariance matrix Σ is in fact a correlation matrix. The simulation plan used here is similar to that of Ganeshanandam and Krzanowski (1990).

Five experimental factors are considered for the simulation of ideal multivariate normal data:

(a) p : the number of variables, considered at 2 levels: $p = 5$, 10.
(b) f : the sample size relative to p , considered at 2 levels: $f =$ small , large. Equal sample sizes used, i.e. $n_1 = n_2 = n_*$ (say), thus for $p = 5$, $n_* = 10$ or 20 and for $p = 10$, $n_* = 20$ or 40.
(c) Δ^2 : the true Mahalanobis squared distance between Π_1 and Π_2, considered at 3 levels: $\Delta^2 = 1.098$ (closed populations), 2.836 (medium separation), and 6.574 (well separated populations).
(d) ν : the dependency factor, considered at 2 levels: $\nu = 0.4$, 0.8 (dependence among variables increases as ν decreases from 1, $0 < \nu \leq 1$).
(e) d : the factor to determine the elements μ_k of $\underline{\mu}$, considered at 2 levels: $d = 0.4$ (large differences among μ_k), 0.8 (small differences among μ_k) and $0 < d \leq 1$.

Hence, the simulation plan is a 2x2x3x2x2 factorial experiment consisting of 48 different combinations. This simulation study plan attempts to generate more realistic data to resemble real life data, and to cover a wide variety of ideal conditions.

4. Generation of the Training Data

Once the values of p and f are fixed, the factor ν determines the eigenvalues λ_i of Σ as $\lambda_i = a\nu^{i-1} + 0.1$ for $i = 1, 2, \ldots, p$ with $a = 0.9p(1-\nu)/(1-\nu^p)$ if $0 < \nu < 1$ or $a = 0.9$ if $\nu = 1$. If $\mathbf{E}$ is the matrix of eigenvectors of Σ and Λ is the diagonal matrix of eigenvalues λ_i, then as we can write $\Sigma = \mathbf{E}\Lambda\mathbf{E}^T$, we only need a random orthogonal matrix $\mathbf{E}$ generated to compute Σ. Having determined the eigenvalues, Lin and Bendel's (1985) algorithm can be used to generate random population correlation matrices with these specified eigenvalues. Factor d is used as an attempt to generate more realistic values for the elements μ_k in the mean vector $\underline{\mu}$, than just the simple case of having zeros in all positions except the first. Then we compute $\mu_i^* = \sqrt{\Omega d^{i-1}}$ for $i = 1, 2, \ldots, p$ and $0 < d \leq 1$, where $\Omega = \Delta^2(1-d)/(1-d^p)$ if $0 < d < 1$ or $\Omega = \Delta^2/p$ if $d = 1$. The elements μ_i are then obtained from $\underline{\mu} = \mathbf{R}\underline{\mu}^*$ where

$\Sigma = \mathbf{R}\mathbf{R}^T$ is given by the Cholesky's decomposition and $\underline{\mu}^* = (\mu_1^*, \ldots, \mu_p^*)^T$. Finally, the desired p-variate observation vector $\underline{\mathbf{x}}$ is obtained by, first generating a vector $\underline{\mathbf{y}}$ of p independent $N(0,1)$ values and then transforming it into $\underline{\mathbf{x}} = \underline{\mu} + \mathbf{R}\underline{\mathbf{y}}$.

5. Calculation of The Actual Error Rate

The *actual error rates* of the linear discriminant function $W(\underline{\mathbf{x}}, \underline{\mathbf{t}})$ are given by

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

Here, P_1 represents the probability of classifying the new individual $\underline{\mathbf{x}}$ in to Π_2 when it is actually belong to Π_1 and P_2 represents the probability of classifying the new individual $\underline{\mathbf{x}}$ in to Π_1 when it is actually belong to Π_2. 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

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

and

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

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

From the expressions above, we can see that the arguments 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.

We generated 50 replicates for each of the 48 sampling situations. The actual error rate AC and the overall error rate estimate from each of R, OS, M, NS, U, $\bar{U}$, JK, ISE, IME, ISC, IMC, FSE, FME, FSC, FMC, ISB, FSB, IMB and FMB, estimators were computed for each replicate. The OPT criterion was then computed as

$$OPT = 2 \sum_{i=1}^{50} \Psi_i,$$

where $\Psi = 1$ if $\hat{P}_i < AC_i$ and 0 otherwise, $\hat{P}_i$ and AC_i are the estimates and the actual of the overall error rates computed from the i-th replicate of a given Monte Carlo sampling situation.

6. Monte Carlo Results and Discussions

First, the effects of the experimental factors p , f , Δ^2, ν and d on the error rate estimators are examined. Recall that the Monte Carlo study plan is a *balanced factorial experimental design*. Since all of the error rate estimators are applied to the same set of simulated training samples, the 19 values of $\hat{P}_i$ are correlated in each of the 50 replicates. Hence, the values of the criterion OPT are correlated. In such a situation, a Repeated Measures Analysis of Variance (Hand and Taylor, 1987) is appropriate, where the error rate estimators can be treated as the repeated measures. Performance of the various error rate estimators are then examined using means of the error rates and the OPT with respect to the significant experimental treatment effects. The statistical computing software SAS was used to carry out the above analysis.

The results of the *repeated measures analysis* is presented in Table 1. Here the levels of the factor *error rate estimation methods*, denoted by *METH*, are the nineteen error rate estimators explained in section 3. In this table, the *ANOVA* of the experimental factors and their interactions are given in the *main plot stratum*, whereas the *repeated* factor *METH* together with its interactions with all experimental factors are given in the *split plot stratum*. For ease of interpretations and to avoid complexity, the order of interactions were kept to 1 among the main plots and to 2 in the split plot stratum. Because of the large number of replicates in the experiment, the F-ratios are also treated as guides to the relative importance of the corresponding treatment effects besides the absolute tests of significance.

The main plot stratum of Table 1 shows that the main and interaction effects p, ν, p x ν, and p x d are important. On the other hand, the split plot stratum shows that the effects of *METH* factor and its interaction with p, f, Δ^2, p x Δ^2, f x Δ^2, ν, p x ν, f x ν and ν x d are all significant. This not only indicates the fact that there are some estimators with significantly different optimism in estimating the actual error rates, but also suggests that any comparison of the estimators must be qualified by the above main and interaction effects of the experimental factors. Here, the influence of the factor ν (dependency of the variables) on optimism of the error rate estimators is much higher than that of the factors Δ^2 , f, and p, as the corresponding F-ratios being 20.13, 11.38, 8.02, and 5.80.

Note that the above results from the repeated measures analysis show the effects of the Monte Carlo experimental factors when *averaged* over the different estimators. However, SAS was also subjected to perform individual *ANOVA*'s separately for each of the estimators, in order to highlight any deviations from the average behaviour of our experimental factors.

Table 1: Main plot and Split plot stratum of the Repeated measures *ANOVA* for the effects of the experimental factors on all methods.

SOURCE	DF	SS	MS	F-ratio	p-value
Main Plot					
p	1	20862.9868	20862.9868	15.01	0.0006
f	1	13.2675	13.2675	0.01	0.9229
$p * f$	1	2022.1096	2022.1096	1.45	0.2383
Δ^2	2	3782.1140	1891.1140	1.36	0.2737
$p * \Delta^2$	2	8799.8158	4399.9079	3.16	0.0583
$f * \Delta^2$	2	2668.6404	1334.3202	0.96	0.3957
ν	1	67706.3202	67706.3202	48.70	0.0001
$p * \nu$	1	9568.1096	9568.1096	6.88	0.0141
$f * \nu$	1	7981.5833	7981.5833	5.74	0.0238
$\Delta^2 * \nu$	2	7775.9562	3887.9761	2.80	0.0788
d	1	1813.3728	1813.3728	1.30	0.2635
$p * d$	1	9490.5307	9490.5307	6.83	0.0145
$f * d$	1	603.6886	603.6886	0.43	0.5155
$\Delta^2 * d$	2	2168.7982	1084.3991	0.78	0.4685
$\nu * d$	1	4500.7412	4500.7412	3.24	0.0832
ERROR	27	37540.8026	1390.4001		
Split plot					
METH	18	80185.5965	4454.7554	335.01	0.0001
METH $* p$	18	1388.9299	77.1628	5.80	0.0001
METH $* f$	18	1920.6491	106.7027	8.02	0.0001
METH $* p * f$	18	187.1404	10.3967	0.78	0.7225
METH $* \Delta^2$	36	5446.7193	151.2978	11.38	0.0001
METH $* p * \Delta^2$	36	1012.0175	28.1116	2.11	0.0002
METH $* f * \Delta^2$	36	739.1930	20.5331	1.54	0.0250
METH $* \nu$	18	4817.5965	267.6443	20.13	0.0001
METH $* p * \nu$	18	1507.8070	83.7671	6.30	0.0001
METH $* f * \nu$	18	693.6667	38.5370	2.90	0.0001
METH $* \Delta^2 * \nu$	36	383.8772	10.6633	0.80	0.7892
METH $* d$	18	161.8772	8.9932	0.68	0.8357
METH $* p * d$	18	230.7193	12.8177	0.96	0.5008
METH $* f * d$	18	196.2281	10.9016	0.82	0.6774
METH $* \Delta^2 * d$	36	437.7018	12.1584	0.91	0.5168
METH $* \nu * d$	18	636.5088	35.3616	2.66	0.0003
ERROR	1486	6462.6140	13.2976		

These *ANOVA*'s are summarized in Table 2. F-ratios associated with significant level ≥ 0.05 have been omitted.

From Table 2 we can see that the effect of factor ν is highly significant for all methods; factor p is important for all estimators except M and JK; and the effect due to the interaction p x ν is significant for all methods except OS, NS, and R. The interaction f x ν has significant effect only for the *bootstrap* estimators, though the effect due to the interaction p x Δ^2 is significant for OS and all the bootstrap error rates except IME, IMC, and FMC. While Δ^2 is important only for OS, NS, R, ISC, and FSC, the sample size factor f seems important only for the interpretation of the R estimator.

Table 2: The F-ratios[a] and their p-values[b] of the effects of the experimental factors on each estimator (cases with p-values > 0.0500 are omitted).

METH	p	f	Δ^2	p x Δ^2	ν	p x ν	f x ν
OS	24.53[a]		4.46	3.57	37.06		
	0.0001[b]		0.0212	0.0421	0.0001		
M					52.81	10.14	
					0.0001	0.0036	
NS	6.88		10.15		6.88		
	0.0142		0.0005		0.0142		
R	11.31	9.75	5.22		4.67		
	0.0023	0.0042	0.0121		0.0396		
U	6.36				53.48	8.78	
	0.0179				0.0001	0.0063	
$\bar{U}$	16.43				47.41	8.87	
	0.0004				0.0001	0.0061	
JK					46.09	7.47	
					0.0001	0.0109	
ISE	13.40			4.13	49.96	8.07	6.70
	0.0011			0.0273	0.0001	0.0085	0.0154
IME	15.04				43.35	6.23	5.36
	0.0006				0.0001	0.0190	0.0284
ISC	14.92		3.37	4.79	51.85	6.19	6.86
	0.0006		0.0493	0.0166	0.0001	0.0193	0.0143
IMC	13.10				41.18	5.77	5.77
	0.0012				0.0001	0.0234	0.0234
ISB	13.23			3.99	47.09	7.57	8.65
	0.0011			0.0303	0.0001	0.0105	0.0066
IMB	18.94			3.46	45.09	7.15	8.18
	0.0002			0.0461	0.0001	0.0126	0.0081
FSE	16.34			4.98	49.47	9.53	8.02
	0.0004			0.0144	0.0001	0.0046	0.0086
FME	15.90			3.38	44.96	5.84	5.00
	0.0005			0.0489	0.0001	0.0227	0.0339
FSC	15.83		3.42	4.65	51.87	6.41	7.45
	0.0005		0.0474	0.0183	0.0001	0.0175	0.0110
FMC	14.06				40.82	5.33	5.33
	0.0009				0.0001	0.0288	0.0288
FSB	14.62			3.95	47 88	7.91	8.27
	0.0007			0.0314	0.0001	0.0091	0.0078
FMB	19.57			3.76	47.21	7.45	8.15
	0.0001			0.0363	0.0001	0.0110	0.0082

Table 3: Mean[a] of error rate and OPT^b for the main effects of the experimental factors p, f, Δ^2 and ν.

METH	p		f		Δ^2			ν	
	5	10	*small*	*large*	1.098	2.836	6.574	0.4	0.8
AC	0.255[a]	0.277	0.282	0.250	0.371	0.269	0.158	0.287	0.244
OS	0.230[a]	0.224	0.237	0.216	0.322	0.231	0.126	0.226	0.227
	62.17[b]	75.25	67.75	69.671	74.25	66.50	65.38	76.75	60.67
M	0.247	0.254	0.269	0.232	0.322	0.231	0.126	0.250	0.251
	54.08	60.33	54.58	59.831	53.25	56.00	62.38	68.33	46.08
NS	0.183	0.176	0.175	0.184	0.260	0.184	0.095	0.179	0.180
	84.58	91.08	90.17	85.50	95.50	85.63	82.38	91.08	84.58
R	0.148	0.149	0.131	0.165	0.229	0.151	0.065	0.149	0.148
	90.75	95.42	95.25	90.92	96.25	91.38	91.63	94.58	91.58
U	0.248	0.251	0.266	0.232	0.359	0.252	0.136	0.248	0.250
	54.58	61.25	56.08	59.75	55.00	57.50	61.25	67.58	48.25
$\bar{U}$	0.273	0.262	0.292	0.243	0.382	0.271	0.151	0.267	0.268
	42.83	56.67	46.25	53.25	46.38	50.13	52.75	61.50	38.00
JK	0.241	0.248	0.259	0.230	0.354	0.247	0.132	0.244	0.245
	56.83	62.25	58.58	60.50	57.75	58.50	62.38	68.75	50.33
ISE	0.230	0.231	0.239	0.222	0.328	0.235	0.129	0.231	0.230
	61.25	70.92	67.25	64.92	70.00	64.25	64.00	75.42	56.75
IME	0.232	0.232	0.241	0.223	0.329	0.237	0.130	0.232	0.232
	59.83	70.58	65.75	64.67	69.13	63.88	62.63	74.33	56.08
ISC	0.227	0.229	0.236	0.220	0.325	0.233	0.127	0.228	0.228
	62.17	72.00	68.00	66.17	71.75	65.00	64.50	76.25	57.92
IMC	0.229	0.230	0.238	0.221	0.326	0.234	0.128	0.229	0.229
	61.50	71.42	67.17	65.75	70.75	65.13	63.50	75.25	57.67
ISB	0.232	0.233	0.242	0.224	0.330	0.237	0.131	0.233	0.233
	60.58	70.17	65.83	64.92	69.50	63.50	63.13	74.42	56.33
IMB	0.234	0.234	0.244	0.225	0.331	0.239	0.133	0.234	0.234
	59.08	70.75	65.67	64.17	69.25	63.38	62.13	73.92	55.92
FSE	0.232	0.232	0.242	0.223	0.330	0.237	0.130	0.232	0.232
	60.17	70.75	66.17	64.75	69.13	63.75	63.50	74.67	56.25
FME	0.233	0.233	0.243	0.223	0.330	0.238	0.131	0.233	0.233
	59.58	70.58	65.50	64.67	69.00	63.75	62.50	74.33	55.83
FSC	0.229	0.230	0.238	0.221	0.326	0.234	0.128	0.230	0.229
	61.92	72.00	67.83	66.08	71.63	65.00	64.25	76.08	57.83

Table 3: Continued.

METH	p		f		Δ^2			ν	
	5	10	*small*	*large*	1.098	2.836	6.574	0.4	0.8
FMC	0.230	0.231	0.239	0.221	0.327	0.235	0.129	0.230	0.230
	61.00	71.42	66.67	65.75	70.50	64.88	63.25	75.08	57.33
FSB	0.235	0.235	0.245	0.225	0.332	0.239	0.133	0.235	0.235
	59.92	70.00	65.00	64.92	69.38	62.88	62.63	74.08	55.83
FMB	0.235	0.235	0.245	0.225	0.332	0.240	0.133	0.235	0.235
	58.83	70.58	65.33	64.08	68.88	63.25	62.00	73.83	55.58

We may conclude from the analysis so far, that the experimental factor d has very little or no effect on the estimation of error rates, while p, f, Δ^2 and ν significantly influence the optimism of the error rate estimators. Hence, further interpretation of the results will be restricted to the above four factors. The *means* of error rate estimates and the means of criterion OPT for the main effects of these four factors are presented in Table 3.

From Table 3, it is very prominent that R and NS are the worst estimators which heavily overoptimistic (about 90%). Hence, these two estimators have been omitted from further analysis. We shall interpret the findings in two folds: among bootstrap estimators only and over all estimators.

Table 3 also shows that, although the balanced bootstrap estimators (*IMB* and *FMB*) outperform the other bootstrap estimators, they all seem to suffer considerably from overoptimism (about 60% for $p = 5$, and about 70% for $p = 10$). For small p, all estimator are overoptimistic except $\bar{U}$. Here, M and U seem to estimate the actual error rate with little overoptimism (about 54.1% and 54.6% respectively), while $\bar{U}$ is slightly (about 42.8%) underoptimistic. For large p, however, $\bar{U}$ becomes the best estimator with the smallest optimism (only about 56.7% of the time), though the estimators M, U and JK also behave better than the bootstrap ones. It also found that all the estimators have evidently larger OPT values all being overoptimistic for large p than that of small p, with a similar behavioural pattern in each case.

As far as the influence of the sample size factor f on the estimators is concerned, Table 3 shows that the difference between means of criterion OPT from the two sample sizes is small, though a clear pattern emerges among the estimators. That is, while all the non bootstrap estimators have larger overoptimism for large samples than for small ones, the bootstrap estimators behave the opposite way except for *FSB* (has the same OPT for both cases). All the estimators are overoptimistic, irrespectives of the sample size, except for $\bar{U}$. $\bar{U}$ is the overall best estimator with the smallest *OPT*imism (about 46.3% and 53.3% respectively) for both cases with small and large sizes of training samples. This is followed by the M estimator for small samples (about 54.6%), and U for large samples (about 56.1%). Among the bootstrap estimators, *FMB* outperforms the rest in all cases except for small sample sizes for which *FSB* becomes the best.

Now consider the behaviour of the estimation methods on the levels of the distance (separation) factor Δ^2. Table 3 shows that the behaviour of the estimators among the different values of Δ^2 is similar to those on the levels of factors p and f. The *OS* estimator together with all the bootstrap ones are all overoptimistic (above 60%), while the best choice is M for $\Delta^2 = 1.098$ (with $OPT = 53.3\%$) and

$\bar{U}$, for $\Delta^2 = 2.836$ (with $OPT = 50.1\%$) and for $\Delta^2 = 6.574$ (with $OPT = 52.8\%$). However, the second best choice goes to $\bar{U}$, M and U respectively for the above cases with corresponding optimism values 46.4%, 56% and 61.3%. Among the bootstrap methods, *FMB* again outperforms the others except when $\Delta^2 = 2.836$ for which *FSB* is the best. Once again all the estimators are overoptimistic in all cases except for $\bar{U}$ for close populations.

Finally, from Table 3 we can easily deduce that all the estimators are overoptimistic when the variables are highly interdependent, though $\bar{U}$ with about 61.5% becomes the best for this case. However, when the variables are almost independent, the *JK* method becomes the best with almost no optimism ($OPT = 50.3\%$) in estimating the actual error rate. Note also in this case that the bootstrap estimators become less overoptimistic (with $< 60\%$) than the other situations. Once again, *FMB* is the best choice among the bootstrap methods. It is also evident here that the optimism involved in estimating the actual error rates is significantly reduced for each estimator when the variables become almost independent from high interdependence.

There are some interesting and peculiar behaviours to be noted from Table 3. The estimator *OS* is the worst with the largest overoptimism in almost every simulated case. The optimism of $\bar{U}$ is peculiar such that it is underoptimistic for data with small number of variables, small sample sizes relative to the number of variables, small separation between populations and almost independent variables; while it is overoptimistic for data with large number of variables, large sample sizes relative to the number of variables, moderate to large separation between populations and highly interdependent variables. An interesting behaviour that we may notice among the bootstrap estimators is that the difference between finite and infinite versions of the estimators due to criterion *OPT* is negligible; while, although the difference between separate and mixture sampling versions also small, estimators based on mixture sampling procedure seem preferable. We also notice that Efron's estimators are slightly superior to Chatterjee's methods.

The presentation of the significant interaction effects of the experimental factors for all estimators is quite cumbersome. Hence, we chose only the estimators, *U*, $\bar{U}$, *OS*, *M*, *FME*, *FMC* and *FMB* for this purpose. The choice here was based on the fact that some of these estimators (eg. $\bar{U}$) outperform the others in particular circumstances with main effects of factors, and the others (eg. *FMC*) are to represent special forms of estimators.

Since only 7 estimators are considered for further interpretation, the choice of the interaction effects to be interpreted also restricted to those interactions which have significant influence on these estimators. From the F-ratios of the repeated measures *ANOVA*'s for the *OPT* values, show that the influence of the interaction *METH* x p x ν is much higher than those of the other interactions. Thus we may choose to interpret only the effect of p x ν on the 7 estimators considered. However, the individual *ANOVA*'s suggests that the interaction p x ν has significant influence only on the *M*, *U*, $\bar{U}$, *FME*, *FMC* and *FMB* estimators. Hence, it would be appropriate to interpret the effect of p x ν only on these 6 estimators.

Result of the analysis shows that all the 6 estimators have similar behaviour. They have smaller *OPT* means when the variables are independent than when they are interdependent, for both levels of p. These differences are greater when $p = 10$ than those when $p = 5$. It was also found that all the 6 estimators are heavily overoptimistic when the data consist of 10 interdependent variables, hence this situation becomes the worst. In general, all the 6 estimators behave less-optimistically when the variables are independent than when they are interdependent.

7. Conclusion

Based on the results of the comparative study under the ideal conditions of multivariate normality with equal covariance matrix, we may deduce some important points as follows. The balanced bootstrap estimators outperform their counter parts and become the best for all situations. The Finite Separate Balanced (*FSB*) estimator becomes the best estimator for cases with large number of variables or with small samples or with medium separation of the populations. For all the other situations, the Finite Mixture Balanced (*FMB*) estimator is the best.

If we compare all estimators together, the best estimator is M for cases with small number of variables or close populations and JK for independent variables case. For cases either with large number of variables or with small or large samples, or with medium or well separated populations, or with interdependent variables, the best choice is the $\bar{U}$ estimator.

References

[1] Anderson, T.W. (1951). Classification by multivariate analysis, *Psychometric*, **16**, 631-650.

[2] Chatterjee, S. and Chatterjee, S. (1983). Estimation of missclassification probabilities, *Commun. Statist-Simula. Computa.*, **12**, 645-656.

[3] Efron, B. (1982). *The Jackknife, The Bootstrap and Other Resampling Plans*, SIAM-CBMS Monograph 38. Philadelphia: S.I.A.M.

[4] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American Statistical Association*, **78**, 316-331.

[5] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problem, *Annals of Eugenics*, **7**, 179-188.

[6] Ganeshanandam, S., and Krzanowski, W.J. (1990). Error-rate estimation in two-group discriminant analysis using the linear discriminant function, *J. Statist. Comput. Simul.*, **36**, 157-175.

[7] Glick, N. (1978). Additive estimators for probabilities of correct classification, *Pattern Recognition*, **10**, 211-222.

[8] Hand, D.J., and Taylor, C.C. (1987). *Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists*, Chapman and Hall, New York.

[9] Lachenbruch, P.A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis, *Biometrics*, **23**, 639-645.

[10] Lachenbruch, P.A., and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis, *Technometrics*, **10**, 1-11.

[11] Lin, S.P., and Bendel, R.B. (1985). Generation of population correlation matrices with specified eigenvalues, *Applied Statistics*, **34**, 193-198.

[12] Mangku, I W. (2007). Balanced bootstrap estimators for the probability of misclassifications in discriminant analysis, *Journal of Mathematics and Its Applications*, **6**, 1, 11-22.

[13] McLachlan, G.J. (1974). An Asymptotic Unbiased Techniques for Estimating The Error Rate in Discriminant Analysis, *Biometrics*, **30**, 239-249.

[14] Okamoto, M. (1963). An Asymptotic Expansion for The Distribution of The Linear Discriminant Function, *Ann. Math. Stat.*, **34**, 1286-1301.

[15] Smith, C.A.B. (1947). Some Examples of Discrimination, *Annals of Eugenics*, **13**, 272-282.